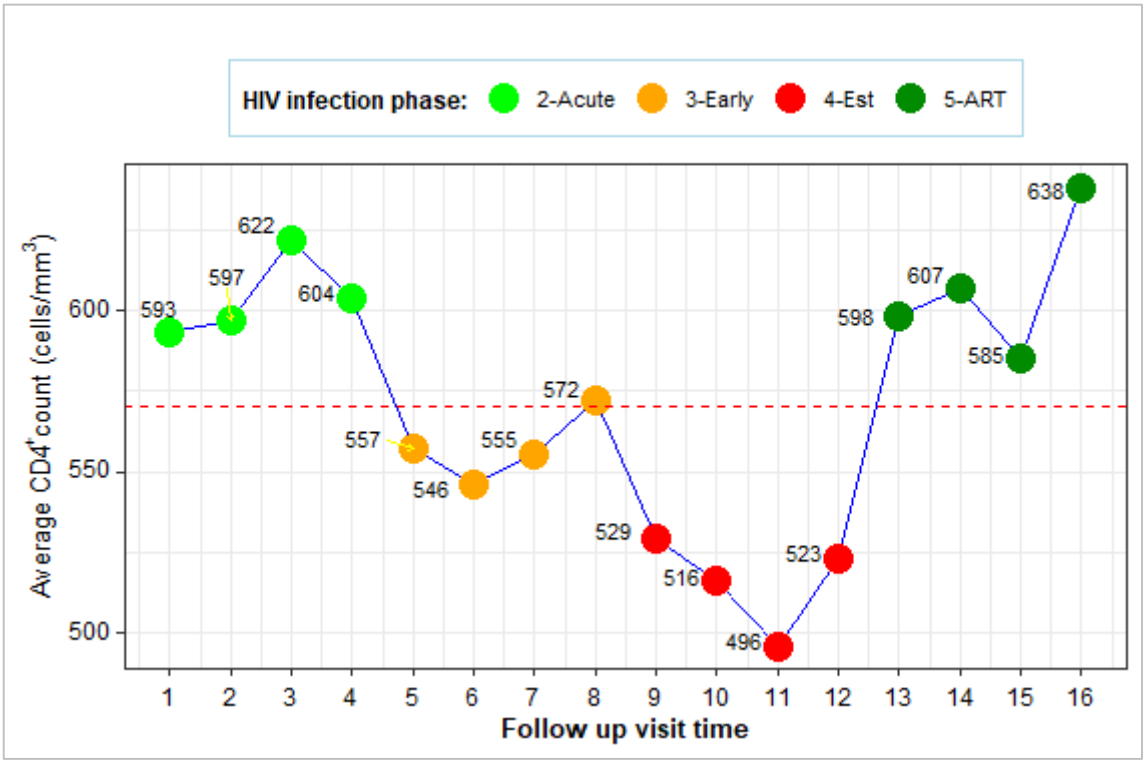
UNIVERSITY OF KWAZULU-NATAL

LONGITUDINAL CLINICAL COVARIATES INFLUENCE ON CD4$^+$ CELL COUNT AFTER SEROCONVERSION



By

PARTSON TINARWO

2019

LONGITUDINAL CLINICAL COVARIATES INFLUENCE ON CD4$^+$ CELL COUNT AFTER SEROCONVERSION

by

PARTSON TINARWO

Submitted in fulfilment of the academic requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

in the

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Westville, Durban, South Africa

Dedication

To my parents,

wife Felistas Nyakusvora-Tinarwo,

daughter Ashley Tinarwo and

son Allan Tinarwo.

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal Westville, under the supervision of Professor Temesgen Zewotir and co-supervised by Professor Delia North.

I, Partson Tinarwo, declare that this thesis is my own and unaided work. It has not been submitted for any academic or examination purpose at any other university.

_____	_____
Partson Tinarwo	Date
_____	_____
Professor Temesgen Zewotir	Date
_____	_____
Professor Delia North	Date

Articles published from this research

1. Tinarwo P, Zewotir T, Yende-Zuma N, Garrett NJ, North D. An Evaluation to Determine the Strongest CD4 count Covariates during HIV Disease Progression in Women in South Africa. *Infect Dis Ther*. 2019:8(2):269–284. https://doi.org/10.1007/s40121-019-0235-4
2. Tinarwo P, Zewotir T, North D. Covariate random effects on the CD4 count variation during HIV disease progression in women. *HIV/AIDS - Research and Palliative Care*. 2019;11:119-131. https://doi.org/10.2147/HIV.S193652

Articles submitted and under review

1. Trends and adaptive optimal set-points of CD4+ count clinical covariates at each phase of the HIV disease progression. *AIDS Research and Treatment*
   [Provisionally accepted with minor corrections and the minor corrections effected and resubmitted on 13 June 2019]

Acknowledgement

First, I would like to thank God for the opportunity to study this far.

Second, again to God be the glory. In just six days You created the Heavens and the Earth (Genesis 1) we are *trying* to understand in small portions over many years through research but even though, still far from full comprehension. All the findings are just estimates and not the exact explanation of creation. Indeed, You are the Mighty God as your creation is extremely complex. Thank you for allowing me the opportunity to have a gist of the understanding of your plan of salvation through studying the wonders of the majestic designs and systems of creation. Amen.

Third, my sincere thanks go to my supervisor Professor Temesgen Zewotir for his guidance from the beginning to the end. From day one when I first approached him as a prospective student for my Master's degree, these were his two words: *"You're welcome"*. Ever since, that stimulated my journey through to the completion of this piece of PhD work. He set the pace and with his patience, I once more acknowledge him for accommodating some of my ad-hoc requests for consultations. My co-supervisor Professor Delia North, thank you for all the encouragement and that was instrumental in giving me hope and staying energised. She was always bubbling with excitement on every bit of my achievements and that kept me motivated as well as oiling the wheels of hard working.

Fourth, this would not have been possible without the support and understanding from my wife Felistas Nyakusvora-Tinarwo for balancing her own study time, work and taking care of the kids. To my kids Ashley Tinarwo ("*When are we going out*?") and Allan Tinarwo ("*What's wrong with you … no pizza for some time*?"), I owe you all the time I spent away from home busy with the research ("*You will understand it later in life*").

Fifth, the Centre for the AIDS Programme of Research in South Africa (CAPRISA team (Nonhlanhla Yende-Zuma and Nigel J. Garrett), your kindness in facilitating the data availability and explaining the technicalities in preparing manuscripts for publication is greatly appreciated and thank you.

Sixth, I would also want to extend my appreciation to our ever-smiling Statistics Department secretary Alershnee Pillay for her willingness and patience when I constantly interrupted her

normal duties to assist with data acquisition (a process she was involved in for nearly 1.5 years) and printing. I thank you for making this piece of work possible.

Seventh and lastly, the Africa Health Research Institute (AHRI) team (Dr Kobus Herbst, Kathy Baisley, Dickman Gareta and Siyabonga Nxumalo), it was a pleasure for the brief moment and thank you for the data management skills I obtained from Pentaho. It was indeed a step ahead on data manipulation in R and SAS for the PhD training.

Abstract

The Acquired Immunodeficiency Syndrome (AIDS) pandemic is a global challenge. The human immunodeficiency virus (HIV) is notoriously known for weakening the immune system and opening channels for opportunistic infections. The Cluster of Difference 4 ($CD4^+$) cells are mainly killed by the HIV and hence used as a health indicator for HIV infected patients. In the past, the $CD4^+$ count diagnostics were very expensive and therefore beyond the reach of many in resource-limited settings. Accordingly, the $CD4^+$ count's clinical covariates were the potential diagnostic tools. From a different angle, it is essential to examine a trail of the clinical covariates effecting the $CD4^+$ cell response. That is, inasmuch as the immune system regulates the $CD4^+$ count fluctuations in reaction to the viral invasion, the body's other complex functional systems are bound to adjust too. However, little is known about the corresponding adaptive behavioural patterns of the clinical covariates influence on the $CD4^+$ cell count. The investigation in this study was carried out on data obtained from the Centre for the Programme of AIDS research in South Africa (CAPRISA), where initially, HIV negative patients were enrolled into different cohorts, for different objectives. These HIV negative patients were then followed up in their respective cohort studies. As soon as a patient seroconverted in any of the cohort studies, the patient was then enrolled again, into a new cohort of HIV positive patients only. The follow-up on the seroconvertants involved a simultaneous recording of repeated measurements of the $CD4^+$ count and 46 clinical covariates. An extensive exploratory analysis was consequently performed with three variable reduction methods for high-dimensional longitudinal data to identify the strongest clinical covariates. The sparse partial least squares approach proved to be the most appropriate and a robust technique to adopt. It identified 18 strongest clinical covariates which were subsequently used to fit other sophisticated statistical models including the longitudinal multilevel models for assessing inter-individual variation in the $CD4^+$ count due to each clinical covariate. Generalised additive mixed models were then used to gain insight into the $CD4^+$ count trends and possible adaptive optimal set-points of the clinical covariates. To single out break-points in the change of linear relationships between the $CD4^+$ count and the covariates, segmented regression models were employed. In getting to grips with the understanding of the highly complex and intertwined relationships between the $CD4^+$ count, clinical covariates and the time lagged effects during the HIV disease progression, a Structural Equation Model (SEM) was constructed and fitted. The results showed that sodium consistently changed its effects at 132mEq/L and 140 mEq/L across all the post HIV infection phases. Generally, the covariate influence on the $CD4^+$ count varied with infection phase and

widely between individuals during the anti-retroviral therapy (ART). We conlude that there is evidence of covariate set-point adaptive behaviour to positively influence the $CD4^+$ cell count during the HIV disease progression.

Keywords: Break-points, HIV infection phases, inter-individual variation, lagged path models, R macros, random smooths, variable selection, voluminous results data wrangling.

Acronyms

Table of Contents

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

The Acquired Immunodeficiency Syndrome (AIDS) pandemic is a global challenge. The condition primarily affects a patient's immune system which is responsible for keeping the body under constant surveillance to defend itself from foreign invasion (Adrian et al. 2016). Among the components of the immune system, Cluster of Difference 4 ($CD4^+$) cells, are responsible for the main defence mechanism. Upon infection, the human immunodeficiency virus (HIV) is notoriously known for targeting and killing these $CD4^+$ cells (Weston and Marett 2009), resulting in a weakened immune system (Beare et al. 2008) that opens channels to opportunistic infections (Wingfield and Wilkins 2010). Prospective cohort studies are a defined approach to following-up on patients (Phair 2009, Lorente et al. 2016) and in the context of the HIV/AIDS, commonly recorded are the repeated measurements of the $CD4^+$ cell count to monitor the HIV disease progression. Hence, the $CD4^+$ cell count is widely regarded as a health indicator in HIV infected patients (Beare et al. 2008). Although the $CD4^+$ count is the most common health indicator for monitoring the HIV disease progression, equally important to focus on are the drivers or covariates that influence the long-term $CD4^+$ cell count.

Understanding of the covariates that influence the long-term $CD4^+$ cell count in monitoring the health status of HIV infected patients, is quite a complex process due to the dynamism surrounding the epidemic. This includes the socio-economic variations associated with HIV patients' attitude towards adherence to health care (Burch et al. 2016, Bunyasi and Coetzee 2017), the rapid mutation of the HIV (Shafer et al. 2007, Cuevas et al. 2015), co-infections (Goud and Ramesh 2014), as well as the biological complexity of the human body. The challenge is further exacerbated by the disease progression from one phase to the other over time (Weston and Marett 2009, Brener et al. 2018). In the North-West Ethiopia region, some determinants of the $CD4^+$ count change were found to be demographic characteristics and disclosure (Seyoum et al. 2017). In addition to this, (Montarroyos et al. 2014) found other factors such as substance abuse, treatment including antiretroviral therapy (ART) and visiting different medical practitioners. Contrary to some of these findings, the $CD4^+$ count change was not related to gender and age according to (Smith et al. 2004) who also found white ethnicity to be a factor. In Iran, insurance coverage, tuberculosis prophylaxis and a higher baseline $CD4^+$ count, were found to be protective factors (Abbastabar et al. 2016). Little attention has to date been given to the clinical covariates, and especially to the clinical covariates that are continuous

in nature. Studies on the exploration to discover patterns in which these continuous clinical covariates influence the long-term $CD4^+$ cell response during the HIV disease progression is limited.

1.1 Motivational background

The ART has been a major milestone in the treatment of HIV positive patients. The case of viral suppression is well known, resulting in designing processes both for monitoring and restoration of patients' $CD4^+$ counts to acceptable levels being relatively easy (Hunt et al. 2003). This also led to an improved survival period of patients (Eholié et al. 2016). However, inasmuch as the untreated HIV disease is life threating, the ART is with no exception due to the side effects thereof, that have been reported to be life threatening and further (Lands 2006) affect the quality of life of some patients (Chen et al. 2013, Tomita et al. 2014). Consequently, the problem of the HIV disease and its treatment is a catch-22 situation, calling for much needed research in this regard. Seemingly, research focused on alternative ways to enhance the $CD4^+$ cell response, serves to be of paramount importance.

In the past, the high costs of the $CD4^+$ count diagnostic devices (Bentwich 2005, Secko 2005) made it difficult for the resource limited settings of the developing world (Manoto et al. 2018) to easily obtain the $CD4^+$ count measurements for monitoring the health status of HIV infected patients during the disease progression. The challenge was exacerbated by overburdened health facilities (Elsa 2016), operational and logistical issues (Leung et al. 2016, Kuupiel et al. 2017), and further in the supply of even essential medicine for the patients (Bateman 2013, Nditunze et al. 2015, Jeffery 2016). Frequent instrument breakdown and poor manufacturer maintenance of the $CD4^+$ count diagnostics were also common in these developing countries (Thairu et al. 2011). Despite all these challenges, keeping abreast of the health status of HIV positive patients, in the absence of the $CD4^+$ count measurements, remained essential. Consequently, different clinical platforms were investigated, either as cost effective $CD4^+$ count surrogates (Alavi et al. 2009, Obirikorang et al. 2012, Kwantwi et al. 2017), or as predictors (Obirikorang and Yeboah 2009, Nzou et al. 2010, Moolla et al. 2015), to pre-treatment assessment and monitoring of therapy of HIV positive patients (Olawumi and Olatunji 2006). At a later stage, the $CD4^+$ count measurements became extraordinarily inexpensive to obtain (CMA Media Inc. 2005, Manabe et al. 2012, Lab Chip 2017), though all the past endeavours to deal with the problem of expensive $CD4^+$ count diagnostics, had already left a long trail of clinical

covariates. From a different angle, it then stands to reason that the introduction of these clinical covariates triggered interest in the attractiveness of their influential effects on the $CD4^+$ cell response.

When recommendations were made to suggest other factors that influence the long-term $CD4^+$ cell response in conjunction with the therapy (Smith et al. 2004), the previous studies had been inclined towards identifying categorical covariates, such as demographic and socio-economic factors (Smith et al. 2004, Montarroyos et al. 2014, Abbastabar et al. 2016, Burch et al. 2016, Yakubu et al. 2016, Bunyasi and Coetzee 2017). In clinical research, categorisation is usually considered due to the simplicity in statistical analysis (Altman and Royston 2006) for it just look at group differences (Myers and Well 2003). However, this has the drawback of discarding information and the cut-off points of such groups are further subject to debate (Altman et al. 1994, Royston et al. 2006), particularly due to the interpretation when seeking a biological meaning (Baneshi and Talei 2010). Since HIV is currently exhibiting the highest known extreme rapid biological mutation in the scientific history (Andrews and Rowland-Jones 2017), it is likely that information is lost in categorising the $CD4^+$ count, say or in estimating the $CD4^+$ count change, using categorical covariates, which even rarely change state in repeated measurements. On the other hand, the "richer" sensitive clinical covariates are more capable of capturing inherent information on the sources of variation in counting the $CD4^+$ cells that are constantly fluctuating in response to the attack by the ever evolving HIV. This is because the measurements of the continuous clinical covariates can be simultaneously recorded at almost any frequency possible to the existing technology, providing an ideal opportunity to capture close to real time evolutionary patterns of the $CD4^+$ cell count, in the face of the HIV rapid mutation during the disease progression.

The clinical covariates that have been suggested come from different clinical platforms such as the blood chemistry components (Voss et al. 1996, Choi et al. 1998, Gomo et al. 2001, Butt et al. 2002, Butt et al. 2002, Volberding et al. 2004, Khaidukov and Litvinov 2005, Fleischbeina et al. 2008, Bani-Sadr et al. 2009, Obirikorang and Yeboah 2009, Semeere et al. 2012, dos Santos and Almeida 2013, Sudfeld et al. 2013, Dusingize et al. 2015, Moolla et al. 2015, Adhikari et al. 2016, Pralhadrao et al. 2016, Shiferaw et al. 2016, Braconnier et al. 2017), full blood count (Shapiro et al. 1998, Alavi et al. 2009, Obirikorang and Yeboah 2009, Sivaram et al. 2012, Leticia et al. 2014, Vanisri and Vadiraja 2016a, 2016b), lipids (Floris-Moore et al. 2006, Iffen et al. 2010, Oka et al. 2012), sugar (Misra et al. 2013, McKnight et al. 2014,

Maganga et al. 2015), and clinical examination measurements (Dannhauser et al. 1999, Palacios et al. 2006, Esposito et al. 2008, Venter et al. 2009, Hsue et al. 2010, Nzou et al. 2010, Manner et al. 2013, Fofana 2016, Kwantwi et al. 2017, Dimala et al. 2018). With the explosion of these clinical covariates fuelled by the advancement in technology, we believe that monitoring only a few strongest ones is worthwhile for resource optimisation during prospective HIV/AIDS cohorts. This consequently allows the available multidimensional viewing lens (statistical models) to zoom into the $CD4^+$ count behavioural patterns in response to the strongest clinical covariates.

1.2 Statement of the problem

The self-regulatory immune system is known for attempting to restore the $CD4^+$ count fluctuations due to the viral infection. Consequently, the strongest covariates are bound to change accordingly too in their influential effects on the $CD4^+$ count. However, given this virally invaded environment, little is known about the corresponding complexity in either the possible adaptive optimal set-points or the interaction among the clinical covariates themselves to influence the $CD4^+$ cell response. In addition, it is common practice in the medical fraternity to administer an average dose across patients where in fact patient-tailored healthcare would have been effective (National Clinical Guideline Centre 2012). As such, there is limited knowledge on whether the effects of the clinical covariates induce variations in the $CD4^+$ count behavioural patterns between patients during the HIV disease progression. It follows that the past endeavours to deal with the obstacle of the expensive $CD4^+$ count diagnostic devices paved the way for another avenue, to consider the clinical covariates as potentially an important and integral part of the $CD4^+$ cell influence during the HIV disease progression. However, ways to harness the benefits of the knowledge of the clinical covariates' influential effects on the $CD4^+$ cell count is still hampered by the limited discovery of patterns in the relationships, particularly in the virally manipulated settings where the body's functional systems tend to deviate from the norm.

1.3 The aim and objectives of the study

This bioinformatics study aims to explore and discover the patterns in which the clinical covariates influence the long-term $CD4^+$ cell response during the HIV disease progression. The main objective in particular is to investigate the continuous clinical covariates that have high information carrying capacity in their sensitiveness, which closely capture the $CD4^+$ cell count variations in response to the corresponding extreme rapid biological mutation, of the HIV.

1.4 The significance of the study

The main goal is to shed more light on the possibilities of integrating and managing the most influential $CD4^+$ count clinical covariates in maintaining the health status of HIV infected patients taking into consideration the infection phase specific relationships. With the explosion of the previously suggested clinical covariates, the identification and understanding of the influential patterns of the few and strong covariates of the $CD4^+$ cell count improves on the resource optimisation in the HIV/AIDS prospective studies. That is, managing fewer covariates becomes cost effective. Further, the study provided benefical insights on the harnessing of the covariate effects as having the potential to optimise the catch-22 challenge arising from the HIV disease and treatment dynamism as well as a component of patient tailored medical care.

1.5 Contributions of the study

Although the variable selection procedures reduce the number of explanatory variables considerably, what can be considered to be the reduced and the most important set of covariates has been found to still be a relatively high-dimensional curse choking further modelling techniques. Hence, this study contributed to the:

i. Improved literature on the model formulation for high-dimensional covariates with a grouping variable, particularly for longitudinal multilevel models and segmented models for multiple covariates with multiple breakpoints.
ii. Techniques for handling excessive output results for presentation and meaningful interpretation. High-dimensional covariates often produce tabular results that are extremely too lengthy to present formally for a meaningful interpretation. In such cases, we proposed a shift towards the use of macros for an automated presentation of multidimensional results.
iii. Use of visual analytics tools and demonstration of graphical displays in interpreting selected results meaningfully in multidimensional data exploration.
iv. The body of knowledge through the articles which are published in reputable scientific journals.

1.6 Thesis layout

The thesis is organised in eight chapters. Chapter 1 gives the introduction and objective of the study. Chapter 2 provides a description of the data and exploratory data analysis that informs or provides evidence of the suitable investigative statistical techniques for Chapters 3 to 7.

Chapter 3 contains variable reduction as an attempt to filter out the unimportant clinical covariates in preparation for further modelling techniques. Chapter 4 is based on multilevel models to understand the $CD4^+$ count variations between patients in response to the covariates. In Chapters 5, generalised additive mixed models (GAMM) focused on random smooths to model the $CD4^+$ count trends and adaptive optimal set-points of the clinical covariates at each phase of the HIV disease progression. Chapter 6 (segmented regression models) we looked at detecting the existence of covariate subintervals within which they are significantly and linearly associated with the $CD4^+$ cell count. Chapter 7 contains the application of structural equation models (SEM) to provide an overview for a more insightful understanding of the highly complex and intertwined relationships between the covariates, as well as the time lagged effects influencing the $CD4^+$ cell influence during the HIV disease progression. Lastly, Chapter 8 provides discussion and conclusions.

CHAPTER 2 DATA AND EXPLORATORY DATA ANALYSIS

2.1 The study design

This study initially enrolled HIV negative patients in different studies at the Centre for the AIDS Programme of Research in South Africa (CAPRISA). The main study was a prospective cohort study (the CAPRISA 002), aimed at documenting acute infection with an extensive follow up to determine the natural history of the HIV-1 subtype C infection. The establishment of the CAPRISA 002 acute infection study was between August 2004 and May 2005 (van Loggerenberg et al. 2008). It was conducted at the Doris Duke Medical Institute (DDMRI) situated at the Nelson R. Mandela School of Medicine of the University of Kwa-Zulu Natal in Durban, South Africa.

Participant recruitment: Community liaison persons assisted in the recruitment and retention efforts for the study, by methods of word of mouth and site visits. The recruitment sites were linked to transport systems into the city, to enable participants to have access to public transport. Participants were reimbursed for time, effort and transport expenses. The selection criteria for women in the community included identifying women at the greatest risk of HIV, that is those older than 18 years, self-identified as female sex workers and those that had at least three sexual partners in the 3 months preceding the recruitment.

Cohort screening and seroconverts: The women initially received voluntary counselling and testing. A urine test was conducted for pregnancy and a blood sample was collected for rapid HIV antibody testing. Seroconverts from the CAPRISA 002 were enrolled a new cohort of HIV positive patients only. Those who further seroconverted from any other CAPRISA study were also enrolled in this cohort. A schematic diagram of the screening and enrolment procedure is shown in Figure 2.1 where eventually, 237 seroconverts were recorded.

Ethical consideration: The local ethics committees reviewed the informed consent documents that were also translated into isiZulu (van Loggerenberg et al. 2008), the local language of the region. These ethics committees include members from the University of KwaZulu-Natal, the Prevention Sciences Review Committee of the Division of AIDS (DAIDS, National Institutes of Health, U.S.A.), the University of Cape Town and the University of the Witwatersrand in Johannesburg.

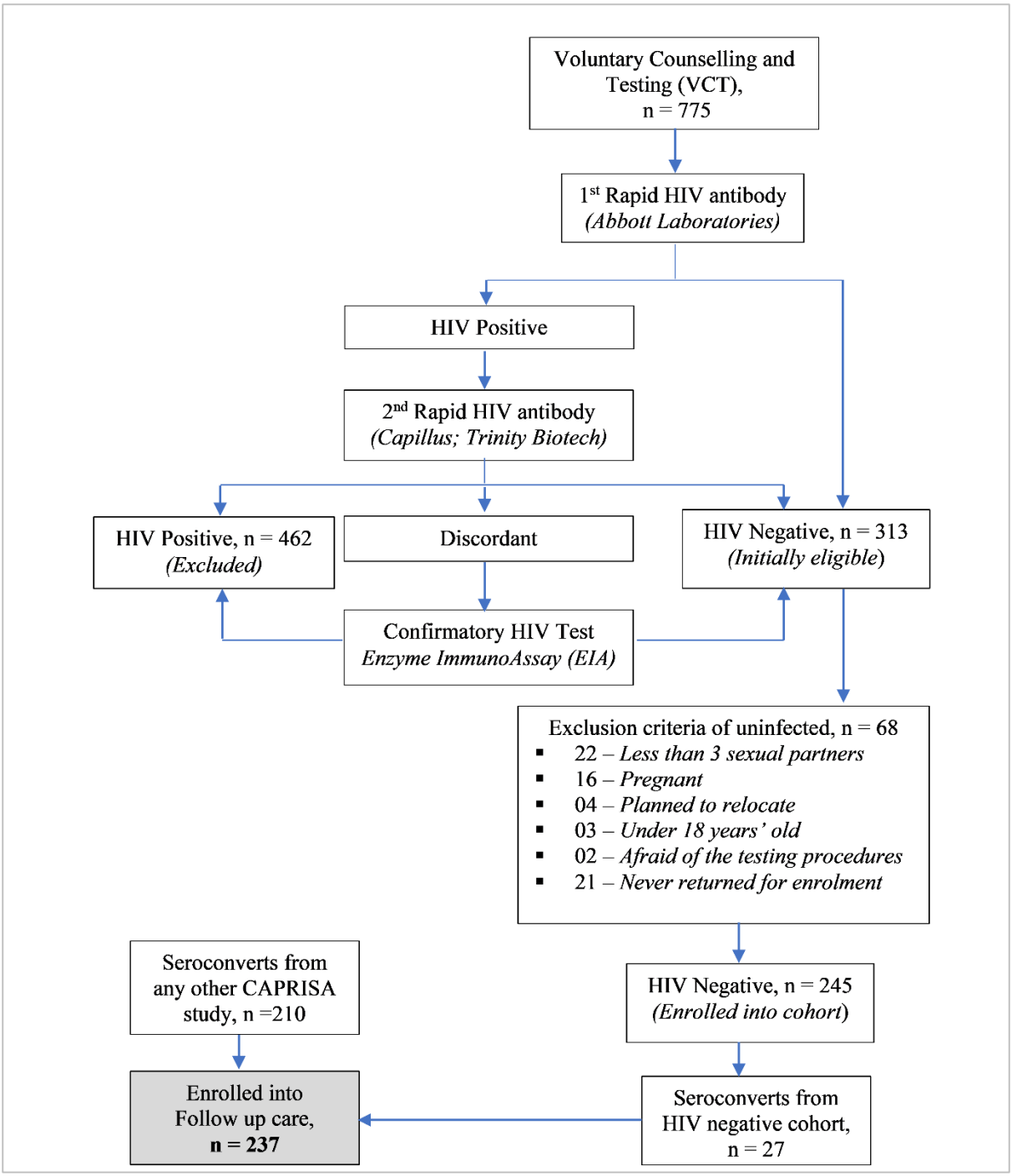


Figure 2.1: The study design.
The CAPRISA 002 HIV negative cohort screening involved 775 voluntary candidates, who could potentially join the study, of which 462 were already HIV positive and 313were therefore initially eligible to join the study. Of the 313 HIV negative candidates, only 245 were enrolled in the study while the rest were excluded for various reasons according to the eligibility criteria. At administrative censoring 27 out of the 245 seroconverted candidates were enrolled in the follow up care. Seroconvertants from other CAPRISA studies (210) were also included into the follow up care, resulting in a total of 237 patients being considered for this study.

The CAPRISA study considered five phases of the HIV/AIDS disease progression and herein briefly explained.

1-Pre-HIV: This is when the patient is not infected and any HIV test will give negative results. Although some other factors may influence the blood chemistry, complete blood count or other measures at this stage, as health indicators, are usually within normal ranges.

2-Acute infection: This was taken as follow up time within 0 - 3 months after infection, where most of the people develop flu-like symptoms that include fever, sore throat, headache, swollen glands, muscle and joint aches. It is the period (acute retroviral syndrome), when large amounts of virus are being produced, with the body naturally responding to the HIV infection (Bellan et al. 2015, Manoto et al. 2018). The virus destroys the $CD4^+$ cells and uses them to replicate itself (Weston and Marett 2009). This lowers the $CD4^+$ cells rapidly, but the immune system will begin to respond by bringing down the virus to a relatively stable level, called the viral set point. This allows the $CD4^+$ count to increase, although it may not be up to the pre-infection levels.

3-Early infection: The time period during which the patients were followed up between 3 to-12 months' post infection.

4-Established infection: The duration of this stage was not fixed, as it depended on the patient's $CD4^+$ count level. It started from 12 months' post infection and ended once the patient has initiated antiretroviral therapy (ART).

5-ART: The patient was on antiretroviral therapy- in this study, it was initiated when the $CD4^+$ count was below 500 cells/mm^3.

The period during which there are no symptoms but the virus is living or developing in the body is referred to as clinical latency. AIDS is when the immune system is badly damaged, the body is vulnerable to opportunistic infections and the $CD4^+$ cell count falls below 200 cells/mm^3 of blood (Wingfield and Wilkins 2010).

2.2 The data

The data were entered onto hard copy case report forms at the study sites and converted into electronic versions by faxing them to the CAPRISA Data Management Centre using a Data Fax system. Although the data management centre at CAPRISA studies recorded many other variables, Table 2.1 shows a summary of the electronic variables that were extracted based on the availability of repeated observations from all 237 seroconvertants, that were in line with this study's objectives.

Table 2.1: An overview of the studied $CD4^+$ count clinical covariates

RESPONSE		BLOOD CHEMISTRY		CLINICAL EXAMINATION	
	$CD4^+$ count	Liver function	ALT(GPT)	Physical	Blood Pressure(systolic)
FULL BLOOD COUNT			AST(GOT)		Blood Pressure (diastolic)
Red blood cells	Red blood cells		Bilirubin total		Pulse
	Haemoglobin (Hb)		ALP		Axillary temperature
	Haematocrit		GGT	Anthropometric	Waist circumference
	MCV	Electrolytes	Calcium		Hip circumference
	MCH		Chloride		Arm right circumference
	MCHC		Magnesium		Triceps skin fold
	RDW		Potassium		Height
	Platelets		Sodium		Weight
White blood cells	Leucocytes	Protein	Total protein		BMI
	Neutrophils		Albumin	**LIPIDS**	
	Lymphocytes		LDH		Cholesterol
	Monocytes	Iron & Vitamins	Fe (Iron)		LDL
	Eosinophils		Folate		Triglycerides
	Basophils		Vitamin B12	**SUGAR**	
			Urea		Glucose

Abbreviations: *mean corpuscular volume (MCV)); mean corpuscular haemoglobin (MCH); mean corpuscular haemoglobin concentration (MCHC); red blood cell distribution width (RDW); Alanine Aminotransferase (Glutamate Pyruvate Transaminase) (ALT(GPT)); Aspartate Aminotransferase (Glutamate Oxaloacetate Transaminase) (AST(GOT)); Alkaline phosphatase (ALP); γ-Glutamyl transferase (GGT); Lactate dehydrogenase (LDH); Body Mass Index (BMI); low density lipoproteins (LDL)*

Due to the scarcity of repeated measurements during the Phase 1 (HIV negative), this thesis is based on the data from Phases 2 to 5. Of primary interest were the last four repeated measurements prior to each of the transition phases. Table 2.2 gives a summary of the studied number of repeated measurements, per individual.

The variables are mostly functional and risk indicators of the heart, kidney or liver. They also give information about parasite invasion, as well as tissue damage in some parts of the body. The normal reference ranges of the measured variables depend on the laboratory and what is considered as "normal", varies with an individual and race (Patients Against Lymphoma 2004). The active infections can further affect the test results, while some other factors that can affect test results include medication taken during the testing period, stress, age and gender (Project Inform 2007). The descriptions summarise the variable effect on the organ functioning, in response to the general deviation from normal, without indicating the specific reference ranges.

Table 2.2: The studied number of repeated measurements per individual

	Phase:	2-Acute				3-Early				4-Est				5-ART			
	Time:	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n
ID	Variable																
01	CD4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
01	c01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
01	c02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
01	C46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
02	CD4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
02	c01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
02	c02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
02	c46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
237	CD4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
237	c01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
237	c02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
237	c46[a]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

[a] *c46 = 46th covariate*

Abbreviations: *Est.-Established; T-Time; c-covariate*

CD4$^+$ count: There are two types of lymphocytes, B or T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012) of which the CD4$^+$ count is a T cell type (Papagno et al. 2004). Around two thirds of the blood's T cells are "helper" cells expressing the surface marker CD4$^+$ (Howard and Hamilton 2002) for fighting against foreign invasion (David and Jordan 2008, NAM 2012).

Full blood count (FBC): All blood components are suspended in plasma liquid. FBC indicates whether a patient is anaemic (low red blood cells), presence of evidence of infection (using white blood cells) or is used to determine the blood flow disturbance (using of platelets).

Red blood cells: Red blood cells are packed with haemoglobin and associated with anaemia (Junqueira et al. 2006). Haemoglobin binds oxygen for transport to tissues and binds tissue carbon dioxide for transporting it back for exhalation (Jensen et al. 1998, Wintrobe and Greer 2009). Haematocrits is the percentage of the total red blood cell volume out of the whole blood volume (Wintrobe and Greer 2009). The CD4$^+$ count is: significantly associated with altered haemoglobin and haematocrit (Vanisri and Vadiraja 2016b); where it is significantly and positively correlated with blood haemoglobin level (Obirikorang and Yeboah 2009) and

haematocrit (Vanisri and Vadiraja 2016a). Hence $CD4^+$ count is significantly predicted by haemoglobin (Nzou et al. 2010). However, some studies found no association between $CD4^+$ count and haematocrits and suggest it to be an unsuitable $CD4^+$ surrogate (Alavi et al. 2009, Emuchay et al. 2014). The average amount of haemoglobin inside a single red blood cell (MCH) (Chhabra 2013) and the average concentration of haemoglobin in a given volume of packed red blood cells (MCHC) (Zheng et al. 2015), reflects diminished oxygenation of tissues (Arika et al. 2016) and anaemia (Project Inform 2007), if their results are low. They are further significantly and positively correlated with $CD4^+$ count (Vanisri and Vadiraja 2016a). Lower haematocrits and MCH are associated with lower $CD4^+$ count, that is below 350 cells/mm^3 (Lumbanraja and Siregar 2018). Mean Corpuscular Volume (MCV) is the average size or volume of individual red blood cells. Deviation (smaller or larger) from the normal range of the red blood cells may indicate anaemia. Red blood cell distribution width (RDW) reflects the degree of variation in erythrocyte size and volume and further classify anaemia (Lippi and Plebani 2014) but similar to red blood cell count, it is insignificantly and negatively associated with the $CD4^+$ count (Vanisri and Vadiraja 2016a). RDW is a marker of cardiovascular risk disease in HIV infection (Al-Kindi et al. 2017), although some studies confirm it to be AIDS related, but begged to differ on its association with the cardiovascular risk (Gallego et al. 2012).

Platelets: Platelets measure the blood clotting condition (Project Inform 2007, NAM 2012, National Institutes of Health Clinical Center 2015, James 2017). Thrombocytopenia (low platelet) occurs more frequently in HIV-infected people, than HIV-negative ones (Omoregie et al. 2009), primarily in the case of those with AIDS, with low $CD4^+$ cell numbers (< 200 cells/mm^3 (De Santis et al. 2011)), and in advanced stages of diseases (Sloand et al. 1992). There is a positive correlation between platelet count and $CD4^+$ count (Leticia et al. 2014) and therefore platelet count is considered a suitable surrogate for monitoring disease progression (De Santis et al. 2011) although elsewhere this relationship was found not to exist (Caso et al. 1999).

White blood cells: Neutrophils fight against bacterial infection (Project Inform 2007, Kimberger and Quast 2013) and an increase of these cells in the blood, is a sign of bacterial infection (Arika et al. 2016). They express endogenous $CD4^+$ and bind HIV (Biswas et al. 2003). The mean white blood cell count was found to be lower in the group of patients with low $CD4^+$ cell counts (De Santis et al. 2011). Lymphocytes: exist in two types, the B or T cells

(Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012) of which the $CD4^+$ count is a type of T cell (Papagno et al. 2004). Around two-thirds of the blood's T cells are "helper" cells expressing the surface marker $CD4^+$ (Howard and Hamilton 2002) for fighting against foreign invasion (David and Jordan 2008, NAM 2012). Some studies found no (Emuchay et al. 2014) or poor (Sen. et al. 2011) association between lymphocytes and $CD4^+$. However, there is enough evidence to suggest that a strong correlation between $CD4^+$ count and total lymphocyte count exists, which is motivation to consequently use lymphocyte count as a surrogate marker for $CD4^+$ count in resource limited settings (Shapiro et al. 1998, Daka and Loha 2008, Alavi et al. 2009, Obirikorang et al. 2012, Abdollahi et al. 2014, Fasakin et al. 2014, Atere et al. 2016, Kwantwi et al. 2017). Monocytes are germ eating cells for fighting of pathogens (Project Inform 2007). Together with $CD4^+$ cells, the monocytes are also infected by HIV (Pasupathi et al. 2008). Eosinophils provide defence against parasites, and their proliferation is induced by HIV infection (Cohen and Steigbigel 1996). They were found to be associated with low nadir $CD4^+$ count (Sivaram et al. 2012) and found to be negatively correlated with $CD4^+$ count (Cohen and Steigbigel 1996), although other studies could not find significant correlations between eosinophil count and $CD4^+$ T cell count (Chorba et al. 2002). Basophils are the least abundant leucocytes (Min et al. 2011) and; control damage to body tissues and inflammation (Project Inform 2007). The direct contact between human basophils and $CD4^+$ T cells can mediate viral trans-infection of T cells through the formation of viral synapses (Jiang et al. 2015, Marone et al. 2016).

Lipids: Measure the risk factors to cardiovascular disease using a common test called lipid profile (The Johns Hopkins Lupus Center 2017). Higher $CD4^+$ lymphocyte counts were found to be associated with higher lipid levels (Floris-Moore et al. 2006). Total cholesterol and other lipids decrease with HIV infection (Wang et al. 2016). Triglycerides, the main body fat storage (NAM 2012), increases the risk of heart disease (The Johns Hopkins Lupus Center 2017), and are affected by malnutrition and were found to be a side effect of anti-HIV drugs (Project Inform 2007). $CD4^+$ lymphocyte count was found to be positively correlated with both high density lipoproteins (HDL) and low density lipoproteins (LDL) (Floris-Moore et al. 2006, Oka et al. 2012) and further inversely related to triglycerides and very low LDL (Iffen et al. 2010). Higher LDL and triglycerides, with lower HDL, were observed among the HIV positive subjects having $CD4^+$ cell counts below 200 cells/mm^3, as compared to the HIV-control group (Iffen et al. 2010).

Sugar: Measured as blood glucose, which is a simple sugar that is obtained from the food consumed and provides energy to the body cells, which cannot function without the glucose (The Johns Hopkins Lupus Center 2017). The hormones insulin and glucagon regulate the sugar levels by decreasing or increasing it, respectively. Glucose was found to be inversely correlated with $CD4^+$ count (McKnight et al. 2014). Also, glucose metabolism disorders are significantly associated with higher $CD4^+$ T-cell counts (Maganga et al. 2015) and this was common in HIV infected patients (McKnight et al. 2014). The relationship between glucose tolerance and $CD4^+$ count was confounded by ethnicity for patients on highly active antiretroviral therapy (Misra et al. 2013).

Blood chemistry: A blood chemistry test measures chemical substances released from body tissues, or are produced during the breakdown (metabolism) of certain substances (Canadian Cancer Society 2017). This helps in determining how well the different organs are functioning, by examining the levels of different elements and waste products in the blood (U.S. Department of Veterans Affairs 2016). An abnormal amount of a substance in the blood can be a sign of disease, or be a side effect of a treatment.

Liver function: Enhanced deterioration of kidney function was found to be associated with a depletion of $CD4^+$ count (Opiyo et al. 2013, Opiyo et al. 2015). Alanine Aminotransferase (Glutamate Pyruvate Transaminase) ALT(GPT) and Aspartate Aminotransferase (Glutamate Oxaloacetate Transaminase) AST(GOT): are the liver enzymes used as indices of HIV disease progression (Pasupathi et al. 2008) and are an indication of the liver damage (Shiferaw et al. 2016). A $CD4^+$ count below 200 cells/mm^3 was found to be associated with elevated ALT and AST (Shiferaw et al. 2016) and were found to be prevalent in HIV positive women (Dusingize et al. 2015). Bilirubin total: is a waste product of the red blood cell breakdown in the liver (Project Inform 2007, The Johns Hopkins Lupus Center 2017). Normally this should be negative in tests such as urinalysis (Burns et al. 2011), though it may increase due to the shortened life span of red blood cells when old ones accumulate at a faster rate than eliminated (The Johns Hopkins Lupus Center 2017). Alkaline phosphatase (ALP) and Gamma Glutamyl transferase (GGT) are liver enzymes used to detect liver health (Whitfield 2001, Beare et al. 2008, Patil et al. 2013). Alkaline phosphatase (ALP) was found to correlate well with the $CD4^+$ count, but enumerate slightly higher counts than the common flow cytometry (Gomo et al. 2001), while in rare cases, it can be affected by drugs to reach abnormally high levels although

this improves the $CD4^+$ count (Fleischbeina et al. 2008). A baseline assessment once showed that the risk of elevated *GGT* was not associated with $CD4^+$ cell count (Bani-Sadr et al. 2009).

Electrolytes: Electrolytes are restored by highly active retroviral therapy in HIV/AIDS patients (Ugwuja and Eze 2007). Sodium and calcium: regulates water balance, blood pressure, blood volume, heart rhythm, and most importantly, the brain and nerve function (Project Inform 2007, James 2017, The Johns Hopkins Lupus Center 2017). Patients with abnormally low sodium levels have been found to have lower $CD4^+$ count and also to have a higher prevalence of AIDS (Braconnier et al. 2017). Serum sodium aberration was observed more frequently in seropositive than in seronegative individuals (Opiyo et al. 2015). Calcium is essential in the blood, muscle contraction, oocyte activation, building strong bones and teeth (Pravina et al. 2013). A change in calcium homeostasis in human $CD4^+$ T lymphocytes has been observed (Khaidukov and Litvinov 2005). Chloride is a measure of pH and fluid balance (Project Inform 2007, The Johns Hopkins Lupus Center 2017) where gastrointestinal problems such as vomiting and diarrhoea can be caused if the levels of chloride in the body are low (James 2017). Magnesium is involved in muscle contractions and protein processing (Project Inform 2007). If the levels of magnesium are too low, then it is known to be associated with a number of chronic diseases (Gröber et al. 2015). Potassium regulates the acid-base chemistry and water balance (The Johns Hopkins Lupus Center 2017), such as nerve impulses and heart muscle (Project Inform 2007, James 2017). Acute infection of $CD4^+$ lymphoid cells by HIV induces an increase in the intracellular concentration of potassium (Voss et al. 1996, Choi et al. 1998).

Protein: The total protein reading is an indication of the level of in the maintenance of normal water distribution between tissues and blood, as well as the acid-base balance (Spectrum 2007). Serum proteins increase with HIV progression (Jemikalajah and Adu 2015) and in response to highly active antiretroviral therapy, are used for HIV/AIDS patients (Ugwuja and Eze 2007). Albumin prevents fluid from leaking out of the body to allow tissue nourishment (James 2017), and its fractions were significantly associated with HIV status, among HIV positive patients, with HIV disease and ART (Sarro et al. 2010, Serpa et al. 2010). A direct relationship between albumin and $CD4^+$ count was observed (dos Santos and Almeida 2013, Pralhadrao et al. 2016), and further considered a predictor of the $CD4^+$ count of less than 200 cells/mm^3, with moderate accuracy (Moolla et al. 2015), so that this is a potential surrogate for $CD4^+$ count in limited resource settings (Olawumi and Olatunji 2006). However, some studies could not find an albumin association with $CD4^+$ count but still concluded that serum concentration can identify

adults initiating ART (Sudfeld et al. 2013). Lactate dehydrogenase (LDH) is a cytosolic enzyme, for enabling the fulfilment of the short-term energy requirements in the absence of sufficient oxygen, at the expense of greater consumption of glucose cells (Valvona et al. 2016). It has been reported that an inverse relationship between $CD4^+$ cell count and LDH level exists (Butt et al. 2002, Butt et al. 2002).

Iron and Vitamins: Fe is a key component of haemoglobin in red blood cells and is essential for the uptake of oxygen and its delivery to tissues (Kelly et al. 2017). Haemoglobin decreases as the HIV disease progresses, and is positively correlated with $CD4^+$ count (Obirikorang and Yeboah 2009). It indicates an increased risk of anaemia for patients with $CD4^+$ count under 200 cells/mm^3 (Volberding et al. 2004). Anaemia was also independently associated with higher mortality (De Santis et al. 2011). Vitamin B12: At a low level, results in megaloblastic anaemia (National Institutes of Health 2016). A low prevalence of Vitamin B12 has been reported in HIV patients, as compared to Vitamin D (Bruno et al. 2017). Studies have shown a higher deficiency of this vitamin amongst HIV patients with TB (Adhikari et al. 2016), with this being common in chronic diseases and is further correlated with mortality (Adhikari et al. 2016). At sub-optimal, Vitamin B12 is further associated with higher $CD4^+$ decline (Semeere et al. 2012). Folate is a B-vitamin that is naturally present in many foods and is needed for cell growth and metabolism (Arya and Kumar 2012, Dieticians of Canada 2014). It improves the $CD4^+$ count but was found to not be associated with HIV progression to AIDS (Adhikari et al. 2016).

Urea is an excretory product from the blood, due to protein catabolism (James 2017). Excess urea might be interpreted as a sign of dehydration, kidney dysfunctional (Project Inform 2007) or other conditions, such as protein diet, hyperbolic conditions, starvation or hepatic injury (NIOS 2012). However, studies on blood urea and $CD4^+$ count association, are currently hard to come by in literature.

Clinical examination: Includes physical examination, the process of evaluating the patient's body anatomic findings (Burns et al. 2011). These measurements show vital signs (Campbell JR and Lynn 1990) that are assessed at an early stage of the physical examination. Morphological measurements provide information on the changes to the body structure in relation to the accumulation of fat in different parts of the body.

Physical: Blood pressure (systolic) is the heart muscle contraction measured by the force of blood on the walls of the arteries (Heart Foundation 2016). A baseline study once showed a negative correlation existing between the systolic blood pressure and $CD4^+$ cell count (Palacios et al. 2006) and further associated this with mortality among patients whose HIV disease was not advanced (Bloomfield et al. 2014). Blood pressure (diastolic) is the left ventricular ejection fraction higher values were found to be independently associated with lower nadir $CD4^+$ T cell count (Hsue et al. 2010). Some other studies found no independent relationship between $CD4^+$ count and hypertension but after adjusting for body mass index (BMI), patients with a $CD4^+$ count of at least 350 cells/mm^3, were more likely to have hypertension (Dimala et al. 2018). The highest proportion of hypertensive patients had both nadir $CD4^+$ cell count below 50 cells/mm^3 and a prolonged ART duration (Manner et al. 2013). Pulse(bpm) is the rhythmic contraction and expansion of the artery due to the heart pumping the blood (Westat Inc. 1993). Axillary temperature change affects metabolism or causes hyperthermia (Kimberger and Quast 2013).

Morphological: Right arm circumference is an indicator of nutritional status (Silva 1999, Yallamraju et al. 2014). There is a significant, but poor correlation between $CD4^+$ cell count and right arm circumference (Venter et al. 2009). Undernourished HIV patients were found to be associated with low $CD4^+$ count, that is under 200 cells/mm^3 (Fofana 2016). Triceps skin fold indicates the body composition, mostly certain fat levels (Cyrino et al. 2003) and an increase in this measure, is associated with lower $CD4^+$ count among HIV patients (Esposito et al. 2008). Height(m) previous studies have not independently looked at height and $CD4^+$ relationships, but rather used this measure as a composite variable. The body mass index (BMI), which was found to have a significant and positive relationship with the $CD4^+$ count (Kwantwi et al. 2017), specifically for cases with $CD4^+$ below 200 cells/mm^3 (Nzou et al. 2010). BMI was also found to be a confounding factor in the relationship between $CD4^+$ count and hypertension (Dimala et al. 2018). Generally, patients with $CD4^+$ below 200 cells/mm^3, tend to have lower anthropometric measurements (Dannhauser et al. 1999). Given this known information about the covariates, we further explore their influential patterns on the $CD4^+$ count in the context of the CAPRISA study design for measurements recorded after seroconversion.

A review of the available literature had indicated that in the midst of all the available clinical covariates, an understanding of the strongest influential effects on the $CD4^+$ count has not been well documented. Further, there is scarce information on the adaptive behavioural patterns of

the covariates, the interconnections amongst the covariates and how their effects cause variation in $CD4^+$ count between patients.

2.3 Exploratory data analysis

Visual analytics enhance the analytic power of data and create visualizations to uncover relevant patterns (SAS Institute Inc 2014) that help to clearly understand the design as well as the properties of the dataset (Liu 2012). Amid the proliferation of information technology, multidimensional data are continuously being collected and hence the real world information is believed to be growing at an exponential rate (Wang et al. 2013). The basic visual exploration techniques have however become limiting. Multidimensional data visualisation for further investigation is hampered by the orientation of the dimensional space. To have an insight into the many dimensions is very important, however difficult (Fisseha and Onana 2017). The difficulty is due to the failure of human intuition about the geometry of high dimensions (Wegman 2001). To augment the multidimensional data visualisation techniques, parallel coordinate plots are considered to be one of the key features in visualising multi-dimensional data, to be easily interpretable (Wegman 2001, Long 2009, Heinrich and Weiskopf 2013).

2.3.1 Parallel coordinates plot

Given m dimensions labelled $x_1, x_2, ..., x_m$, an m-dimensional geometry with m equidistant copies of the $x\text{-}axis$ or $y\text{-}axis$, can be obtained depending on preference. The most common implementation is vertical, with the dimension axes having the same positive orientation, parallel to the $y-axis$ (Heinrich 2013). The dimensions are represented as vertical lines parallel to each other, instead of being orthogonal (see Figure 2.2).

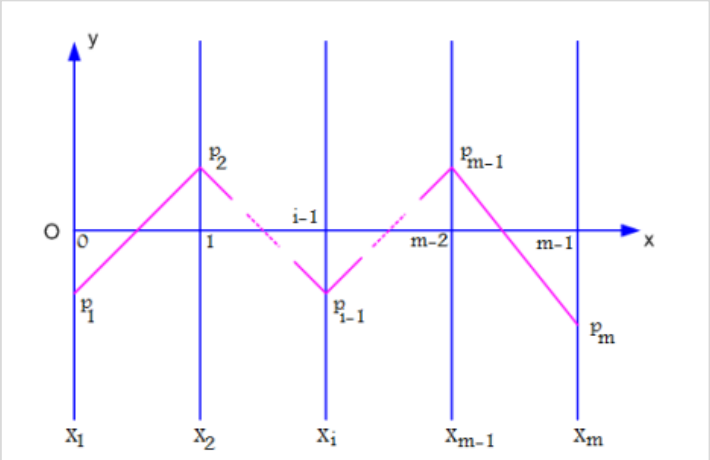


Figure 2.2: The implementation of the vertical parallel coordinates plot

Source : (Long 2009)

A single record (row) in a data frame, depicts a point $P = (p_1, p_2, ..., p_m)$ on a parallel coordinates plot with m dimensions such that a polygonal line can be drawn joining the vertices P_k, $k = 1,..., m$, at each k^{th} dimension. Basically each row in a data table is mapped as a line, or profile, and each row becomes a point on the line resembling a line chart, but the translation is very different (TIBCO-Spotfire® 2014). The values for each variable can also be normalised, to have 0% corresponding to the lowest value in a column, with the highest value in that same dimension, being set to 100%. Hence, the lower and upper points on the vertical lines may correspond to the minimum (0%) and maximum (100%) values of a dimension respectively. Scales across the different variables are separate and heights of the variables are incomparable. However, the SAS JMP software is capable of plotting each axis with its own scale (Gregg 2011), thereby allowing the visual detection of the outliers using the original values.

Parallel coordinates plots are a generalisation of the two-dimensional plot, and as such, their display allows for interpretations of statistical data analogous to two-dimensional Cartesian scatter plots. Hence, correlations between multiple variables can be revealed by connecting lines between pairs of neighbouring dimensions (Few 2006, Wang et al. 2013). Too few crossing lines is an indication of lack of correlation (Dasgupta and Kosara 2010). Also, entities with a multivariate profile are visually determined. Parallel coordinates derive their strength in modelling relations among multivariate datasets by transforming them into a 2-D pattern recognition problem (Inselberg 1997). Points in high dimensional spaces, such as three or more dimensions, can be displayed. The technique can be applied to any kind of data without much computational resources, and it is further easy to interpret, without necessarily requiring the need for high level expertise.

The display of parallel coordinate plot patterns depends on the variable scale and axes. The series of connected values form poly lines, that do not indicate change but rather indicate meaningful multivariate patterns and allow for comparisons, when used interactively for analysis (Few 2006). This allows easy investigation when interacting with data in many ways, giving rise to the presentation of different combinations of results as data trends become clearly visible (Cuzzocrea and Zall 2013). However, achieving all the possible combinations can be time consuming and inefficient, as the corresponding axes do not appear beside each other.

There is a tendency of excessive line crossing and overlapping in the parallel coordinates plot, so that reducing this visual clutter, is very important (Zhou et al. 2009). Brushing with hue and saturation, allow clusters or subsets to be isolated by colour paint settings with conditional plots or animations. Same way of using different colours, enables the analyst to track coherent clusters, through different representations, whereas animation colouring follows the data subsets through the time evolution of the animation. Dimension reordering, use of curved urges, opacity bands and edge bundles, can further be used to solve the problem of cluttering, by visually inspecting distinguishing groups.

Outliers detection

Detection of outliers and how they are treated varies, depending on the context and in some cases, common sense or investigators' experience, is needed (Van den Broeck et al. 2005). Undesirable values need attention before a detailed investigation of a dataset. All the entries in error, that are sometimes referred to as outliers, can be treated differently from the onset of data exploration, by removing them (Osborne and Overbay 2004) as single cases only, without dropping the entire records, as this leads to the loss of power and systematic deviation from true population parameters (Nakagawa 2015). Although the genuine outliers provide some insights into the underlying data structure, they still cause a threat to model accuracy, as they tend to distort distributional assumptions (Chen and Mu Liu 2015) and induce biased parameter estimates (Van den Broeck et al. 2005).

The non-error outliers though "dubious in the eyes of the analyst" (Dixon 1950) could be genuine values, simply representing extreme cases (Ghosh and Vogt 2012). Those observations that appeared to be inconsistent from the remainder of the data set (Johnson and Wicherm 2007), are potentially fruitful based on the biological context, usually had their fate delayed until the end of pattern detection, but they could possibly have been a source of discovery (Ben-Gal 2005, Johnson and Wicherm 2007, Ngatia 2012).

Outlier treatment methods include capping, where an outlier that takes the smallest value is assigned a value of the 5^{th} percentile and similarly, any outlier well above the rest of the data set, is replaced with the 95^{th} percentile value. In this case, the actual observation is unknown in as much as the one which was originally not there and both, are as good as missing. This imputation by capping procedure does not consider the missing mechanism and leads to biased results. A common way of getting rid of such extreme cases is by deletion (Van den Broeck et al. 2005, Johnson and Wicherm 2007), using the rule of thumb that any value above upper

quartile, or below the first quartile by 1.5 x Inter-quartile range, is an outlier (Moore et al. 2009, Sunitha et al. 2014). However, this generates missing values, that will require treatment.

2.3.2 Missingness patterns

For the cases of as little as 5% or more missing values, the need for some data imputation is recommended (Graham 2009). Analysis using a complete case data set by either case deletion, imputation or augmentation, would be ideal for producing parameter estimates that are close to the population true values (Nakagawa 2015). The data imputation considers the proportion of missing values which affects the quality of statistical inferences (Dong and Peng 2013) and requires the missing data mechanism that greatly influences the research results (Tabachnick and Fidell 2012). The missing data mechanism could either be missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR), (Howell 2008). The missingness mechanism is a statistical relationship between the variables observed and the probability of missing data (Nakagawa 2015). When the probability of missing data is related to some other variable(s) in the data set, such that they (not the variable with missingness), can predict the missingness, the data is MAR. On the other hand, MCAR, is when no relationship exists between missing values and the variable with missingness, or the other variables in the data set. In the case of the missingness, being predicted by the unobserved value of the variable missing, the scenario is referred to as MNAR. By understanding the missingness mechanism, the appropriate imputation method can be chosen to increase the accuracy of the results. The MCAR can be tested using the Chi-square test statistic, the most appropriate method for quantitative variables (Little 1988, Beaujean 2012). This test is more convenient than the several t-tests for each variable, which makes the simultaneous inferences difficult, due to the complex correlated structure of the corresponding t-statistics (Little 1988). There is however no statistical test and neither a visual technique, to decide whether the missingness will be MAR or MNAR because their corresponding probability distributions differ only in that MNAR, depends on unobserved values, of which in any case, there is no way to know what they were (McKnight et al. 2007, van Buuren 2012). Hence, if the missingness is confirmed not to be MNAR, the assumption of the MAR as the missingness mechanism, is considered to hold.

The imputation could be single (one complete data set), or multiple (more than one complete data sets). The latter requires the separate analysis of the data sets and then pooling the results whereas, in the former, analysis is done using the conventional procedures. Single imputation

could be any of the common methods such as mean, median, simple regression, hot and cold deck, and last observation carried forward (LOCF) or next observation carried backwards (NOCB). However, these tend to produce biased parameter estimates (McKnight et al. 2007). The most attractive and unbiased single imputation is by stochastic regression, which also uses the assumption of MAR (Gelman and Hill 2007, Enders 2010). The **R** library multivariate imputation by chained equations (**MICE)** implements the Markov chain Monte Carlo (MCMC) procedures under the MAR assumption too and allows for random selection of one of those complete data sets from multiple imputations (van Buuren and Groothuis-Oudshoorn 2011). To increase prediction, each variable has its own imputation model and the built-in models can handle continuous, binary and categorical data (van Buuren and Groothuis-Oudshoorn 2017).

2.4 The results of exploratory data analysis

The results of the overall distribution of each variable showed the existence of some extreme cases (see Figure 2.3). The $CD4^+$ count did not have any outliers from an overall perspective. Among the anthropometric covariates, outliers were common in the triceps skinfold only. The majority of the outliers were observed in the full blood count components and the liver functions tests. It was further revealed that the red blood cell group had outliers in both extremes, whereas the white blood cells and the liver function indicators tended to have extremely high values only. Of all the studied variables, a wide distribution was observed in potassium, LDH and folate measurements. The white blood cells and the liver function indicators were mostly below the 50^{th} percentile.

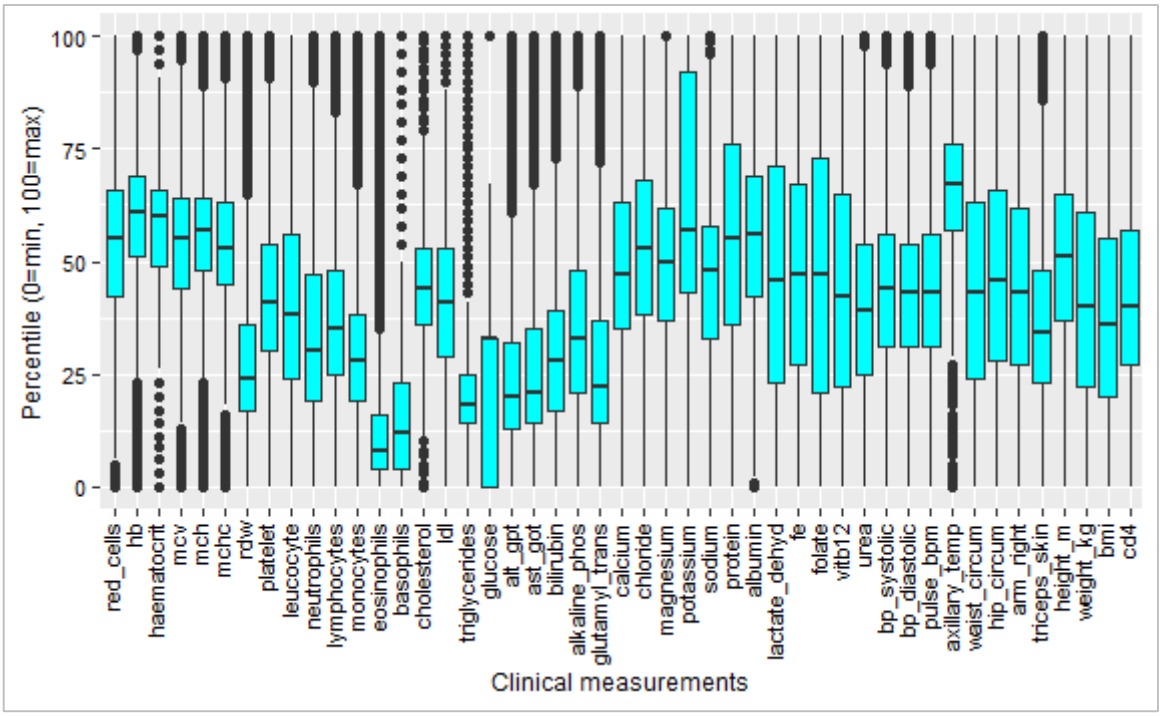


Figure 2.3: Boxplots of the overall variable distribution

A comparison of the variable distributional properties within the HIV infection phases is shown in Figure 2.4. The data showed that an overview of the variable distributions masked the outliers within the phases. As the HIV disease progressed, there was a notable shift in the distributions of some of the covariates. The distributions of clinical measurements associated with phases 2 and 3 were fairly similar. However, the clinical examination measurements of phase 4 were generally higher than those of phase 3. This was also observed in the red blood cell indices (MCHC, MCH, Hb), ALP, GGT, potassium, LDH and the $CD4^+$ count were the measurements recorded during the ART phase, were generally higher than those recorded

during the established phase. This warranted a further investigation into a possible drift in the homeostatic levels of the $CD4^+$ count clinical covariates, using the application of advanced statistical methods.

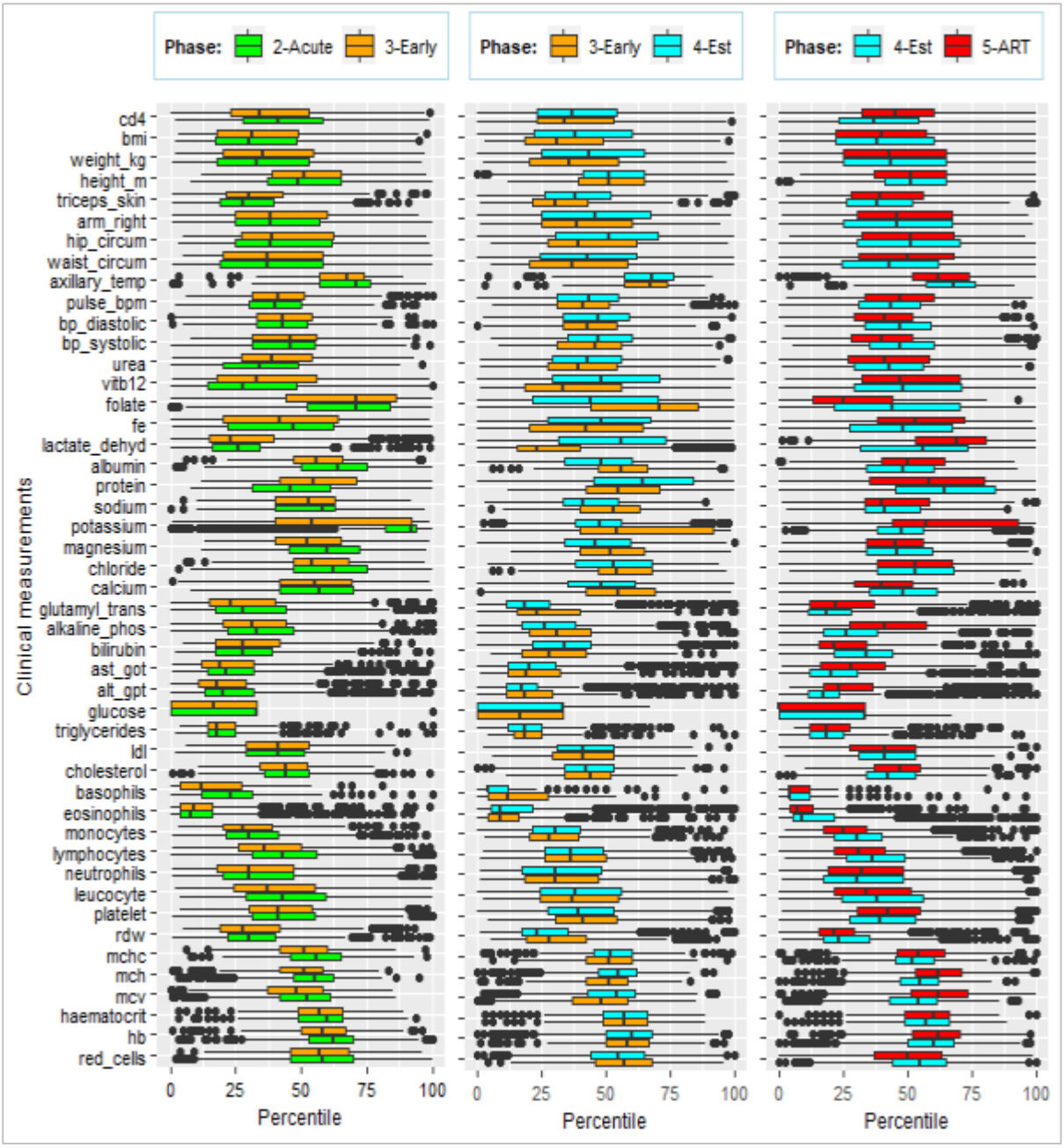


Figure 2.4: Boxplots of variable distribution within each phase

The data were also displayed using the parallel coordinates plot to augment the boxplots (see Figure 2.5).

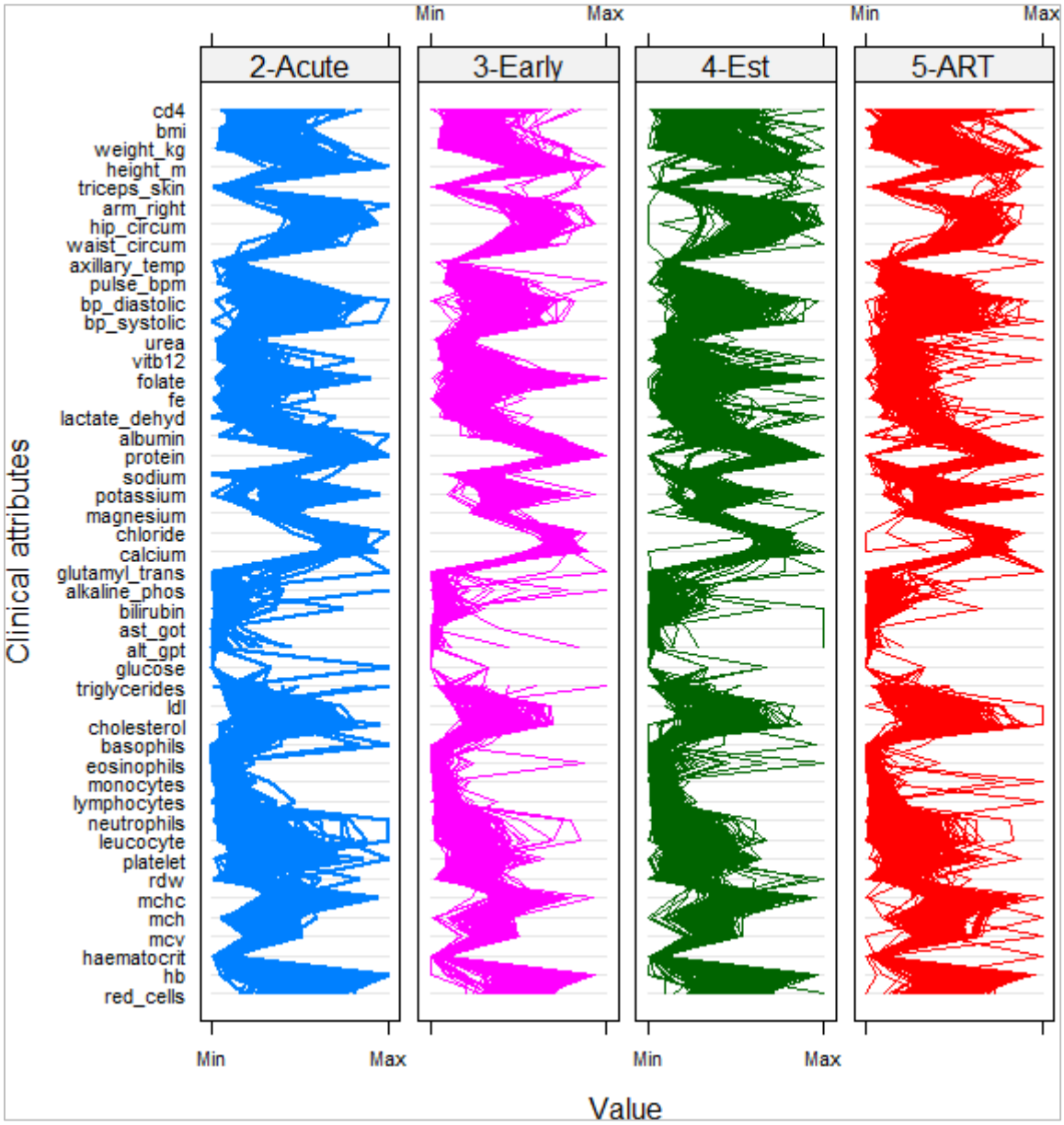


Figure 2.5: The results of parallel plot indicating outliers within each phase

The outliers can clearly be seen to disturb the distributional assumptions and the potential to affect the calculations of summary statistics, such as the means and variances. They were dropped out per infection phase using the rule of thumb that any value above the upper quartile or below the first quartile by 1.5 x Inter-quartile range, is an outlier. This further introduced some missing values. Figure 2.6 shows an overview of the cleaned data set, that could provide some realistic mean and variance values.

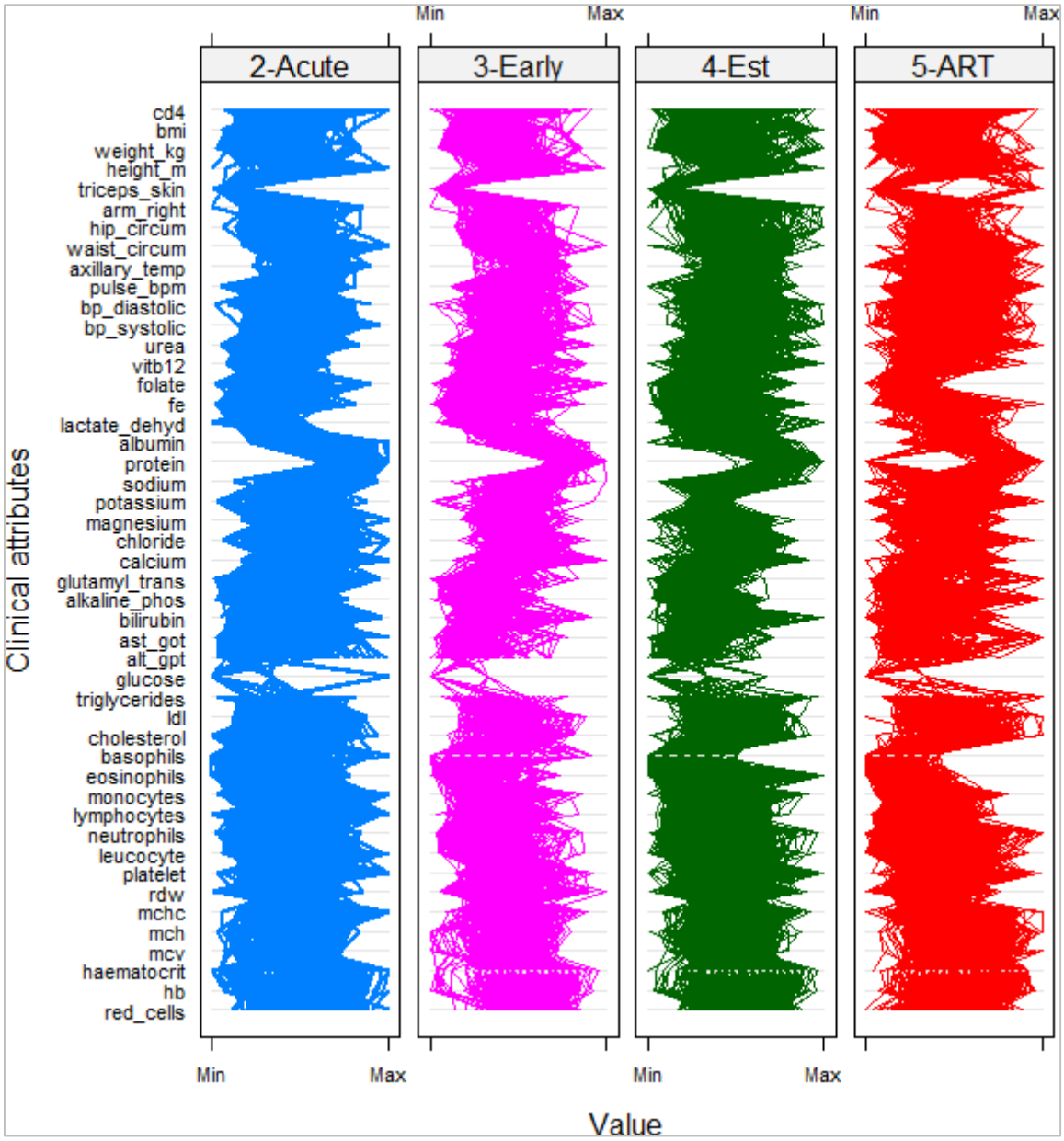


Figure 2.6: The results of parallel plot indicating the treated outliers

Both the error and genuine extreme cases were eventually treated as missing values giving rise to a final cleaned data set with a 34.49% proportion of missingness, in a 3 792x52 data matrix. The missingness test produced a statistic $\chi^2\ value = 14\ 040.997$, $DF = 7733$ and $p - value < 0.001$ for the 34.49% missing proportion, suggesting that the evidence that the study data was MCAR. The missingness pattern in Figure 2.7 clearly shows that the weakness was due to some blocks of missing values which are likely to have resulted in the none randomness.

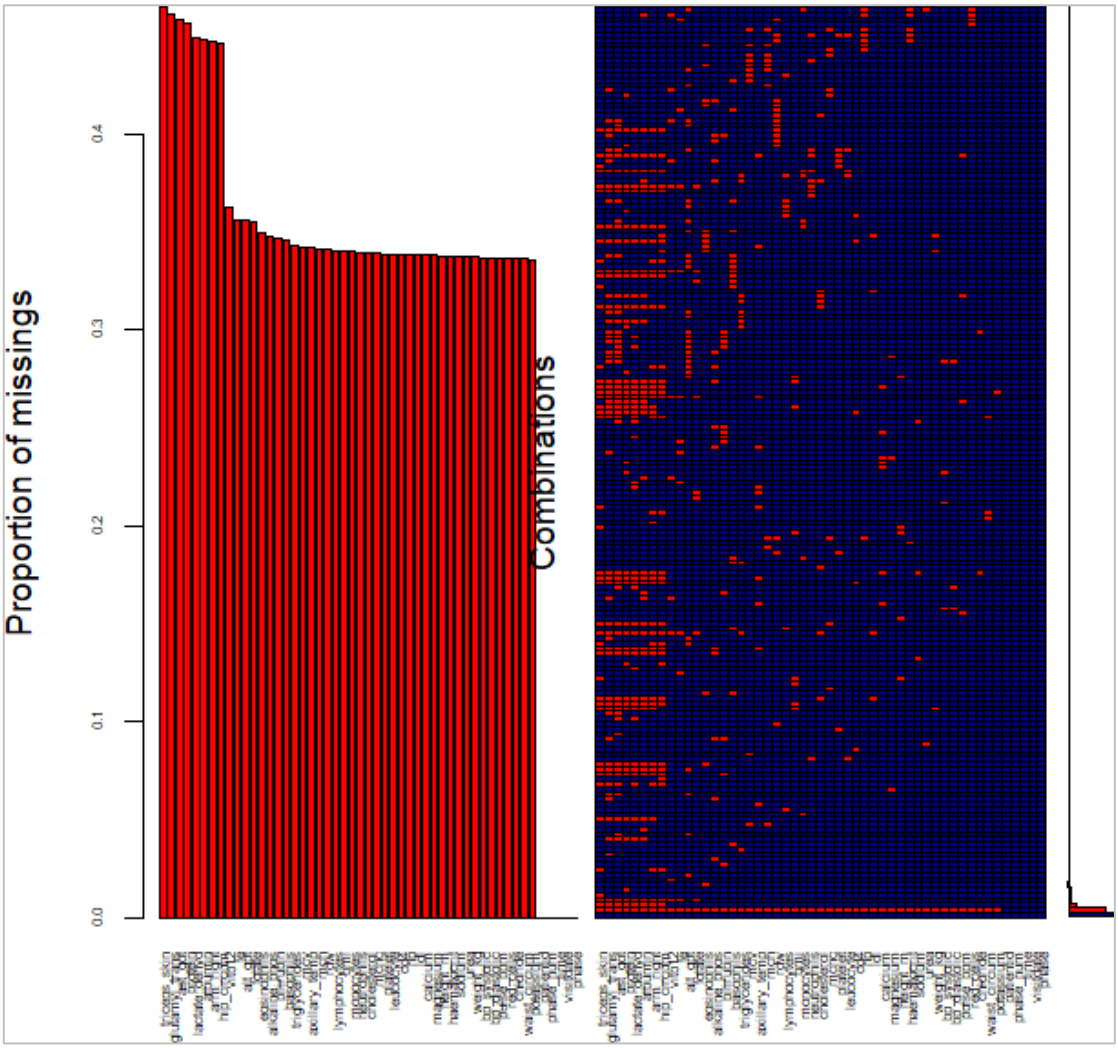


Figure 2.7: The results of missingness pattern in the data

Prior to data imputation, the last four measurements per variable from each phase were selected and multiple imputations with chained equations were considered as an appropriate approach, since it takes into account the missingness mechanism. The missing value imputation considered the multilevel structure of the data, where both random and fixed effects were incorporated. The imputation procedure produced five datasets and their distributions were compared against the observed variables (see Figure 2.8). The diagnostic results showed that the distribution of all the imputed data sets were identical (blue), and very similar to, the observed values. This further strengthens the confidence to select a single dataset to use, with more investigative techniques.

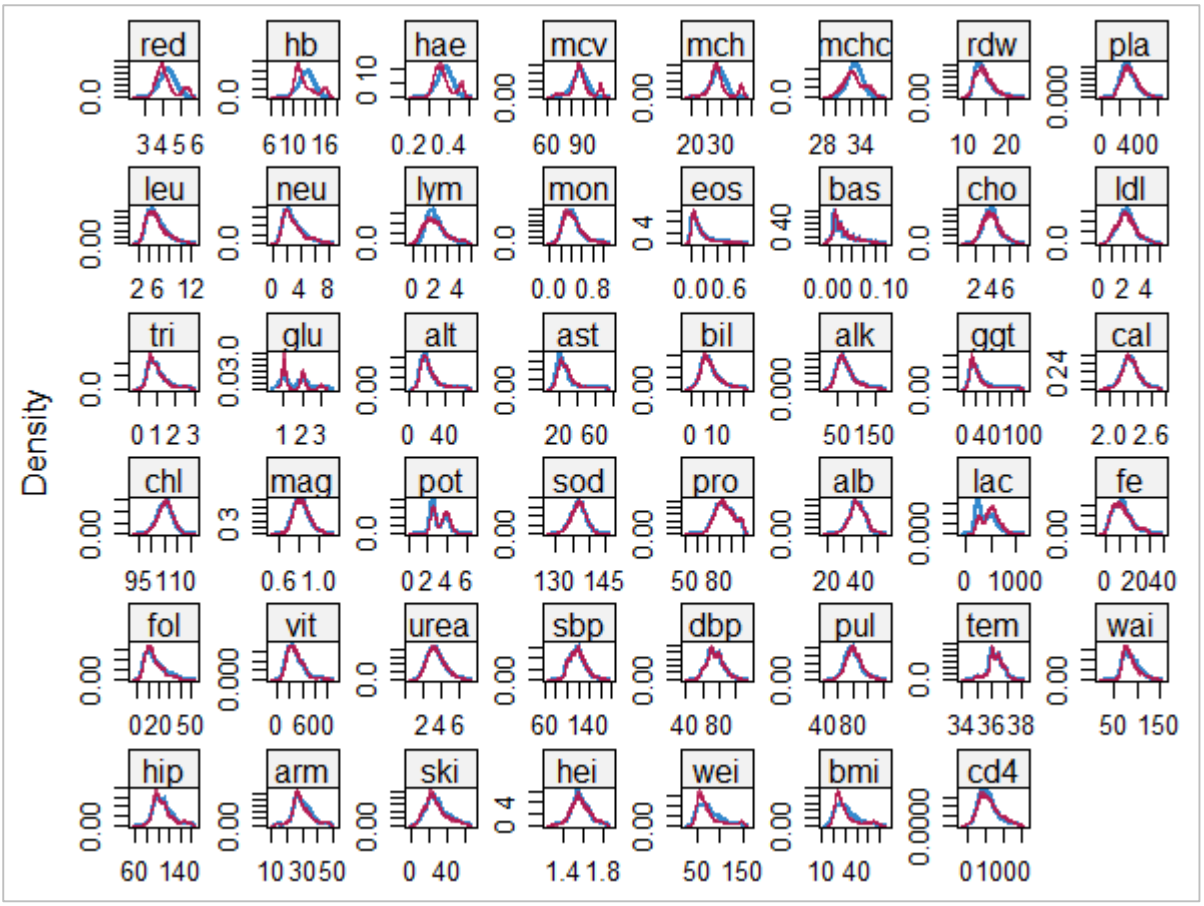


Figure 2.8: The results of data imputation diagnostics
Blue represents the observed data whilst red is the imputed data

Abbreviations**: **red**-* red blood cells; ***hae- haematocrit, ***pla***- platelet, ***lym***- lymphocytes, ***mon***-monocytes, ***bas***-basophils, ***glu***-glucose, ***alk***-alkaline phosphatase, ***cal***-calcium, ***mag***-magnesium, ***pot***-potassium, ***sod***-sodium, ***pro***-protein, ***alb***-albumin, ***lac***- lactate dehydrogenase, ***fol***-folate, ***leu***-leucocyte, ***neu***- neutrophils, ***eos***-eosinophils, ***cho***-cholesterol, ***tri***-triglycerides, ***alt***-ALT(GPT), ***ast***-AST(GOT), ***bil***-bilirubin, ***ggt***- glutamyl transferase, ***chl***-chloride, ***vit***- vitamin B12, ***sbp***-sytolic blood pressure, ***dbp***-diastolic blood pressure, ***pul***-pulse(bpm), ***tem***-axillary temperature, ***wai***-waist circumference, ***hip***-hip circumference, ***arm***-right arm circumference, ***ski***-triceps skin, ***hei***-height(m), ***wei***-weight(kg)

Table 2.3 shows the descriptive statistics of the variables within each phase as well as those from an overall point of view, that considers the entire follow-up period. During the entire follow up period, at least 50% of the $CD4^+$ cell count repeated measurements were above 540 cells/mm^3, averaging 570±240 cells/mm^3, with an overall variation of 41.8% around the cohort's average $CD4^+$ cell count. Although the average $CD4^+$ cell counts varied with the infection phase, the maximum $CD4^+$ cell count ever recorded during each phase was 1 400 cells/mm^3. The greatest variation of the covariates' repeated measurements around the mean was observed in eosinophils (at least 98%), basophils (at least 64%) and Gamma glutamyl transferase (at least 57%).

Table 2.3: The results of descriptive statistics of the data

	2-Acute (n=948)	3-Early (n=948)	4-Est (n=948)	5-ART (n=948)	Overall (n=3792)
CD4+ Count (cells/mm^3)					
Mean±SD(CV%)	600±240(39.7)	560±240(43.6)	520±230(44.1)	610±230(38.2)	570±240(41.8)
Median[Min;Max]	570(53;1400)	520(54;1400)	490(45;1400)	590(45;1400)	540(45;1400)
RBC (x10^6cells/mm^3)					
Mean±SD(CV%)	4.4±0.47(10.8)	4.3±0.50(11.8)	4.3±0.52(12.2)	4.2±0.45(10.8)	4.3±0.49(11.5)
Median[Min;Max]	4.3(2.7;5.6)	4.2(3.0;5.6)	4.2(2.7;5.6)	4.2(2.7;5.6)	4.2(2.7;5.6)
Haemoglobin (g/dL)					
Mean±SD(CV%)	12±1.5(12.3)	12±1.7(13.9)	12±1.7(13.7)	12±1.5(11.8)	12±1.6(13.0)
Median[Min;Max]	12(7.2;16)	12(7.2;16)	12(7.3;16)	12(8.0;16)	12(7.2;16)
Haematocrit (Hct/100)					
Mean±SD(CV%)	0.37±0.041(10.9)	0.37±0.043(11.8)	0.37±0.042(11.5)	0.38±0.039(10.5)	0.37±0.042(11.2)
Median[Min;Max]	0.37(0.23;0.48)	0.36(0.24;0.48)	0.37(0.25;0.47)	0.38(0.25;0.47)	0.37(0.23;0.48)
MCV (fL)					
Mean±SD(CV%)	86±6.5(7.5)	86±7.3(8.5)	87±6.7(7.7)	90±6.6(7.4)	87±7.0(8.0)
Median[Min;Max]	86(64;110)	86(64;110)	87(65;110)	90(67;110)	87(64;110)
MCH (pg/cell)					
Mean±SD(CV%)	28±2.6(9.2)	28±3.1(10.9)	29±2.9(10.0)	30±2.7(9.2)	29±2.9(10.0)
Median[Min;Max]	29(20;37)	28(19;37)	29(20;37)	30(20;37)	29(19;37)
MCHC (g/dL)					
Mean±SD(CV%)	33±1.3(3.9)	33±1.4(4.2)	33±1.4(4.1)	33±1.3(3.9)	33±1.3(4.0)
Median[Min;Max]	33(29;37)	33(29;37)	33(29;37)	33(29;37)	33(29;37)
RDW (%)					
Mean±SD(CV%)	15±1.8(11.7)	15±2.1(13.8)	15±1.8(12.5)	14±1.7(11.9)	15±1.9(12.7)
Median[Min;Max]	15(11;22)	15(11;22)	14(11;22)	14(11;22)	14(11;22)
Platelet (x10^9/L)					
Mean±SD(CV%)	300±81(27.0)	300±81(27.1)	280±83(29.2)	300±78(26.1)	300±81(27.4)
Median[Min;Max]	290(75;590)	290(43;590)	280(43;590)	300(60;590)	290(43;590)
leucocytes (x10^9/L)					
Mean±SD(CV%)	5.6±1.7(30.4)	5.3±1.7(32.3)	5.2±1.8(34.1)	5.2±1.7(33.4)	5.3±1.7(32.6)
Median[Min;Max]	5.3(2.1;12)	5.0(2.2;12)	4.8(1.9;12)	4.9(1.9;11)	5.0(1.9;12)
Neutrophils (x10^9/L)					
Mean±SD(CV%)	2.7±1.3(46.3)	2.7±1.3(47.4)	2.7±1.4(50.0)	2.9±1.4(47.1)	2.8±1.3(47.8)
Median[Min;Max]	2.4(0.59;7.8)	2.4(0.82;7.8)	2.3(0.69;7.8)	2.7(0.59;7.8)	2.4(0.59;7.8)
Lymphocyte (x10^9/L)					
Mean±SD(CV%)	2.2±0.76(35.2)	1.9±0.74(38.3)	1.8±0.72(40.0)	1.7±0.64(37.4)	1.9±0.74(38.8)
Median[Min;Max]	2.1(0.37;4.4)	1.9(0.37;4.4)	1.7(0.37;4.4)	1.6(0.37;4.4)	1.8(0.37;4.4)
Monocytes (x10^9/L)					
Mean±SD(CV%)	0.43±0.16(37.5)	0.42±0.15(34.9)	0.42±0.16(37.7)	0.39±0.15(37.4)	0.41±0.15(37.1)
Median[Min;Max]	0.41(0.10;1.0)	0.40(0.12;1.0)	0.40(0.070;1.0)	0.37(0.12;1.0)	0.39(0.070;1.0)
Eosinophils (x10^9/L)					
Mean±SD(CV%)	0.16±0.16(100.5)	0.14±0.15(100.5)	0.15±0.15(101.9)	0.13±0.13(98.6)	0.15±0.15(101.1)
Median[Min;Max]	0.10(0.0;0.80)	0.10(0.0;0.78)	0.10(0.0;0.80)	0.090(0.0;0.80)	0.10(0.0;0.80)
Basophils (x10^9/L)					
Mean±SD(CV%)	0.031±0.020(64.0)	0.025±0.019(74.0)	0.016±0.013(76.9)	0.019±0.013(68.1)	0.023±0.017(75.7)
Median[Min;Max]	0.030(0.0;0.090)	0.020(0.0;0.090)	0.010(0.0;0.090)	0.020(0.0;0.090)	0.020(0.0;0.090)
Cholesterol (mmol/L)					
Mean±SD(CV%)	3.8±0.89(23.4)	3.8±0.87(22.8)	3.7±0.83(22.2)	4.0±0.88(22.0)	3.8±0.87(22.7)
Median[Min;Max]	3.8(1.1;7.0)	3.8(1.6;6.4)	3.7(1.1;7.0)	3.9(2.0;7.0)	3.8(1.1;7.0)
LDL (mmol/L)					
Mean±SD(CV%)	2.3±0.71(30.4)	2.3±0.72(31.3)	2.3±0.67(29.2)	2.3±0.74(32.0)	2.3±0.71(30.7)
Median[Min;Max]	2.3(0.60;4.7)	2.3(0.60;4.4)	2.2(0.60;4.3)	2.3(0.60;4.8)	2.3(0.60;4.8)
Triglycerides (mmol/L)					
Mean±SD(CV%)	0.99±0.45(45.7)	0.95±0.43(45.1)	0.93±0.45(48.4)	1.0±0.49(49.0)	0.97±0.46(47.2)
Median[Min;Max]	0.90(0.10;2.8)	0.90(0.0;2.8)	0.80(0.0;2.8)	0.90(0.20;2.7)	0.90(0.0;2.8)

Continued

Table 2.3: The results of descriptive statistics of the data (continued)

	2-Acute (n=948)	3-Early (n=948)	4-Est (n=948)	5-ART (n=948)	Overall (n=3792)
CD4+ Count (cells/mm^3)					
Mean±SD(CV%)	600±240(39.7)	560±240(43.6)	520±230(44.1)	610±230(38.2)	570±240(41.8)
Median[Min;Max]	570(53;1400)	520(54;1400)	490(45;1400)	590(45;1400)	540(45;1400)
Glucose					
Mean±SD(CV%)	1.4±0.56(41.0)	1.6±0.67(41.1)	1.7±0.67(38.5)	1.7±0.67(38.9)	1.6±0.66(41.0)
Median[Min;Max]	1.0(1.0;3.0)	2.0(1.0;3.0)	2.0(1.0;3.0)	2.0(1.0;3.0)	2.0(1.0;3.0)
ALT(GPT) (U/L)					
Mean±SD(CV%)	21±9.5(45.5)	20±9.3(46.0)	21±9.5(45.7)	23±9.2(39.9)	21±9.4(44.4)
Median[Min;Max]	19(6.0;64)	18(4.0;64)	18(7.0;62)	21(9.0;63)	19(4.0;64)
AST(GOT) (U/L)					
Mean±SD(CV%)	26±9.4(35.4)	26±8.9(34.1)	28±11(37.9)	29±11(38.4)	27±10(36.9)
Median[Min;Max]	24(13;72)	24(13;72)	25(13;72)	26(13;71)	25(13;72)
Bilirubin (mmol/L)					
Mean±SD(CV%)	7.4±3.2(43.7)	7.5±3.3(44.7)	8.1±3.5(44.1)	6.4±3.1(47.8)	7.3±3.3(45.7)
Median[Min;Max]	7.0(1.0;20)	7.0(0.0;20)	7.3(0.0;20)	6.0(0.0;20)	7.0(0.0;20)
ALP (IU/L)					
Mean±SD(CV%)	69±20(29.1)	68±20(28.7)	66±20(30.3)	79±24(30.0)	71±22(30.5)
Median[Min;Max]	65(30;150)	64(28;160)	62(30;160)	75(36;160)	66(28;160)
GGT (U/L)					
Mean±SD(CV%)	25±14(57.7)	24±15(62.6)	21±14(66.6)	24±16(69.2)	24±15(64.2)
Median[Min;Max]	22(6.0;94)	20(4.0;90)	17(6.0;90)	19(6.0;92)	19(4.0;94)
Calcium (mmol/L)					
Mean±SD(CV%)	2.3±0.11(4.7)	2.3±0.11(4.7)	2.3±0.10(4.5)	2.3±0.092(4.1)	2.3±0.11(4.7)
Median[Min;Max]	2.3(2.0;2.7)	2.3(1.9;2.7)	2.3(1.9;2.7)	2.3(1.9;2.6)	2.3(1.9;2.7)
Chloride (mEq/L)					
Mean±SD(CV%)	110±2.7(2.5)	110±2.7(2.6)	100±2.9(2.7)	100±2.8(2.7)	110±2.8(2.7)
Median[Min;Max]	110(96;120)	110(95;120)	110(95;110)	100(95;110)	110(95;120)
Magnesium (mmol/L)					
Mean±SD(CV%)	0.86±0.082(9.6)	0.83±0.077(9.3)	0.81±0.076(9.5)	0.80±0.068(8.5)	0.82±0.079(9.6)
Median[Min;Max]	0.85(0.62;1.1)	0.83(0.58;1.1)	0.80(0.58;1.1)	0.80(0.58;1.1)	0.82(0.58;1.1)
Potassium (mmol/L)					
Mean±SD(CV%)	3.7±0.64(17.4)	3.2±0.83(25.9)	2.8±0.69(24.8)	3.1±0.84(27.3)	3.2±0.82(25.7)
Median[Min;Max]	3.8(0.65;6.0)	3.0(0.65;6.0)	2.6(0.86;5.8)	2.7(0.74;6.0)	2.9(0.65;6.0)
Sodium (mEq/L)					
Mean±SD(CV%)	140±2.1(1.6)	140±2.3(1.7)	140±2.4(1.8)	140±2.3(1.7)	140±2.3(1.7)
Median[Min;Max]	140(130;150)	140(130;150)	140(130;140)	140(130;150)	140(130;150)
Total protein (g/L)					
Mean±SD(CV%)	81±6.6(8.1)	83±7.3(8.8)	86±8.1(9.4)	86±8.2(9.6)	84±7.8(9.3)
Median[Min;Max]	81(63;99)	83(63;99)	86(56;99)	86(63;99)	83(56;99)
Albumin (g/L)					
Mean±SD(CV%)	39±4.2(10.7)	38±4.5(11.7)	37±4.6(12.5)	37±4.3(11.5)	38±4.5(11.9)
Median[Min;Max]	39(21;52)	38(21;52)	37(23;51)	37(25;50)	38(21;52)
LDH (U/L)					
Mean±SD(CV%)	340±140(41.2)	380±160(42.5)	500±140(28.9)	530±150(28.4)	440±170(38.7)
Median[Min;Max]	290(27;930)	320(120;1000)	510(120;1100)	540(150;1000)	440(27;1100)
Fe(Iron) (mcg/dL)					
Mean±SD(CV%)	11±5.7(52.6)	11±6.1(55.6)	12±6.2(51.0)	13±6.2(49.5)	12±6.1(52.5)
Median[Min;Max]	11(1.7;29)	10(1.0;31)	11(1.8;34)	12(1.7;34)	11(1.0;34)
Folate (nmol/L)					
Mean±SD(CV%)	19±8.8(46.7)	18±9.3(50.2)	13±7.0(52.4)	12±6.2(51.3)	16±8.5(54.0)
Median[Min;Max]	18(1.2;49)	17(1.2;49)	12(1.4;45)	11(1.2;37)	14(1.2;49)
Vitamin B12 (ng/mL)					
Mean±SD(CV%)	260±110(44.3)	280±120(43.4)	320±130(40.5)	330±130(38.6)	300±130(42.7)
Median[Min;Max]	240(72;810)	250(72;810)	300(0.0;830)	300(110;820)	270(0.0;830)
Urea (mmol/L)					
Mean±SD(CV%)	3.3±0.94(28.4)	3.4±0.93(27.4)	3.5±0.91(25.9)	3.5±1.0(29.4)	3.4±0.95(27.9)
Median[Min;Max]	3.2(1.1;6.5)	3.3(1.3;6.8)	3.4(1.4;6.8)	3.3(1.2;6.8)	3.3(1.1;6.8)

Continued

Table 2.3: The results of descriptive statistics of the data (continued)

	2-Acute (n=948)	3-Early (n=948)	4-Est (n=948)	5-ART (n=948)	Overall (n=3792)
CD4+ Count (cells/mm^3)					
Mean±SD(CV%)	600±240(39.7)	560±240(43.6)	520±230(44.1)	610±230(38.2)	570±240(41.8)
Median[Min;Max]	570(53;1400)	520(54;1400)	490(45;1400)	590(45;1400)	540(45;1400)
BP(systolic) (mmHg)					
Mean±SD(CV%)	120±14(11.5)	120±14(11.5)	120±14(11.4)	120±14(12.0)	120±14(11.6)
Median[Min;Max]	120(74;170)	120(84;170)	120(80;170)	120(79;170)	120(74;170)
BP(diastolic) (mmHg)					
Mean±SD(CV%)	75±9.4(12.5)	75±9.8(13.1)	76±9.7(12.8)	75±9.7(13.0)	75±9.7(12.9)
Median[Min;Max]	75(47;110)	75(46;110)	75(50;110)	74(46;110)	75(46;110)
Pulse (bpm)					
Mean±SD(CV%)	79±8.7(10.9)	81±9.9(12.3)	81±9.8(12.1)	83±11(13.7)	81±10(12.4)
Median[Min;Max]	79(52;110)	80(56;120)	81(48;120)	82(55;120)	80(48;120)
Ax.Temp (D.Celsius)					
Mean±SD(CV%)	36±0.44(1.2)	36±0.47(1.3)	36±0.47(1.3)	36±0.58(1.6)	36±0.49(1.4)
Median[Min;Max]	36(34;38)	36(34;38)	36(34;38)	36(34;38)	36(34;38)
Waist circum (cm)					
Mean±SD(CV%)	85±16(18.5)	86±16(18.7)	86±16(18.3)	91±17(18.5)	87±16(18.7)
Median[Min;Max]	83(31;150)	82(31;150)	84(40;150)	89(32;150)	84(31;150)
Hip circum (cm)					
Mean±SD(CV%)	110±15(13.8)	110±15(14.0)	110±14(13.2)	110±14(13.0)	110±15(13.5)
Median[Min;Max]	110(70;160)	100(79;160)	110(64;160)	110(64;160)	110(64;160)
Arm(right)circum (cm)					
Mean±SD(CV%)	30±5.2(17.7)	30±5.4(18.3)	30±5.4(18.1)	30±5.3(17.6)	30±5.4(17.9)
Median[Min;Max]	29(14;47)	28(14;47)	29(14;47)	30(14;47)	29(14;47)
Triceps skin fold (mm)					
Mean±SD(CV%)	26±10(40.8)	26±10(38.6)	28±11(37.2)	29±11(36.4)	27±11(38.6)
Median[Min;Max]	24(5.0;61)	25(6.0;61)	27(5.0;60)	28(6.0;60)	25(5.0;61)
Height (m)					
Mean±SD(CV%)	1.6±0.080(5.1)	1.6±0.080(5.0)	1.6±0.079(5.0)	1.6±0.085(5.4)	1.6±0.081(5.1)
Median[Min;Max]	1.6(1.3;1.8)	1.6(1.4;1.8)	1.6(1.4;1.8)	1.6(1.3;1.8)	1.6(1.3;1.8)
Weight (kg)					
Mean±SD(CV%)	72±21(28.4)	72±21(29.4)	74±21(28.4)	76±21(27.4)	74±21(28.5)
Median[Min;Max]	67(42;150)	67(41;150)	69(39;150)	73(42;150)	69(39;150)
BMI (kg/m^2)					
Mean±SD(CV%)	29±7.7(26.6)	29±8.2(28.5)	30±8.3(28.0)	31±8.1(26.3)	30±8.1(27.4)
Median[Min;Max]	27(18;57)	26(18;60)	28(16;61)	29(16;58)	28(16;61)

There are some covariates that indicated that the quantities remained fairly constant throughout the disease progression. For example, MCHC, calcium, chloride, magnesium, sodium, axillary temperature and height. Their coefficients of variation around the means were barely under 5%. The average measurements of ALP, LDH and Vitamin B12 tended to fluctuate depending on the infection phase. However, ALP and Vitamin B12 indicated that the variation of the measurements remained fairly the same at around the respective averages within each infection phase. The ALP revealed approx. 30% measurement spread about the mean whereas the Vitamin B12 was approx. 40%. The LDH variation was approximately the same during the acute and early phases at around 40%, narrowly varying during the established and ART phases, where the variation was around 30% around the respective phase means. Generally, most of the covariates seem to show slight variation in the phase averages, with moderate variations of the measurements around their respective phase averages. Although the

descriptive statistics provided some insight, due to the relatively high-dimensional nature, it was not easy to relate the average changes and variations with the $CD4^+$ cell count calling for pattern discovery using graphical visualisations.

An overview of how the cohort's average $CD4^+$ count changed over the follow-up period is shown in Figure 2.9. Soon after HIV infection, the cohort's average $CD4^+$ count was around 600 cells/mm^3 and generally kept on declining in the subsequent visits, during the early and up to the established phases. The cohort's average $CD4^+$ counts at each visit during the early and established phases were all below the cohort's grand average $CD4^+$ count (570 cells/mm^3) for the entire follow-up period. During the entire follow-up period, the cohort's lowest average $CD4^+$ count was recorded during the established phase and this was restored during the uptake of medication back to around 600 cells/mm^3 of average $CD4^+$ count. This indicated that infection phase was a contributing factor to the changes in the $CD4^+$ count.

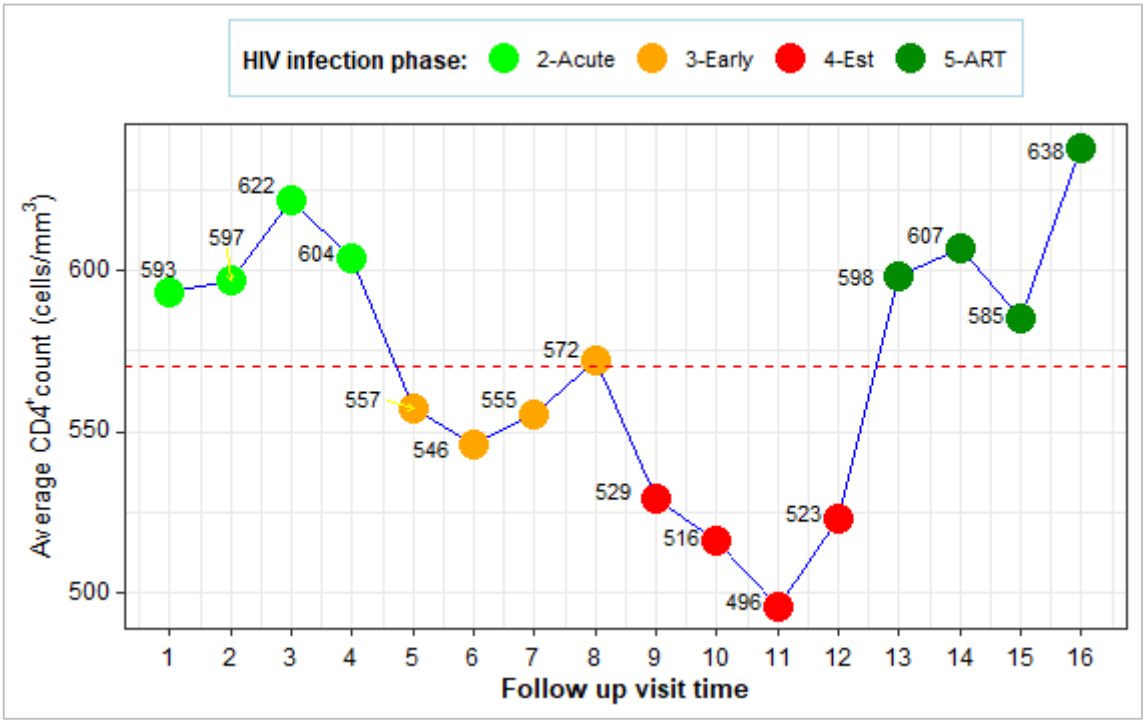


Figure 2.9: The results of the cohort's average $CD4^+$ count at each visit time
The horizontal broken red line represents the cohort's average $CD4^+$ count during the follow-up period

Although the $CD4^+$ cell count behaviour at each patient visit gave some meaningful insights on the HIV disease progression, it was not easy to have an overview of the overall relationships between all the variables due to an increased number of summary statistics. This argument arose from the fact that the patient's health status is a resultant of other body's complex systems that operate in harmony with the immune system (Klein and Zion 2015). The $CD4^+$ count is an

indicator of the immune status whilst the clinical covariates of the $CD4^+$ count are also indicators of the other body's functional systems. For example, the red blood cells play a role in the respiratory system (Jensen et al. 1998, Wintrobe and Greer 2009). As such there is need to consider the clinical covariates as facets to provide a holistic overview of the patient's health status during the HIV disease progression. The parallel coordinate plot provided an enhanced multidimensional visualisation in a consolidated display. The plot confirmed that the $CD4^+$ count mean was affected by the infection phase. The clinical covariates dynamically took different mean levels accordingly, as the mean $CD4^+$ count changed from one phase to the other (Figure 2.10).

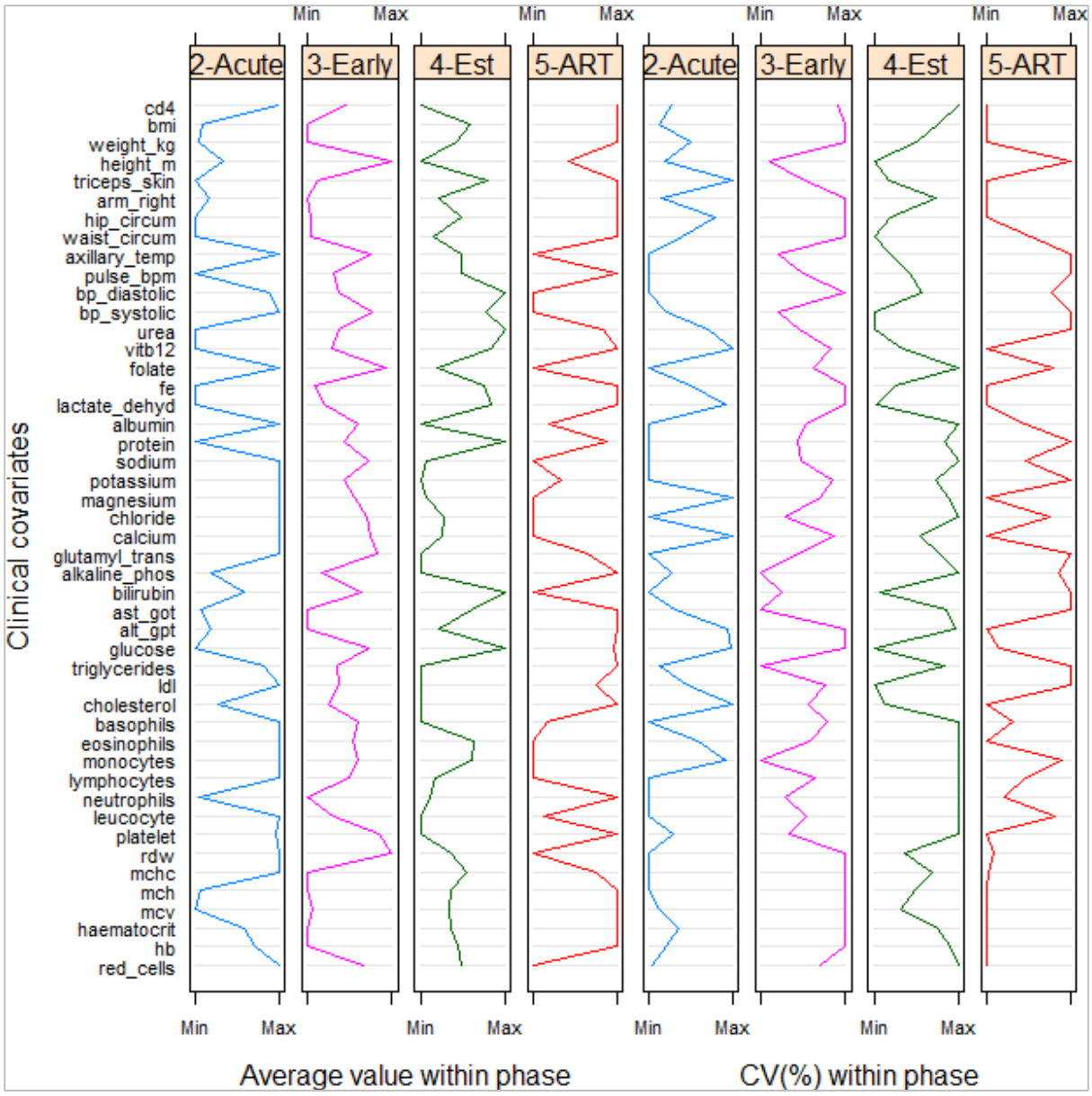


Figure 2.10: The results of the parallel plots of the mean and CV distributions
The CVs give information about the spread of the repeated measurements around the mean.

Echoing Figure 2.9 plot, the results showed that the highest mean $CD4^+$ counts were observed during the acute and ART phases, whereas the lowest mean $CD4^+$ counts were recorded during the early and established phases. However, it was during the acute and ART phases where the $CD4^+$ count variation was the least, and on the other hand, very high at lower $CD4^+$ counts. As the mean $CD4^+$ count changed from one phase to the next, the covariates also took different means, but in a more complex fashion. Taking for example BMI, the cohort's mean $CD4^+$ count was relatively high during the acute phase, but the BMI was the lowest. This was vice versa during the established phase. On the other hand, both the $CD4^+$ count and BMI averages were low during the established phase, but both were high during the ART. When the mean $CD4^+$ count was very low during the established phase, and the mean BMI was very high, it followed that both registered a very high degree of variation around their means. When both had high means relative to the other phases during the ART, they also both indicated the least variations around their means, relative to the other infection phases. Given the collective behavioural patterns of all the covariate influence on the $CD4^+$ count, the interrelationships were shown to be highly complex.

It was further worthwhile to explore for the existence of inter-individual variations in the $CD4^+$ count in response to the covariate changes (see Figures 2.11 and 2.12). For example, taking haematocrit in Figure 2.11, there was a noticeable general upward trend in the $CD4^+$ count as the haematocrit increases. However, the $CD4^+$ counts seem to vary widely at extremely lower and higher haematocrit. More so, patients with low $CD4^+$ counts at much lower haematocrit tended to have higher $CD4^+$ counts at elevated haematocrit levels and vice versa. In Figure 2.12, we note that protein at lower levels shows a high variation in the $CD4^+$ counts, but as the protein level increases, all the patients tended to have less variations in the $CD4^+$ counts, due to the $CD4^+$ count values seemingly approaching a common lower $CD4^+$ count. Time wise, the patients tended to enter each infection phase with high variations in the $CD4^+$ counts and further exited with such variations still at a high level. Generally, all the variables indicated that there was a point at which the variation in the $CD4^+$ count was minimal. This was observed either at extremely low values (for example, eosinophils) or mostly at the median (for example, LDH). Due to the multidimensional curse, visualising the individual profiles per infection, was incredibly overwhelming. Nevertheless, the general trends clearly indicated that the patients had different projections as either time, or the covariates change. It has already been established that the infection phase played a role in the $CD4^+$ count changes (Figure 2.9). This called for

an investigation to establish whether the covariate induced inter-individual variations were indeed significant within each phase. The application of multilevel modelling was deemed necessary to achieve this objective.

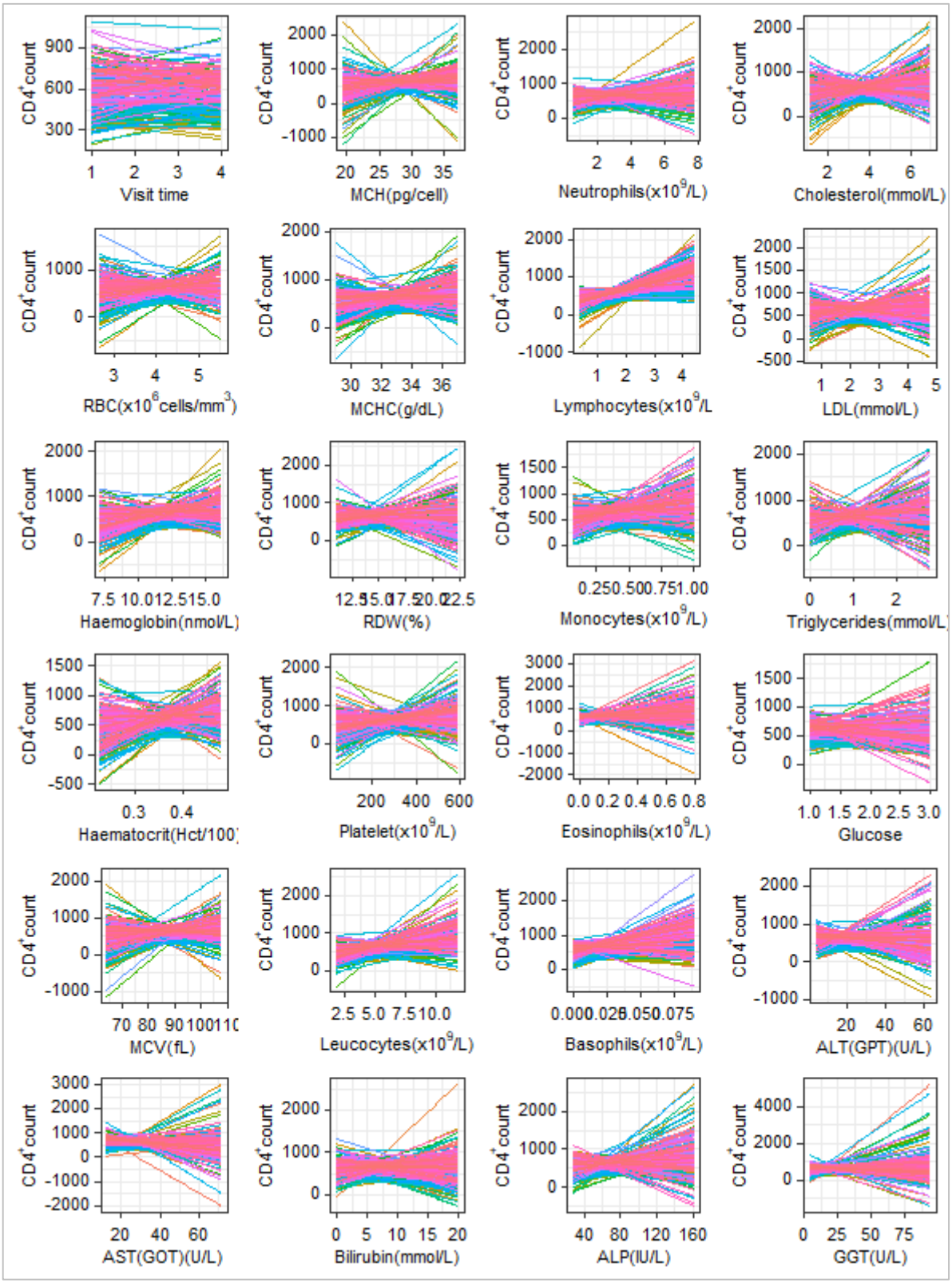


Figure 2.11: Individual linear profiles of the CD4⁺ count in response to the covariates (a)

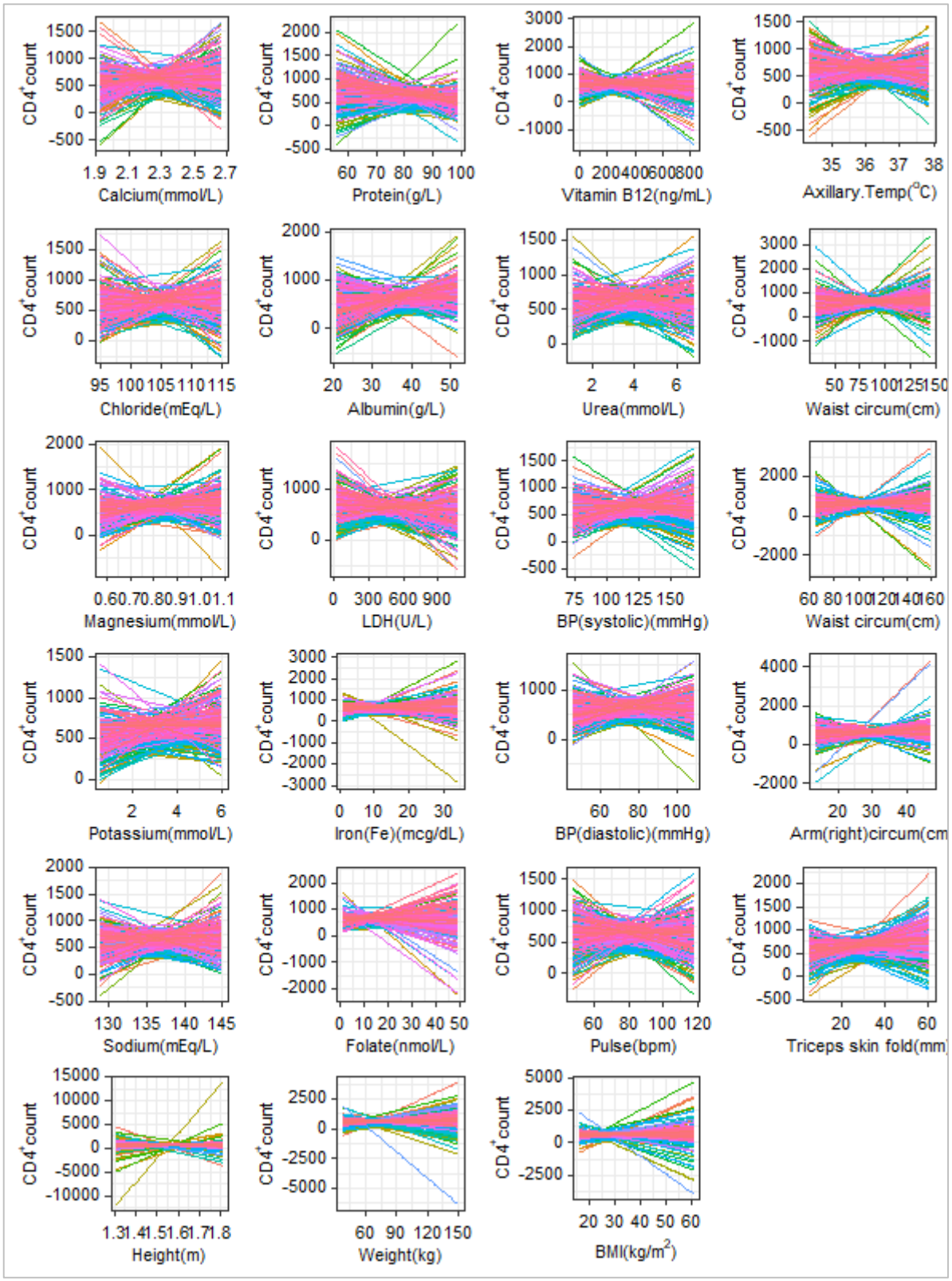


Figure 2.12: Individual linear profiles of the CD4⁺ count in response to the covariates (b)

An attempt to relax the linearity assumption of the individual profiles, by considering locally weighted scatterplot smoothers, produced clumsy visualisations with no meaningful patterns that could be interpreted. However, the smooth curves are known to discover some unimaginable patterns, which are of interest due to the continuous nature of the variables under investigation. The loess visualisations are data-driven and capable of giving insights into complex trends. The results of the loess trends tended to vary by infection phase (Figures 2.13 and 2.14). Some of the $CD4^+$ count and predictor patterns indicated either a general downward or upward trend. For example, lymphocytes and haematocrit showed upward trends whereas protein and folate showed downward trends. The patterns also revealed the existence of some highly non-linear relationships between the $CD4^+$ count and the predictors, for example, ALT and chloride. The smooth trends within each phase seem to trace the $CD4^+$ cell response curves, showing desirable ranges of the covariates. For example, the $CD4^+$ count responded well to platelet count increase within $<400x10^9$/L in all the infection phases (Figure 2.12). According to Figure 2.9, the early phase reported lower average $CD4^+$ counts, while the smooth curves indicated that it was during this phase where there was better $CD4^+$ count influence due to platelet count increase.

In light of the loess results, more advanced data-driven random smooths of additive models were considered appropriate to model the trends in the $CD4^+$ count deviations from average, in response to the covariates. There are also some segments of the loess trends showing that the $CD4^+$ count was linearly dependent on the covariates at certain intervals. The additive models are known to be explicitly visual, with no clear details on the breakpoints in the segmented relationships. As such, complementing the additive models with segmented regression models would provide some meaningful biological breakpoints in interpreting the trends from the purely graphical random smooths.

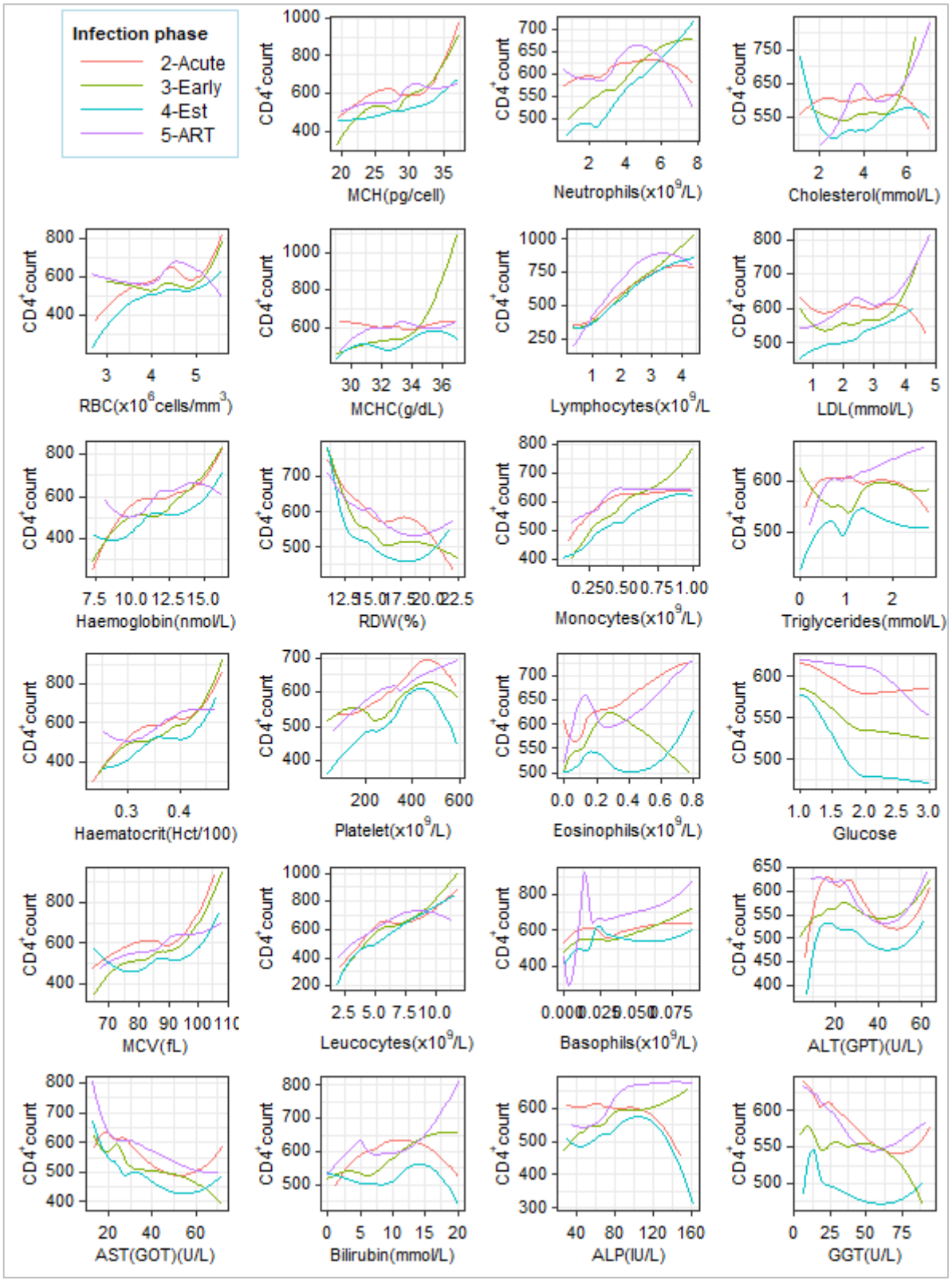


Figure 2.13: LOESS of the CD4$^+$ count against the covariates (a)

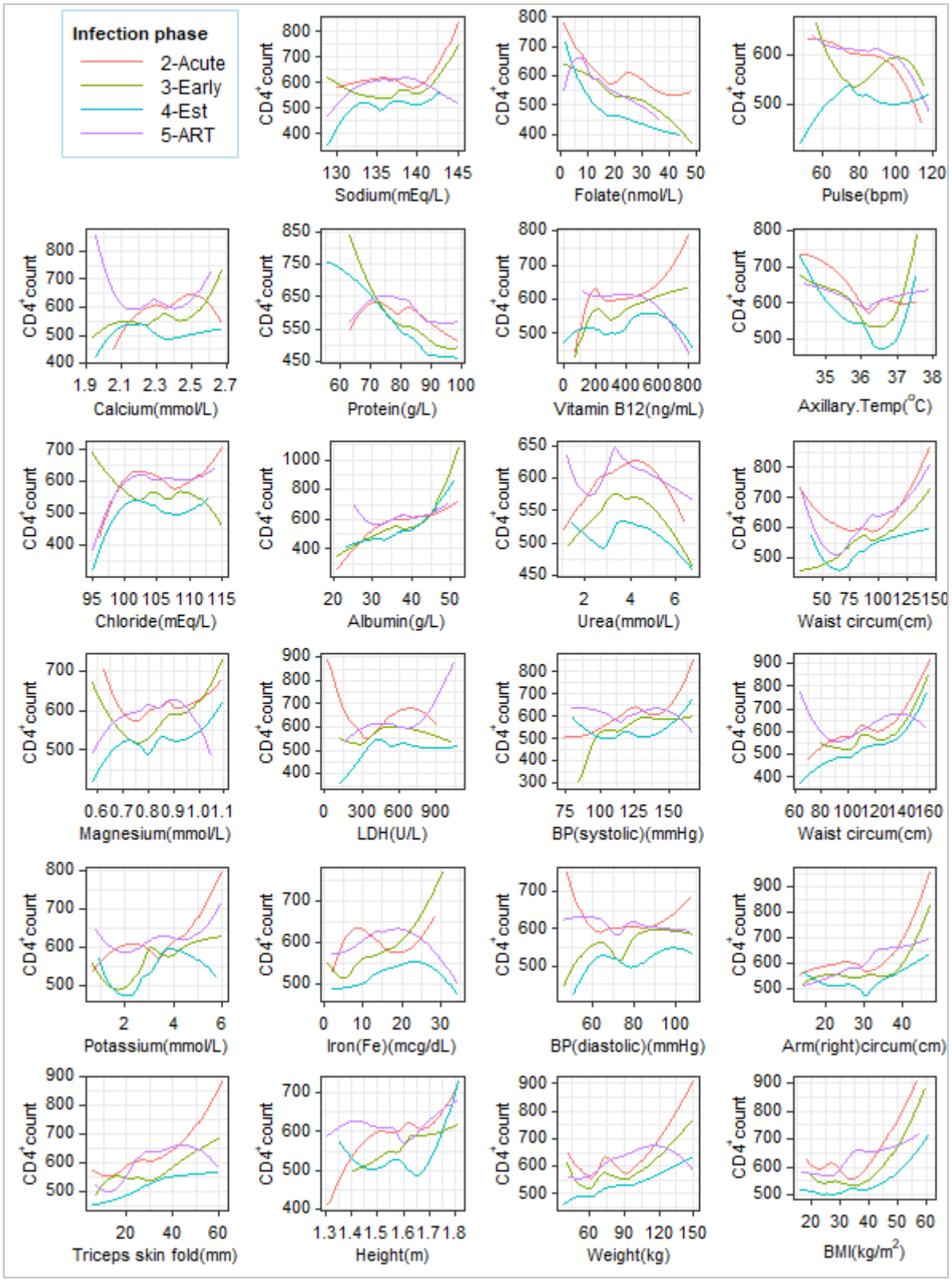


Figure 2.14: LOESS of the CD4⁺ count against the covariates (b)

Visualising the overall correlations between the variables further revealed the complexity in the relationships between the covariates themselves and $CD4^+$ count (see Figure 2.15).

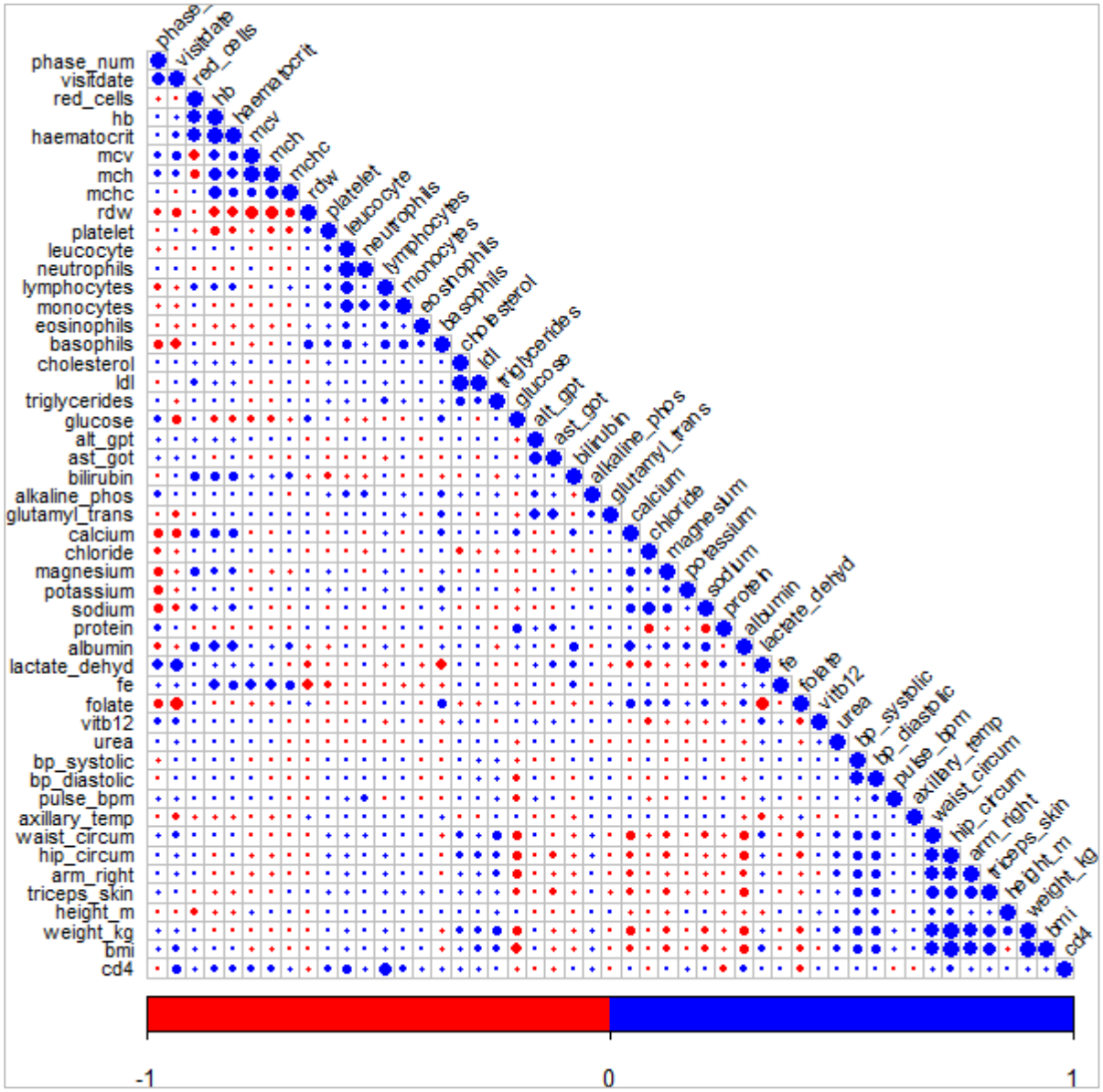


Figure 2.15: The results of the overall variable correlations
The blue diagonal circles represent a correlation of 1. All the other correlations are interpreted relative to the diagonal size. The smaller the circle, the weaker the correlation between the covariates.

The results showed that $CD4^+$ count was highly correlated with the total lymphocytes and to some extent, with the other white blood cells. The correlation results showed that the RDW was negatively correlated with the other red blood cell components. The RDW, basophils, calcium and folate decreased as the HIV disease progressed. Calcium and albumin were negatively related to the anthropometric measurements, whereas blood pressure was positively related to the anthropometric measurements. The anthropometric measurements were observed to be highly correlated among themselves. Also the electrolytes tended to decrease over time.

Taking the correlation plot's associations between $CD4^+$ count and covariates, and fitting it into the parallel coordinate plot puzzle, revealed even much more complex relationships. The covariates from different clinical platforms indicated interconnected relationships that further influence the $CD4^+$ count. In addition to the direct influence of the clinical covariates on the $CD4^+$ count, there are some underlying theoretical constructs (latent or hidden variables), formed by the observed clinical covariates that cannot be directly measured, yet they also play a crucial role in the monitoring of the patient's health status. The body's harmonic systems that support life are highly complex due to the relationships between the observed and latent clinical covariates, as well as the time lagged effects during the disease progression. Hence, the HIV invasion into the patient's body does not only affect the $CD4^+$ cell count but further disrupts the harmonic nature of the body's functional systems, as evidenced by several symptoms (Yerly and Hirschel 2012, Hoenigl et al. 2016). Such complex relationships were many questions are usually raised, Structural Equation Models (SEM) become the choice for handling this task.

The correlation plots further showed that some of the correlations were too high, for example, the intersection between the haemoglobin and haematocrit had a large blue circle, indicating that feature redundant selection was needed. The dropped redundant features are shown in Table 2.4.

Table 2.4: The results of redundant feature selection based on correlation

	Hb[a]	MCH[a]	Leucocytes[a]	Cholesterol[a]	Hip circumference[a]	Weight (kg) [a]	BMI[b]
Haematocrit	0.9499	-	-	-	-	-	-
MCV	-	0.9211	-	-	-	-	-
Neutrophils	-	-	0.8464	-	-	-	-
LDL	-	-	-	0.8636		-	-
Waist circumference	-	-	-	-	0.7580	0.8158	0.7810
Hip circumference	-	-	-	-	-	0.8884	0.8414
Arm(right) circumference	-	-	-	-	0.7867	0.8006	0.7925
Weight(kg) [a]	-	-	-	-	-	-	0.9149

[a] *Dropped redundant feature. Highly correlated (r >0.75) covariates of the $CD4^+$ count were dropped*
[b] *Intuitively included redundant feature for further investigation*

The results showed that haemoglobin (Hb), MCH, leucocytes, cholesterol, hip circumference, weight and BMI were highly correlated with the other covariates and were therefore suitable candidates to be dropped out from further analysis. The anthropometric measurements were the most highly correlated among themselves but the BMI, although marked as a redundant feature, was intuitively included for further investigation.

The volume of the suggested continuous $CD4^+$ count clinical covariates has increased tremendously in the recent years where it is now coined 'big data' (Demchenko 2013, Hersh 2014) often existing in many disjointed datasets, which are usually brought together by relational databases. This is owing to the new era of information technology, where patient electronic health records are being stored at a faster pace and are relatively cheaper to store than in the past (Sun 2013). Clearly, the issue of the multidimensional case was noted in the exploratory data analysis. Previously, the associations of these covariates at hand with the $CD4^+$ count have been analysed with statistical methods that ranged from Pearson or Spearmen correlation analysis, (Chorba et al. 2002, Lumbanraja and Siregar 2018) sensitivity, specificity and positive prediction, (Olawumi and Olatunji 2006, Sen. et al. 2011) linear regression (Daka and Loha 2008, Fasakin et al. 2014) multivariate regression, (Floris-Moore et al. 2006) logistic regression, (Butt et al. 2002, Fofana 2016) Chi Square tests, (Cohen and Steigbigel 1996, dos Santos and Almeida 2013, Moolla et al. 2015) non-parametric tests, (Shiferaw et al. 2016, Braconnier et al. 2017) independent samples t-tests, (Abdollahi et al. 2014, Atere et al. 2016) confidence intervals, (Dannhauser et al. 1999) analysis of variance (Opiyo et al. 2013, Adhikari et al. 2016) to generalized estimating equations (Sudfeld et al. 2013). Their limitations include the inability to give the covariates an opportunity to compete in a single multidimensional model to identify the most influential ones. On the other hand, given the massive HIV patient electronic health records, humans have difficulty in visualising the data in more than two dimensions (Davenport 2013) posing another challenge to approach this multidimensional problem. As such, an evaluation to determine the strongest candidates of these $CD4^+$ count covariates was considered to be the first step moving forward. Since the explored data structure is multilevel in nature, an application of the variable selection techniques for longitudinal data was called for.

2.5 Summary

Chapter 2 provided an understanding of the study variables including their roles as indicators to the functional status of some vital body organs. The data cleaning process removed some erroneous values that had the potential to mask some pattern discovery and avoided the challenges associated with the violation of distributional assumptions in further investigative statistical modelling approaches. Highly correlated covariates were also dropped as a first step in dealing with redundant features. The visual displays discovered some data patterns which gave some insights into the clinical covariate influence on the $CD4^+$ cell count, which tended

to vary with the infection phase. Although the discovered data patterns proved to be useful, they were limited in providing a better comprehension of the clinical covariate influence on the $CD4^+$ cell count. Nevertheless, the data exploration was instrumental in highlighting the best course of action in understanding the problem at hand, although this is still challenged by the high-dimensional curse. Hence, any step from the data exploratory analysis demanded that only the strongest candidates of these $CD4^+$ count clinical covariates should be dealt with. As such, the next chapter focuses on the variable selection in high-dimensional longitudinal data.

CHAPTER 3 VARIABLE SELECTION IN HIGH-DIMENSIONAL LONGITUDINAL DATA

To enhance the parsimonious description between the response and covariates, it is a common practice to include only the important variables (Fan and Li 2004). Not all the covariates are actually important in explaining the response and as such, they have to be dropped from the model resulting in a reduced number of variables (Tobias 1995, Maitra and Yan 2008, Aggarwal and Kosian 2011). All the $CD4^+$ cell count covariates under consideration in this study are repeated measurements that are continuous in nature. If the interest is not in the understanding of the response at the individual level, such repeated measurements data are commonly analysed using generalised estimation equations (GEE). Currently of interest is variable selection, and it suffices to reason that GEE inclined variable selection procedures are attractive. It is important to note that a number of variable selection procedures have been developed, though some are without a special focus on variable screening in high-dimensional longitudinal data. Reviews of such techniques have been conducted (Fan and Lv 2010, Degenhardt et al. 2019) and are not relevant for the longitudinal data at hand. Explaining the influential effects on $CD4^+$ cell count with just a few clinical covariates, will not only improve the efficiency of the subsequent models but also improve on resource optimisation during the data collection in the HIV/AIDS prospective cohorts. As time is a key component of such data from prospective studies, one of the most cited advantages of the GEE is that the modelling techniques provide consistent estimates, even if the correlation structure of the repeated measurements is miss-specified (Inan and Wang 2017). On the other hand, and unlike the model-driven GEE, time-varying coefficient models that are more data-driven, have also been proposed for variable selection in longitudinal data where B-Splines are central to the modelling of the response variable in a semi-parametric fashion (Chu et al. 2016). As such, the GEE and the time-varying coefficient models have been some of the basis for developments in the variable selection techniques in high-dimensional longitudinal data. These two approaches could be viewed as variable selection applications in high-dimensional longitudinal data, in conjunction with regression parameter estimation. In addition to these techniques, the problem of variable selection has also been realised in non-regression estimation problems, such as multi-omics data integration, where there is an increasing demand for bringing at least two data sets together whilst simultaneously filtering out the unimportant variables (Wu et al. 2019). The data integration can be with, or without, an outcome from the two data sets, and in the

latter case, an unsupervised variable selection method is applicable. Hence, the variable selection techniques in the longitudinal data sets, appear to have been developed in three main categories: the ones with GEE as an underlying model, the time-varying coefficient models and multi-level omics data integration. The variable selection techniques that extend the GEE are characterised by penalised functions whereas those solving the omics data integration problems, have the advantage of incorporating both the penalised function and an unsupervised approach. Under these categories, there are many other variants of the variable selection methods developed for specific solutions. However, it has been reiterated that there is no method that can dominate the rest (Wu et al. 2019). The following sections discuss the most recent developments in each of these main categories.

3.1 Variable selection with time-varying coefficient models in high-dimensional longitudinal data

In general, the time-varying coefficient models use B-Splines to calculate the time varying error variance for each covariate. Several other variable screening approaches have been developed but did not incorporate the within-subject correlation, dynamic error variance and demographic baseline variables (Chu et al. 2016, Inan and Wang 2017). Suppose n denotes the number of subjects where each i^{th} subject has been followed and visited at t_{ij} times for $j = 1,\ldots,n_i$ repeated number of observations for both the response (y) and the p time-varying covariates ($x's$), then the response for the i^{th} subject at time t can be denoted by $y_i(t)$. Similarly, the observed value for the k^{th} covariate of the i^{th} subject at time t can be denoted as $x_{ik}(t)$, for $k = 1,\ldots,p$ and a low dimensional time-invariant baseline record for the l^{th} covariate z_l at time t, is denoted by $z_l(t)$ for $l = 1,\ldots,q$. In general, the time t is assumed to be bounded within an interval of positive real numbers. The relationship between the $y_i(t)$ and the time-varying $x_{ik}(t)$ and time-invariant $z_l(t)$) covariates can be expressed as

$$y_i(t) = \beta_0(t) + \sum_{l=1}^{q} \beta_l(t) z_{il}(t) + \sum_{k=1}^{p} \gamma_k(t) x_{ik}(t) + \varepsilon_i(t), \qquad (3.1)$$

where $\beta_l(t)$ and $\gamma_k(t)$ are nonparametric smooth coefficient functions. The $\varepsilon_i(t)$ is the error term, that can be assumed to have a variance across time, independent between subjects (across

i) and correlated within the same subject (across t). The importance of the k^{th} covariate can be measured by considering a single χ-covariate nonparametric regression model

$$y_i(t_{ij}) = \beta_{0k}^*(t_{ij}) + \sum_{l=1}^{q} \beta_{lk}^*(t_{ij}) z_{il}(t_{ij}) + \gamma_k^*(t_{ij}) x_{ik}(t) + \varepsilon_i^*(t_{ij}) \ , \tag{3.2}$$

where β_{lk}^* and $\gamma_k^*(t_{ij})$ are the smooth coefficient functions, that are approximated using cubic B-splines, such that the relationship can be reduced to a linear regression model

$$y_i(t_{ij}) \approx \sum_{m=1}^{M_{0n}} \eta_{0m} B_{0m}(t_{ij}) + \sum_{l=1}^{q} \sum_{m=1}^{M_{ln}} \eta_{lm} B_{lm}(t) z_{il}(t_{ij}) + \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t) x_{ik}(t_{ij}) + \varepsilon_i^*(t_{ij}), \tag{3.3}$$

where $B(.)$ are the set of B-splines that may differ across the basis functions m and $\boldsymbol{h}$. The M_{0n}, M_{ln} and L_{kn} represent the number of basis functions for $\beta_{0k}^*(t_{ij})$, $\beta_{lk}^*(t_{ij})$ and $\gamma_k^*(t_{ij})$ respectively. The problem at hand has a single response variable, the $CD4^+$ count and p covariates that were repeatedly recorded at $j = 1,...,16$, where $j = 1,...,4$, $j = 5,...,8$, $j = 9,...,12$ and $j = 12,...,16$ are time points for the different infection phases 2(acute) to 5(ART). Hence in our study, the $z_l(t)$ is not invariant across j, for each subject. The term $\sum_{l=1}^{q} \sum_{m=1}^{M_{ln}} \eta_{lm} B_{lm}(t) z_{il}(t_{ij})$ is not applicable and consequently dropped from (3.3) to model

$$y_i(t_{ij}) \approx \sum_{m=1}^{M_{0n}} \eta_{0m} B_{0m}(t_{ij}) + \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t) x_{ik}(t_{ij}) + \varepsilon_i^*(t_{ij}), \tag{3.4}$$

Since the $\varepsilon_i^*(t_{ij})$ are assumed to be independent between subjects, correlated within subject and time-varying, increasing the screening accuracy would also require the incorporation of the error variance (Chu et al. 2016). This attracts the application of generalised estimation equations or weighted least squares, but due to some misspecifications and computational costs, the covariance matrix of the errors $\varepsilon_i^*(t_{ij})$ is replaced by a working variance function for the $\varepsilon_i^*(t_{ij})$ which is a smooth function of t and estimated by

$$\hat{V}(t_{ij}) = \sum_{h=1}^{H_n} \hat{\phi}_h B_{hn}(t_{ij}). \tag{3.5}$$

This is used for constructing weighted least squares estimates for θ_{kh} in (3.3). Given the working variance function $\hat{V}(t_{ij})$, an $n_i \times n_i$ working correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ for the i^{th} subject allows for the use of a parametric model in which the $\boldsymbol{\alpha}$ fully characterises the correlation structure. The autoregressive (AR), stationary and non-stationary M-dependent correlation structures are the commonly used correlation structures. Given the $\boldsymbol{\alpha}$ parameters, the working variance and correlations structures, the weight matrix for the generalised estimation equation is obtained by $\boldsymbol{W_i} = n_i^{-1}\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{-1/2}\boldsymbol{R}_{\boldsymbol{i}}^{-1}(\hat{\boldsymbol{\alpha}})\hat{\boldsymbol{V}}_{\boldsymbol{i}}^{-1/2}$. The weighted least squares estimate for the k^{th} covariate is then calculated as

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} = \left(\frac{1}{n}\sum_{\boldsymbol{i=1}}^{n}\boldsymbol{U}_{\boldsymbol{ki}}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{U}_{\boldsymbol{ki}}\right)^{-1}\left(\frac{1}{n}\sum_{\boldsymbol{i=1}}^{n}\boldsymbol{U}_{\boldsymbol{ki}}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{i}}\boldsymbol{y}_{\boldsymbol{i}}\right). \tag{3.6}$$

The response value of the i^{th} subject can be estimated by the k^{th} covariate, using the relationship $\hat{y}_i^{(k)} = \boldsymbol{U}_{ki}\hat{\boldsymbol{\theta}}_k$ where $\hat{u}_{nk} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{k})}\right)^T\boldsymbol{W}_{\boldsymbol{i}}\left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{k})}\right)$, are the weighted mean squared errors. Sorting the $\hat{u}_{nk}$ in increasing order, allows for the variable selection, where the smaller the $\hat{u}_{nk}$, value, the stronger the k^{th} covariate relationship with the response or the stronger the marginal association between the k^{th} covariate and the response.

3.2 Variable selection with penalized GEE in high-dimensional longitudinal data

The GEE calculate penalised regression estimates and the model is consistent even if the working correlation structure is misspecified. In the case of the GEE, for $j = 1,..., n_i$, the j^{th} time response observation of the i^{th} subject is denoted by y_{ij}, where $i = 1,..., n$. Assumed are two marginal moments of y_{ij} such that $E(y_{it} \mid \boldsymbol{X}_{it}) = \mu(\theta_{it})$ and $Var(y_{it} \mid \boldsymbol{X}_{it}) = \dot{\mu}(\theta_{it})$ where $\mu(\theta_{it}) = \boldsymbol{X}_{it}^T\boldsymbol{\beta}$ and the unknown regression coefficients to be estimated are given by $\boldsymbol{\beta} = (\beta_1,..., \beta_p)^T$ (Liang and Zeger 1986). Instead of the time-varying covariance as in (3.4), the GEE marginal covariance is obtained by $\mathbf{V}_i = Var(\mathbf{Y}_i \mid \boldsymbol{X}_{it})$ estimated via a correlation structure given by $\mathbf{V}_i = \boldsymbol{A}_i^{1/2}R_{n_i}(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$ where $\boldsymbol{A}_i$ is an $n_i \times n_i$ diagonal matrix of marginal

variance of responses and $R_{n_i}(\boldsymbol{\alpha})$ is a working correlation matrix which is also an $n_i \times n_i$ matrix (Liang and Zeger 1986). The unknown regression parameters, $\boldsymbol{\beta} = (\beta_1,..., \beta_p)^T$ are usually estimated by

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{A}_i^{1/2} \hat{R}^{-1}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{3.7}$$

where the equations are solved by a modified Fisher scoring algorithm in which the $\boldsymbol{\alpha}$ can be estimated by a residual-based moment method (Hardin and Hilbe 2003).

In the case of high-dimensional covariates where the $\boldsymbol{\beta} = (\beta_1,..., \beta_p)^T$ can be assumed to be sparse, that is, some of the regression estimates are exactly zero, $\boldsymbol{S}(\boldsymbol{\beta})$can have some of the components penalised to clearly distinguish between the zero and non-zero parameters. With the penalised generalised estimation equations (PGEE), the new penalised estimates are obtained by solving $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{\beta}) - \boldsymbol{q}_\lambda(|\boldsymbol{\beta}|) \circ sign(\boldsymbol{\beta})$ (Wang et al. 2012). The $\boldsymbol{q}_\lambda(|\boldsymbol{\beta}|)$ is the penalty function whose k^{th} component is given by $q_\lambda(|\beta_k|)$ and the $sign(\boldsymbol{\beta})$ is the Hadamard product of the two vectors in $\boldsymbol{q}_\lambda(|\boldsymbol{\beta}|)$. For a large value of $|\beta_k|$, the $q_\lambda(|\beta_k|) = 0$ and if $|\beta_k|$ is small, the $q_\lambda(|\beta_k|)$ becomes large. This implies that the smaller the $|\beta_k|$, the more the $S(\beta_k)$ is shrunken towards zero with $\beta_k < 10^{-3}$ being considered a candidate for exclusion in the variable selection process (Cho and Qu 2013). A comprehensive review of the developed penalty functions has been detailed by (Wu and Ma 2015) and (Fan and Lv 2010). Among these, the least absolute shrinkage and selection operator (LASSO) L_1 by (Tibshirani 1996) was the most popular but (Inan and Wang 2017) recently suggested that a nonconvex smoothly clipped absolute deviation penalty by (Fan and Li 2001) be considered for it avoids over penalizing of the large coefficients. The smoothly clipped absolute deviation is given by

$$q_\lambda(t) = \lambda \left\{ I(t < \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t \geq \lambda) \right\}, \tag{3.8}$$

where $\lambda \geq 0$ and $a = 3.7$ (Fan and Li 2001) found to work well. The tuning parameter λ is obtained by partitioning the data into non-overlapping S sub-samples of equal sizes where the s^{th} sub-sample is left out as a test data set whilst the remaining data is used for training with

the penalized general estimation equations. The s^{th} sub-sample used for the test data can be used to evaluate the prediction error given by $PE_{-s}(\lambda)$. In practice, a grid of λ values are provided and for each λ, each s^{th} sample is given a chance as a test data set such that the overall cross-validated predictor error is given by $CV(\lambda) = \frac{1}{S}\sum_{s=1}^{S} PE_{-s}(\lambda)$. The λ that yields the smallest $CV(\lambda)$ is chosen as the tuning parameter. Given the optimal λ, the penalised estimated equations in (3.7) can be solved by minorization-maximization and Newton-Raphson algorithm to obtain $\hat{\boldsymbol{\beta}}$ (Johnson et al. 2008) which can then further be used as an ingredient to a sandwich formula for estimating the asymptotic covariance matrix of the penalised general estimation equations denoted by $Cov(\boldsymbol{\beta})$ (Inan and Wang 2017).

3.3 Variable selection using high-dimensional multi-level omics data integration

Data coming from high throughput technologies such as transcriptomics, metabolomics and proteomics have given birth to the term "omics". Bringing together such multivariate omics data sets for analysis has led to the need for specialised statistical methods with a focus on data exploration, variable selection and visualisation. With an explosion of the $CD4^+$ cell count covariates from different clinical platforms, the omics data integration techniques incorporating variable selection become very relevant. A recent comprehensive review of the techniques developed specifically for variable selection in the multilevel omics data integrations has been provided (Wu et al. 2019) where the existing integrative analyses methods have been described as either supervised, semi-supervised or unsupervised techniques. The supervised methods are characterised by regression analysis whereas the unsupervised are exploratory in nature. The problem at hand requires the identification of the strongest covariates of the $CD4^+$ cell count and hence the unsupervised methods become more attractive for the exploration. The unsupervised techniques are also considered as optimization problems with functions that seek to achieve different objectives and their sparse properties have been found to be sought after. Among the most common unsupervised feature selection techniques include the partial least squares (PLS), principal component analysis (PCA), canonical correlation analysis (CCA), clustering and co-inertia analysis (CIA). In the variable selection literature, the shrinking of the estimates of the unimportant variables to zero is termed sparsity. As such, these unsupervised variable selection techniques also incorporate the sparsity

with different loss functions where the sparse PLS (SPLS) has a covariance-based loss, sparse PCA (approximation loss), sparse CCA (correlation-based loss), sparse clustering (within-cluster sum of squares) and sparse CIA (co-inertia-based loss). The sparsity is achieved by imposing different penalty functions to the loss functions. The commonly used penalties in this area are the least absolute shrinkage and selection operator (LASSO) , fused LASSO (Tibshirani et al. 2005), adaptive LASSO (Zou 2006), ridge penalty, elastic net (Zou 2005), network penalty and the minimax concave penalty (Zhang 2010). The CCA is interested in the maximum correlation between two omics measurements, whereas the PCA *partly* focuses on the singular value decomposition of the covariate matrix only. On the other hand, the PLS is closely related to the PCA algorithms (Garthwai 1994) and employs the singular value decomposition but has the advantage that both the response and covariate matrices are simultaneously decomposed. CIA is a generalisation of the PLS and the CCA whereas clustering minimises the within cluster sum of squares. The current problem in our data consists of a single response variable data set and a high-dimensional covariate set which boils down to solving a problem that maximises the variation in both the response and covariate data sets. This is similar to dealing with a pair-wise omics dataset that requires the strength of the SPLS. Since the PLS is based on the PCA, the following sections discuss the PCA, PLS and SPLS in that order.

3.3.1 Principal component analysis (PCA)

PCA is a multivariate technique first introduced by Pearson in 1901 and also independently developed by Hotelling in 1933, with its main objective being to reduce a large number of variables by transforming them into a few new set of uncorrelated variables called principal components. The first few components capture most of the variation in the original variables and this is represented as eigenvalue-eigenvectors. However, there are some cases where the last few components can be potentially useful (Jollife 1986). The variable reduction accelerates model building with acceptable accuracy and generalisation without losing much information from the original data.

Given a response matrix $\mathbf{Y}_{N\times r}$ to be explained by a covariate matrix $\mathbf{X}_{N\times p}$ where N is the number of observations in a data set with r responses and p covariates, there exists $M < p$ readily interpreted principal components (composite variables) that can be extracted from $\mathbf{X}_{N\times p}$ alone. The PCA approach decomposes the matrix $\mathbf{X}_{N\times p}$ without taking $\mathbf{Y}_{N\times r}$ into

consideration. The original p random variables can be represented as a vector $\mathbf{X}$ such that p variances and $\frac{1}{2}p(p-1)$ covariances or correlations can be obtained. The PCA is most sensitive to the variances which are usually the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ (if known or $\mathbf{S}$ if unknown) and the off diagonal elements are the covariances.

For principal components $1,...,M$ the computation of the first (1) principal component with the maximum variation in $\mathbf{X}$, considers that for each x_k, $k=1,...p$ variable there exist a constant φ_{1k} such that a linear function $\varphi_{1k}x_k$ captures the variance in the x_k variable. This follows that a combination of φ_{1k}'s that maximizes the total variance for the first principal is the sum of all the linear functions, $\sum_{k=1}^{p}\varphi_{1k}x_k$ which can further be denoted in matrix form as $\varphi_1'\mathbf{X}$ where the vector φ_1 represents constants $\varphi_{11},\varphi_{12},...,\varphi_{1p}$ for the p random variables such that $\varphi_1'\mathbf{X}$ is given by

$$\varphi_1'\mathbf{X} = \varphi_{11}x_1 + \varphi_{12}x_2 + ... + \varphi_{1p}x_p = \sum_{k=1}^{p}\varphi_{1k}x_k \ . \tag{3.9}$$

The second principal component is denoted by $\varphi_2'\mathbf{X}$ and is uncorrelated to $\varphi_1'\mathbf{X}$. The variation captured by the m^{th} principal component is then given by $\varphi_m'\mathbf{X}$. The variance of the first principal $\varphi_1'\mathbf{X}$ is given by $Var\left[\varphi_1'\mathbf{X}\right] = \varphi_1'\boldsymbol{\Sigma}\varphi_1$, where $\varphi_1'\varphi_1 = 1$, an imposed normalisation constraint to achieve the maximization for the finite φ_1 (Jollife 1986). Maximising the first principal component variance, $\varphi_1'\boldsymbol{\Sigma}\varphi_1$ requires the Lagrange multipliers that maximize

$$\varphi_1'\boldsymbol{\Sigma}\varphi_1 - \lambda\left(\varphi_1'\varphi_1 - 1\right) \tag{3.10}$$

where λ is a Lagrange multiplier. Differentiating (3.10) with respect to φ_1 and equating to zero gives

$$\left(\boldsymbol{\Sigma} - \lambda\mathbf{I_p}\right)\varphi_1 = 0 \ , \tag{3.11}$$

.

where $\mathbf{I_p}$ is a $p \times p$ identity matrix and then λ becomes the eigenvalue of $\boldsymbol{\Sigma}$. Simplifying (3.11) gives $\varphi_1^{'}\boldsymbol{\Sigma}\varphi_1 = \kappa$ such that $Var\left[\varphi_1^{'}\mathbf{X}\right] = \varphi_1^{'}\boldsymbol{\Sigma}\varphi_1 = \kappa_1$, where κ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}$. Hence, $Var\left[\varphi_m^{'}\mathbf{X}\right] = \varphi_m^{'}\boldsymbol{\Sigma}\varphi_m = \kappa_m$ for $m = 1,..., M$.

The correlation matrix of $\mathbf{X}$ can also be used to obtain the principal components. If σ_{kk} is the variance of x_k, the vector $\mathbf{X}$ can be standardised to $\mathbf{X}^*$ where the k^{th} element of $\mathbf{X}^*$ is given by $\frac{x_k}{\sqrt{\sigma_{kk}}} = \frac{x_k}{w_k}$, $w_k = \sqrt{\sigma_{kk}}$ for $k = 1,..., p$. The advantage of the correlation matrix approach over the covariance matrix is that measurements with different units are standardised and comparable (Jollife 1986, Abdi and Williams 2010). Also if not standardised, the covariance matrices are affected by large differences between variances which dominate the first few principal components (Jollife 1986). Since in a correlation matrix each variable contributes a variance of 1, the total eigenvalues of the correlation matrix are equal to the number of variables (p) in the vector $\mathbf{X}$ and this can be summarised as $\sum_{m=1}^{M} \kappa_m = p$. Hence, the proportion of variance accounted for by the m^{th} principal component is then given by $\frac{\kappa_m}{p}$ and the total proportion of variance accounted for by the first m principal components becomes $\frac{\sum_{m=1}^{M} \kappa_m}{p}$.

3.3.2 Partial Least Squares (PLS)

The concept of PLS was introduced by (Wold 1966) and is represented in Figure. 3.1, where the orthogonal decomposition (detailed in Section 3.3.1) of $\mathbf{X}_{N\times p}$ gives the scores $\mathbf{T}_{N\times m}$ where M is the number of extracted factors from $\mathbf{X}_{N\times p}$. On the other hand maximum redundancy analysis or reduced rank regression can derive $\mathbf{U}_{N\times M}$ an orthogonal matrix with M factors extracted from $\mathbf{Y}_{N\times r}$. The PLS's goal is to simultaneously decompose $\mathbf{X}$ and $\mathbf{Y}$ to maximize the covariance between $\mathbf{T}$ and $\mathbf{U}$. Since $\mathbf{X}$ may suffer from the problem of multicollinearity and most often associated with a large number of explanatory variables that result in a complex model (Garthwai 1994), it is substituted by fewer explanatory variable matrix $\mathbf{T}$, which is no

longer collinear (orthogonal). $\mathbf{T}$ is also effective in the sense that it accounts for the variation in the covariates controlling (using $\mathbf{U}$) for the variation accounted for by the response variables. Hence, this allows for the identification of a direction within the $\mathbf{X}$ space that explains the maximum amount of variation in the $\mathbf{Y}$ space (Sawatsky et al. 2015). Eventually, the most important matrices that underpins the PLS procedure are $\mathbf{T}$ and $\mathbf{Y}$, where the regression of $\mathbf{Y}$ on $\mathbf{T}$ is referred to as the partial least squares regression (PLSR). The terms *factors*, *components*, *latent variables* and *x-scores* are interchangeably used and denoted by $\mathbf{T}$.

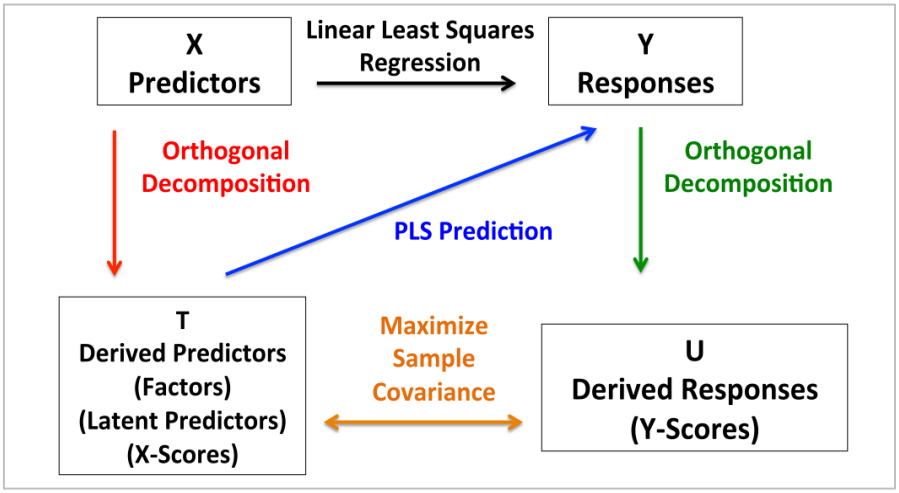


Figure 3.1: The concept behind the PLS approach.
Source: Predictum Inc. (Cai 2012)

The partial least squares regression requires that the explanatory variables be continuous probably due to the matrix manipulation whereas the response variables can either be categorical or continuous and although not imperative, the data should be relatively normal (Sawatsky et al. 2015). In the case of the categorical responses, it is called the partial least squares discriminant analysis. This study deals with a single continuous response variable and hence only the PLS algorithms for one continuous dependent variable are discussed. The derivation of $\mathbf{T}$ is an iterative procedure where the factors are extracted one at a time. Suppose $\mathbf{X} = \mathbf{X}_1$ and $\mathbf{Y} = \mathbf{Y}_1$ are the centred (mean subtracted from each observation) and scaled (divided by standard deviation) matrices for the explanatory and response variables respectively, their covariance matrix is given by $\mathbf{X}_1^{'}\mathbf{Y}_1$ which the PLS procedure intends to reduce by singular value decompose. The most commonly used algorithms are Nonlinear Iterative Partial Least Squares (NIPALS) and Statistically Inspired Modification of the PLS (SIMPLS) and both produce identical results for a single response variable (SIMPLS preferred for multivariate). The NIPALS works with the residuals whereas the SIMPLS deals directly

with scaled scores for both explanatory and response matrices. For a single response (as in this study), the NIPALS was preferred as it gives the same results as the SIMPLS. The NIPALS is also handled with microcomputers and most importantly allow a good understanding of the PLS (Geladi and Kowalski 1986). The NIPALS iterations are summarised in Table 3.1.

Table 3.1: The NIPALS iterative procedure for factor extraction

Iteration		Covariate matrix	Response matrix	Subject to
Factor 1	Decompose	$\hat{\mathbf{X}}_1 = \mathbf{TG}' + \boldsymbol{\varepsilon}_{1x}$	$\hat{\mathbf{Y}}_1 = \mathbf{UQ}' + \boldsymbol{\varepsilon}_{1y}$	$\mathbf{T}'\mathbf{T} = \mathbf{1}$ $\mathbf{U}'\mathbf{U} = \mathbf{1}$ $\mathbf{T}'\mathbf{U}$ *is maximal*
	Where	$\mathbf{T} = \mathbf{X}_1\mathbf{W}$ $\mathbf{G}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}_1$	$\mathbf{U} = \mathbf{Y}_1\mathbf{C}$ $\mathbf{Q}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}_1$	$\mathbf{U} = \mathbf{bT} + \boldsymbol{\varepsilon}_{inner}$
	Descriptions	$\mathbf{T} = X$ *- scores* $\mathbf{G} = X$ *- loadings* $\boldsymbol{\varepsilon}_{1x} = X$ *- errorterms* $\mathbf{W} = X$ *- weights*	$\mathbf{U} = Y$ *- scores* $\mathbf{Q} = Y$ *- loadings* $\boldsymbol{\varepsilon}_{1y} = Y$ *- errorterms* $\mathbf{C} = Y$ *- weights*	$\mathbf{b} = beta\ coefficients$ $\boldsymbol{\varepsilon}_{inner} = Random\ error$
Factor 2	Decompose	$\mathbf{X}_2 = \mathbf{X}_1 - \hat{\mathbf{X}}_1$	$\mathbf{Y}_2 = \mathbf{Y}_1 - \hat{\mathbf{Y}}_1$	
		•		
		•		
		•		
Factor M	Decompose	$\mathbf{X}_M = \mathbf{X}_{M-1} - \hat{\mathbf{X}}_{M-1}$	$\mathbf{Y}_M = \mathbf{Y}_{M-1} - \hat{\mathbf{Y}}_{M-1}$	

The $\mathbf{X}_2$ and $\mathbf{Y}_2$ are called the deflated $\mathbf{X}_1$ and $\mathbf{Y}_1$ blocks after the subtraction (partial out) of the first factor. Ignoring the error terms $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_y$, the process is repeated until the matrices $\mathbf{X}$ and $\mathbf{Y}$ are deflated to null matrices. The decomposed explanatory and response matrices give their outer relationship whereas $\mathbf{U} = \mathbf{bT} + \boldsymbol{\varepsilon}_{inner}$ gives the inner relationship between the score matrices and the regression coefficients of the inner relationship given by $\mathbf{b}$ (Jun et al. 2009).

The PLS is equivalent to the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_{pls} + \varepsilon$ where the PLS beta coefficients ($\boldsymbol{\beta}_{pls}$) are given by $\boldsymbol{\beta}_{pls} = \mathbf{W}(\mathbf{G}'\mathbf{W})^{-1}(\mathbf{T}'\mathbf{T})^{-1}(\mathbf{T}'\mathbf{Y})$ and ε the random error term. The $\beta_{pls}^{(k)}$ is interpreted as an increase of a particular y-variable as a change of the x_k-variable when the other x-variables are fixed. The last iteration results in a set of M derived orthogonal covariates and p beta coefficients. For the extracted matrix $\mathbf{T}_{N\times M}$ we seek to find

the first $m^* < M$ optimal number of factors that are just enough to explain $\mathbf{Y}$ without over fitting.

3.3.3 The sparse Partial Least Squares (SPLS)

The sparse partial least squares (SPLS) is a two omics unsupervised model approach (Rohart et al. 2017). In addition to the traditional PLS, during the singular value decomposition, the SPLS include a soft-thresholding penalization, LASSO L_1 which is a commonly used method for analysing high dimensional data (Tibshirani 1996). The sparse PCA improves the limitations of the PCA (Zou et al. 2006) which underlies the SPLS. In the case of the SPLS, the LASSO is applied on both loadings $\mathbf{G}$ and $\mathbf{Q}$ of the explanatory $\mathbf{X}$ and response $\mathbf{Y}$ matrices (refer to Table 3.1) respectively (Lê Cao et al. 2008). In general, given $\mathbf{H} = \mathbf{X}'\mathbf{Y}$ such that $\mathbf{H} = \mathbf{G}\Delta\mathbf{Q}'$ where Δ contains singular values as a diagonal matrix, the SPLS tries to optimize.

$$\min_{\boldsymbol{g}_m, \boldsymbol{q}_m} \|\boldsymbol{H}_m - \boldsymbol{g}_m, \boldsymbol{q}_m'\|_F^2 + L_{\lambda_1 m}(\boldsymbol{g}_m) + L_{\lambda_2 m}(\boldsymbol{q}_m), \tag{3.12}$$

where

F = Frobenius norm

$\boldsymbol{g}_m$ $= \left(g_{l^{(g)},m}\right)_{l^{(g)}}$ loading vectors of $\mathbf{G}$, $l^{(g)} = 1, \ldots, G$. At each iteration $\boldsymbol{g}_m$ is alternatively fixed while $\boldsymbol{q}_m$ is constrained to be of unit-norm for $m = 1, \ldots, M$ component scores or latent variables

$\boldsymbol{q}_m$ $= \left(q_{l^{(q)},m}\right)_{l^{(q)}}$ loading vectors of $\mathbf{Q}$, $l^{(q)} = 1, \ldots, Q$. At each iteration $\boldsymbol{q}_m$ is alternatively fixed while $\boldsymbol{g}_m$ is constrained to be of unit-norm for $m = 1, \ldots, M$ component scores or latent variables

$\boldsymbol{H}_m$ $= \left(h_{l^{(g)}l^{(q)},m}\right)_{l^{(g)}l^{(q)}} = \mathbf{X}_m'\mathbf{Y}_m$ for $m = 1, \ldots, M$

$\boldsymbol{g}_m$ $= \left(g_{l^{(g)},m}\right)_{l^{(g)}}$ for $m = 1, \ldots, M$

$\boldsymbol{q}_m$ $= \left(q_{l^{(q)},m}\right)_{l^{(q)}}$ for $m = 1, \ldots, M$

$L_{\lambda_1 m}(\boldsymbol{g}_m)$ $= \sum_{l^{(g)}=1}^{G} 2\lambda_1^m \left|g_{l^{(g)},m}\right|$ penalize the loadings $\boldsymbol{g}_m$

$L_{\lambda_2 m}(\boldsymbol{q}_m)$ $= \sum_{l^{(q)}=1}^{Q} 2\lambda_2^m \left|q_{l^{(q)},m}\right|$ penalize the loadings $\boldsymbol{q}_m$

The algorithms of the SPLS can use either the NIPALS or the SIMPLS but in the case of weak signals of the relevant variables the SPLS-NIPALS is known to choose the correct set of relevant variables (Chun and Keles 2010). As a result, the SPLS produce correlated (Liquet et al. 2012) variables from both the explanatory $\mathbf{X}$ and response $\mathbf{Y}$ matrices that are indicated in the sparse (Liquet et al. 2016) loading vectors.

Model validation

Although the ordinary least squares fit the observed data better than the PLS regression, they have a tendency of overfitting (Hawkins 2004) the observed data especially in the presence of many covariates. It has been reiterated that the quality of data should not dictate the number of extracted factors. The optimal factor selection should be rather based on how well the model fits the observations that have not been involved in the model selection (validation) using fewer underlying factors from the original explanatory matrix (SAS Institute Inc 2009).

A portion of the data is selected (*training set*) to build the model whilst the remaining part (*test set*) is held out for measuring the performance of the model from the training set. The training or calibration set is usually taken to be two-thirds of the original data set (Borovicka et al. 2012). However, this test set validation is useful when there is enough data to partition. When the data set happens to be small to make sizable sets, more than one divisions of the data set can be made to obtain several training data sets and test data sets, a process called cross-validation. Depending on the nature of the data, the test set can be held out as one observation at a time, leave-one-out, successive blocks (blocked validation), successive groups of widely separated observations (split-sample) or groups of randomly sampled observations (random sample cross validation). Although the blocked and split-samples are not computationally intensive as the leave-one-out, their application is of limited use in the face of serially correlated data (SAS Institute Inc 2009). The random sample also employs the blocking system in a way and in addition, the observation mix in the blocks is improved to avoid the chances of highly correlated observations in a single block. However, the same seed is required for different researchers to arrive at the same results. The SPLS uses the leave-one-out approach with an option to split the data for a repeated number of times (Lê Cao et al. 2008).

Selection of the optimal number of components

This applies the principle of parsimony (Vandekerckhove et al. 2014) where the preferred model is the one with the best explanation obtained from the simplest model with a fewer number of covariates. Irrespective of the validation method, the optimal number of factors are the ones that minimise the predicted residual sum of squares (PRESS) which is given by: $PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Some other applications use a scaled down version of PRESS called the root mean square error (RMSE) or root mean PRESS is given by $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ and essentially gives the same results. The PRESS will not necessarily give the optimal number of factors to be selected as its statistics may be marginally larger than the absolute minimum and hence different models are compared (van der Voet 1994). The marginal contribution of each principal component in the case of the regression context is given by

$$Q_m^2 = 1 - \frac{\sum_{k=1}^{p} PRESS_{km}}{\sum_{k=1}^{p} RSS_{k(m-1)}} \geq \left(1 - 0.95^2\right) = 0.0975. \qquad (3.13)$$

Hence, a component is considered to be significantly contributing to the prediction if $Q_m^2 \geq 0.0975$ (Lê Cao et al. 2008).

Variable selection in SPLS

Not all the variables are important in the explanation of the response variable and as such the most influential ones in the model building are selected (Chong and Jun 2005). Variable Importance in the Projection (VIP) and the beta coefficients are useful techniques in determining the explanatory variables to include in the SPLS regression model. For the k^{th} covariate, whose weight to the m^{th} principal component is φ_{km}, its x_k VIP score is given by $VIP_k = \sqrt{\frac{M\sum_m \varphi_{km}^2 b_m^2 \mathbf{T}_m'\mathbf{T}_m}{\sum_m b_m^2 \mathbf{T}_m'\mathbf{T}_m}}$ where b_m is the regression coefficient of the inner relationship of the score matrices for the m^{th} principal component. The higher the VIP score the more influential it is in the SPLSR model for both covariates and responses. The cut-off point varies with literature and some of the suggestions are 0.8 (Sawatsky et al. 2015), 1.0 (Chong and Jun 2005) and 1.5 (Jun et al. 2009). It is recommended that the regression coefficients for the

centred and scaled data with mean 0 and standard deviation of 1 be used for variable selection (Cox and Gaudard 2013) as this allows all the coefficients to be compared on an equal footing. To reach a compromise between the VIP > 1.0 rule, say and that of the corresponding standardised regression coefficients that are far from zero, a scatter plot of regression weights against VIP can be used. Depending on the possibility of the regression coefficients to be influential, the scatter plot can be used to allow for the flexibility to err and retain the variables whose VIP is slightly lower than the chosen cut off. However, covariates with low VIP and small regression coefficients are confidently removed (Sawatsky et al. 2015). The receiver operating curve and LASSO can also be used to assess the performance of the SPLS regression coefficients in selecting the relevant covariates (Palermo et al. 2009). However, since the regression coefficients of the standardised variables represent the prediction power of a covariate to the response, the closer the coefficient to zero the lesser it is important in variable selection. Hence, covariates with relatively small coefficients in absolute value are simply dropped (SAS Institute Inc 2014).

3.4 Data analysis for the variable selection approaches and software

All the analysis was performed in the open source R software, version 3.5.3. To determine the strongest covariates of the $CD4^+$ cell count for further investigation with the other proposed statistical models, the time-varying coefficient models were fitted using the **VariableScreening** package using the function `screenLD` assuming an AR(1) correlation structure. Currently, the package does not accommodate a time-varying grouping variable where each subject can take more than one factor level over time except for the baseline variables that are time-invariant per subject. Hence, the models were fitted for each phase as well as the overall data. The variable screening was set to model cubic splines (`degree = 3`) and the degrees of freedom allowed to take the number of inner knots as `length(knots)`. The variable *time* was scaled in each case as recommended for the time-varying coefficient models. The penalised GEE was applied with the aid of the **PGEE** package using the function `PGEE` also assuming an AR(1) correlation structure. Prior to PGEE fitting, a 4-fold cross-validation was used to select an optimal tuning parameter lambda from an arbitrary range of 0.1 to 5 with increments of 0.2. Similarly, the authors of the **PGEE** package (Inan and Wang 2017) could not accommodate the time-varying categorical variable per subject such that the variable selection procedure was applied per each infection phase and the overall data as well. The **mixOmics** package for the multi-level omics data integration used the `spls` function for the SPLS approach to model building. The library **mixOmics** is capable of handling the

complex structure of repeated measurements without analysing the data per phase. The package also incorporates a design matrix to account for variation in the multilevel structure of the longitudinal data where the time points and the infection phases are considered as two factors in the design matrix. It handles multicollinearity and a very large number of variables in longitudinal data and ranks the covariates from strongest to weakest allowing variable selection and consequently dimension reduction. The leave-one-out cross validation was used as recommended by (Mevik and Cederkvist 2004, Olson et al. 2014) with 20 folds. Since the SPLS is a multidimensional analysis technique, graphical displays of the results were vital to comprehensively visualize the variable selection process with the aid of the instrumental R libraries: the **ggplot2** and **ggrepel**. All the R codes are presented in Appendix C3.

3.5 The results of variable selection

3.5.1 The results of variable selection with time-varying coefficient models

The results of the sum of squared errors and their ranks for both the overall and within infection phase variable selection are shown in Table 3.2. The covariate names in the first column are ranked top to bottom in order of strength in the marginal association with the $CD4^+$ cell count as suggested by the fitted model for the overall data. Although all the models ranked the covariates in order of their strength in the marginal association with the $CD4^+$ cell count, the voluminous nature of the results makes it challenging to compare the extent of the magnitude to which each covariate performed. To have a better picture of the comparative magnitude of the sum of squared errors, the results were graphically displayed. The overall variable selection is graphically displayed in Figure 3.2 showing that the lymphocytes had the strongest marginal association with the $CD4^+$ cell count. The lymphocytes' sum of squared errors was significantly smaller than any of the other competing covariates. Although the errors could provide some form of ranking, only the top 10 variables say, showed some noticeable significant differences in the magnitude of the errors. Hence, the sum of the squared errors approach seems not to be very sensitive in showing the extent to which the covariate marginal associations with the response differ.

Table 3.2: The results of variable screening based on the sum of squared errors

	2-Acute		3-Early		4-Est		5-ART		Overall	
Covariate, k	SSE, $\hat{u}_{nk}$	Rank	SSE, $\hat{u}_{nk}$	Rank	SSE, $\hat{u}_{nk}$	Rank	SSE, $\hat{u}_{nk}$	Rank	SSE, $\hat{u}_{nk}$	Rank
Lymphocytes	732.6976	1	616.4395	1	658.7154	1	637.0272	1	2633.6361	1
Haematocrit	915.2696	2	861.3563	2	898.4525	2	914.1520	3	3611.3872	2
Albumin	931.2192	8	887.2026	3	902.6692	3	938.4982	20	3666.3311	3
Monocytes	928.6214	7	900.3725	6	908.5251	5	933.9995	11	3680.3492	4
Basophils	937.0654	15	918.2028	8	927.1424	14	904.9222	2	3681.2811	5
MCV	933.7848	11	888.6040	4	926.5886	12	929.8569	8	3684.5670	6
Total protein	938.3947	18	915.2888	7	908.2053	4	933.5506	10	3702.5536	7
Platelet	923.5494	5	932.5760	14	912.2777	6	935.8100	15	3711.5482	8
MCHC	946.1970	37	891.9692	5	928.6709	17	943.0221	27	3718.0157	9
Triceps skin fold	923.0612	4	939.8723	27	926.4540	11	922.7305	5	3719.7242	10
AST(GOT)	928.2812	6	935.5530	21	926.2984	10	919.6229	4	3722.6686	11
ALP	946.0393	36	933.5793	15	928.6032	16	924.9192	6	3731.3513	12
Folate	937.1897	16	927.1688	13	916.4533	8	931.2381	9	3732.1292	13
Neutrophils	941.9480	21	924.1729	10	914.8815	7	943.1416	28	3732.2197	14
Waist circum	942.0426	22	938.4587	26	926.0138	9	925.7271	7	3741.8180	15
Potassium	942.3428	23	926.3916	12	929.2855	18	941.8951	24	3744.2961	16
RDW	938.0563	17	926.3556	11	926.6601	13	935.1467	14	3744.8056	17
Red blood cells	933.9083	12	935.1783	19	929.5359	19	936.5557	16	3749.0187	18
Eosinophils	918.2135	3	944.7860	40	946.9631	38	937.2756	18	3749.1474	19
Arm(right) circum	931.5624	10	934.5133	18	934.3717	23	934.6025	12	3750.1001	20
BMI	931.5485	9	933.7916	16	933.4231	21	937.6208	19	3750.9169	21
LDL	945.2458	33	944.2675	38	937.1565	25	935.1061	13	3762.6372	22
Fe(Iron)	946.7302	38	921.4980	9	936.8767	24	943.0150	26	3763.9456	23
Glucose	944.8873	31	943.9330	36	927.7222	15	945.9707	34	3763.9520	24
BP(systolic)	934.6347	13	935.6373	22	944.4534	33	941.0109	23	3767.3377	25
Magnesium	943.3971	27	935.4538	20	940.6982	27	944.3100	31	3770.6807	26
Height	942.5940	24	937.7285	24	943.6759	31	946.5425	38	3770.7244	27
Calcium	936.7913	14	936.7317	23	943.2199	30	947.5185	40	3772.6487	28
Axillary Temp	942.9761	25	940.8773	30	930.7408	20	946.8891	39	3773.8225	29
Bilirubin	942.9922	26	934.1208	17	943.9164	32	944.0395	30	3774.1126	30
Sodium	943.8420	29	940.5834	29	943.2188	29	943.2799	29	3776.8810	31
Triglycerides	947.7712	40	944.0616	37	945.0930	35	940.6276	22	3777.3231	32
GGT	941.1627	20	944.5341	39	933.7655	22	940.0874	21	3778.5258	33
BP(diastolic)	945.1784	32	938.3865	25	945.7776	37	945.0255	33	3780.6964	34
ALT(GPT)	943.4429	28	942.5043	33	945.6974	36	936.9974	17	3780.8513	35
Urea	945.3904	34	941.3528	31	947.3040	39	942.5917	25	3781.5653	36
LDH	940.6826	19	940.1424	28	940.8591	28	946.4028	36	3782.0402	37
Vitamin B12	947.1212	39	942.2840	32	940.5432	26	946.2196	35	3783.9795	38
Chloride	945.7649	35	942.7814	34	947.4555	40	946.5015	37	3786.2738	39
Pulse	944.8472	30	942.9799	35	944.8055	34	944.6024	32	3787.2178	40

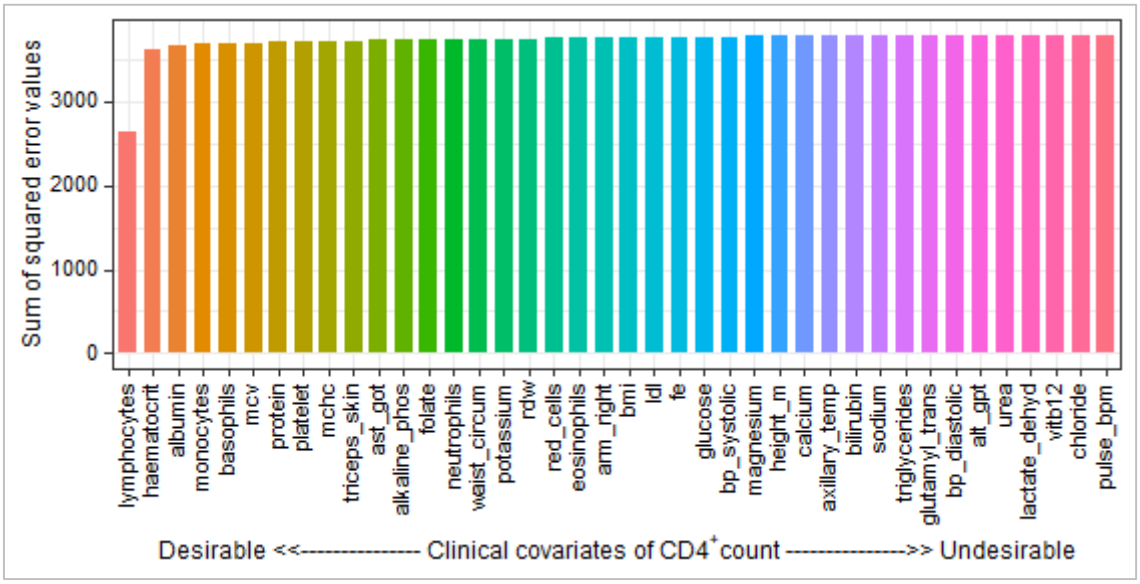


Figure 3.2: The visual display of the overall variable selection using the sum of squared errors

The results of variable screening based on the sum of squared errors are presented in Table 3.2. These results revealed that the covariates were not consistently ranked in the same position across the different phases. Again, a comparative analysis of the covariate ranking positions across the phases was not easy without some visual aids. Figure 3.3 shows a graphical display of how the variables performed rank-wise from one phase to the other as well as from an overall point of view. Generally, the covariates that seem to have indicated the strongest marginal association with the $CD4^+$ cell count during the entire period of the disease progression were also the strongest within the phases. Lymphocytes were consistent throughout as the highly ranked covariate to have the strongest marginal association with the $CD4^+$ cell count across all the HIV infection phases. Similarly, all the other covariates that performed better are characterised by ranking positions that appeared in the top left hand corner of Figure 3.3.

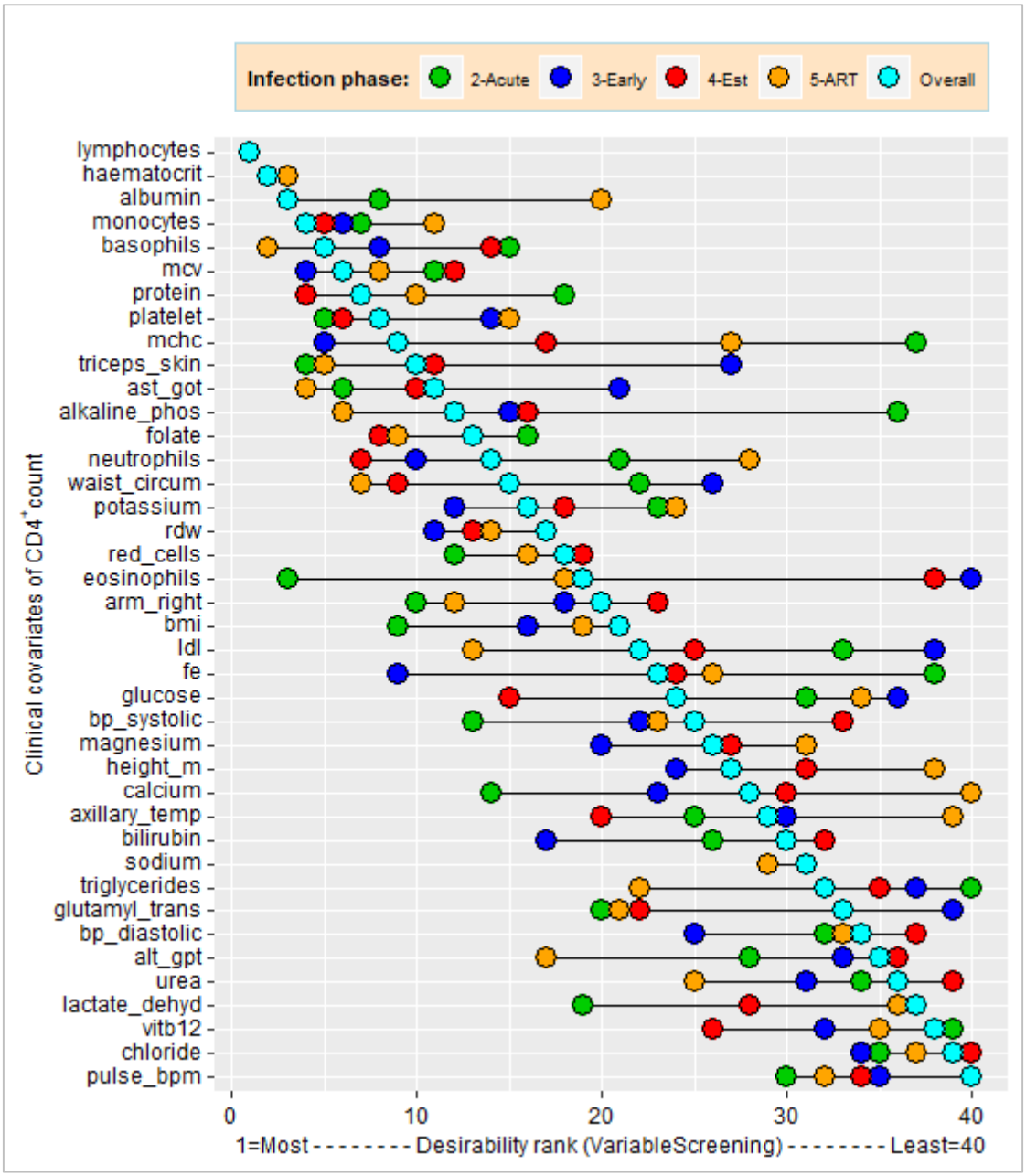


Figure 3.3: Comparison of variable importance by infection phase using the sum of squared errors

The closer are the ranking positions for each covariate is the more consistent the covariate in having either a poor or strong marginal association with the $CD4^+$ cell count. For example, in addition to the lymphocytes, the haematocrit, monocytes, basophils, MCV, total protein, platelet count and folate were among those in the top left hand corner and also having ranking positions in close proximity. They indicated some consistency in having the strongest marginal associations with the $CD4^+$ cell count and consequently candidates for variable selection. The ranking positions of the covariates like Vitamin B12 and pulse appeared in the bottom right hand corner and also in close proximity showing their consistency in providing higher errors and hence candidates for exclusion from subsequent model fitting.

However, there were some covariates that revealed some extreme cases of ranking positions in a particular phase only showing either a very good or bad performance in that phase. Taking, for example, the eosinophils, they presented themselves to have a strong marginal association with the $CD4^+$ cell count during the high viral load (acute phase) but being a candidate for exclusion during the other phases. On the other hand, albumin and ALP showed weak marginal associations with the $CD4^+$ count only during the uptake of medication and the acute phase respectively.

3.5.2 The results of variable selection with penalised GEE

Unlike the time-varying coefficient models that simply rank the covariates and leaving up to the discretion of the researcher to determine the number to select, the PGEE comes with a cut-off point on the number to select. Table 3.3 shows the number of selected covariates at each infection phase together with the corresponding model tuning parameters to achieve the results. Generally, all the other infection phases selected fairly the same number of covariates between 12 and 17 except during the early phase where 34 out of 40 covariates were considered relevant explanatory variables for the $CD4^+$ cell count.

Table 3.3: A summary of variable selection using the PGEE

	Infection phase				
	2-Acute	3-Early	4-Est	5-ART	Overall
Tuning parameter, λ	4.7	1.5	4.9	4.9	4.9
Number of selected covariates (out of 40)	17	34	12	14	12

The penalised regression estimates for each phase and the overall data are presented in Table 3.4. The covariates are also in the order they were ranked top to bottom based on the variable selection in the overall data. It is evident from the results that the estimates are either very large or very close zero as expected. Taking, for example, the highly ranked (Rank 1) lymphocytes had the largest penalised estimates of at least 118 in all the phases whereas the least ranked (Rank 40) covariates had penalised estimates of at most $1.1x10^{-17}$. These covariate ranks based on the absolute penalised estimates revealed that the lymphocytes were the most important covariate to consistently better explain the $CD4^+$ cell count within each infection phase.

Table 3.4: The results of variable selection based on the penalized estimates

Covariate, k	2-Acute		3-Early		4-Est		5-ART		Overall	
	Penalised Estimate, $U(\beta_k)$	Rank	Penalised Estimate, $U(\beta_k)$	Rank	Penalised Estimate, $U(\beta_k)$	Rank	Penalised Estimate, $U(\beta_k)$	Rank	Penalised Estimate, $U(\beta_k)$	Rank
Lymphocytes	118.31472	1	141.076	1	128.3174	1	143.233	1	136.582	1
Total protein	-34.8871	4	-44.548	2	-37.1036	3	-21.193	10	-38.568	2
Monocytes	-22.34633	10	-27.551	8	-36.2003	4	-41.749	3	-34.876	3
Albumin	30.236124	6	38.6804	4	38.57276	2	-3E-19	36	32.7918	4
MCV	43.554001	2	26.4756	9	28.89126	7	44.5182	2	29.2119	5
Folate	-19.39606	13	-24.197	10	-30.7003	5	-29.423	5	-26.571	6
Neutrophils	19.100556	14	28.304	7	27.82752	8	18.4181	12	23.5427	7
Haematocrit	1.277E-18	19	32.2071	5	4.11E-18	22	-2E-19	38	22.279	8
Platelet	22.730344	9	21.1084	11	29.8699	6	6.4E-19	29	21.4346	9
ALP	-5.53E-19	30	11.5329	20	19.7541	10	28.4424	6	18.8533	10
RDW	-8.27E-19	28	11.0495	21	-8.3E-18	12	2.2E-19	37	-7E-28	11
LDL	1.193E-19	37	-8.371	28	3.9E-18	23	-6E-19	30	3E-28	12
Arm(right) circum	1.014E-20	40	-12.083	18	1.11E-18	35	19.3014	11	1.8E-28	13
Red blood cells	1.067E-18	23	-28.315	6	2.54E-18	26	27.3392	8	-2E-28	14
Triceps skin fold	30.023151	7	5.9785	31	2.6E-18	25	1.7E-19	39	-1E-28	15
Triglycerides	3.789E-19	33	-16.759	13	-1.2E-17	11	9.4E-19	21	8.8E-29	16
Glucose	-1.21E-18	20	-9E-17	33	-4.7E-18	20	6.6E-19	28	7.9E-29	17
Eosinophils	18.90892	15	6.30214	30	-1.7E-18	32	-8E-19	23	7.8E-29	18
Calcium	-2.95E-20	39	-4E-17	36	-2E-18	30	-8E-19	25	-7E-29	19
Bilirubin	1.565E-18	17	5.75172	32	6.57E-18	14	-1E-18	20	-6E-29	20
Basophils	-1.38E-18	18	-6E-17	34	-2.1E-18	29	1.6E-18	16	-6E-29	21
Sodium	-1.98E-19	36	-8.4326	27	5.07E-19	39	-4E-19	35	-5E-29	22
MCHC	-1E-18	24	10.3136	23	3.78E-21	40	-1E-18	19	-3E-29	23
AST(GOT)	-3.43E-19	34	-10.088	24	-3.9E-18	24	-1E-18	18	2.8E-29	24
Magnesium	5.183E-19	31	13.5257	16	2.41E-18	27	6.1E-20	40	2.7E-29	25
BP(diastolic)	-7.13E-19	29	1.5E-17	39	-1.9E-18	31	-2E-18	13	-3E-29	26
Potassium	4.197E-19	32	3.8E-17	35	5.71E-18	15	4.5E-19	33	-2E-29	27
LDH	1.783E-18	16	14.8586	15	5.44E-18	17	1.3E-18	17	2.2E-29	28
Waist circum	-33.58308	5	-10.404	22	7.3E-19	38	27.6597	7	1.7E-29	29
Vitamin B12	8.336E-19	27	15.8768	14	7.89E-18	13	8.1E-19	22	-2E-29	30
Height	25.173145	8	20.6395	12	-1.6E-18	33	7.2E-19	26	-1E-29	31
BP(systolic)	21.575894	11	1.1E-17	40	1.31E-18	34	-2E-18	15	1.3E-29	32
Pulse	-3.24E-19	35	2.8E-17	38	-7.3E-19	37	-2E-18	14	1.3E-29	33
Fe(Iron)	3.452E-20	38	9.63578	25	5.34E-18	18	5.3E-19	31	7.9E-30	34
Urea	8.365E-19	26	-8.8753	26	-1.1E-18	36	-7E-19	27	-7E-30	35
GGT	-19.43521	12	-12.394	17	-5.6E-18	16	-21.495	9	6.3E-30	36
Axillary Temp	-9.99E-19	25	-7.6752	29	-2.2E-18	28	-4E-19	34	6.3E-30	37
BMI	41.242228	3	43.5471	3	20.546	9	-35.211	4	6.3E-30	38
Chloride	-1.08E-18	22	-12.027	19	-4.2E-18	21	-8E-19	24	5.5E-30	39
ALT(GPT)	-1.09E-18	21	-3E-17	37	-4.7E-18	19	-5E-19	32	-8E-31	40

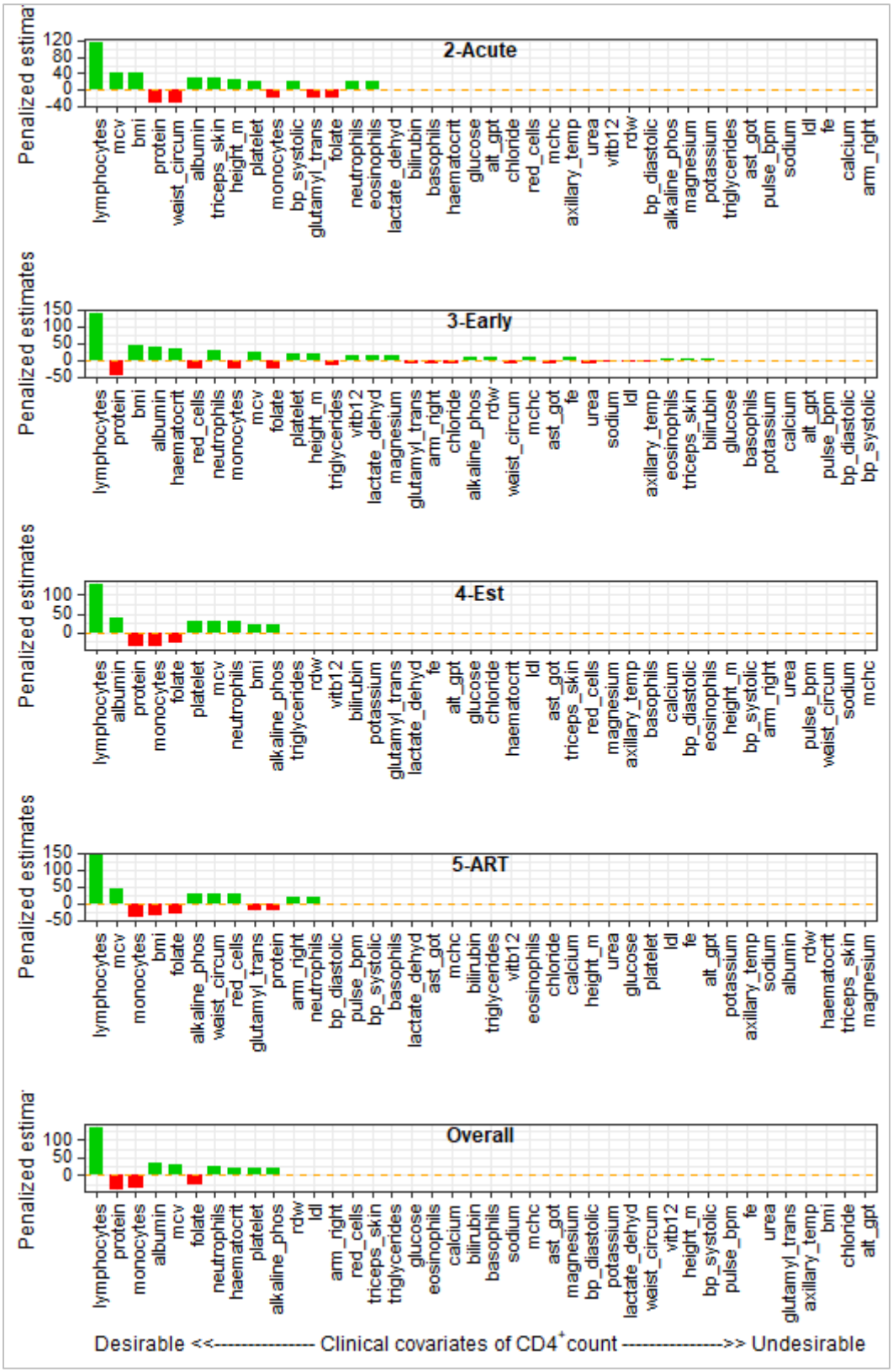


Figure 3.4: Variable importance and effects by infection phase based on the penalized estimates

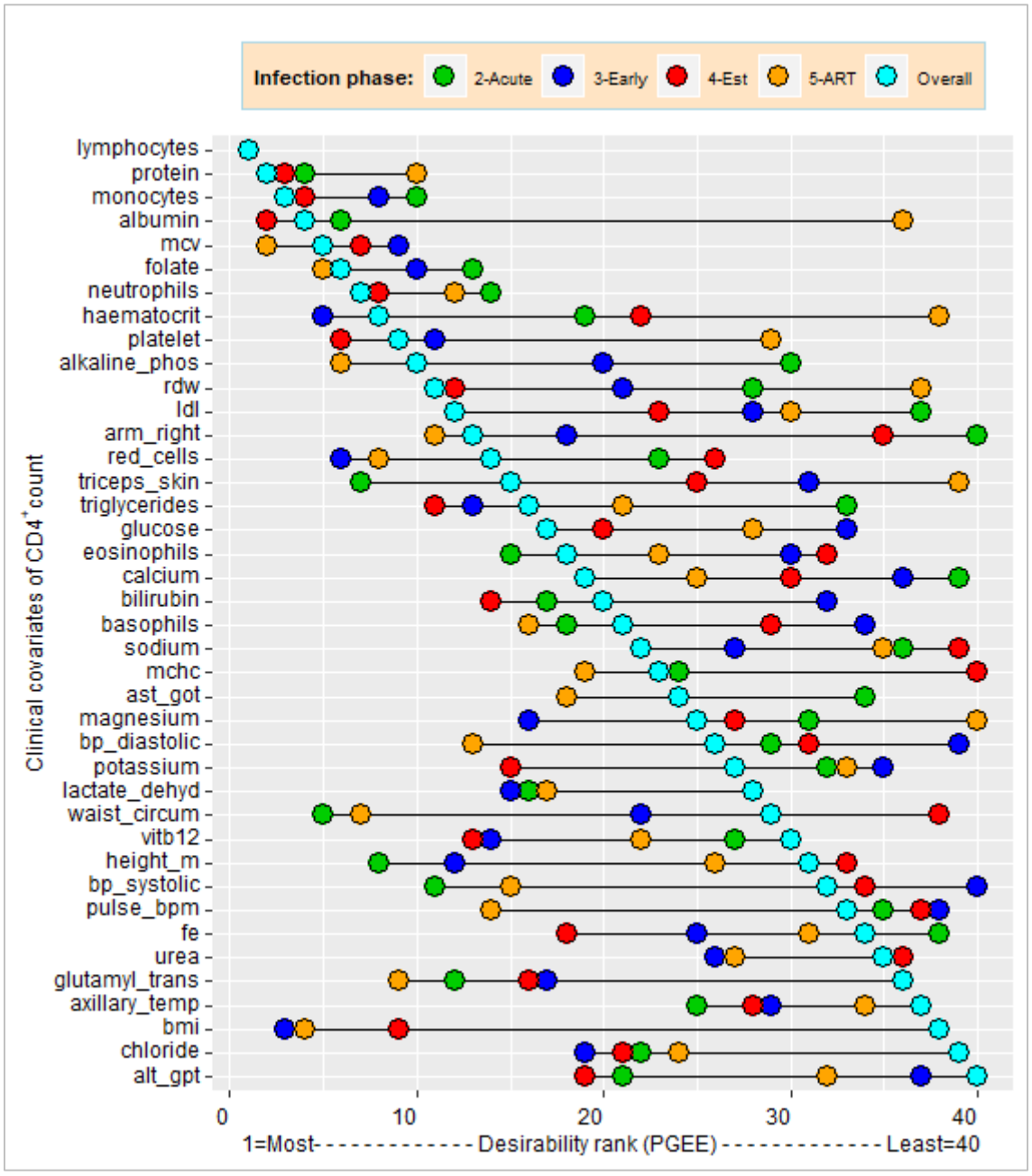


Figure 3.5: Comparison of variable importance by infection phase using the penalized estimates

In addition to the ranks, the PGEE results have the advantage of showing the effect of the selected variables as compared to the time-varying coefficient model's variable selection procedure which indicates the model fit due to each covariate. Figure 3.4 gives a visual display of both the effects of the selected covariates and the order of their importance in explaining the variation in the $CD4^+$ cell count.

Although they did not take the same rank across the phases, the covariates monocytes, protein and folate were consistently appearing in the variable selection process at each phase and the results also indicated that their increase negatively impacts on the $CD4^+$ cell count at all times.

A comparative analysis of the covariate ranks across the phases revealed that the importance of each covariate varied widely based on the PGEE (Figure 3.5) except for the highly ranked lymphocytes, protein, monocytes, MCV, folate and neutrophils. The covariates such as the axillary temperature and urea took fairly the same low ranks across the phases where they consistently showed poor performance in explaining the $CD4^+$ cell count. However, albumin, protein and haematocrit effect on the $CD4^+$ cell count was affected by treatment, the only period during which these covariates were ranked as unimportant.

3.5.3 The results of variable selection with multi-level omics data integration

The results of selecting the SPLS's optimal number of principal components

The model allowed the variation in the response to be captured by 40 components (the number of available covariates) and their performance measurements are shown in Table 3.5 and visually displayed in Figure 3.6 indicating that there was almost no deflation after the 5^{th} component. The results showed that only the first three components captured the variation in the response variable ($CD4^+$ cell count). The first component is characterised by explaining a 100% variance in the response but however suffers from poor fitting as indicated by a very high PRESS and very low R^2. An attempt to increase the R^2, decreased the marginal contribution of each principal component in the case of the regression context. Hence, the 2^{nd} component could only reach a compromise between the measures and was selected where it explained 68.95% of the variance in the response ($CD4^+$ count).

Table 3.5: Results of selecting the SPLS optimal number of principal components

Component, m	MSEP	PRESS	R^2	Q^2.Total	X.Expl.Var	Y.Expl.Var
1	0.6928	2625.7139	0.3070	0.3074	0.0815	1.0000
2	0.6128	2307.7032	0.3871	0.1124	0.0824	0.6858
3	0.5728	2151.0881	0.4271	0.0590	0.0558	0.6030
4	0.5602	2098.9020	0.4398	0.0124	0.0531	0.5606
5	0.5557	2078.6532	0.4443	-0.0024	0.0526	0.5470
6	0.5531	2078.7978	0.4468	-0.0113	0.0308	0.5422
7	0.5527	2071.3475	0.4472	-0.0124	0.0392	0.5397
8	0.5526	2080.5303	0.4474	-0.0177	0.0285	0.5392
9	0.5524	2098.7642	0.4475	-0.0270	0.0179	0.5390
10	0.5524	2077.9628	0.4475	-0.0171	0.0271	0.5389
11	0.5524	2070.1708	0.4475	-0.0134	0.0352	0.5389
12	0.5524	2076.8485	0.4475	-0.0167	0.0272	0.5389
13	0.5524	2082.6263	0.4475	-0.0195	0.0272	0.5389
14	0.5525	2107.1127	0.4475	-0.0315	0.0137	0.5388
15	0.5525	2105.8314	0.4474	-0.0309	0.0106	0.5388
16	0.5526	2074.0205	0.4474	-0.0154	0.0188	0.5388
17	0.5526	2069.7570	0.4474	-0.0134	0.0205	0.5388
18	0.5526	2069.2021	0.4474	-0.0131	0.0155	0.5388
19	0.5526	2067.3282	0.4474	-0.0122	0.0207	0.5388
20	0.5526	2066.2956	0.4474	-0.0117	0.0220	0.5388
21	0.5526	2065.4561	0.4474	-0.0113	0.0186	0.5388
22	0.5526	2064.2988	0.4474	-0.0107	0.0129	0.5388
23	0.5526	2063.3217	0.4474	-0.0102	0.0191	0.5388
24	0.5526	2062.2302	0.4474	-0.0097	0.0167	0.5388
25	0.5526	2061.2100	0.4474	-0.0092	0.0166	0.5388
26	0.5526	2059.9493	0.4474	-0.0086	0.0161	0.5388
27	0.5526	2058.9381	0.4474	-0.0081	0.0150	0.5388
28	0.5526	2057.8901	0.4474	-0.0076	0.0171	0.5388
29	0.5526	2056.6976	0.4474	-0.0070	0.0166	0.5388
30	0.5526	2055.6400	0.4474	-0.0065	0.0161	0.5388
31	0.5526	2054.5465	0.4474	-0.0059	0.0149	0.5388
32	0.5526	2053.3562	0.4474	-0.0053	0.0172	0.5388
33	0.5526	2052.1716	0.4474	-0.0048	0.0166	0.5388
34	0.5526	2050.9202	0.4474	-0.0041	0.0140	0.5388
35	0.5526	2049.7665	0.4474	-0.0036	0.0121	0.5388
36	0.5526	2048.6710	0.4474	-0.0030	0.0171	0.5388
37	0.5526	2047.3804	0.4474	-0.0024	0.0156	0.5388
38	0.5526	2046.1449	0.4474	-0.0018	0.0155	0.5388
39	0.5526	2044.8952	0.4474	-0.0012	0.0168	0.5388
40	0.5526	2043.6531	0.4474	-0.0006	0.0151	0.5388

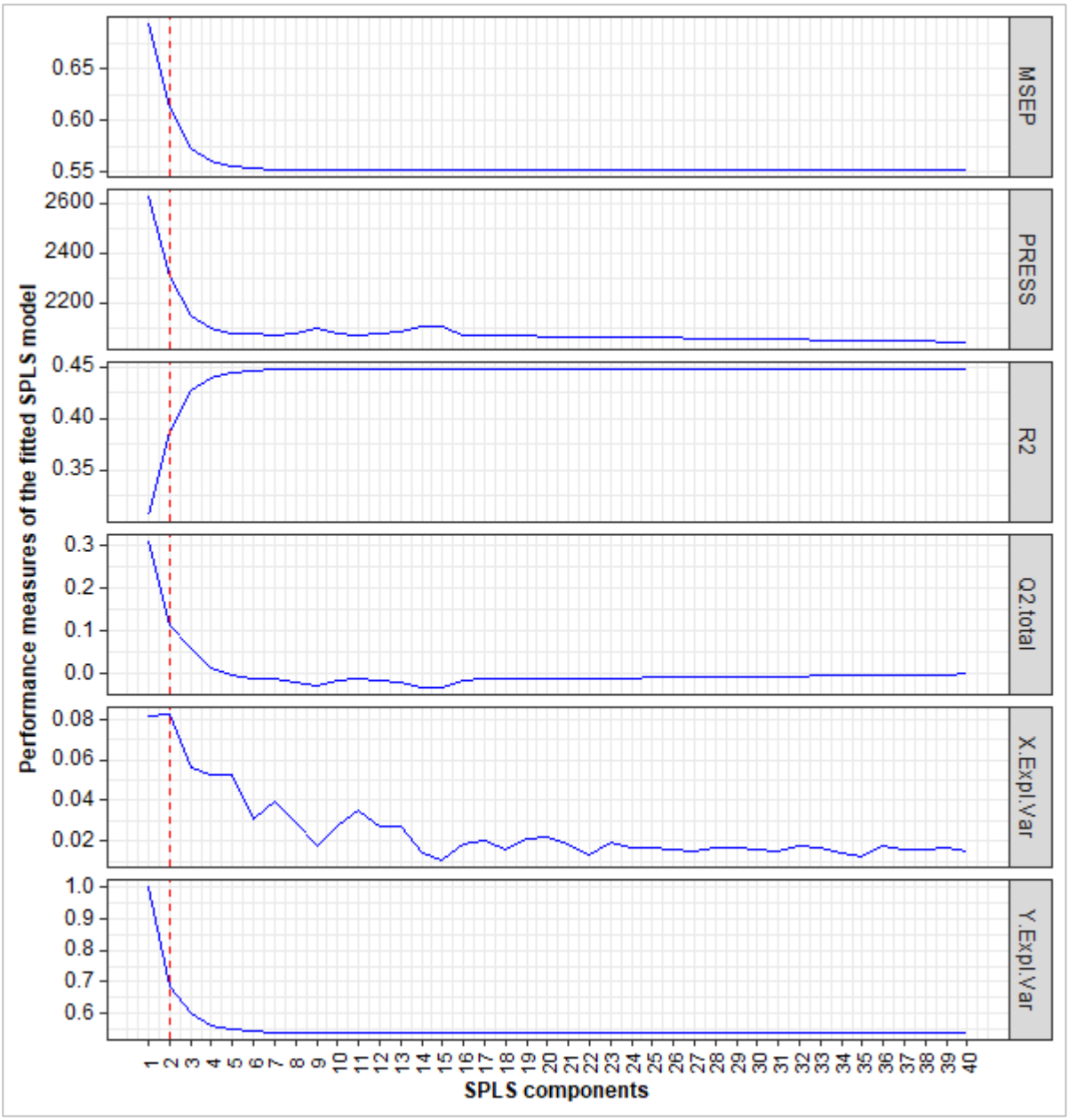


Figure 3.6: The visualisation of the SPLS model diagnostic performance measures. The amount of variation captured in the CD4⁺ count

The results of the SPLS variable selection

Table 3.6 shows the results of the variable selection based on the optimal 2nd component. The strongest $CD4^+$ count covariates based on the VIP only are shown in Figure 3.7. The variable selection by simultaneously considering both the VIP and the loadings as well as the VIP against the regression coefficients are shown in Figures 3.8 and 3.9 respectively.

Table 3.6: Results of variable selection based on the second component ($m = 2$)

Covariate, k	VIP	Loadings φ_{mk}	Coefficients $\beta_{pls}^{(k)}$	Correlations	
				r_k	p-value
Lymphocytes	3.5600	0.2891	0.0799	0.5421	0.0000
Basophils	1.8256	-0.1154	-0.1050	0.1353	0.0000
Albumin	1.4831	-0.0992	-0.1659	0.1694	0.0000
Haematocrit	1.4078	-0.0241	0.0495	0.2337	0.0000
ALP	1.3550	0.2003	0.2671	0.1154	0.0000
MCV	1.3474	0.2857	0.3178	0.1698	0.0000
Platelet	1.1072	0.0944	0.0922	0.1550	0.0000
Potassium	1.0856	-0.1724	-0.1570	0.1245	0.0000
Monocytes	1.0742	-0.1393	0.0376	0.1547	0.0000
Total protein	1.0604	0.0818	0.1587	-0.1740	0.0000
LDH	1.0310	0.2782	0.3498	0.0569	0.0005
Folate	1.0191	-0.3478	-0.3958	-0.1530	0.0000
Magnesium	1.0144	-0.2472	-0.2736	0.0724	0.0000
Glucose	0.9634	0.3335	0.1679	-0.1233	0.0000
Calcium	0.9428	-0.2665	-0.2985	0.0288	0.0762
MCHC	0.8859	0.0132	0.0968	0.0970	0.0000
Red blood cells	0.8579	-0.2173	-0.1659	0.1081	0.0000
Sodium	0.8555	-0.2325	-0.2944	0.0348	0.0321
Vitamin B12	0.6805	0.2107	0.2505	0.0167	0.3042
Triceps skin fold	0.6374	0.1755	0.2691	0.1369	0.0000
Triglycerides	0.6326	-0.0676	0.0996	0.0399	0.0140
Neutrophils	0.6097	0.0467	0.1166	0.1260	0.0000
AST(GOT)	0.5931	0.0289	0.0405	-0.1355	0.0000
Eosinophils	0.5799	-0.0300	-0.0070	0.0810	0.0000
Height	0.5519	-0.0069	0.0854	0.0606	0.0002
Chloride	0.5477	-0.1453	-0.2344	-0.0062	0.7049
Waist circum	0.4887	0.0721	0.2721	0.1428	0.0000
LDL	0.4548	-0.0978	0.0476	0.0786	0.0000
BMI	0.4082	0.0912	0.2299	0.1181	0.0000
BP(systolic)	0.3606	-0.0910	-0.0303	0.0540	0.0009
Bilirubin	0.3568	-0.0937	-0.0949	0.0467	0.0040
Arm(right) circum	0.2988	0.0191	0.1812	0.1231	0.0000
Fe(Iron)	0.2842	0.0692	0.1616	0.0711	0.0000
GGT	0.2745	-0.0498	-0.0415	-0.0601	0.0002
BP(diastolic)	0.2586	-0.0527	-0.0028	0.0187	0.2499
RDW	0.2348	-0.0412	-0.2120	-0.1332	0.0000
Pulse	0.1842	0.0547	0.0989	-0.0178	0.2729
Urea	0.1841	0.0167	0.0927	0.0232	0.1537
ALT(GPT)	0.1570	0.0324	0.0672	-0.0349	0.0315
Axillary Temp	0.1193	-0.0401	-0.0475	-0.0629	0.0001

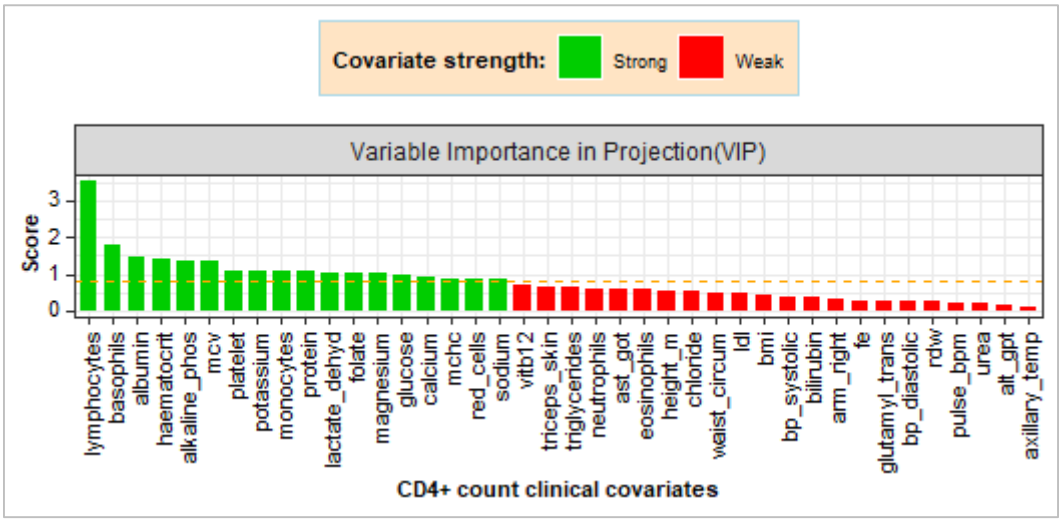


Figure 3.7: The strongest CD4⁺ count covariates based on VIP only.

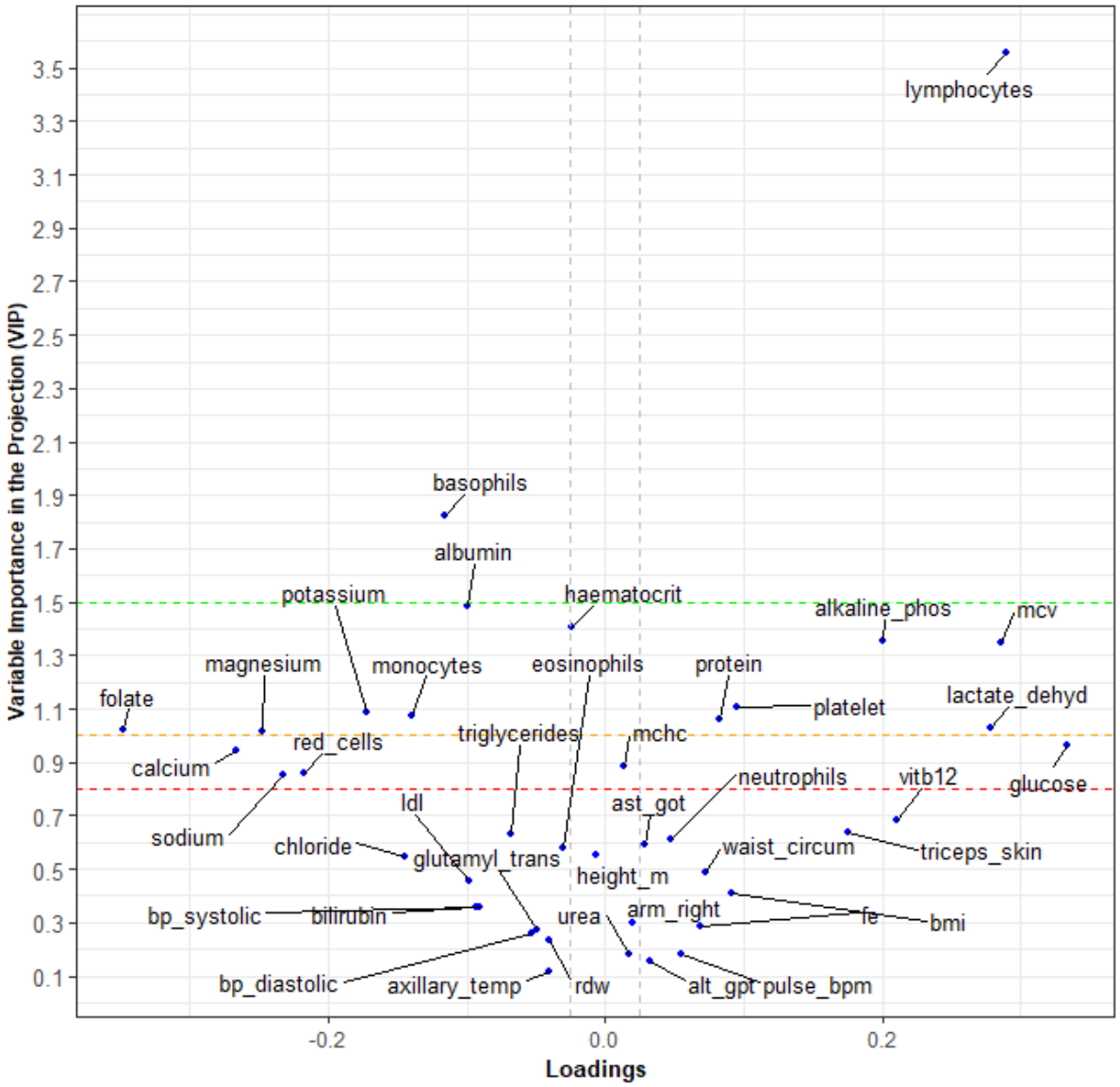


Figure 3.8: The scatter plot of VIP against the loadings

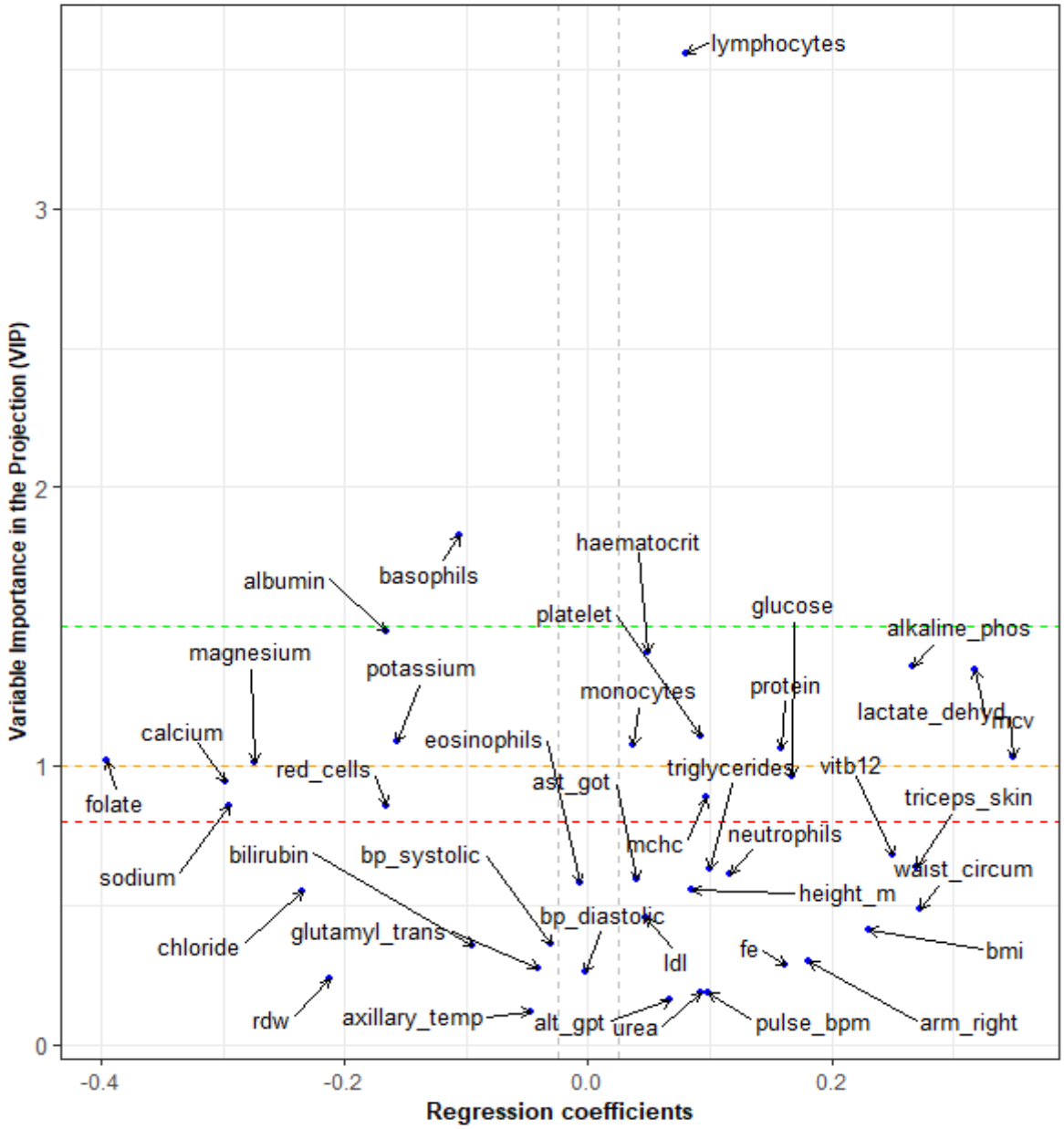


Figure 3.9: The scatter plot of VIP against the regression coefficients.
All three VIP cut off points are shown 0.8 (red), 1.0 (orange) and 1.5 (green). The vertical broken grey lines indicate $\left|\beta_{pls}^{(k)}\right| < 0.025$ *, the region within which the variables have less explanatory power even if they have a higher VIP score.*

All the three VIP cut off points suggested in the literature were presented where a cut-off point of 1.5 can be considered as a strict selection, 1.0 being moderate and the 0.8 as being lenient. A stricter variable selection process selected two covariates, moderate (13) and the lenient (18) out the 40 non-redundant features available for the study. We developed an interest in all the 18 strongest covariates as selected by the lenient cut off point as shown in Figure 3.7. Figure 3.10 provides a list of all the 40 covariates from the strongest to the weakest significance as well as their behavioural patterns in the explanatory power (coefficients), component

construction (loadings) and independent association (correlation) with the $CD4^+$ count together with the associated p-values.

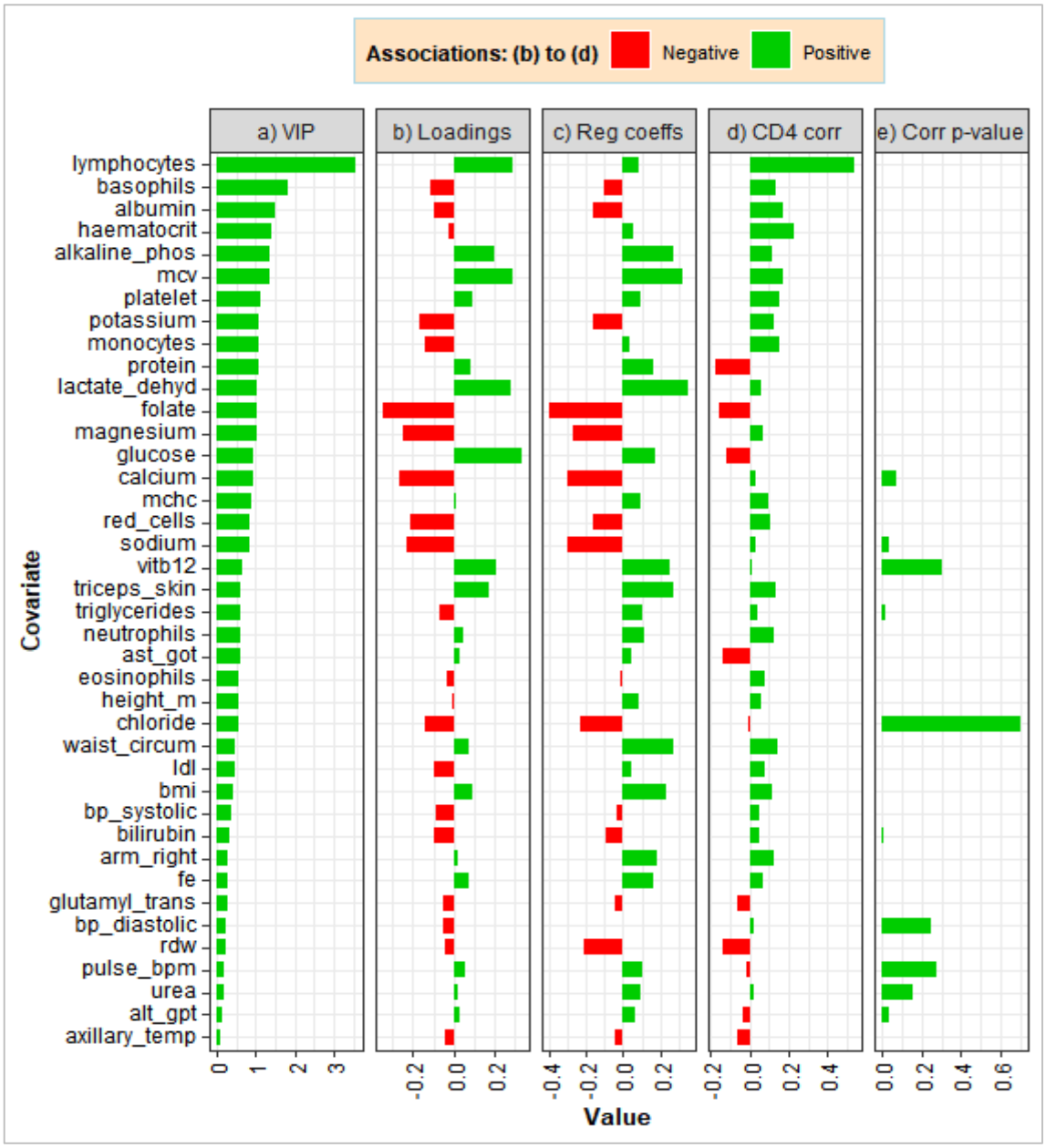


Figure 3.10: Variable importance.
Also shown are the related loadings, standardised regression coefficients and the correlations of each covariate with the response variable ($CD4^+$ count).

In this study, the lymphocytes had the highest direct independent positive correlation with the $CD4^+$ (r = 0.5421, p-value <0.0001) followed by haematocrit (r = 0.2337, p-value <0.0001). On the other hand, protein had the highest negative correlation (r = -0.1740, p-value <0.0001) with the $CD4^+$ count followed by folate (r = -0.1530, p-value <0.0001). The results showed that the top eight of the 18 selected covariates were positively and independently associated with the $CD4^+$ count. Out of all the investigated 40 non-redundant covariates, red blood cell

distribution width (RDW), pulse, urea, Alanine Aminotransferase (Glutamate Pyruvate Transaminase) ALT(GPT) and axillary temperature were the least important. A look at the significant variables by clinical category is shown in Figure 3.11.

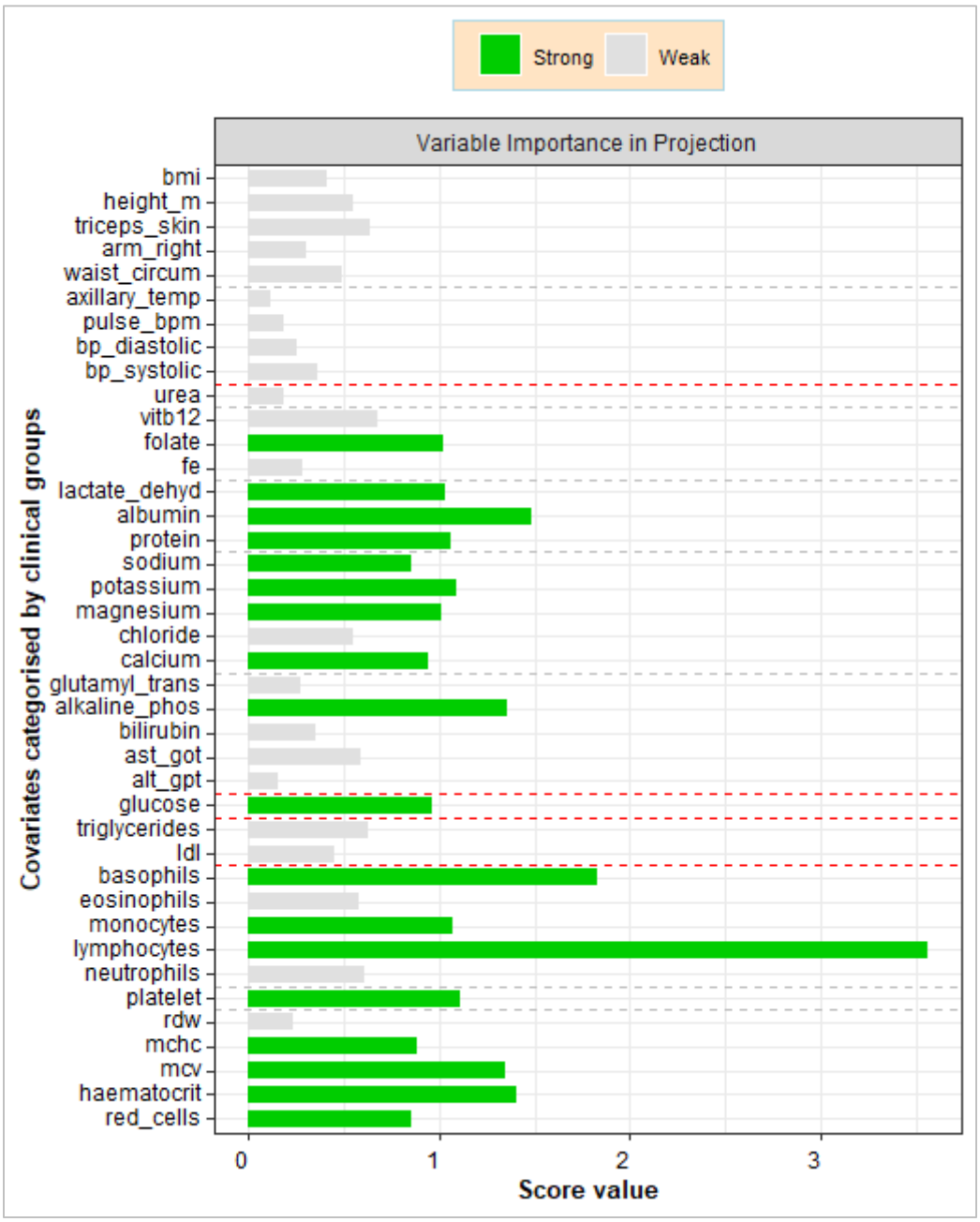


Figure 3.11: Variable importance by clinical category.
The broken red horizontal lines divide the major groups. From the top, the groups are clinical examination, blood chemistry, sugar, lipids and full blood count. The horizontal broken grey lines divide the major groups

The results revealed that there was no significant variable selected from lipids, physical examination and anthropometric measurements. Folate was the only significant variable in its category and similarly, alkaline phosphatase only among the liver function indicators. The SPLS suggested chloride and RDW as the only insignificant $CD4^+$ count covariates among the

electrolytes and red blood cells respectively. Given the lymphocytes, basophils and monocytes, the significant covariates within the white blood cells group, the lymphocytes were dominantly significant. Generally, most of the significant $CD4^+$ count covariates were selected from electrolytes, proteins and red blood cells.

3.5.4 A comparison of the results of the variable selection methods

The approach of the time-varying coefficient models allowed a single variable into the model and then captured the model's goodness of fit based on the sum of squared errors. The technique does not explicitly provide a cut off point for the number of covariates to be selected. Nevertheless, the underlying B-spline modelling technique is data driven which gave the technique a competitive edge. Both the time-varying coefficient models and the PGEE provided an insight into the variable importance within each phase. The PGEE is model driven which may not be as powerful as the time-varying coefficient models but explicitly indicate both the cut off point for the variable selection as well as the variable importance rank. The four infection phases involved in our study was a fairly small number of factor levels where different models could be fitted to each level. However, in practice, such levels could be exceedingly high resulting in a complex data structure that might make the separate models difficult to work with or derive meaningful information for variable selection. With the SPLS design matrix, this is easily achievable where the important variables across all the factor levels are filtered out. In addition, the SPLS explains the variation in both the covariates and response. Hence, the variable selection based on the SPLS was adopted for identifying the most salient clinical covariates.

3.6 Clinical interpretation of the SPLS model results

In this present chapter, we evaluated a list of $CD4^+$ count clinical covariates that were available at CAPRISA to determine the strongest candidates that can potentially become an important integral part of the HIV treatment process. The intention of this chapter was to select the clinical covariates that contributed to the greatest variation in the $CD4^+$ count from an overall perspective after seroconversion. The evaluated covariates were already known to be associated with the $CD4^+$ count based on other statistical methods that were limited in some way or suffered from information loss due to grouping and details given in the introduction section. The predictive nature of the selected covariates was beyond the scope of this chapter as our focus was on variable selection yet paving the way for other statistical modelling

techniques with streamlined and richer clinical covariates of the $CD4^+$ count. We hereby provide a brief summary of the functions of the selected and strongest 18 (out of 46) covariates according to our SPLS model. This will serve to point out the direction in which the next models can explore in assessing the feasibility of incorporating the clinical covariates in the HIV treatment process for influencing long-term $CD4^+$ cell response. On our list of the selected clinical covariates, the lymphocytes were as expected to be the strongest because the $CD4^+$ cells are a T cell type (Papagno et al. 2004) whereas the lymphocytes are either B or T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012). Our results also showed the lymphocytes to have the highest independent positive correlation with the $CD4^+$ count ($r = 0.5421$, p-value <0.0001). Hence, efforts to improve the $CD4^+$ cell response seem to be similar to that of the lymphocytes and the results obtained hereby serves to give an assurance of the effectiveness of our statistical methodology. In light of the other selected variables, our results pointed out to pay much attention to the white blood cells (basophils and monocytes) and platelet count. Basophils and monocytes control damage to body tissues and inflammation, and fight pathogens respectively (Project Inform 2007). Platelet count measures the blood clotting condition (Project Inform 2007, NAM 2012, National Institutes of Health Clinical Center 2015, James 2017). Although they are the least abundant leucocytes (Min et al. 2011), our study has found basophils to explain the greatest variation in the $CD4^+$ count following the lymphocytes. However, direct contact between human basophils and $CD4^+$ T cells is known to mediate viral trans-infection of T cells through the formation of viral synapses (Jiang et al. 2015, Marone et al. 2016). Also, the presence of basophils and other white blood cells in the blood is affected by underlying infection (Siracusa et al. 2013). Areas of potential consideration in the blood chemistry group included the potassium, sodium, calcium, magnesium, ALP and folate. Potassium regulates the acid-base chemistry and water balance (The Johns Hopkins Lupus Center 2017), nerve impulses and heart muscle (Project Inform 2007, James 2017). Potassium effect on the $CD4^+$ count is affected by underlying comorbidities (Collins et al. 2017). Sodium and calcium regulate water balance, blood pressure, blood volume, heart rhythm and most importantly the brain and nerve function (Project Inform 2007, James 2017, The Johns Hopkins Lupus Center 2017). Changes in the sodium concentration are known to create an osmotic gradient between extracellular and intracellular fluid in cells (Shu et al. 2018) suggesting that a proper balance is essential. Magnesium is involved in muscle contractions and protein processing (Project Inform 2007), ALP in detecting liver health (Whitfield 2001, Beare et al. 2008, Patil et al. 2013) and folate for cell growth and metabolism (Arya and Kumar 2012, Dieticians of Canada 2014). Red blood cells indices (haematocrit, MCV, mean corpuscular

haemoglobin concentration (MCHC) and red blood cells) are related to haemoglobin (Junqueira et al. 2006) which binds oxygen for transport to tissues and binds tissue carbon dioxide for transporting it back for exhalation (Jensen et al. 1998, Wintrobe and Greer 2009). The indices indicate volume, concentration and proportions of the red blood cells (Wintrobe and Greer 2009, Arika et al. 2016). In line with the red blood cell indices, our results revealed that LDH also needed attention in the $CD4^+$ cell influence. LDH is a cytosolic enzyme, for enabling the fulfilment of the short-term energy requirements in the absence of sufficient oxygen at the expense of a greater consumption of glucose cells (Valvona et al. 2016). Proteins (total protein, albumin and LDH) were included in the selected list for the maintenance of normal water distribution between tissues and blood, as well as acid-base balance (Spectrum 2007).

3.7 Summary

Of the three techniques evaluated for variable selection in high-dimensional longitudinal data, the time-varying coefficient models had the advantage of being data driven whereas the model-driven PGEE provided both the covariate effects and the actual cut-off number of the covariates to be selected. The two approaches can be helpful if variable selection within a given factor levels is of interest. However, with an increasing number of such levels, it becomes a tedious process and the SPLS is recommended for it incorporates the levels as part of the longitudinal design matrix. This also allows for variable selection that are important across all the levels of a given factor. Hence, the results of the SPLS proved to be effective in paving the way for further investigation of the clinical covariates on the $CD4^+$ cell response. The data exploration revealed the existence of variation in the repeated measurements between patients. This suffices to be a point of departure in investigating inter-individual $CD4^+$ cell count variation in response to each of the selected few strongest covariates. Hence, the next chapter deals with the application of longitudinal multilevel models as the most suitable method to handle the task.

CHAPTER 4 LONGITUDINAL MULTILEVEL MODELS

Despite the existence of measurement reliabilities (Bajpai and Bajpai 2014, Mohajan 2017) in recording patient information, repeated measurements from the same individual are bound to vary and so are those from different subjects. Irrespective of these variations, the healthcare fraternity generally administer an average dose of medication to patients who are likely to have differences in either the body tolerances, preferred medical treatment or specific needs. There ought to be some medical measurements that are likely to remain fairly the same across patients whilst others greatly fluctuating to bring about the individual or time uniqueness. The strongest clinical covariates selected in Chapter 3 were believed to be no exception in differently influencing the $CD4^+$ cell count between the HIV positive patients. Among them, should be some whose influential effects bring about or induce wide variations in the $CD4^+$ count between the patients. Identifying these clinical covariates will give an insight on how to streamline the management of the HIV disease using a tailor-made patient care for influencing the $CD4^+$ cell count.

Given this task, the regular regressions are not suitable for they treat the within-subject regression as one data dataset for independent subjects leading to Type I error due to sample size dependence when testing the statistical significance of the results (Huta 2014). Methods that account for the longitudinal studies produce more efficient estimators than cross-sectional designs and also provide information about individual change (Hedeker 2004, Hedeker and Gibbons 2006). In our data, we intend to capture the different sources of variation at individual level and then assess how these vary between the individuals. Analysis of variance has been the conventional analysis method for repeated measurements but suffers from sphericity assumption, design effect or sampling hierarchy and the requirement for complete designs and datasets (Quene and van den Bergh 2004). The other approaches to longitudinal analysis include covariance pattern models, generalised estimation equations models and transition models, but lack the ability to capture individual change over time (Hedeker 2004, Hedeker and Gibbons 2006). Another less well known method for longitudinal data analysis is called functional data analysis which models fluctuation patterns undergone by a variable over time (Ramsay and Silverman 2002, Ramsay and Silverman 2005). The repeated measurements from the same individual are more likely to be correlated to some degree and that change across time is also not likely to be the same. This is usually ignored by linear models in their application to the longitudinal data. Growth mixture modelling also models longitudinal data but it is

person-centered and aims to classify subjects into groups based on their responses across a given set of variables (Wang and Bodner 2007, Jung and Wickrama 2008).

Multilevel regression models were developed to deal with this problem (Goldstein 1995) as they consider individual-specific effects into the model in order to account for data dependency including the capturing of change or variation between the subjects (Hedeker 2004). They have been known as either mixed models (Longford 1987, Wolfinger 1993, Pinheiro and Bates 2000, Bates 2010), random coefficient models (De Leeuw and Kreft 1986), random regression models (Gibbons et al. 1988, Bock 1989), variance component models (Dempster et al. 1981, Longforg 1993), hierarchical models (Raudenbush and Bryk 1986, Bryk and Raudenbush 1992), random-effects models (Laird and Ware 1982), two-stage models (Bock 1989), empirical Bayes models (Hui and Berger 1983, Strenio et al. 1983), nested models or growth models (Hedeker 2004). Multilevel models are an extension of the standard linear models (Paterson and Goldstein 1991, Huta 2014) and basically a special case of the general linear models (McCullagh and Nelder 1989). The advantages of the multilevel models over the general linear models is that they: allow proper specification and computation of random effects which are specific individual effects; allow correlation of errors giving more flexibility in modelling the error covariance structure; and non-constant variability in the error terms is allowed to provide more flexibility in modelling the dependent variable. Multilevel models also allow study effects to vary by groups which are usually ignored by regular regression. The data for multilevel models require measurements on at least two levels of a system (Raudenbush and Bryk 2002, Shieh and Fouladi 2003) and the procedure estimates group level averages. Several ways are used to express the multilevel models where separate equations can be written at multiple levels (Singer 1998). These can further be substituted in to arrive at a single equation which specifies multiple sources of variation. In multilevel models, model selection is more complicated than the usual regression models due to the fixed and random effects (Snijders 1996). That is, multilevel modelling is the only technique that provides different coefficients across groups with separate regression equations for each higher level group, reports variance of each coefficient across groups and also correlations between lower level coefficients across groups (Huta 2014). Basically, given a very large number of subjects, the longitudinal multilevel model is able to consolidate all the variations in the individual linear projections into a variance-covariance structure. The individual trajectory parameters, that is the actual intercepts and slopes are not of interest but rather their variations from one subject to the next is important in the longitudinal multilevel modelling. The aim of this chapter is to

investigate the variation in $CD4^+$ count averages (intercepts) and trend directions (slopes) between HIV patients in response to the strongest clinical covariates at each phase of the HIV disease progression.

4.1 The formulation of the longitudinal multilevel model

The n_i repeated measurements of the i^{th} individual are referred to as level 1 source of variation (within-variation) and the n individuals representing level 2 source of variation (between subject variation). Hence, the multilevel models require the formulation of equations at each level. To clearly demonstrate the concept of multilevel modelling, first, a model with a single covariate and continuous response is presented showing the level 1 and level 2 equations. Second, a categorical (grouping) variable is introduced. The groups do not necessarily need to have the same number of members or measurements in the case of longitudinal data. Also, independence across groups is allowed but not within each group or subject. Our data is a balanced design with each of the $n = 237$ individuals having $n_i = 16 (= m)$ repeated measurements. Third and lastly, the number of covariates is increased to more than one.

4.1.1 Single covariate

Given a covariate x and $CD4^+$ count y for the i^{th} individual, the measured values x_{ij} and y_{ij} at follow up time j, are related as $y_{ij} = \beta_{0i} + \beta_{1i} x_{1ij} + \varepsilon_{ij}$ where $\beta_{0i} = \beta_0 + u_{0i}$, $\beta_{1i} = \beta_1 + u_{1i}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The parameters β_0 and β_1 are the population intercept and slope respectively. The error term ε_{ij} is assumed to be normally distributed with mean 0 and variance σ^2. Furthermore, the u_{0i} and u_{1i} are assumed to be bivariate normal and they are independent of the error term. This can be concisely presented as level 1 and level 2 equations, where

Level 1 : $y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \varepsilon_{ij}; \; i = 1, \ldots, n; \; j = 1, \ldots, m$

Level 2 : $\begin{aligned} \beta_{0i} &= \beta_0 + u_{0i} \\ \beta_{1i} &= \beta_1 + u_{1i} \end{aligned}$

4.1.2 Single covariate and single categorical covariate with G levels

Since our interest is to examine the relationship between the covariate x and the CD4$^+$ count, y, across the four phases after seroconversion, and if $g = 1,\ldots,G$ are the levels of the categorical covariate where G is the reference (ref) group, for the level 1 and level 2 equations, we model

Level 1 : $y_{ij} = \sum_{g=1}^{G}\left(\beta_{0gi} + \beta_{1gi}x_{gij}\right) + \varepsilon_{ij}$

Level 2: $\beta_{0gi} = \beta_{0g} + u_{0gi} \quad where\ \beta_{0g} = \beta_0 + b_{0g}$
$\beta_{1gi} = \beta_{1g} + u_{1gi} \quad where\ \beta_{1g} = \beta_1 + b_{1g}$

Where

β_0 = the population intercept of the reference group G

β_1 = the population slope of the covariate x within the reference group G

β_{0g} = the population intercept of the g^{th} group

β_{1g} = the population slope of the covariate x within the g^{th} group

b_{0g} = difference between the population intercept of the reference group G and the population intercept of the g^{th} group

b_{1g} = difference between the x population slope of the reference group G and the x population slope of the g^{th} group

u_{0i} = the i^{th} individual's intercept deviation from the population's intercept within the reference group G

u_{1i} = the i^{th} individual's covariate x slope deviation from the population's covariate x slope within the reference group G

u_{0gi} = the i^{th} individual's intercept deviation from the population's intercept within the g^{th} group

u_{1gi} = the i^{th} individual's covariate x slope deviation from the population's covariate x slope within the g^{th} group

4.1.3 General p covariates and single categorical covariate with G levels

In this study, 18 covariates were selected for investigation into their influence in the $CD4^+$ cell count including the time as another covariate. Hence $p = 19$ covariates are to be modelled, where

Level 1: $y_{ij} = \sum_{g=1}^{G}\sum_{k=1}^{p}\left(\beta_{0gi} + \beta_{1_k gi}x_{kij}\right) + \varepsilon_{ij}$

Level 2: $\beta_{0gi} = \beta_{0g} + u_{0gi} \quad where\ \beta_{0g} = \beta_0 + b_{0g}$
$\beta_{1_k gi} = \beta_{1_k g} + u_{1_k gi} \quad where\ \beta_{1_k g} = \beta_{1_k} + b_{1_k g}$

Where

β_0 = the population intercept within the reference group G

β_{1_k} = the population slope of the k^{th} covariate x_k within the reference group G

β_{0g} = the population intercept of the g^{th} group

$\beta_{1_k g}$ = the population slope of the k^{th} covariate x_k within the g^{th} group

u_{0i} = the i^{th} individual's intercept deviation from the population's intercept within the reference group G

$u_{1_k i}$ = the i^{th} individual's k^{th} covariate x_k slope deviation from the population's k^{th} covariate x_k slope within the reference group G

$u_{0_k gi}$ = the i^{th} individual's intercept deviation from the population's intercept due to the k^{th} covariate x_k within the g^{th} group

$u_{1_k gi}$ = the i^{th} individual's k^{th} covariate x_k slope deviation from the population's covariate x_k slope in the g^{th} group

4.1.4 Variance-covariance structures

The random effects can assume different covariance structures. Table 4.1 gives a summary of the covariance structures common in SAS software.

Table 4.1: Covariance structures common in SAS software

Structure	Description
Variance components	Which are standard variance components without covariances
Autoregressive AR(1)	These are homogenous variances that decline exponentially with distance.
Compound symmetry	Variances are also homogenous but the correlation between two separate measurements is assumed to be constant regardless of how far the measurements are.
Toeplitz	Similar to AR(1) but the correlations do not necessarily have the pattern
Heterogeneous	Diagonal elements of the variances do not have to be the same.
Unstructured	This is the most liberal structure and requires fitting the most parameters $(p(p+1)/2)$ of any structure for p covariates

In this study we adopted the autoregressive structure for the repeated measurements and an unstructured covariance structure for the random effects because all the correlations were assumed to be different. The autoregressive (AR) process of order a denoted by AR(a) is usually combined with moving average (MA) process of order q to form an autoregressive moving average (ARMA(a, q)). The values of a and q are chosen with the aid of Akaike Information Criterion (AIC) that assesses the model fit for the different value combinations. The smallest AIC provides the best fit.

The $\beta's$ are parameters that are the same (fixed) for all the subjects. The $\mu's$ parameters are the individual's deviation from the population that are allowed to vary (random) over subjects and are normally distributed with mean zero and having a variance-covariance $\boldsymbol{\tau}$ of which is assumed to be unstructured in this study. In the case of a single covariate x with no grouping factor, the distribution of the individual intercept and slope deviations is given by

$$\begin{pmatrix} \mu_{0i} \\ \mu_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{11}^2 \end{pmatrix}\right]$$

where

τ_{00}^2 = variability in the individual intercepts of the n individuals

τ_{01} = covariance between the individual intercepts and slopes of the n individuals

τ_{11}^2 = variability in the individual slopes of the of the n individuals

If τ_{00}^2 and τ_{11}^2 are close to zero, it indicates that the averages and slopes of the individual repeated measurements did not deviate much from the population average and slope respectively. In other words, as τ_{00}^2 and τ_{11}^2 increases, the individuals exhibit much heterogeneity and vice versa (Hedeker 2004). The interpretation of the intercept-slope covariance (τ_{01}) has been elaborated by (Pillinger 2018) where the covariance can possibly take the following values:

$$\tau_{01}\begin{cases} > 0 \\ \approx 0 \\ < 0 \end{cases}$$

For $\tau_{01} > 0$ implies that subjects with larger intercepts also have larger slopes whereas $\tau_{01} < 0$ is an indication that subjects with larger intercepts have smaller slopes or smaller intercepts corresponding to larger slopes. In the case of no relationship between the intercepts and slopes, $\tau_{01} \approx 0$.

Allowing such a covariate x to take different distributions within each of the four infection phases, a typical distribution of the $\mu's$ parameters in the g^{th} group is given by:

$$\begin{pmatrix} \mu_{0gi} \\ \mu_{1gi} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00g} & \tau_{01g} \\ \tau_{01g} & \tau^2_{11g} \end{pmatrix}\right]$$

Consequently, in the case of several covariates $x_1,...,x_p$ and a grouping effect with G levels to be considered, the covariances can be allowed to change from one group to the other. Such a typical distribution of the $\mu's$ parameters in the g^{th} group for the k^{th} covariate is given by

$$\begin{pmatrix} \mu_{0_k gi} \\ \mu_{1_k gi} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00_k g} & \tau_{01_k g} \\ \tau_{01_k g} & \tau^2_{11_k g} \end{pmatrix}\right]$$

Where

$\tau^2_{00_k g}$ = variability in the individual intercepts due to the k^{th} covariate within the g^{th} group

$\tau_{01_k g}$ = covariance between the individual intercepts and slopes due to the k^{th} covariate in the g^{th} group

$\tau^2_{11_k g}$ = variability in the individual slopes due to the k^{th} covariate in the g^{th} group

A comparison of the variations in the $\tau^2_{00_k g}$, $\tau_{01_k g}$ and $\tau^2_{11_k g}$ due to the k^{th} covariate across the G groups can be achieved by expressing them as percentages of the total variation including that of ARMA and common variance. For example, the proportion of the intercept variation due to the k^{th} covariate in the g^{th} group is given by

$$\frac{\tau^2_{00_k g}}{\sum_{g=1}^{G}\left(\tau^2_{00_k g} + \tau_{01_k g} + \tau^2_{11_k g}\right) + \rho_k + \gamma_k + \sigma^2_k} \times 100\%,$$

where ρ_k is the variation captured by the AR, γ_k the variation captured by the MA and σ^2_k be the residual or common variance for all the individuals.

4.2 Data analysis and software

The statistical analysis for the multilevel model was done using **PROC MIXED** in System Analysis Software (SAS) 9.4 where $g = 1$ was considered as Phase 2-Acute, $g = 2$ (Phase 3-Early), $g = 3$ (Phase 4-Est) and the reference group $g = G = 4$ was the Phase 5-ART. The time within each phase was also considered as a variable and the total covariates were $p = 19$. The restricted maximum likelihood (REML) was used for providing efficient estimates and correct standard errors (Laird 1988). **PROC HPMIXED** was used to provide the overall contribution of the covariates in the fixed effects. The SAS memory capacity was very low and this was solved by upgrading the memory limit in the `sasv9.cfg` file. The model with an unstructured variance was appropriate to estimate the intercept-slope covariance and the repeated measurements took an autoregressive moving average correlation structure of [ARMA ($a = 1, q = 1$)]. Each covariate was mean centered to obtain the intercepts as the average for each patient and scaled for estimates comparison. The variance-covariance estimates were expressed as proportions of the total variation in order to compare variations across the G groups (4 infection phases). The SAS results were visualised with the aid of the library **ggplot2** in the open source R software version 3.5.3, by the R Core Team. SAS was powerful in modelling the relationships but presenting the voluminous results was a challenge. All the results were saved as `csv` files and exported to R. A high degree of data manipulation was required to merge and reshape the results data. The formal way of presenting the results in tables was not feasible owing to the extremely long tables. The R software comes handy in all these tasks and in some cases sample results were tabulated only and the rest graphically displayed. All the SAS and R codes are presented in Appendix C4.

4.3 The results of fitting the longitudinal multilevel models

Figure 4.1 shows the diagnostic plots for the variable *time.* The residuals were randomly distributed around zero suggesting that their mean was approximately zero. The histogram was following an approximate normal distribution indicating a constant variance which was also confirmed by the Q-Q plot that did not show heavy tails. Hence the fulfilment of the assumption that the error term ε_{ij} was normally distributed with mean 0 and variance σ^2. A similar pattern of diagnostic plots was observed in the other covariates and the fit statistics are summarised in Table 4.2.

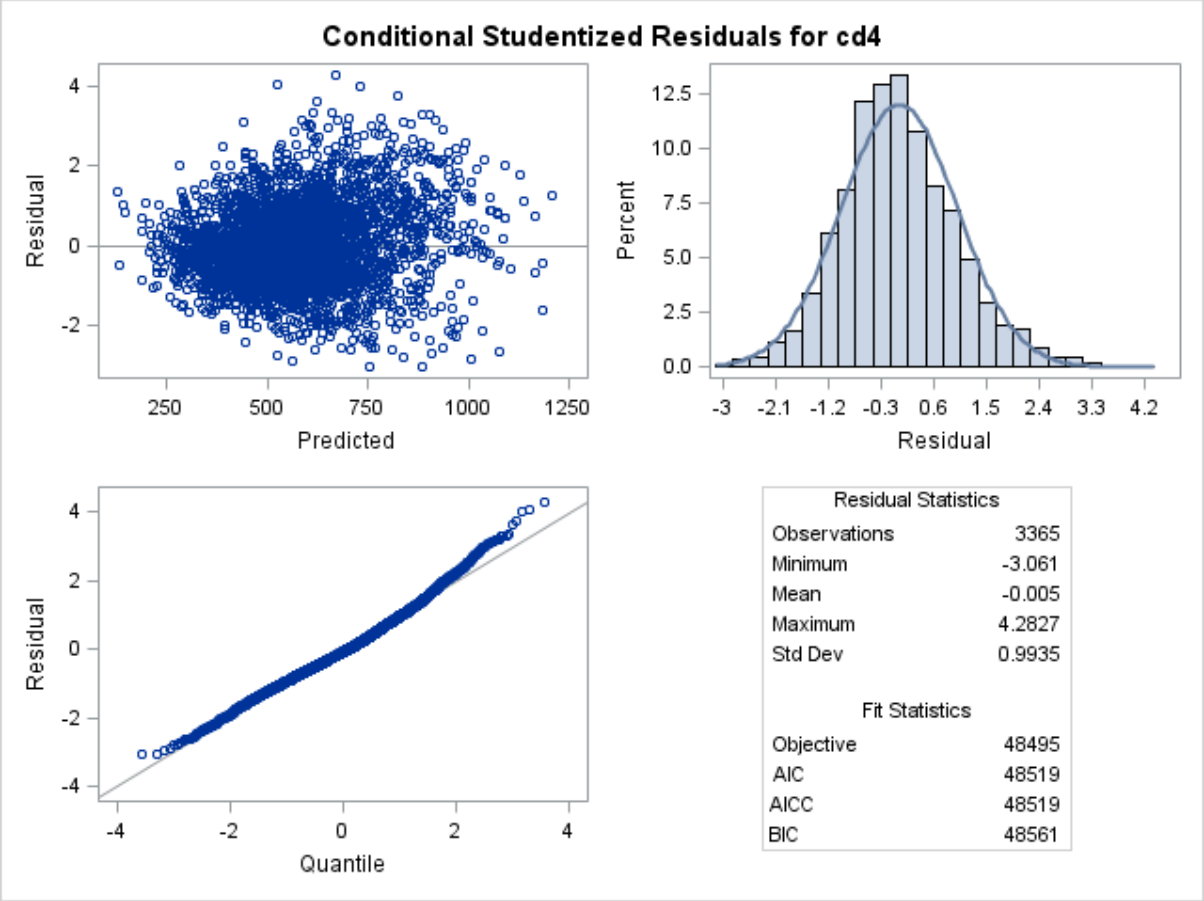


Figure 4.1: Diagnostic plots to test for the normality assumption

Table 4.2: The fit statistics for the covariate effect models

Covariate	-2 Res Log Likelihood	AIC	AICC	BIC
Lymphocytes	48330.7	48360.7	48360.8	48412.7
Platelet	48458.4	48486.4	48486.5	48534.9
Haematocrit	48483.7	48511.7	48511.8	48560.2
LDH	48483.8	48513.8	48514.0	48565.8
MCHC	48485.8	48513.8	48513.9	48562.4
Red blood cells	48489.7	48517.7	48517.8	48566.3
Time	48494.9	48518.9	48519.0	48560.5
ALP	48495.8	48521.8	48521.9	48566.8
Albumin	48502.3	48530.3	48530.4	48578.8
Magnesium	48503.7	48531.7	48531.8	48580.2
MCV	48506.1	48532.1	48532.2	48577.2
Sodium	48509.8	48537.8	48537.9	48586.4
Calcium	48511.3	48539.3	48539.4	48587.8
Basophils	48515.4	48543.4	48543.5	48591.9
Folate	48516.5	48544.5	48544.6	48593.0
Potassium	48528.5	48554.5	48554.6	48599.6
Monocytes	48527.7	48555.7	48555.8	48604.2
Total protein	48528.9	48556.9	48557.0	48605.4
Glucose	48545.6	48569.6	48569.7	48611.3
Fixed	48585.1	48591.1	48591.1	48601.5

The results of the fit statistics are arranged in ascending order of the AIC. They also revealed that the lymphocytes had the best model fit with the least AIC = 48 360.7 in confirmation to the variable selection procedures. This was followed by the platelet count with AIC = 48 486.4. The rest of the covariates were having AIC ≈ 48 500 indicating similar levels of goodness of fit.

4.3.1 The results of the general trends within phase

CD4⁺ count general trends against each covariate within phase: Table 4.3 shows the results of the mixed model in which the marginal (fixed) effects indicate the cohort's general CD4⁺ count linear trajectories in response to the covariates within each phase.

Table 4.3: Fixed effects – The cohort's general CD4⁺ count trajectories within each phase

		2-Acute ($g=1$)	3-Early ($g=2$)	4-Est ($g=3$)	5-ART ($g=4$)
		Estimate(Pr > |t|)	Estimate(Pr > |t|)	Estimate(Pr > |t|)	Estimate(Pr > |t|)
		b_{01}	b_{02}	b_{03}	$\beta_0 = \beta_{04}$
		-52.6627(**0.0053**)	-53.3344(**0.0013**)	-34.3791(**0.0339**)	(ref)616.0300(<.**0001**)
Effect	k	β_{1_k1}	β_{1_k2}	β_{1_k3}	β_{1_k4}
Time	1	8.9971(**0.0376**)	0.6577(0.8770)	-12.1440(**0.0008**)	5.1856(0.1572)
Red blood cells	2	38.2319(0.3902)	18.2958(0.6231)	14.2265(0.6069)	-1.9031(0.9609)
Alkaline phosphatase	3	-2.9375(0.7416)	10.6827(0.1402)	25.1476(**0.0002**)	14.3366(**0.0044**)
Basophils	4	-0.4112(0.9439)	0.0747(0.9905)	7.2633(0.3959)	16.8130(**0.0454**)
Calcium	5	8.1087(0.4206)	-6.3913(0.3409)	-9.4794(0.1370)	-5.9208(0.3670)
Folate	6	-43.7866(<.**0001**)	-20.1953(**0.0075**)	-16.2681(**0.0214**)	-46.6532(<.**0001**)
Glucose	7	2.7933(0.7398)	3.4673(0.5911)	-1.0686(0.8407)	4.8128(0.4096)
Haematocrit	8	-20.6596(0.6311)	-12.1918(0.7407)	-2.3535(0.9300)	3.1107(0.9308)
LDH	9	-9.6857(0.3913)	10.6586(0.1616)	-0.2509(0.9706)	-1.7578(0.8041)
Lymphocytes	10	102.5100(<.**0001**)	127.1300(<.**0001**)	128.1800(<.**000**1)	165.7000(<.**0001**)
Magnesium	11	4.0267(0.6886)	3.5717(0.6022)	-9.2187(0.1098)	13.8217(0.0507)
MCHC	12	-13.5170(0.0728)	12.5703(**0.0409**)	-5.2952(0.3453)	16.1918(**0.0095**)
MCV	13	52.9572(0.1077)	56.5794(**0.0388**)	30.7462(0.1373)	11.3666(0.6600)
Monocytes	14	-3.2578(0.6058)	-10.1268(0.1212)	-18.5394(**0.0018**)	-18.7442(**0.0016**)
Platelet	15	28.4224(**0.0002**)	12.7530(0.0773)	36.6385(<.**0001**)	16.1257(**0.0291**)
Potassium	16	-1.8457(0.8039)	-3.2011(0.4461)	7.0560(0.3034)	1.6780(0.6404)
Protein	17	-30.8203(**0.0015**)	-39.5359(<.**0001**)	-29.0654(<.**0001**)	-13.3719(**0.0394**)
Albumin	18	25.9973(**0.0044**)	29.4461(<.**0001**)	29.6052(<.**0001**)	11.7354(0.0725)
Sodium	19	-19.2748(**0.0148**)	-14.6409(**0.0177**)	-7.2344(0.1810)	6.0560(0.2789)

All the significant trends are in bold. Lymphocyte increase was associated with an improved CD4⁺ count throughout the phases of the HIV disease progression whereas folate and protein

increase resulted in a decline of the $CD4^+$ count at each phase. Before treatment, an increase in albumin improved the $CD4^+$ count by almost the same magnitude, whereas basophils increase could only have a significant positive effect on the $CD4^+$ count during therapy. The $CD4^+$ count improved with increase in alkaline phosphatase (ALP) during the established and ART phases with more improvement at the established phase. Contrary to ALP behaviour, it was during the established and ART phases where the monocytes indicated a negative impact on the $CD4^+$ count. The platelet count showed positive effects on the $CD4^+$ count in all the stages except the early phase. Our results also showed that it was in this early phase only where the mean corpuscular volume (MCV) increase significantly improved the $CD4^+$ count. The mean corpuscular haemoglobin concentration (MCHC) also indicated a positive association with the $CD4^+$ count in the early phase and then during the ART as well. The results revealed that an increase in sodium content soon after HIV infection (acute and early phases) was associated with a $CD4^+$ decline. Our data shows that over time within the acute phase, the $CD4^+$ count increased by 8.9971 cells/mm^3 (p-value = 0.0376) at each visit and dropped by 12.1440 cells/mm^3 (p-value = 0.0008) at each visit during the established phase. Our mixed model estimated that the ART phase records were on average of 616.03 cells/mm^3 of $CD4^+$ count and those from the acute phase being 52.6627 cells/mm^3 below that of the ART average. Table 4.4 is a multiple comparison of all the infection phases. It shows that the ART phase was at least 45 cells/mm^3 of $CD4^+$ count above that of any other investigated phase. All the average $CD4^+$ counts from the other phases before therapy (acute to established) were found not to be significantly different from each other.

Table 4.4: Least squares means and differences

Least squares means

Effect	Phase	Estimate	Standard Error	DF	t Value	Pr > |t|
phase	5-ART	623,81	11,5931	3712	53,81	**<.0001**
phase	4-Est	563,43	9,1055	3712	61,88	**<.0001**
phase	3-Early	563,68	8,3091	3712	67,84	**<.0001**
phase	2-Acute	576,86	12,9588	3712	44,52	**<.0001**

Least squares means differences

Effect	phase	_phase	Estimate	Standard Error	DF	t Value	Pr > |t|
phase	5-ART	4-Est	60,3735	14,7414	3712	4,1	**<.0001**
phase	5-ART	3-Early	60,1262	14,2633	3712	4,22	**<.0001**
phase	5-ART	2-Acute	46,9455	17,3876	3712	2,7	**0.0070**
phase	4-Est	3-Early	-0,2473	12,3269	3712	-0,02	0.9840
phase	4-Est	2-Acute	-13,428	15,8379	3712	-0,85	0.3966
phase	3-Early	2-Acute	-13,1807	15,3939	3712	-0,86	0.3919

4.3.2 The results of the random effects due to each covariate

We further investigated the random effects due to each covariate by allowing each patient to have own $CD4^+$ count trajectory with intercept and slope. This improved the Akaike Information Criterion (AIC) in the modelling of the $CD4^+$ count (Table 4.2).

Time within each phase was also considered as a covariate. The variations in the intercepts (intr) and slopes of individual patient's $CD4^+$ counts against *time* are presented in Table 4.5.

Table 4.5: Sample covariance parameter tests of time ($k=1$) effect and the proportions

Covariate	Phase	Covariance parameter						
$k=1$	g	$\tau^*_{**_k g}$	Name	Estimate	Estimate(%)	Standard Error	Z Value	p-value
Time	2-Acute	$\tau^2_{00_i1}$	Intr	8 011.37	17.9147	2637.70	3.04	**0.0012**
Time	2-Acute	τ_{01_i1}	Intr-Slope	-2 565.41	-5.7367	991.25	-2.59	**0.0097**
Time	2-Acute	$\tau^2_{11_i1}$	Slope	1 864.63	4.1696	518.92	3.59	**0.0002**
Time	3-Early	$\tau^2_{00_i2}$	Intr	0.00	0.0000	-	-	-
Time	3-Early	τ_{01_i2}	Intr-Slope	-1 384.33	-3.0956	451.95	-3.06	**0.0022**
Time	3-Early	$\tau^2_{11_i2}$	Slope	1 925.76	4.3063	441.40	4.36	**0.0001**
Time	4-Est	$\tau^2_{00_i3}$	Intr	0.00	0.0000	-	-	-
Time	4-Est	τ_{01_i3}	Intr-Slope	78.61	0.1758	443.99	0.18	0.8595
Time	4-Es	$\tau^2_{11_i3}$	Slope	682.38	1.5259	381.74	1.79	**0.0369**
Time	5-ART	$\tau^2_{00_i4}$	Intr	7 107.74	15.8941	2244.07	3.17	**0.0008**
Time	5-ART	τ_{01_i4}	Intr-Slope	40.32	0.0902	555.11	0.07	0.9421
Time	5-ART	$\tau^2_{11_i4}$	Slope	0.00	0.0000	-	-	-
Time	AR	ρ_1	Rho	0.94	0.0021	0.01113	84.79	**0.0001**
Time	MA	γ_1	Gamma	0.44	0.0010	0.02703	16.20	**0.0001**
Time	-	σ^2_1	Residual	28 957.00	64.7526	1340.19	21.61	**0.0001**
					100.0000			

NOTE: This table is an extract of 285 rows for the results of 19 covariates
AR =Autoregressive, MA = Moving average

Also shown are the relationships between the patients' intercepts and slopes within each phase. The variations were then expressed as percentages of the total variation captured by the model. The average $CD4^+$ counts varied widely upon entering the acute and ART phases with proportions 17.9147% (p = 0.0012) and 15.8941% (p = 0.0008) respectively. The intercepts and the slopes were negatively related at the acute and early phases in which the $CD4^+$ counts had upward trends of 11.6459 and 3.3582 respectively. This suggests that over time all the patients' $CD4^+$ count trajectories during the acute and early phases approached a higher focal

level. This phenomenon indicates that the patients who entered the acute and early phases at lower $CD4^+$ count had their counts increasing at a faster rate than those who entered with a higher $CD4^+$ count already. Eventually all the patients' $CD4^+$ counts approached the same higher $CD4^+$ count level. Similar estimate proportions of the intercepts and slope relationships for the other covariates are presented in Figure 4.2 where the intercepts represent the average $CD4^+$ counts at the mean covariate value (mean centred). The trajectory slopes are the rates of $CD4^+$ count change as the values of the covariates measurements increase.

Red blood cells changes within each phase showed that the average $CD4^+$ counts also varied widely during medication at 15.5160% (p = 0.0001) in response to an average red blood cell count (Table 4.6).

Table 4.6: Sample covariance parameter tests of red blood cells ($k=2$) effect and the proportions

Covariate	Phase	Covariance parameter						
$k=2$	g	$\tau^*_{**_k g}$	Name	Estimate	Estimate(%)	Standard Error	Z Value	p-value
Red blood cells	2-Acute	$\tau^2_{00_2 1}$	Intr	3 029.97	6.0309	1442.50	2.10	**0.0178**
Red blood cells	2-Acute	$\tau_{01_2 1}$	Intr-Slope	657.46	1.3086	915.94	0.72	0.4729
Red blood cells	2-Acute	$\tau^2_{11_2 1}$	Slope	3 490.44	6.9475	1065.33	3.28	**0.0005**
Red blood cells	3-Early	$\tau^2_{00_2 2}$	Intr	0.00	0.0000	-	-	-
Red blood cells	3-Early	$\tau_{01_2 2}$	Intr-Slope	-1 169.27	-2.3273	537.66	-2.17	**0.0297**
Red blood cells	3-Early	$\tau^2_{11_2 2}$	Slope	2 024.44	4.0295	629.50	3.22	**0.0007**
Red blood cells	4-Est	$\tau^2_{00_2 3}$	Intr	1 191.09	2.3708	1007.84	1.18	0.1186
Red blood cells	4-Est	$\tau_{01_2 3}$	Intr-Slope	-527.81	-1.0506	548.18	-0.96	0.3356
Red blood cells	4-Est	$\tau^2_{11_2 3}$	Slope	1 676.12	3.3362	615.98	2.72	**0.0033**
Red blood cells	5-ART	$\tau^2_{00_2 4}$	Intr	7 845.55	15.6160	1817.21	4.32	**0.0001**
Red blood cells	5-ART	$\tau_{01_2 4}$	Intr-Slope	2 483.56	4.9433	902.10	2.75	**0.0059**
Red blood cells	5-ART	$\tau^2_{11_2 4}$	Slope	720.56	1.4342	632.97	1.14	0.1275
Red blood cells	AR	ρ_2	Rho	0.93	0.0018	0.01	71.14	**0.0001**
Red blood cells	MA	γ_2	Gamma	0.43	0.0009	0.03	16.02	**0.0001**
Red blood cells	-	σ^2_2	Residual	28 817.00	57.3581	1333.77	21.61	**0.0001**
					100.0000			

NOTE: This table is an extract from 285 rows for the results of 19 covariates
AR =Autoregressive, MA = Moving average

The second highest source of variation was observed in the rates of change of the $CD4^+$ count in response to increase in the red blood cells during the acute phase. This had the proportion of

6.9475% (p = 0.0005). During the ART phase, there was also a considerable variation in the rates of change of the $CD4^+$ count (4.9433, p = 0.0059) as the red blood cells increase from the average. In all the covariates, the greatest proportion of variation was observed in the residuals but this was not of interest in this study.

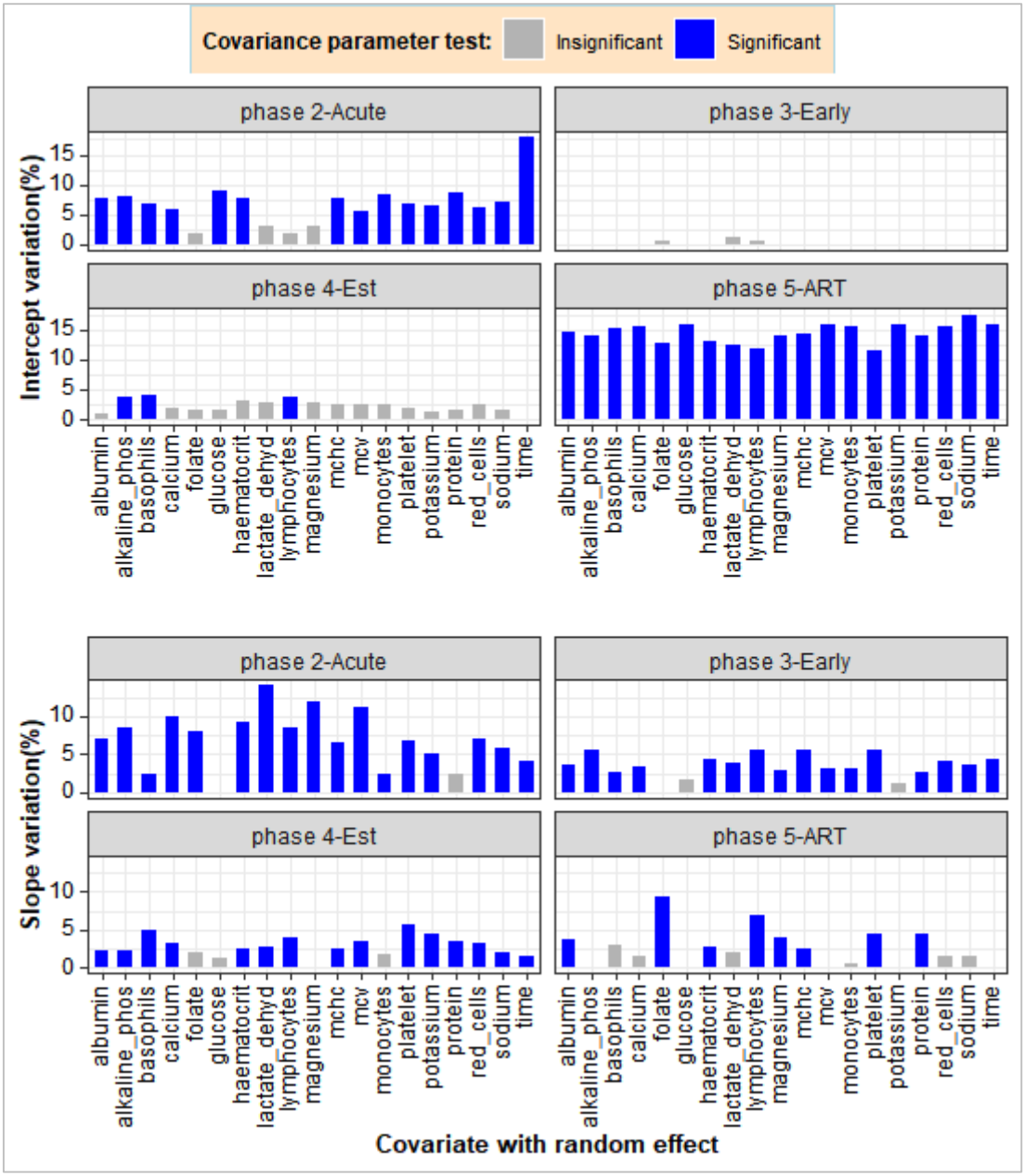


Figure 4.2: Proportion of variation in intercepts and slopes.

The fixed effects parameters were identical and each covariate at a time was allowed to have a random effect. Different variance parameter estimates were obtained for each phase (group) and these were expressed as a percentage of the total variation including the ARMA (1,1) and residuals.

Variations in the average $CD4^+$ counts as induced by each covariate: Figure 4.2 (intercept variation) shows that given the average values of folate, LDH, lymphocytes and magnesium at

the acute phase, there was no significant difference in the $CD4^+$ counts for all the 237 patients under study. The same phenomenon was also observed during the early phase where there was no significant difference in the average $CD4^+$ counts for all the patients in response to each of studied covariates. This was almost the same situation at the established phase except for the significant differences in the patients' average $CD4^+$ counts at the mean values of alkaline phosphatase, basophils and lymphocytes. The results also show that upon taking medication, all the patients' average $CD4^+$ counts were significantly different from each other. Generally, the patients' average $CD4^+$ counts did not vary too much in response to the covariates during the early and established phases. Wide variations in the average $CD4^+$ counts were observed during the acute and ART phases.

Variations in the rate of $CD4^+$ count change as induced by each covariate: We further explored the variations in the rates of the $CD4^+$ count change in response to the increase in the values of each covariate. Figure 4.2 (slope variation) shows that the rate of $CD4^+$ count change in response to each of the covariates varied among the patients mostly from the acute to the established phase. The acute phase was characterised by no significant difference in the $CD4^+$ count rate of change in response to the increase in glucose and protein. Similarly, in the early phase, folate, glucose and potassium did not induce any differences in the rate of change of $CD4^+$ count among the patients. During the established phase, an increase in the folate, glucose, magnesium and monocytes resulted in no significant difference in the $CD4^+$ count rate of change among all the patients. However, upon taking medication, more than half of the covariates were associated with similar rates of $CD4^+$ count change among the patients. The greatest variation in the rate of $CD4^+$ count change was observed soon after infection (acute phase) in which an increase in the LDH induced the widest variations in the $CD4^+$ count rate of change between the patients. This was followed by folate during the ART phase.

Correlation between random intercepts and slopes of $CD4^+$ count trajectories: Throughout the post HIV infection follow up period, there was a positive relationship ($r > 0.80$) between the intercepts and slopes of the $CD4^+$ count trajectories against lymphocytes (Table 4.7). In Figure 4.2 the fixed effects parameters were identical and each covariate at a time was allowed to have a random effect. Different covariance parameter estimates were obtained for each phase (group) and their covariance tests were used to assess the significance of correlation between the random intercepts and slopes.

Table 4.7: Correlation between intercept and slope

k^{th} Covariate	2-Acute ($g = 1$) Corr(covtest)	3-Early ($g = 2$) Corr(covtest)	4-Est ($g = 3$) Corr(covtest)	5-ART ($g = 4$) Corr(covtest)
Time	-0.6638(**0.0097**)	0.0000(**0.0022**)†	0.0000(0.8595)	0.0000(0.9421)
Red blood cells	0.2022(0.4729)	-	-0.3736(0.3356)	1.0000(**0.0059**)
Alkaline phosphatase	0.4280(0.0930)	0.0000(0.4897)	0.5497(0.2165)	0.0000(0.8407)
Basophils	0.3238(0.3120)	0.0000(0.5522)	0.8030(**0.0422**)	0.8775(**0.0104**)
Calcium	-0.1085(0.7203)	0.0000(0.2056)	-0.8544(0.0905)	0.4093(0.3003)
Folate	0.8445(0.0840)	0.0000(0.1124)	-0.1260(0.8364)	0.2060(0.4763)
Glucose	0.0000(0.6623)	0.0000(0.3932)	0.5831(0.4188)	0.0000(0.2206)
Haematocrit	0.0846(0.7095)	0.0000(0.2133)	0.0680(0.8709)	0.5053(0.0776)
LDH	0.5580(0.1187)	1.0000(**0.0050**)	-0.5951(0.2352)	-0.0552(0.8932)
Lymphocytes	0.8498(**0.0060**)	1.0000(**0.0033**)	0.9767(**0.0005**)	0.9803(**0.0001**)
Magnesium	0.1499(0.6507)	0.0000(0.9944)	-1.0000(0.1406)	-0.1519(0.5945)
MCHC	0.0528(0.8257)	0.0000(0.3265)	-0.0588(0.8901)	0.1009(0.7069)
MCV	0.0189(0.9448)	0.0000(0.6571)	0.0730(0.8570)	0.0000(0.4749)
Monocytes	0.2954(0.3582)	0.0000(0.6132)	0.3425(0.5015)	1.0000(**0.0247**)
Platelet	-0.0176(0.9416)	1.0000(0.2117)	0.9625(**0.0097**)	0.1234(0.5933)
Potassium	0.1456(0.7031)	0.0000(0.6158)	0.1396(0.8302)	0.0000(0.9515)
Protein	0.1995(0.6021)	0.0000(0.5694)	-0.1578(0.7586)	-0.3955(0.1034)
Albumin	0.0222(0.9354)	0.0000(0.0589)	-0.9695(0.1951)	0.2702(0.2788)
Sodium	-0.2855(0.3131)	0.0000(0.0879)	-0.7696(0.2077)	0.8205(**0.0203**)

Notes: †The intercept variation in Table 4 was zero but covariance significant, hence the intercept and slope correlation zero. Bold p-value indicates the significant correlation between the intercept and slope.

The results indicate that at each phase of the HIV disease progression, an increase in lymphocytes resulted in the patients whose average $CD4^+$ counts that were already high to increase at a faster rate than those whose average $CD4^+$ counts were lower. The $CD4^+$ count trajectories against red blood cells (ART phase), LDH (early phase), basophils (established and ART phases), platelets (established phase) and sodium (ART phase) showed an upward trend with positive intercept-slope correlations. This means that, as these covariates increase within the indicated phases, the patients with higher average $CD4^+$ counts had their $CD4^+$ counts increasing at a faster rate than those who had lower $CD4^+$ counts. The cohort's $CD4^+$ count trajectory against monocytes was heading downwards during the ART phase with positive intercept-slope relationships. This indicated that as the monocytes increased, patients with lower average $CD4^+$ counts became worse than those with higher average $CD4^+$ counts. On the other hand, there was a negative relationship (covtest, p = 0.0297, Figure 4.3) between the average $CD4^+$ counts and their rate of change with red blood cells during the early phase. This early phase's $CD4^+$ count and red blood cells trajectories followed a general upward trend suggesting that as the red blood cells increase, all the patients' $CD4^+$ counts approached a

common higher $CD4^+$ count level than the cohort's average. That is, red blood cell increase during the uptake of medication, resulted in the patients whose $CD4^+$ count that was higher to increase even faster than those whose count was lower.

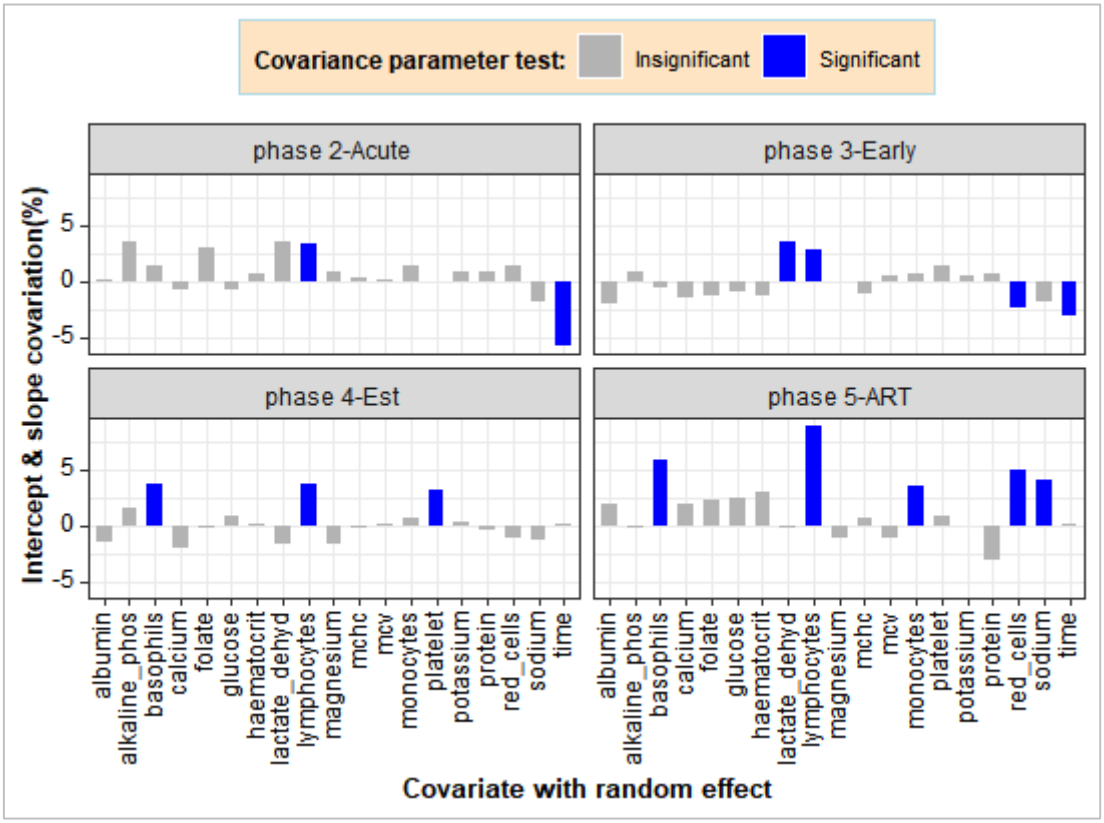


Figure 4.3: Proportion of variation in intercept and slope covariations.

4.4 Clinical interpretation of the results

The investigated data from the CAPRISA studies showed complex relationships and variations in the $CD4^+$ count and its covariates during the different phases of the HIV disease progression. The cohort's repeated measurements for the $CD4^+$ count varied widely around their mean within the established phase and narrowly during the ART phase. All the red blood cell count components were also found to narrowly vary during the ART phase as compared to the other phases.

There was great variation in the patients' average $CD4^+$ counts upon entering the acute and ART phases explaining the patients' immune responses to viral invasion and treatment respectively. This is likely to be attributed to the high level of inter-individual diversity of the human system which is also affected by different factors.(Adrian et al. 2016). During the acute

phase, the mean values of folate, lactate LDH, lymphocytes and magnesium corresponded to similar $CD4^+$ count levels for all the patients. These results revealed that on average the patients' $CD4^+$ counts were not affected by the demand for cell growth and metabolism (folate (Arya and Kumar 2012, Dieticians of Canada 2014)), glucose conversion (LDH (Valvona et al. 2016)) and muscle contractions and protein processing (magnesium (Project Inform 2007)). The $CD4^+$ cells are T cells (Papagno et al. 2004) which are also part of the lymphocytes and the results showed that on average the $CD4^+$ count did not significantly differ between patients during the acute and the early phases, given the average lymphocytes count. However, during the established and ART phases, our data showed that the average lymphocytes count (total B and T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012)), were associated with significantly different average $CD4^+$ levels among the patients. With the exception of the early phase, the indicators of damage to body tissues and inflammation (basophil (Project Inform 2007)) and liver health (ALP (Whitfield 2001, Beare et al. 2008, Patil et al. 2013)) were on average significantly inducing $CD4^+$ count variations among the patients. Our data shows that it is not only during therapy where treatment interferes with the biochemical properties among HIV patients as found by (Mgogwe et al. 2012) during a six-month treatment period study. We observed that liver damage was one of the most common biochemical associated with significant $CD4^+$ count variation throughout the HIV disease progression except during the early phase. Hence, it then turns out that on average, tissue damage indicators were associated with the $CD4^+$ count variation in most of the phases. However, our data further revealed that upon taking medication which significantly improved the $CD4^+$ count than any other phase, all the 18 covariates induced wide variations in the patients' average $CD4^+$ counts. HIV treatment is known to affect the clinical attributes (Ibeh et al. 2013) which could consequently be the attributing factor to the $CD4^+$ count variations in response to all the 18 covariates in our data during the ART phase. This is because treatment has proved to be effective but also increasingly complex due to new developing syndromes (Montessori et al. 2004).

Our results showed that all the patients' $CD4^+$ counts changed at different rates in response to each of the covariates upon taking medication. An increase in glucose and protein did not bring about variation in the rate of change of the $CD4^+$ counts between patients during the acute phase. Early phase $CD4^+$ counts also changed at the same rate when either of folate, glucose or potassium increase. Similarly, folate, glucose, magnesium and monocytes increase at the established phase gave rise to the same rate of the $CD4^+$ count change. Most of the covariates

induced wide variations in the rate of the $CD4^+$ count change during the acute phase. Our data showed that lymphocytes increase in every phase resulted in patients whose $CD4^+$ count was already higher increasing even faster than those patients with lower average $CD4^+$ counts. Similarly, patients with higher $CD4^+$ counts were found to have their count increasing at a much faster rate as the following covariates increase in certain phases: LDH (early phase), basophils (established and ART phases), platelets (established phase) and sodium (ART phase). During the early phase of the disease progression, the patients whose average $CD4^+$ count that was lower, increased at a faster rate in response to the red blood cells increase such that all the patients' $CD4^+$ counts eventually approached a common higher $CD4^+$ count. However, upon taking medication, an increase in the red blood cell count resulted in those individuals whose $CD4^+$ count which was already higher to become even much better as compared to the ones that were lower. Our data shows that red blood cells that are packed with haemoglobin (Junqueira et al. 2006) and plays a role in the respiratory process (Jensen et al. 1998, Wintrobe and Greer 2009) were associated with $CD4^+$ count improvement. Monocytes increase during medication (ART phase) resulted in $CD4^+$ counts that were lower to become much worse than for those patients whose average $CD4^+$ counts that were higher. Although monocytes are infected together with the $CD4^+$ T cells (Pasupathi et al. 2008), our data show that during therapy, monocytes were spared more than the $CD4^+$ T cells.

4.5 Summary

The few strongest $CD4^+$ count clinical covariates were found to induce either wide variations in the patients' average $CD4^+$ counts in some infection phases and show no effect in the others. An increase in the measurements or quantities of the covariates was found to either improve the $CD4^+$ count at a faster rate for those patients whose average $CD4^+$ was already high or worsen the $CD4^+$ level which was already lower than that of the other patients. In some cases, the increase in the covariates values caused the patients' $CD4^+$ counts to approach a common level which was lower or higher than that of the general cohort. Tissue damage indicators were the most common covariates associated with $CD4^+$ count variation between patients. Patients who entered either the acute or early phases with lower average $CD4^+$ counts had their count increasing at a faster rate than those who entered with a higher $CD4^+$ count already resulting in the cohort approaching a common higher $CD4^+$ count. In addition to other treatment measures, the manipulation of selected $CD4^+$ count covariates for patients within a specific phase can usefully augment tailored methods for monitoring HIV patients using the $CD4^+$

count. Generally, the studied covariates induced wide variations in the $CD4^+$ count between HIV positive patients during the ART phase. The multilevel model indicated that given a large number of subjects, it is possible to assess the variations between the individual linear trajectories. However, a visual display of all the subjects becomes difficult. More so, the linear relationships have the tendency of masking some sources of discovery due to the model-driven trajectories. To allow data-driven response trajectories, the next chapter discusses the additive models with the same concept of the mixed models but smoothing the trends for each phase with respect to the covariates.

CHAPTER 5 GENERALISED ADDITIVE MIXED MODELS

In Chapter 4 we have examined the parametric individual linear trajectories of each covariate on the $CD4^+$ count within each infection phase. We also pooled all the repeated measurements from every individual to look at the overall linear fixed effects of the covariates on the $CD4^+$ count. Given a large number of individual linear profiles to be assessed, the longitudinal multilevel model comes to hand with its variance-covariance structure that consolidates the relationships of all the individual trajectories. However, the failure to visualise the behavioural patterns of the $CD4^+$ count trends results in some of the covariates' dubious linearity being silently unaccounted for. On the other hand, polynomials of any order could have been easily incorporated in order to account for the non-linearity but their interpretation is often a challenge. Accordingly, we adopted the generalized additive model (GAM) that are more flexible than the linear models. They also come with the best transformations that are determined simultaneously without parametric assumptions associated with their form (Faraway 2006). They impose some penalties to the wiggliness that makes the polynomial interpretation difficult. Additive models derive their strengths from the ability to deal with highly non-linear and monotonic relationships between the response and the explanatory variables (Yee and Mitchell 1991). This is achieved by automatically determining the smoothness that prevents overfitting and treating random effects as smooths trends too. The additive models are not affected by a high dimensional curse because the smooths are applied as univariates (Fan et al. 2013). Both the parametric and non-parametric components can be incorporated in which case they are referred to as semi-parametric models (Zuur et al. 2007). The semi-parametric models are useful in the sense that they compromise the restrictive nature of parametric models and too much flexibility that comes with non-parametric models (Fan and Li 2004). Although it is difficult to interpret the non-parametric terms of the additive models (Hastie and Tibshirani 1990), they provide a very suitable platform for data set exploration and visualisation of the relationship between the response and covariates. This enables them to uncover the structure in the relationship that might otherwise be missed (Xiang 2001). Their data-driven approach allows the data to determine the shape of the response with a wider range of response curves (Yee and Mitchell 1991) as compared to the limited shapes in the parametric class that assume pre-determined model shapes.

In the additive models, linear predictors in the linear models are basically replaced by smooth functions of the covariates which are related to the response variable (Wood 2016). In this case,

the possibly transformed response variable is related to non-linear smooth functions of the covariates. That is, to establish a relationship between the mean of the response variable and the smooth function of the explanatory variables, the additive models use a link function (Guisan et al. 2002). The link function transforms (link) the response mean to lie on the plane of the covariate space (Yee and Mitchell 1991). Gaussian and non-Gaussian distributions can easily be modelled (Xiang 2001) allowing the additive models to be extended to a wide range of distribution families (Hastie and Tibshirani 1990). As such, they are also estimated in the same way as the general linear models (Wood 2017).

5.1 Model specification

5.1.1 GAM specification

Given a response variable CD4$^+$ count, y_{ij} and covariates x_k, for $k = 1,\ldots,p$, a standard linear regression model has the parametric form

$$y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \varepsilon_{ij}\,, \tag{5.1}$$

where the random errors ε_{ij} are independent and identically distributed with a mean of zero and a constant variance, σ^2, $\varepsilon_{ij} \sim N(0,\sigma^2)$. The regression parameters β_k 's are obtained by the least squares approach. On the other hand, the GAM relaxes the linearity in $\beta_k x_k$ to capture the response curves using the non-parametric smooth function $s_k(x_k)$ for $k = 1,\ldots,p$ such that (5.1) takes the form

$$y_{ij} = s_0 + \sum_{k=1}^{p} s_k(x_k) + \varepsilon_{ij}\,, \tag{5.2}$$

where s_0 is the parametric intercept, $s_k(x_k)$ replacing the $\beta_k x_k$ and $\varepsilon_{ij} \sim N(0,\sigma^2)$. According to (Hastie and Tibshirani 1986) we can combine (5.1) and (5.2) to model a semi-parametric of the form

$$y_{ij} = s_0 + \sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} s_k(x_k) + \varepsilon_{ij}\,. \tag{5.3}$$

As in the general linear models, the GAM is also made up of random and additive components that are related by a link function where the random component is assumed to have an exponential family density given by

$$f_Y(y;\boldsymbol{\theta},\boldsymbol{\varphi}) = \exp\left\{\frac{y\boldsymbol{\theta} - (b\boldsymbol{\theta})}{a(\boldsymbol{\varphi})} + c(y,\boldsymbol{\varphi})\right\}, \tag{5.4}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are natural and scale parameters respectively (Chen 2000). Commonly used statistical distributions in this exponential family are the Binomial, Poisson and Normal distributions (Guisan et al. 2002). If the y_{ij} mean response is denoted by $\boldsymbol{\mu}$ which is related to the covariates x_k's through a link function L then

$$L(\boldsymbol{\mu}) = \boldsymbol{\eta}, \text{ where } \boldsymbol{\eta} = s_0 + \sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} s_k(x_k). \tag{5.5}$$

The $\boldsymbol{\eta}$ models the overall response curves against the covariates. If a smoothed trajectory is required for every level of a factor say, the separate trends are referred to as random smooths. Generalised additive mixed models (GAMM) are used for fitting such relationships.

5.1.2 GAMM specification

GAMM are applicable to clustered, spatial and hierarchical data as well as capable of modelling correlation between observations (Lin and Zhang 1999). Given smooth factor interactions z_r for $r = 1,...,q$, the GAM in (5.5) can be extended to GAMM by adding the term $s_k(x_k, z_r)$ to model

$$\boldsymbol{\eta} = s_0 + \sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} s_k(x_k) + \sum_{k=1}^{p}\sum_{r=1}^{q} s_k(x_k, z_r). \tag{5.6}$$

Two issues will arise from our data for the application of the GAMM. First, smoothing individual profiles within infection phase would be limited due to the fewer data points (4 visits). Second, if we are to ignore the infection phase and smooth the 16 data points for each subject, the larger number of individuals (n = 237) would produce twisted trends that could not easily be visualised for interpretation. Instead the interest should be in the phase trends of the $CD4^+$ count in response to the covariates and our problem will then involve one random factor

of z_r where $r=1$ in (5.6). Hence, we intend to model the semi-parametric with random smooths given by

$$\boldsymbol{\eta} = s_0 + \sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} s_k(x_k) + \sum_{k=1}^{p} s_k(x_k, z). \tag{5.7}$$

In the case of parametric fit, either linear or polynomial smoothing is used whereas the non-parametric can take either the locally weighted scatterplot smoother or splines. Different scatter plot smoothers with also different amount of smoothing can be used on different covariates (Faraway 2006).

5.1.3 Scatter plot smoothers

The scatter plot smoother $s_k(x_k)$ can either be a running mean, running median, running least squares line, kernel estimate or spline (Yee and Mitchell 1991). However, the running mean is not considered a satisfactory smoother due to the creation of large biases at the end points (Hastie and Tibshirani 1986). It also has the problem of not generally reproducing straight lines in the case of the data that lie exactly along a straight line. The running lines also pose some weakness in the GAM as the choice of the span may either result in high variance which is not smooth or may be too smooth as in the least squares regression line such that the underlying function does not pick up the curvature. Nevertheless, choosing a span size between $\frac{1}{n}$ and 2 has been recommended to trade-off the bias and variability of the estimate, making the GAM a much stronger technique in the modelling arena (Hastie and Tibshirani 1986). Cubic splines have been considered special among the other possible scatter plot smoothers (Wood 2000, Wood 2017). These splines are real piecewise-defined polynomial functions that are generally estimable under identifiability constraints. They are confounded via the intercept s_0 such that $\sum_{k=1}^{p} s_k(x_k) = 0$ with the $s_k(x_k)$ and $s_k(x_k, z)$ creating centred smooths (Wood 2010). Hence, the dependent axis of the smoothed curves are mean centred values that are around zero in graphical displays (Fan et al. 2013, Clark 2014). They are difficult to interpret using the original response scale but provide good prediction or interpolation in exploring the functional nature of a response. Negative smooth values will then indicate a drop below the average and similarly positive showing that the original response was above the intercept.

The number of knots (points) $M-1$, say at which the piecewise-defined polynomial functions (basis functions) connect is usually pre-set and an extra basis dimension is required for the intercept to meet the identifiability constraint on the smooths (Wood 2017). Hence, for a given smooth function, the basis dimension is given as M and the adequacy can be determined by an $M\ index$. This is the ratio of differenced near neighbours residual variance and residual variance (Wood 2016). The smooth function $s_k(x_k)$ for the k^{th} covariate is composed of a sum of $M-1$ basis functions with the m^{th} basis function denoted by $b_m(x_k)$. The relationship between the smooth function and the basis functions for the k^{th} covariate is represented as

$$s_k(x_k) = \sum_{m=1}^{M-1} b_m(x_k)\beta_m , \tag{5.8}$$

where β_m's are the corresponding unknown regression coefficients. In other words, the response y_{ij} linearly depends on the unknown β_m's in the smooth functions $s_k(x_k)$ (Wood 2017). In the presence of a smooth factor interaction z with levels $g = 1,\ldots,G$, equation (5.8) becomes $s_k(x_k, z) = \sum_{m=1}^{MG} b_m(x_k)\beta_m z$. The $M-1$ and MG are the maximum complexities of the smooths $s_k(x_k)$ and $s_k(x_k, z)$ respectively. M is usually pre-set for the model and is subject to tuning (Wood 2016).

5.1.4 Model estimation

GAM estimation

In the case of GAM, the mean response estimated by (5.5) can be represented as

$$\boldsymbol{\eta} = s_0 + \sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} s_k(x_k) = \boldsymbol{X\beta} \tag{5.9}$$

where $\boldsymbol{X}$ is the covariate matrix and $\boldsymbol{\beta}$ the unknown linear parameters to be estimated. The residual autocorrelation can then be incorporated to modify (5.9) and model

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{e} \tag{5.10}$$

where $\boldsymbol{e}$ is the autoregressive correlation matrix usually of order 1 in the R package **mgcv** for estimating GAMM (Wood 2016). To express (5.10) in the form of (5.9), the correlation matrix, $\boldsymbol{e}$ is banded with a Choleski factor $\boldsymbol{C}$ such that the random error $\varepsilon = \boldsymbol{Ce}$ (Wood 2017). We then model the response accounting for the autocorrelation by $\boldsymbol{Y} = \boldsymbol{X\beta} + \varepsilon$ which is solved by minimising

$$\left\|\boldsymbol{Y} - \boldsymbol{X\beta}\right\|^2. \tag{5.11}$$

In practice, the longitudinal data are presented in chronological order and the degree of correlation ρ between the repeated measurements is manually estimated as an average. The roughness or "wiggliness" in (5.11) is controlled by adding a penalty to the least squares which restricts the freedom of coefficients to vary (Wood 2016). The regularization of the smoothness requires the use of a penalised likelihood maximisation technique with a penalty function of the form

$$\sum_{k=1}^{p} \lambda_k \boldsymbol{\beta}^T \boldsymbol{S_k} \boldsymbol{\beta}, \tag{5.12}$$

where the k^{th} covariate smoothing parameter λ_k controls the trade-off for $s_k\left(x_k\right)$ and $\boldsymbol{S_k}$ is a suitable matrix of known coefficients depending on the chosen basis (Faraway 2006). The least squares parameter estimation will then attempt to minimise

$$\left\|\boldsymbol{Y} - \boldsymbol{X\beta}\right\|^2 + \sum_{k=1}^{p} \lambda_k \boldsymbol{\beta}^T \boldsymbol{S_k} \boldsymbol{\beta}. \tag{5.13}$$

The penalised least square estimator of $\boldsymbol{\beta}$ for a specific value of optimal λ is given by $\hat{\boldsymbol{\beta}}_\lambda = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The idea is to maximise $\log(\boldsymbol{\beta}) - \sum_k \boldsymbol{\beta}_k^T \boldsymbol{S}_k \boldsymbol{\beta}_k$ where the $L(\boldsymbol{\beta})$ is the likelihood with respect to $\boldsymbol{\beta}$ and the λ_k s. When $\lambda_k \approx 0$, the regression spline estimate is not penalised and is wiggly. Whereas $\lambda_k \to \infty$, the curve gets closer to a linear fit, which is a characteristic of over smoothing. The optimal amount of smoothing for each covariate λ_k is selected by generalised cross-validation, $v_g(\lambda) = \dfrac{n\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right\|^2}{\left\{n - tr\left(\boldsymbol{F}_\lambda\right)\right\}^2}$

where $\boldsymbol{F}_\lambda = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{S}_\lambda = \sum_{k=1}^{p}\lambda_k\boldsymbol{S}_k$ and $tr\left(\boldsymbol{F}_\lambda\right)$ is the effective degrees of freedom (edf). The effective degrees of freedom are the actual degrees of freedom used by a smooth for the pre-set basis dimension and selected smoothing parameter λ_k. They measure the complexities of the penalised smooth terms and if they are well below the basis dimension, suggests that increasing the basis dimension will have no effect on smoothing the response curvature (Pedersen et al. 2019). On the other hand, larger effective degrees of freedom indicate the potential to increase the wiggliness of the smooth term.

GAMM estimation

The GAMM is estimated just as a general linear mixed model where (5.7) becomes a linear model whose mean response is given by

$$\begin{aligned} s_0 + \sum_{k=1}^{p}\beta_k x_k + \sum_{k=1}^{p} s_k\left(x_k\right) + \sum_{k=1}^{p} s_k\left(x_k, z\right) &= \boldsymbol{X\beta} + \sum_{k=1}^{p} s_k\left(x_k\right) + \boldsymbol{Z\gamma} \\ &= \boldsymbol{X\beta} + \boldsymbol{Z\gamma} \end{aligned} . \qquad (5.14)$$

where $\boldsymbol{Z}$ are the random effects with the coefficients of the random effects $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$. Each smooth in (5.14) is regarded as the unpenalized fixed effects component that is absorbed in the $\boldsymbol{X\beta}$ and the random effects being the penalized that are absorbed in the $\boldsymbol{Z\gamma}$ (Wood 2017). The $\boldsymbol{Z\gamma}$ can be treated like a smooth with a penalty $\boldsymbol{\gamma}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}$ such that writing $\boldsymbol{\Omega}^{-1}$ in the form $\sum_{m=1}^{M}\lambda_m\boldsymbol{S}_m$ will have the GAMM being exactly like a GAM (Wood 2010).

Algorithms

The parameters of the linear function of the GAM that are estimated by maximum likelihood have models that assume a linear or some form of parametric for the covariates. These include the normal linear regression model, logistic regression model for binary data and the Cox's proportional hazards for survival data (Hastie and Tibshirani 1986). Although Cox's proportional hazards method is also used among these, it is not in the exponential family as required for the link function for GAM (Hastie and Tibshirani 1986). The mathematical calculations require the aid of the application of numerical optimisation in the presence of Gaussian general linear models. The Gaussian models are used in many statistical applications including the modelling of discrete responses (Abe 1999) where the normal distribution may

not be adequate (Xiang 2001). In the case of known scale parameter, Mallow's Cp or Unbiased Risk Estimator (UBRE) can be used.

An iteration is required to estimate the $s_k(x_k)$ for the additive model (Hastie and Tibshirani 1986). The additive models with any regression-type fitting mechanisms can be fitted by a backfitting algorithm which is a general algorithm (Zeger and Diggle 1994). Alternatively, the smoothing technique can use an iterative procedure called a local scoring algorithm (Xiang 2001) which is applicable to any likelihood-based regression model (Hastie and Tibshirani 1986). The local scoring method is faster and gives the GAM approach an added advantage. The fitting procedure can also be described as comprised of two parts that entail the estimation of the additive predictor first by solving a system of normal equations and then secondly, linking the predictor to a function $s_k(x_k)$.

The additive part can be fitted in at least three ways in R software packages (of interest in this study) namely: Mixed GAM Computation Vehicle (**mgcv)**; Generalised Additive Model (**gam)** and General Smoothing Splines (**gss**). The **mgcv** has the ability to automatically choose the smoothing amount and also has the advantage of a wider functionality (Wood 2016). In GAMM, any of the penalised regression smoothers can be written as components of the mixed model by treating the smoothing parameters as variance component parameters that are estimated by likelihood, restricted maximum likelihood or penalised quasi-likelihood (Wood 2017). The `bam` function in **mgcv** uses the fast restricted maximum likelihood.

Model diagnostics

The model with the least generalised cross-validation is the best fit and in some cases where there are challenges of over dispersion, selection of the best model can be achieved by comparing deviances. Deviance is one of the measures of goodness of fit and is a generalisation of the residual sum of squares in ordinary regression (Yee and Mitchell 1991). The application of the Akaike Information Criterion (AIC) is another measure of best fit and is structured as follows: $AIC = 2[-Max\{\log(likelihood)\} + p]$, where p is the number of parameters. The model with the smallest AIC is chosen (Guisan et al. 2002). The F-test is used for comparing over dispersion, whilst the Chi-square suitable if over dispersion does not exist. The GAM model also provides confidence intervals for any predicted amount from the fitted model. These appear to be reliable whereas component-wise intervals only give an idea of the uncertainty of the components (Wood 2016). Visual model diagnostics include histogram of

residuals that are supposed to be approximately normally distributed. A Q-Q plot with heavy tails indicates the presence of outliers and as such the scatter plot of sample quantiles versus theoretical quantiles should form approximately a straight line at 45^0 angle. Non-constant error variance is checked by residuals against the linear predictor that are supposed to be randomly distributed around zero. The observed response against the fitted values are to be highly correlated with the scatterplots very close to the line at 45^0 angle. If the effective degrees of freedom (edf) of a given covariate are so small, suggest that the smooth of that particular covariate is insignificant (Wood 2017). For M indices close to 1 will be an indication that it is less likely that there are missing patterns in the residuals (Wood 2016). If the GAMM hypothesis is to be tested at α level of significance, a $p-value<\alpha$ suggests that there is significant difference in the trends of the random smooths. The nature of the differences in the random smooths requires graphical displays which is the main aim of this chapter.

5.2 Data analysis and software

The analysis was performed in the open source R software, version 3.5.3 of the R Core Team. Since Figure 2.3 boxplot of the $CD4^+$ count did not show departure from normality, we assumed that L takes a Gaussian identity link function. A GAMM was fitted using a function called `bam` (for large datasets) that can also be used to fit the GAMM with factor smooth interactions basis `fs` in the library **mgcv** (Wood 2019). Random smooths produced by `fs` already include random intercepts and random slope effects. This allowed random smooths for each phase of the HIV disease progression where the smooth factor Z levels were $g=1$ (phase 2-Acute), $g=2$ (Phase 3-Early), $g=3$ (Phase 4-Est) and $g=4$ (Phase 5-ART). Hence $G=4$. The basis dimension was set to $M=5$ assuming that the curvatures were slightly complex than cubic splines. Note that the `bam` syntax in R uses $k\equiv M$ in this case). The penalty argument was set to 1 to give a heavier penalty for the smooth moving away from zero, causing shrinkage to the mean. The random smooths were explored using the function `inspect_random` which was embedded in macros that significantly reduced the lengthy codes required to obtain the outputs. The testing of model terms was conducted at 5% level of significance with a single exception of a term at 10%. An autoregressive of order 1 (AR(1)) structure using the library Interpreting Time Series and Auto-correlated Data Using GAMMs (**itsadug)**, was incorporated to reduce the effects of autocorrelation in the individual repeated observations (van Rij 2016). All the R codes are presented in Appendix C5.

5.3 Results of the generalised additive mixed model

5.3.1 The results of model selection and fit diagnostics

Our intention was to model a semi-parametric model for efficiency and focus on the factor smooth interactions to visualize the random smooths. Instead several factor smooth interaction models were fitted to assess if other terms were indeed necessary (Table 5.1).

Table 5.1: Model selection

Model	Response curve, η	DF	AIC	R^2	Dev.Expl	Rho
1	$s_0 + s(x,z)$	153.5817	49 945.30	0.4818	0.5013	
2	$s_0 + s(x,z) + AR$	141.0442	49 513.71	0.4707	0.4889	0.2861
3	$s_0 + s(x) + s(x,z) + AR$	**104.4118**	**49 494.92**	0.4704	0.4826	0.2861
4	$s_0 + x + s(x) + s(x,z) + AR$	111.6214	49 516.34	0.4697	0.4816	0.2861
5	$s_0 + s(x) + s(x,z) + AR$	123.0018	49 513.14	0.4710	0.4855	0.2861

AR-residual autocorrelation corrected

Initially a strictly non-parametric factor smooth interaction model of interest was modelled (Model 1). Upon inspection, if this model was suffering from any residual autocorrelation, a decaying autocorrelation pattern was observed. The strongest average correlation was at lag 1 and estimated to be $\rho = 0.2861$ (Figure 5.1 top row). Incorporating the average correlation into Model 2 showed no residual patterns in the corrected ACF plot (Figure 5.1 bottom left). The Model 2 results showed that the correction of the autocorrelation improved the model fit as characterised by a drop in both the degrees of freedom (141.0042) and the AIC (49 513.71). We added the effect of overall covariate smooth (Model 3) and this further improved the model parsimony too with reduced degrees of freedom (104.4118) as well as a lower AIC (49 494.92) than Model 2. However, an attempt to consider parametric terms (Model 4), the model fit was poor than Model 3, where both the degrees of freedom (111.6214) and AIC (49 516.34) increased. Table 5.2 presents a summary of Model 3 terms.

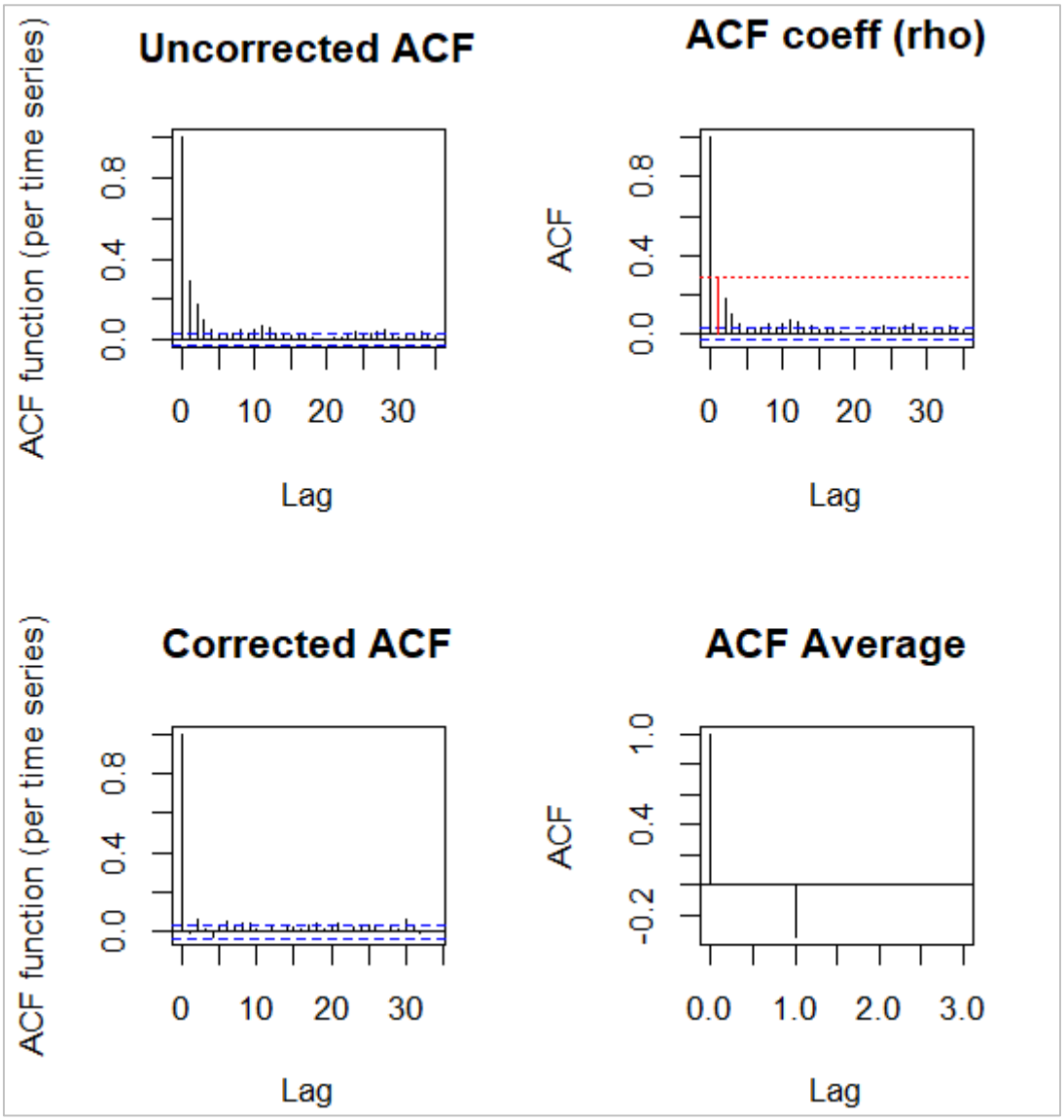


Figure 5.1: Inspection and correction of residual autocorrelation

The effective degrees of freedom for all the random smooth terms were far below the set $MG = 20$ suggesting that the fitted random smooth curves for each infection phase were adequately explaining the $CD4^+$ count trends. Nevertheless, s(LDH,phase) with edf = 10.1688 showed the greatest complexity in the penalised random smooth terms. To some extent, this was followed by s(haematocrit,phase) with edf = 6.8664, s(lymphocytes,phase) with edf = 6.4017, s(red blood cells,phase) with edf = 5.7216 and s(ALP,phase) with edf = 5.1570. For the set basis dimensions $M = 5$ for the overall smooth terms in Model 3, some of the effective degrees of freedom were close to $M - 1 = 4$ showing that the penalty complexities could possibly be increased. In Model 5, the basis dimensions were doubled ($M = 10$) for s(lymphocytes) whose edf was 3.9278, s(total protein) with edf = 3.6558, s(MCV) with edf = 3.4599, s(folate) with edf = 3.4566, s(albumin) with edf = 3.3584, s(basophils) with edf =

3.2409 and s(platelet) with edf = 3.1121. Both the degrees of freedom (123.0018) and AIC (49 513.14) increased indicating that Model 5 had a poor fit than Model 3.

Table 5.2: A summary of Model 3

S_0	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	574.3461	23.6848	24.2496	<**0.001**

k	*Overall,* $s_k(x_k)$	M-1	M Index	edf	F	p.value
1	s(Red blood cells)	4	0.97	0.0005	0.0001	0.2407
2	s(Haematocrit)	4	0.88	2.3867	1.4682	**0.0002**
3	s(MCV)	4	0.98	3.4599	10.9677	**0.0000**
4	s(MCHC)	4	0.96	0.0005	0.0001	0.3866
5	s(Platelet)	4	0.99	3.1121	7.0860	**0.0000**
6	s(Lymphocytes)	4	0.98	3.9278	89.0726	**0.0000**
7	s(Monocytes)	4	0.98	2.8098	6.8695	**0.0000**
8	s(Basophils)	4	0.80	3.2409	5.1025	**0.0001**
9	s(ALP)	4	0.95	1.8215	1.2221	**0.0010**
10	s(Calcium)	4	0.83	0.0002	0.0000	0.7152
11	s(Magnesium)	4	0.80	0.0001	0.0000	0.7418
12	s(Potassium)	4	0.94	0.0014	0.0003	0.3348
13	s(Sodium)	4	0.89	1.8233	1.4124	**0.0022**
14	s(Total protein)	4	0.89	3.6558	16.7466	**0.0000**
15	s(Albumin)	4	0.86	3.3584	8.5429	**0.0000**
16	s(LDH)	4	0.91	0.0056	0.0014	**0.0167**
17	s(Folate)	4	0.79	3.4563	18.2167	**0.0000**
k	*Phase specific,* $s_k(x_k, z)$	*MG*				
1	s(Red blood cells,phase)	20	0.97	5.7216	0.5062	**0.0232**
2	s(Haematocrit,phase)	20	0.88	6.8664	0.8138	**0.0008**
3	s(MCV,phase)	20	0.98	0.0006	0.0000	0.3808
4	s(MCHC,phase)	20	0.96	2.7234	0.1963	0.1484
5	s(Platelet,phase)	20	0.99	4.0657	0.4305	**0.0205**
6	s(Lymphocytes,phase)	20	0.98	6.4017	1.1305	**0.0001**
7	s(Monocytes,phase)	20	0.98	1.2916	0.0823	0.2127
8	s(Basophils,phase)	20	0.80	0.0011	0.0000	0.4206
9	s(ALP,phase)	20	0.95	5.1570	0.6649	**0.0022**
10	s(Calcium,phase)	20	0.83	0.9796	0.0587	0.2741
11	s(Magnesium,phase)	20	0.80	0.0011	0.0001	0.4274
12	s(Potassium,phase)	20	0.94	0.0003	0.0000	0.6605
13	s(Sodium,phase)	20	0.89	3.0043	0.2508	**0.0652**
14	s(Total protein,phase)	20	0.89	3.9438	0.4838	**0.0099**
15	s(Albumin,phase)	20	0.86	4.0936	0.3857	**0.0381**
16	s(LDH,phase)	20	0.91	10.1688	1.6168	**0.0001**
17	s(Folate,phase)	20	0.79	0.0003	0.0000	0.7075

Model 3 was chosen as the best fit and the diagnostic plots are shown in Figure 5.2.

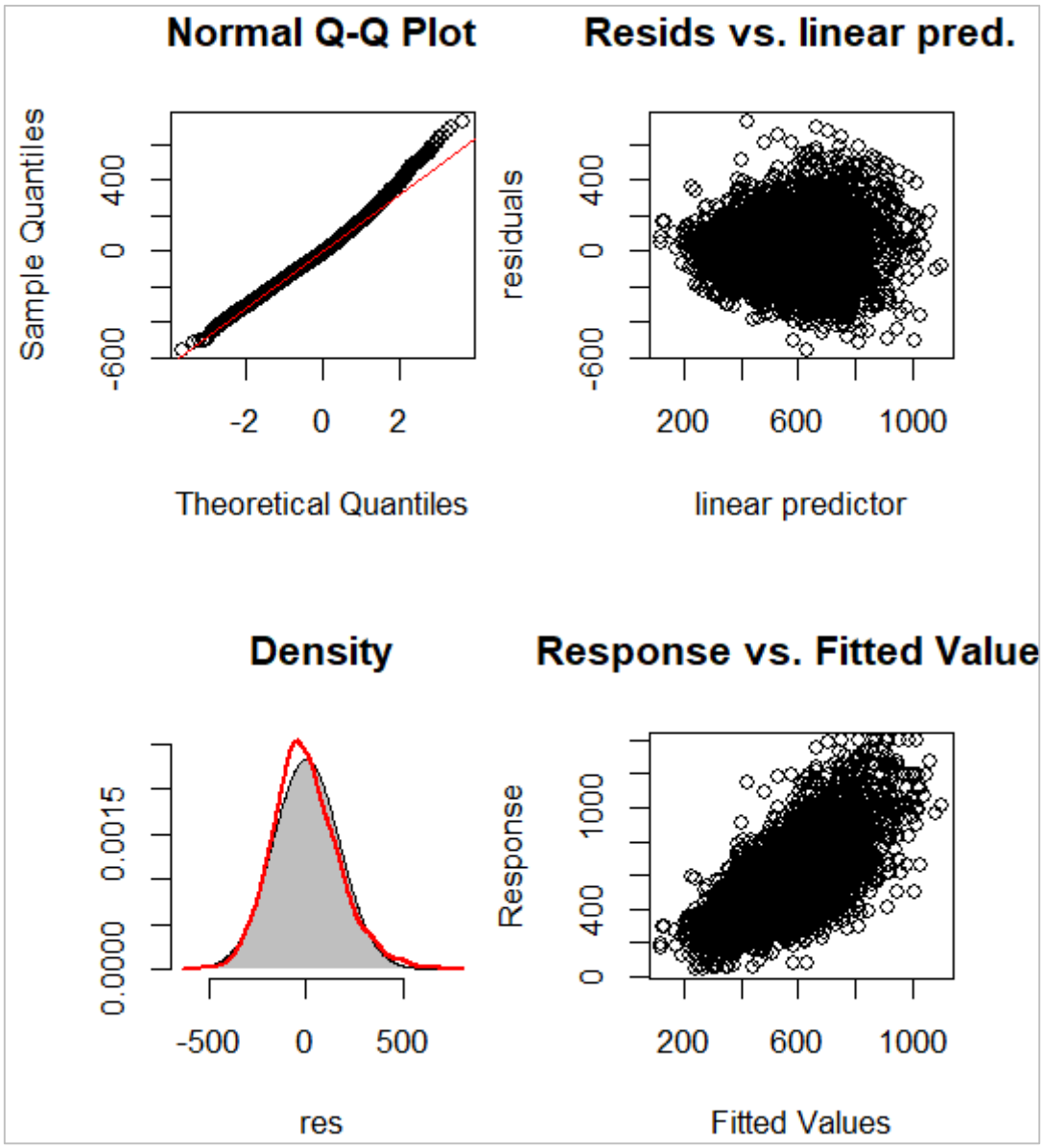


Figure 5.2: The diagnostic plots of the best fit GAMM
The red curve represents the residual distribution of the selected Model 3

All the patterns were acceptable with the Q-Q plot approximately at 45^0 to the horizontal with less indication of heavy tails for outliers. The residuals were approximately randomly distributed around zero with no clearly defined shape for suspecting non-constant error variance. This was confirmed by all the M-indices (Table 5.2) that were close to 1 indicating that it is less likely that there were missing patterns in the residuals. The residual behaviour in the density plot showed an approximate normal distribution as expected. Although slightly wide, the actual response versus the fitted was on the diagonal line as expected too.

The Table 5.2 summary of the selected Model 3 also shows the significance of the model terms where the intercept was estimated to be $s_0 = 574.3461$ cells/mm^3, $p < 0.001$, a value very close to the observed overall cohort average of 570 cells/mm^3 (Table 2.3). The overall covariate smooths terms for s(MCHC) with p = 0.3866, s(calcium) with p = 0.7152, s(magnesium) with p = 0.7418 and s(potassium) with p = 0.3348 were insignificant contributors to the $CD4^+$ count changes. Consequently, all the $CD4^+$ count trends within the separate HIV infection phases were also not significantly different from each other in response to these covariates. This was also confirmed by their smooth terms showing very small effective degrees of freedom. There was no sufficient evidence to suggest that the random smooth terms s(MCV, phase) with p = 0.3808, s(monocytes, phase) with p = 0.2127, s(basophils, phase) with p = 0.4206 and s(folate, phase) with p = 0.7075 were significantly different from each other. However, their overall smooth terms contributed to the $CD4^+$ count changes during the follow up period.

The following section provides detailed graphical displays of the response trends showing some possible optimal set points for influencing the $CD4^+$ cell count within the infection phases.

5.3.2 The results of visualising the trends and optimal set-points

Significant difference between the random smooths

General upward trends: Figure 5.3 shows the covariates that positively influenced the $CD4^+$ cell count overall and having different trends within the HIV infection phases. Generally, an increase in lymphocytes, haematocrit, platelets, albumin and ALP was associated with an improved $CD4^+$ cell count. With the exception of ALP, they showed an almost overall direct relationship with the $CD4^+$ count although the rates of change were fairly low. An increase in ALP approx. between 60 – 100 IU/L, resulted in a sharp increase in the overall $CD4^+$ count and then levelled off thereafter. The upward $CD4^+$ count trends exceeded the cohort's average at approx. lymphocytes count > $2x10^9$/L, haematocrit > 35%, platelet count > $350x10^9$/L, albumin > 42 g/L and ALP > 70 IU/L.

The behavioural patterns of the random smooths were quite complex. Recalling that the average observed $CD4^+$ counts for the early and established phases were below the cohort's average (Figure 2.9), the GAMM plot showed that during the early phase, the $CD4^+$ count remained below the cohort average despite the increase in the lymphocytes. During the acute

and established phases, the $CD4^+$ count declined in response to the lymphocytes increase when approx. < $2x10^9$/L. At lymphocytes count approx. < $2.5x10^9$/L, ART supported a direct influence on the $CD4^+$ count and this relationship diminished as lymphocytes increased beyond the $2.5x10^9$/L but the $CD4^+$ cell counts remained well above average. Above this point, the $CD4^+$ cell counts in the pre-treatment phases were below the average. In response to haematocrit, the $CD4^+$ cell count was staggering below the average during the established phase, the period during which the lowest $CD4^+$ counts were recorded. The $CD4^+$ counts increased with an increase in the haematocrit during the acute and early phases whilst declining during medication (ART). The interaction with medication showed that the $CD4^+$ count dropped to below average at haematocrit approx. > 40%. Since the $CD4^+$ count was negatively related to the haematocrit during the ART phase whilst positive in both the acute and early phases, the plot indicated that maintaining the haematocrit within the neighbourhood of approx. 40% improved the $CD4^+$ count to above average in all the three phases (acute, early and ART).

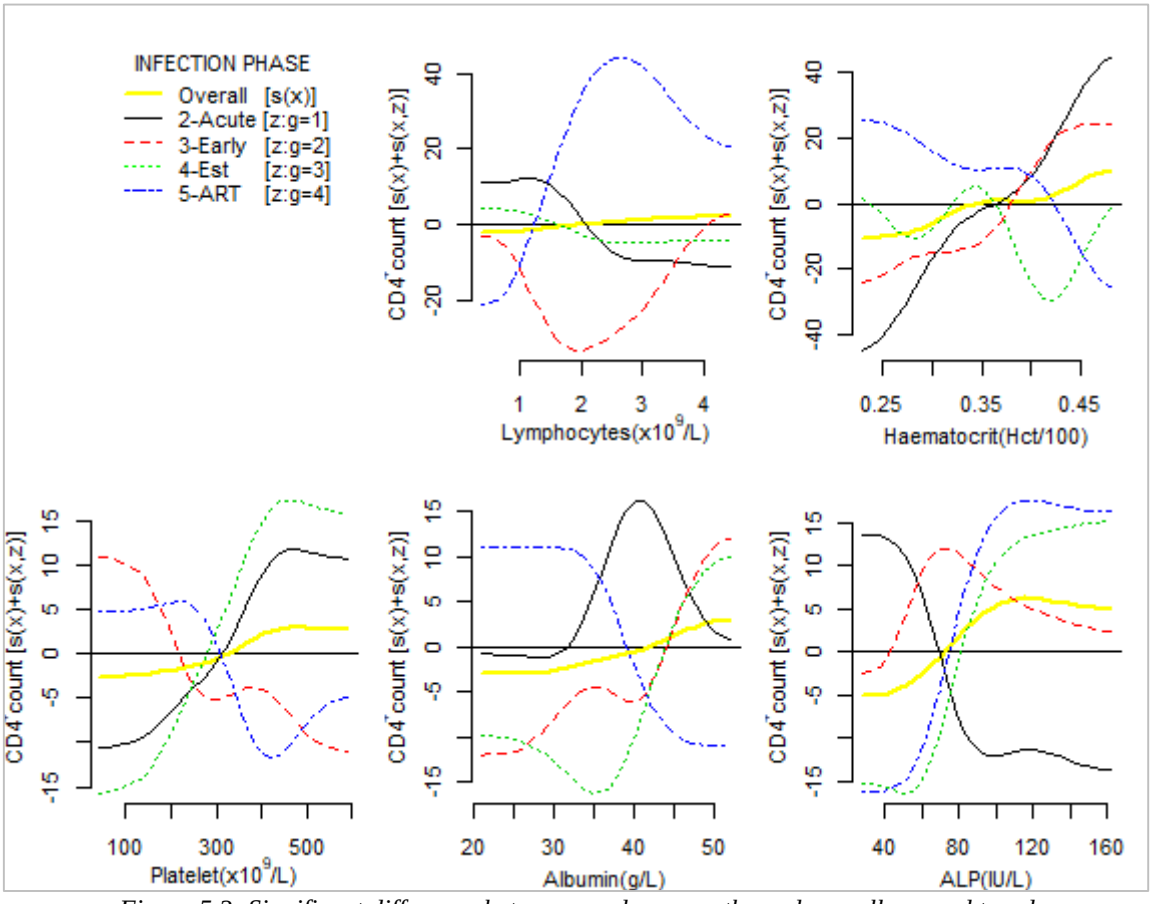


Figure 5.3: Significant difference between random smooths and overall upward trend

According to our data recorded at the lowest levels of $CD4^+$ counts (established phase) and during high viral load (acute phase), an increase in the platelet count positively influenced the

$CD4^+$ counts. Desirable linear effects were at platelet count approx. > $275x10^9$/L and levelled off at approx. > $450x10^9$/L. The trends showed that the rate of such platelet influence was higher during the established phase than the acute phase. These trends were opposite to those observed in the early and ART phases where the influence on the $CD4^+$ cell count to above average was only at lower platelet count approx. < 200 $x10^9$/L. During the acute phase, desirable $CD4^+$ counts were observed at almost the entire range of the recorded albumin measurements (20 – 50 g/L) with the most desirable effects in the neighbourhood of approx. 40g/L. The ART trend behaved oppositely to those of both the early and established phases in response to the albumin. During the ART, the albumin desirably influenced the $CD4^+$ cell count at lower levels of approx. < 37g/L yet the early and the established phases required that the albumin be approx. > 45g/L. The random smooths shapes of ALP effect on the $CD4^+$ count were almost the same during the established and ART phases with the ART showing a slightly better influence on the $CD4^+$ cell count. With the exception of the acute phase, all the other infection phases showed that at ALP approx. > 80IU/L the $CD4^+$ count cell count was above average with the early phase having desirable effects during the entire range of the recorded measurements (40 – 160IU/L). The ALP and $CD4^+$ count cell count were inversely related during the acute phase with the desirable effects of ALP at approx. < 60IU/L.

General downward trends: Generally, the cohorts' total protein and sodium were negatively related to the $CD4^+$ cell count (Figure 5.4).

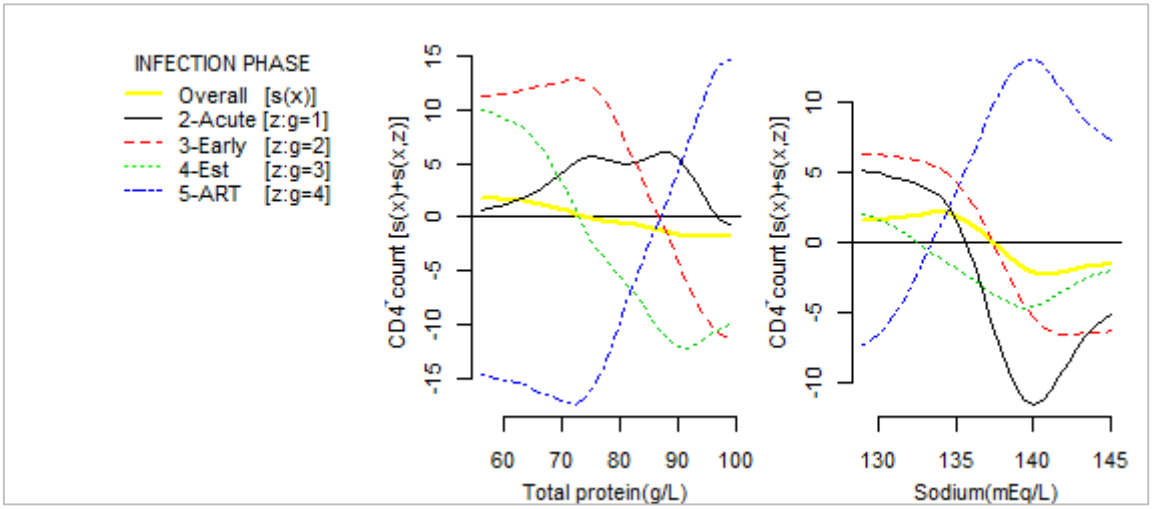


Figure 5.4: Significant difference between random smooths and overall downward trend

The interaction between HIV treatment and these covariates showed a positive influence on the $CD4^+$ count and the most desirable effects were observed at total protein approx. > 85g/L and sodium approx. > 134mEq/L. Although elevated sodium levels influenced $CD4^+$ count to above

average, the trend nosed down at approx. > 140mEq/L during this period of medication uptake. However, the $CD4^+$ counts remained high above average at the sodium levels approx. > 140mEq/L. During the acute phase, the $CD4^+$ count remained above average in response to all the recorded total protein levels (60 – 100g/L). However, at lower $CD4^+$ counts (early and established phases), an increase in the protein levels negatively impacted on the $CD4^+$ count with the lowest $CD4^+$ counts (established phase) being the hardest hit. Desirable effects of the total protein on the $CD4^+$ count were observed at approx. < 75g/L during the established phase whereas at approx. < 85g/L in the early phase. The sodium had negative effects on the $CD4^+$ count during all the pre-treatment phases with the established phase being the most affected again. The plot indicated that all the pre-treatment phases would generally influence the $CD4^+$ cell count to above average at optimally lower sodium levels of approx. < 135mEq/L with the established phase desirable effects restricted to approx. < 132mEq/L.

Irregular trends (more complex): An increase in the LDH and red blood cells produced complex trends in both the overall and the within phase $CD4^+$ count trends (Figure 5.5).

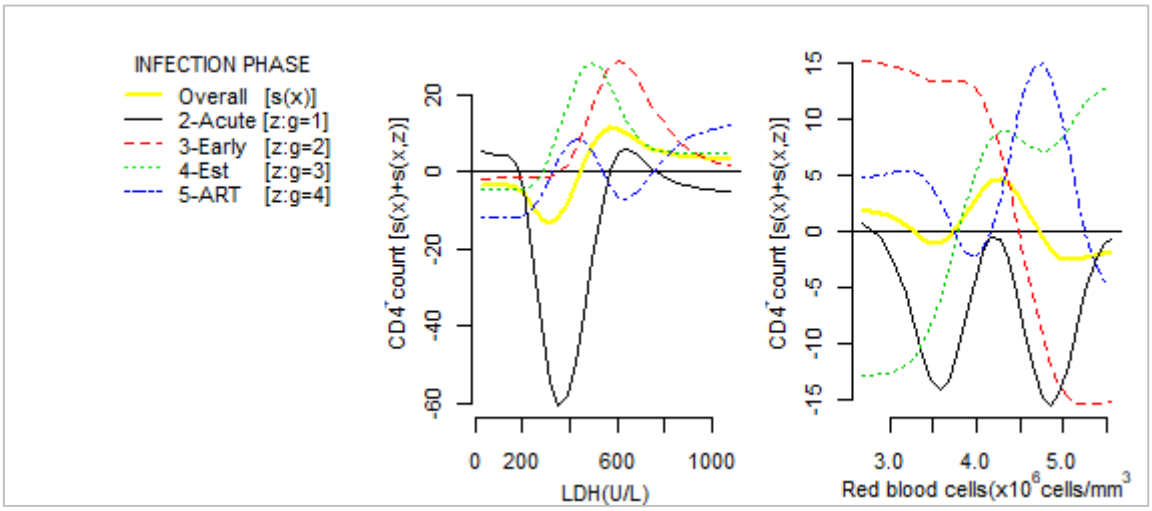


Figure 5.5: Significant difference between random smooths and overall irregular trends

Although fluctuations existed in the overall $CD4^+$ count trend in response to LDH, the $CD4^+$ count remained fairly constant and above average at approx. > 500U/L of LDH. On the other hand, the overall $CD4^+$ count trend in response to the red blood cells fluctuated around the mean. The effects of medication caused the $CD4^+$ count trends in the ART phase to also fluctuate in response to both the LDH and red blood cells. Both covariates were hardly associated with $CD4^+$ counts above average during the acute phase. At lower records of the $CD4^+$ counts (early and established phases), the LDH of approx. > 300U/L showed desirable

effects on the $CD4^+$ count. In response to the red blood cells during these early and established phases, only the lowest records of the $CD4^+$ counts (established phase) positively responded to the red blood cell increase. The plot revealed that the optimal red blood cell count for both the early and established phases could be set in the neighbourhood of approx. $4.2x10^6$ cells/mm^3.

Insignificant differences between the random smooths

General upward trends: Although the random smooths for MCV and basophils show different shapes (Figure 5.6), these trends were found to be statistically insignificantly different. However, the overall $CD4^+$ count trends showed a statistically significant increase in response to a unit increase in these covariates. The plot showed that the cohort's overall MCV supported the $CD4^+$ count to be above average at approx. > 90fL. Generally, the increase in the basophils corresponded to an increase in the $CD4^+$ count but fluctuating very closely to the cohort's average.

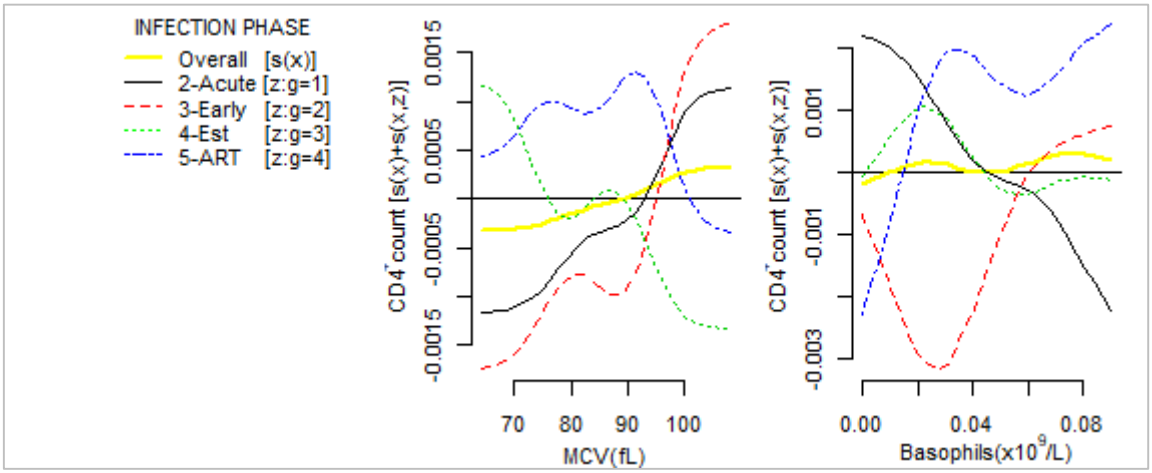


Figure 5.6: Insignificant difference between random smooths and overall upward trend

General downward trends: Similarly, there was no significant difference in the effect of monocytes and folate on the $CD4^+$ count across the HIV infection phases (Figure 5.7). However, the general increase in these covariates was associated with a significant decline in the $CD4^+$ cell count. The overall trends indicated that the monocytes count and folate showed desirable effects on the $CD4^+$ cell count at measurements of approx. < $0.5x10^9$/L and < 15nmol/L respectively.

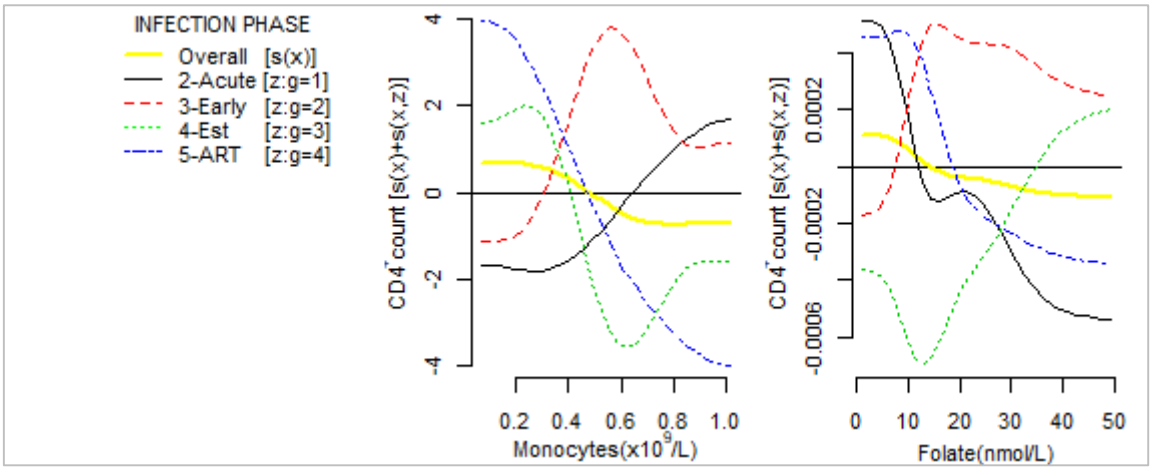


Figure 5.7: Insignificant difference between random smooths and overall downward trend

Insignificant general trends: Despite the different shapes of the $CD4^+$ count trends either overall or within the infection phases, potassium, magnesium, calcium and MCHC had no significant effect on the $CD4^+$ count behavioural changes (Figure 5.8).

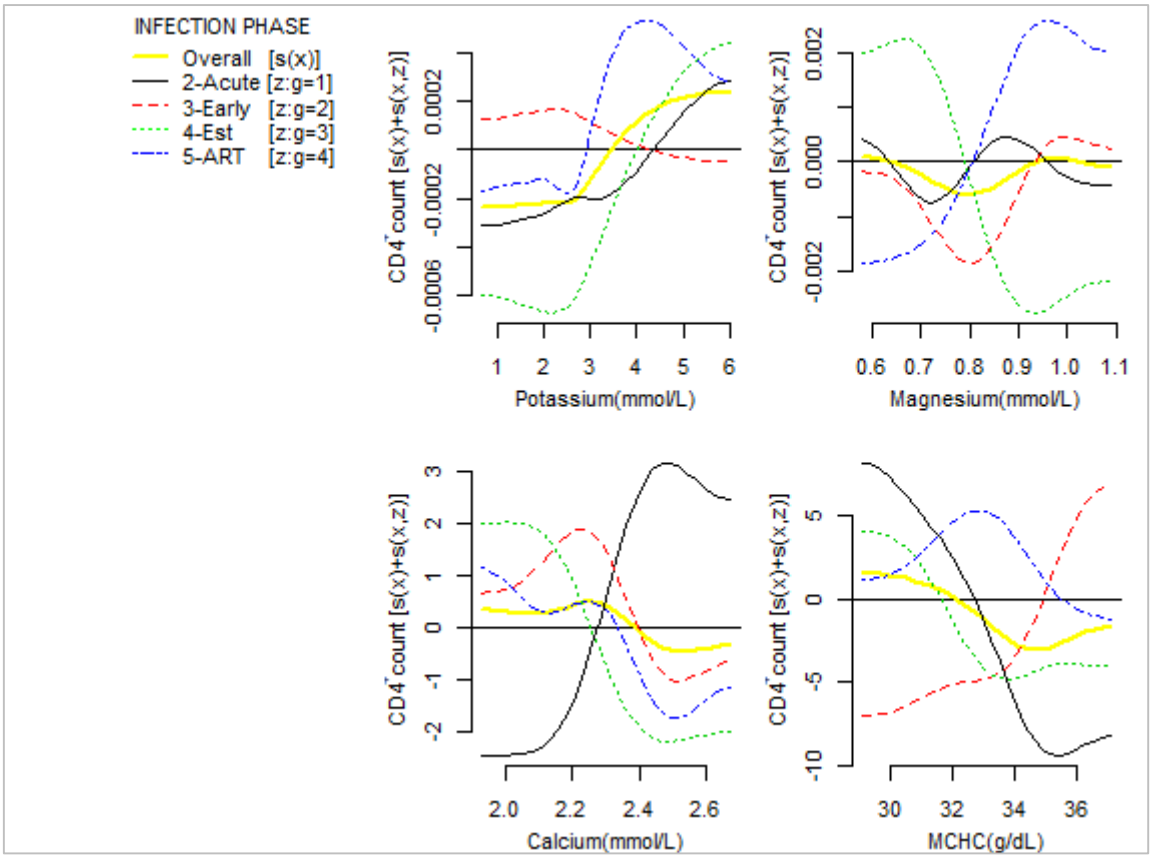


Figure 5.8: Insignificant terms

5.4 Clinical interpretation of the results

This chapter visually examined the $CD4^+$ count trends in response to the strongest clinical covariates in an attempt to discover possible covariate adaptive optimal set-points for positively influencing the $CD4^+$ cell count in HIV infected patients. Among the selected as the strongest $CD4^+$ count covariates are the lymphocytes that are B or T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012) which also consists of the $CD4^+$ cells, a T cell type (Papagno et al. 2004). Hence we found the overall linear relationship between the lymphocytes and $CD4^+$ count. Since HIV is known to mainly attack the $CD4^+$ cells (Weston and Marett 2009), this suggests the decline in the $CD4^+$ count during the pre-treatment phases of our data despite the lymphocytes increase. The suppression of HIV during the ART (Hunt et al. 2003) had consequently seen the high number of $CD4^+$ cells being spared during this treatment phase. These findings on the $CD4^+$ count behaviour in response to lymphocytes were a confirmation of the expected results giving confidence to the accuracy of the fitted model. Monocytes for fighting against pathogens (Project Inform 2007) have been reported to be infected by HIV (Pasupathi et al. 2008) such that their count was supposed to be similarly affected as the $CD4^+$ count. However, we observed a paradox in our data where there was a general inverse relationship between the monocytes and the $CD4^+$ cell count. The damage to body tissues and inflammation as indicated by basophils (Project Inform 2007) was only observed from an overall point and likely due to the basophils being the least abundant leucocytes (Min et al. 2011). A study by (Sloand et al. 1992) found that a low blood clotting condition (platelet count (Project Inform 2007, NAM 2012, National Institutes of Health Clinical Center 2015, James 2017)) was associated with a low $CD4^+$ count. Our data confirmed the same relationship but only holding during the period of high viral load, the acute phase (Bellan et al. 2015, Manoto et al. 2018) and established phases where the lowest $CD4^+$ counts were recorded. During these two phases, the optimal set-point for the platelet count was observed to be approx. $> 450 \times 10^9$/L which was higher than the normal reference range of $178 - 454 \times 10^9$/L (Lawrie et al. 2009, Omuse et al. 2018).

The general increase in the $CD4^+$ cells in response to the tissue oxygenation based on haematocrit and MCV was also observed in a study by (Vanisri and Vadiraja 2016b). This is likely because these two clinical covariates are both red blood cell indices for determining the level of tissue oxygenation (Jensen et al. 1998, Wintrobe and Greer 2009, Arika et al. 2016). The indices' contribution to the $CD4^+$ count was greatly affected during high viral load, the

acute phase. The red blood cells are responsible for oxygen transportation (Jensen et al. 1998, Wintrobe and Greer 2009) and LDH catalyses the compensation of energy levels during insufficient oxygen (Valvona et al. 2016). Both were associated with lower than average $CD4^+$ counts during the acute phase. This high viral replication phase (Weston and Marett 2009) has been reported to have complex relationships with oxygen effects (Morinet et al. 2015) which may also suggest the twisted $CD4^+$ count trends in response to the LDH and red blood cell in our data. Aerobic endurance is referred to as the functional state of the oxygen transport system (Baquet et al. 2003) and has been reported to be reduced in HIV positive patients than negative ones (Cade et al. 2003, Oursler et al. 2006, Chisati and Vasseljen 2015). Our results based on the LDH suggested that aerobic endurance was associated with a negative impact on the $CD4^+$ count mostly during the acute phase.

The acid-base and normal water balance (total protein (Spectrum 2007)) supported $CD4^+$ cell counts above average at high viral loads for almost all the recorded measurements of the total protein between 60-100g/L. The normal total protein range is known to be between 60-80g/L (Busher 1990) and correspond to the range in which our results indicated $CD4^+$ count above average for the early and established phases. Also revealed is that the longer the patient has been leaving with the virus without medication, the less responsive was the $CD4^+$ cell count to protein levels. However, our data showed that during treatment, the normal protein range had negative effects on the $CD4^+$ cell count. At total protein levels approx. > 75g/L during ART, a positive linear relationship with $CD4^+$ count was observed and the $CD4^+$ counts exceeded the average at approx. > 90g/L of total protein. This confirmed the report by (Ugwuja and Eze 2007) that the serum protein increases with highly active antiretroviral therapy which also enhances the $CD4^+$ cell count (Hunt et al. 2003). Albumin which is also a type of protein (Buzanovskii 2017) helps with tissue nourishment (James 2017). Both total protein and albumin results were consistent in positively influencing the $CD4^+$ count to above average at almost all their measurements during the acute phase. The albumin normal reference range is considered to be between 35 - 50g/L (Vincent et al. 2003) and was associated with desirable $CD4^+$ counts at elevated viral load in our data. However, this range corresponded to a sharp decline in the $CD4^+$ count during medication. It has been reported that serum albumin concentrations increase significantly on ART initiation (Chong et al. 2015). To positively influence the $CD4^+$ cell count in response to albumin during ART, the data suggested that albumin levels be lower than normal (approx. < 35g/L) whilst higher albumin levels (approx. > 45g/L) being favourable for the early and established phases. The general direct positive

relationship between albumin and the $CD4^+$ count concurred with the studies by (dos Santos and Almeida 2013, Pralhadrao et al. 2016).

The normal ALP is known to be in the range 30-120 IU/L (Kratz et al. 2004, Park et al. 2013), the range in which our data showed an inverse relationship with the $CD4^+$ count in the presence of a high viral load (acute phase). During the acute phase, $CD4^+$ cell count improved to above average at lower ALP (approx. < 70IU/L). After the acute phase, the immune system is known to fight back to restore the $CD4^+$ count (Manoto et al. 2018). Within 3-12 months of infection (early phase), the $CD4^+$ cell count responded well to normal ALP and remained above average. The results further showed that as the immune system continued to fight back with (ART phase) or without treatment (established phase), the ALP showed a strong positive linear relationship with the $CD4^+$ cell count. From at least 3 months of infection, the ALP's positive linear association with the $CD4^+$ count diminishes beyond the normal ALP upper limit of 120 IU/L but still supporting the $CD4^+$ count to above average. However, at such elevated ALP levels, it is known to be an indication of liver damage (Pasupathi et al. 2008). Sodium also like calcium plays a crucial role in the regulation of water balance, blood pressure, blood volume, heart rhythm and the brain and nerve function (Project Inform 2007, James 2017, The Johns Hopkins Lupus Center 2017). Under normal circumstances, it operates between 135-145mEq/L (NIOS 2012, Strazzullo 2014). The results also indicated that there was a shift in the sodium optimal range where measurements below the normal range during the pre-treatment phases were associated with an improved $CD4^+$ count above average. This may suggest the changes in the osmotic gradient between the extracellular and intracellular fluid in cells due to sodium (Shu et al. 2018) in the presence of viral infection before treatment. Upon viral suppression during ART, there was a direct relationship between the sodium and $CD4^+$ count. A similar positive correlation was observed by (Braconnier et al. 2017) among HIV positive patients but without considering infection phase. Our data further revealed that the positive correlation during ART tails off at approx. > 140mEq/L of sodium but still influencing the $CD4^+$ count to reach levels above average. Many foods naturally contain folate, a B-vitamin which is needed for cell growth and metabolism (Arya and Kumar 2012, Dieticians of Canada 2014). Contrary to (Adhikari et al. 2016) that it improves the $CD4^+$ count, our data showed that generally a unit increase in the folate was associated with a drop in the $CD4^+$ cell count.

5.5 Summary

The general $CD4^+$ count trends in response to the covariates were either upwards, downwards or irregular. The phase specific trends were mostly taking various shapes indicating different $CD4^+$ count trends in response to the covariates with the infection phases. Their optimal set points tended to drift and adapt to either new ranges or overlapped with the known reference ranges to positively influence the $CD4^+$ cell counts. Recommendation for phase specific $CD4^+$ cell count influence in adaptation to HIV invasion includes the monitoring of the strongest covariates related to dietary conditions (sodium, albumin and total protein), tissue oxygenation (red blood cells and its haematocrit) and hormonal control (LDH and ALP). The GAMM has shown to be effective in modelling highly non-linear relationships that can only be interpreted by visualisation of the trends. The trends clearly bring into perspective the nature of the response curves that cannot be imagined. Although we are not able to extract all the information such as the actual response values due to the transformation of the dependent variable, the most likely nature of the behavioural patterns of interest is well understood. In biological terms, the turning points in the non-linear trends often have a meaning and if their close estimates are known, facilitates the monitoring process in patient care. The GAMM showed such turning points in the trends of the $CD4^+$ count changes where the visual inspection was not easy in approximating the values. More so, the GAMM approach is not equipped to provide the rate of change between the change-points. The understanding of such break-points calls for segmented regression which is the central focus of the following chapter.

CHAPTER 6 SEGMENTED REGRESSION MODELS

In the previous chapter, we observed the non-linear trajectories of the $CD4^+$ count in response to the covariates. In confirmation to (Yee and Mitchell 1991), the non-linear models captured a wide range of shapes in the response curves. These came with number of or fewer turning points, an indication that we succeeded in smoothing the curves as recommended by (Hastie and Tibshirani 1990, Muggeo 2003). This is because too many turning points are questionable for interpretation purposes (Seber and Wild 1989). Further, as reiterated by (Xiang 2001), some patterns that could not easily be imagined were discovered. Also, in line with (Hastie and Tibshirani 1990), the turning points were not easy to interpret such that approximate values were considered based on visual inspection. However, in biological applications, the turning points are usually either cut-off points for classification purposes (Muggeo 2003) or locations where the effect on the response changes abruptly and associated with a biological meaning (Robbins et al. 2006). Medical practitioners rely on laboratory reference ranges of clinical covariates as a decision making tool when monitoring the health status of patients (Horn and Pesce 2003, Segolodi et al. 2014). For example sodium concentration < 35mmol/L is a common electrolyte abnormality in clinical practice that is referred to as hyponatremia (Braconnier et al. 2017). To some extent the additive model can attempt to provide information on the turning points but the drawback of this procedure is that they require bootstrapping techniques (Molinari et al. 2001). Further, the basis splines induce pseudo-variables that are related to the beta parameters and not necessarily measuring the response effect (Hastie and Tibshirani 1990). Other classical methods such as polynomial regression can be used but are also limited to accounting for the non-linear effects only. As such, to augment the powerful data visualisation and pattern discovery from additive models (Yee and Mitchell 1991, Wood 2000, Xiang 2001, Wood 2017), we adopt the segmented regression models that are useful in extracting the turning point information of the non-linear trajectories (Feder 1975, Kuchenhoff 1997). These locations or cut-off points have also been called with other different names such as threshold-points (Molinari et al. 2001), switch-points, change-points, break-points, transition-points (Seber and Wild 1989, Stasinopoulos and Rigby 1992). We will simply use the term “break-point” throughout this chapter as it seems more appealing in this context. The segmented regression models that are applied to detect the break-points are also either called broken-lines (Robbins et al. 2006), piecewise (Ertel and Fowlkes 1976, Tishler and Zang 1981) or multi-phase (Beckman and Cook 1979). Also preferred in this context is the term “segmented” regression model. Our interest is in the break-points and the slopes of the

segments. As observed in the previous chapter, the transformed response in the graphical displays of the additive models could not easily be projected to the original response scale for the slope interpretation. The segmented models are able to preserve the original response scale.

The segmented regression models are non-linear models that consist of hybrid functions defined by multiple sub-functions in the form of straight lines, curves or both having different parameters and the lines or curves connected by the break-points. The exploratory loess trends in Figures 2.13 and 2.14, as well as the findings from Chapter 5, suggest that straight line multiple sub-functions are most appropriate. Standard maximisation procedures cannot be applied to the non-linear break-point parameters (Seber and Wild 1989). Further, log-likelihood is piecewise differentiable complicating standard likelihood-based inference such that the classical regularity conditions are not met in the estimation of the break-point parameters (Feder 1975, Seber and Wild 1989, Kuchenhoff and Ulm 1997). A number of methods have been proposed to handle the estimation of the break-point parameters. For example, if the break-points are fixed, they can be separately estimated by linear models or calculated from scatter smooth plot (Vermont et al. 1991, Kunst et al. 1993). In a similar case, specific algorithms or grid-search can also be used (Hawkins 1976, Ulm 1991, Rigby and Stasinopoulos 1992). However, the resultant beta parameters of such break-points are considered not completely true parameters. In the case where the break-point is interpreted as no effect in a biological application and the present distribution not accommodating such no-effect break-point, the response probability distribution can be modified to include that no-effect break-point parameter. This approach is known to be poor and difficult to generalize (Cox 1987). The segmented relationship can also be estimated by a continuous differentiable function on the neighbourhood of the unknown break-point (Tishler and Zang 1981) or the entire range of the covariate (Bacon and Watts 1971, Griffiths and Miller 1973, Tishler and Zang 1981). This way, extra parameters are estimated in the modelling process (Bacon and Watts 1971) which tends to investigate only the Gaussian model such that the break-point of interest is no longer the parameter to be estimated (Tishler and Zang 1981). It can be assumed that there exists a linear effect before a given break-point and a quadratic effect after that point again. This allows the regression model to have continuous first derivatives (Pastor and Guallar 1998). However, the estimation of the linear-quadratic segmented model requires a non-standard maximisation algorithm, induce high correlation among the parameter estimates which are also not well approximated by the Gaussian model (Muggeo 2003).

Bayesian Markov chain Monte Carlo (MCMC) segmented regression methods do not require the assumption of differentiability but demand heavy computational effort with even simple models due to a large number of iterations (Gossl and Kuchenhoff 2001).

The common limitations of the above-mentioned segmented regression models include the detection of single break-points only and in few cases of multiple parameters, the break-points are not treated as true parameters. There is a tendency of constraining the model to the response variable probability distribution and associated with a high demand on the computational effort that is likely to increase the model complexity (Muggeo 2003). To counter for the computational effort and to bypass the non-linear and non-differentiable problems, (Muggeo 2003) proposed a much simpler iterative method that requires starting values only for fitting the segmented linear models. The break-points and slopes of the segments are estimated together with standard errors and also providing the plot of the results. This procedure is currently implemented in R software, the **segmented** library (Muggeo 2017) and has the capability to deal with an unknown number of break-points although smoothed scatter plots are recommended at the initial stage of the data exploration for they give a good sense of the existence of a certain number of break-points (Muggeo 2003). They provide reasonable starting points. It is known that given such initial values, algorithms always reach exact solutions and minimises the problem of local maxima (Julious 2001). The algorithm in the **segmented** package is less sensitive to the starting values because it implements the bootstrap starting (Wood 2001). At convergence, the algorithm indicates the existence of significant break-points together with difference-in-slope parameters that are tested and confidence intervals also shown. This method can be applied virtually to any regression model having one or more segmented relationships including the handling of several covariates. Basically the unknown multiple break-points in many covariates are automatically detected and simultaneously providing inferences of all the segmented model parameters (Muggeo 2017). Although the method might have its own difficulties, say if either there exist high correlations among the break-points or having break-points near the edge of the covariate leading to wider confidence intervals, it outweighs all the previously discussed methods for modelling segmented relationships (Muggeo 2003).

6.1 The segmented model specification

In our data, both the response, $CD4^+$ count denoted by y_{ij} for $i = 1,..,n$ individuals each with $j = 1,...,T$ repeated measurements and the clinical covariates x_k's for $k = 1,...,p$ are numeric

variables. Since the segmented regression model relies on updating a general linear model (Muggeo 2017) and in line with the previous chapter, we will also ignore the measurement dependence by fitting phase specific models of the form $y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ijk} + \varepsilon_{ij}$, where the random error $\varepsilon_{ij} \sim N(0, \sigma^2)$. The parameterisation to the segmented regression model for the relationship between the response and covariates will then follow using an iterative approach (Muggeo 2003). The updated segmented model is expected to show a better model fit than the general linear model. The deviance of the segmented regression model should be smaller than that of the general linear model (Coelho et al. 2013) and the complexities of the models also compared using the AIC. For easy understanding, the sections below will first illustrate the model formulation of a single break-point and then followed by multiple break-points. The model parameterisation focuses on the relationship between the break-points and the slopes of the associated segments.

6.1.1 A singe break-point and a single covariate

A single break-point simply produces two line segments and two slopes. Given a response variable y that non-linearly depends on the covariate x, the break-point for the two line segments can be denoted by $\psi \in x$, say. If the slope of the left line segment is β_0 for $x \le \psi$, and $(\beta_0 + b_1)$ for $x > \psi$, is the slope of the right line segment, the segmented regression model is given by

$$\beta_0 x + b_1 (x - \psi)_+, \tag{6.1}$$

where $(x-\psi)_+ = (x-\psi) \times I(x > \psi)$ such that $I(x > \psi) = 1$ if $x > \psi$ is true and b_1 is the difference between the right and left slopes, usually referred to as the difference-in-slopes. If $|b_1| > 0$, it implies that a break-point exists. The segmented regression model (6.1) is non-linear and non-differentiable with the log-likelihood also not differentiable at $x = \psi$. If $\psi^{(0)} \in x$ is the estimated starting value of the break-point, the segmented regression model can be fitted by means of linearization with a relevant first-order Taylor's expansion around the $\psi^{(0)}$ and the $(x-\psi)_+$ in (6.1) can be represented as

$$(x-\psi)_+ = (x-\psi^{(0)})_+ + (\psi-\psi^{(0)})(-1)I(x>\psi^{(0)}), \tag{6.2}$$

where $(-1)I(x>\psi^{(0)})$ is the first derivative of $(x-\psi)_+$ that is assessed in $\psi^{(0)}$. The equation (6.2) can be reduced to

$$(x-\psi)_+ = U^{(0)} + \gamma^* V^{(0)}, \tag{6.3}$$

where $U^{(0)} = (x-\psi^{(0)})$, $\gamma^* = (\psi-\psi^{(0)})$ and $V^{(0)} = (-1)I(x>\psi^{(0)})$. Substituting (6.3) into (6.1) gives an iterative model

$$\begin{aligned}\beta_0 x + b_1(U^{(0)} + \gamma^* V^{(0)}) &= \beta_0 x + b_1 U^{(0)} + b_1\gamma^* V^{(0)} \\ &= \beta_0 x + b_1 U^{(0)} + \gamma V^{(0)}, \ where\ \gamma = b_1\gamma^*.\end{aligned} \tag{6.4}$$

Consequently the s^{th} iteration step requires that the break-point $\psi^{(s)}$ be fixed and the model to be fitted becomes

$$\beta_0 x + b_1 U^{(s)} + \gamma V^{(s)}. \tag{6.5}$$

If follows that the model fitted at the next $(s+1)^{th}$ iteration step with the break-point $\psi^{(s+1)}$ is given by

$$\beta_0 x + b_1 U^{(s+1)} + \gamma V^{(s+1)}. \tag{6.6}$$

The improvement of the break-point $\psi^{(s)}$ to $\psi^{(s+1)}$ is measured by γ and is calculated as $\psi^{(s+1)} - \psi^{(s)} = \frac{\hat{\gamma}}{\hat{b}_1}$. When $\hat{\gamma} \approx 0$, it is an indication that there is no improvement in the iteration, the algorithm stops and $\hat{\psi} \equiv \psi^{(s)}$. The 95% Wald-based confidence interval (McCullagh and Nelder 1989) for the $\hat{\psi}$ is given by $\hat{\psi} \pm 1.96 \times SE(\hat{\psi})$ where the standard error of $\hat{\psi}$ is given by

$$SE(\hat{\psi}) = \sqrt{\frac{\mathrm{var}(\hat{\gamma}) + \mathrm{var}(\hat{b}_1)\left(\frac{\hat{\gamma}}{\hat{b}_1}\right)^2 + 2\left(\frac{\hat{\gamma}}{\hat{b}_1}\right)\mathrm{cov}(\hat{\gamma}, \hat{b}_1)}{\hat{b}_1^{\,2}}}.$$

6.1.2 Multiple break-points and a single covariate

The exploratory plots in Figure 2.13 and findings from Chapter 5 have indicated the existence of at least one break-point in the covariates. In the case of multiple break-points, if there exist M break-points in the range of a given covariate x, they produce $M+1$ sub-intervals of x, $M+1$ line segments and $M+1$ slopes. If the M break-points are $\psi = (\psi_1, \psi_2, ..., \psi_M)^T$ and $(\psi_1, \psi_2, ..., \psi_M)^T \in x$, it also implies that there exist a categorical indicator variable W with M levels such that $\psi = \psi_1 W_1 + \psi_2 W_2 + ... + \psi_M W_M$ where

$$W_m = \begin{cases} 1, \ for\ observations\ in\ m^{th}\ group/\ sub-interval \\ 0, \ otherwise \end{cases}.$$

It is important to note that the m^{th} sub-interval W_m is actually referring to the sub-interval to the right of ψ_m which is denoted by $\psi_m < x \le \psi_{m+1}$. Hence, the segmented regression model for the multiple break-points is given by

$$\sum_{m=1}^{M} \beta_0 \{xW_m\} + \sum_{m=1}^{M} b_m (\{xW_m\} - \psi_m)_+ = \beta_0 x + \sum_{m=1}^{M} b_m (\{xW_m\} - \psi_m)_+ , \qquad (6.7)$$

where $\beta_0 x$ models the first slope in the sub-interval $x \le \psi_1$ before the first break-point ψ_1. The relationship (6.7) can be reduced to (6.8)

$$\beta_0 x + \sum_{m=1}^{M} b_m U_m + \sum_{m=1}^{M} \gamma_m V_m , \qquad (6.8)$$

where for the s^{th} iteration step $U^{(s)} = (\{xW_m\} - \psi_m^{(s)})_+$ and $V^{(s)} = (-1) I(\{xW_m\} > \psi_m^{(s)})$ for $m = 1, 2, ..., M$. Successive iterations in the m^{th} sub-interval to improve the break-point $\psi_m^{(s)}$ to $\psi_m^{(s+1)}$ is given by $\psi_m^{(s+1)} = \frac{\hat{\gamma}_m}{\hat{b}_m} + \psi_m^{(s)}$. Since β_0 is the first slope of the segmented lines in the first subinterval $x \le \psi_1$ before the first break-point ψ_1, the m^{th} slope of the sub-interval $W_m = 1$ within $\psi_m < x \le \psi_{m+1}$ is given by

$$\beta_m = \beta_0 + \sum_{m=1}^{M} b_m . \qquad (6.9)$$

where b_m is the m^{th} difference-in-slope of x .

6.1.3 Multiple break-points and multiple covariates

Our data consists of $p=18$ covariates such that a segmented regression model with several covariates x_k's for $k=1,\ldots,p$ and multiple break-points for each covariate can be fitted by generalising (6.8) for each of the x_k's to

$$\sum_{k=1}^{p}\beta_{0k}\,x_k+\sum_{m=1}^{M}\sum_{k=1}^{p}b_{mk}U_{mk}+\sum_{m=1}^{M}\sum_{k=1}^{p}\gamma_{mk}V_{mk}\,, \tag{6.10}$$

where

β_{0k} = the first slope of the k^{th} covariate x_k for $k=1,\ldots,p$.

$U_{mk}=\left(\{x_kW_{mk}\}-\psi_{mk}\right)_+$ for $m=1,2,\ldots,M$ and $k=1,\ldots,p$.

$V_{mk}=(-1)I\left(\{x_kW_{mk}\}>\psi_{mk}\right)$ for $m=1,2,\ldots,M$ and $k=1,\ldots,p$.

ψ_{mk} = the m^{th} break-point of the k^{th} covariate x_k for $k=1,\ldots,p$.

W_{mk} = the m^{th} sub-interval $\psi_{mk}<x_k\le\psi_{(m+1)k}$ of the k^{th} covariate x_k for $k=1,\ldots,p$.

γ_{mk} = the coefficient measuring the difference between two fitted lines connected by ψ_{mk}

It follows from (6.10) that if β_{0k} is the first slope of segmented line in the first subinterval $x_k\le\psi_{1k}$ before the first break-point ψ_{1k}, the m^{th} slope of the sub-interval $W_{mk}=1$ within $\psi_{mk}<x_k\le\psi_{(m+1)k}$ is given by

$$\beta_{mk}=\sum_{k=1}^{p}\beta_{0k}+\sum_{m=1}^{M}\sum_{k=1}^{p}b_{mk}\,. \tag{6.11}$$

where b_{mk} is the m^{th} difference-in-slope of the k^{th} covariate, x_k .

6.2 Data analysis and software

The data analysis was conducted in R software, version 3.5.3. With the aid of the loess breakpoints, the breakpoints and slope parameters of the segmented linear relationships (Muggeo 2003) were estimated using the `segmented` function in **segmented** R library. We adopted the subset of data for each level of the grouping variable (infection phase) as implemented by the piecewise procedure in SAS. The results were further visualised using the function `plot.segmented` and the plots panelled with the aid of `par` function. With the increased number of programming codes, it was necessary again to write macros for simplifying the tasks. Although the loess would guide on the most probable number and position of the breakpoints, at times the algorithm tended to be caught up in a localised turning point giving the results not as expected. It was important to try several initial points with small increments keeping a close eye on the exploratory loess in order to obtain a meaningful and the most accurate results. Voluminous results were also a challenge and the function `multiplot` in R was used to panel the plots in a concise way. All the R codes are presented in Appendix C6.

6.3 Results of fitting the segmented model

6.3.1 The results of model diagnostics

Table 6.1 is a summary of the fitted models. The null represents the model with the intercept

Table 6.1: The results of model diagnostics

Phase	Model	Null	Deviance	AIC
2-Acute	GLM	54 486 557.99	36 558 533.63	12 741.25
	Segmented	54 486 557.99	31 266 439.87	12 713.02
3-Early	GLM	55 944 592.70	27 734 647.71	12 479.39
	Segmented	55 944 592.70	25 085 078.98	12 500.20
4-Est	GLM	49 013 192.22	26 075 699.56	12 420.92
	Segmented	49 013 192.22	23 432 088.96	12 423.58
5-ART	GLM	50 792 042.48	26 982 511.12	12 453.32
	Segmented	50 792 042.48	23 169 290.10	12 420.88

only whereas the deviance shows how well the model improved by the addition of the covariates. All the updated models (segmented) explained the $CD4^+$ count better than the classical general linear models. For example, in the acute phase, the general linear model had a deviance of 36 558 533.63 which dropped to 31 266 439.87 upon modelling the $CD4^+$ count

per segment. However, our results showed that the segmentation may too simplify the model complexity as observed in the acute and ART phases.

The random errors ε_{ij} in the underlying general linear models were assumed to follow a normal distribution with a mean of zero and variance σ^2. The Q-Q plots in Figure 6.1 indicate the fulfilment of the assumption where all the quantile plots were approximately straight along the diagonal line. That is, the standardised residuals which are a measure of the distance between the observed points and the predicted points, were very close to the theoretical quantiles.

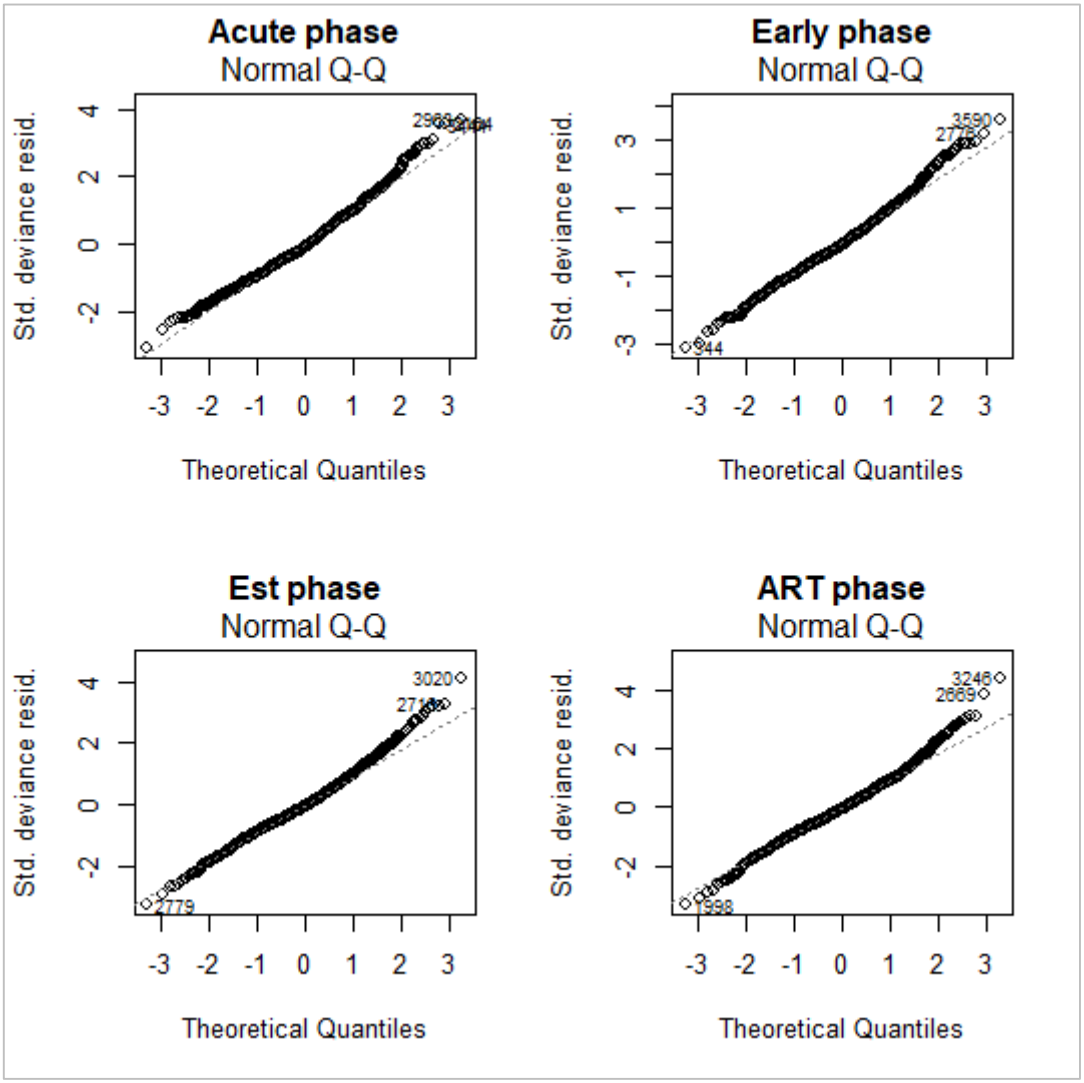


Figure 6.1: Q-Q plots: The test for the normality assumption

6.3.2 The results of the detected break-points

Mostly two breakpoints were detected to signify commonly three possible subintervals of each covariate within which the covariates had possibly different effects on the $CD4^+$ count. These breakpoints (see Table 6.2) echoed the loess trends (Figures 2.12 and 2.13).

Table 6.2: The detected breakpoints of the covariates and their confidence intervals

Covariate	Point	2-Acute	3-Early	4-Est	5-ART
k	m	ψ_{mk} (95%CI)	ψ_{mk} (95%CI)	ψ_{mk} (95%CI)	ψ_{mk} (95%CI)
Albumin	1	36.50(22.17 ; 50.83)	36.58(32.33 ; 40.82)	40.90(38.10 ; 43.70)	28.13(24.63 ; 31.63)
Albumin	2	-	39.49(36.70 ; 42.29)	-	-
ALP	1	97.25(76.97 ; 117.53)	80.00(43.28 ; 116.72)	116.39(74.46 ; 158.32)	75.58(70.09 ; 81.06)
ALP	2	110.98(92.42 ; 129.54)	-	-	77.00(67.58 ; 86.41)
Basophils	1	0.04(0.01 ; 0.06)	0.04(-0.01 ; 0.09)†	0.04(-0.01 ; 0.09)†	0.04(0.03 ; 0.04)
Basophils	2	0.08(0.07 ; 0.09)	-	-	0.05(0.04 ; 0.06)
Calcium	1	2.52(2.44 ; 2.60)	2.10(1.97 ; 2.23)	2.23(2.02 ; 2.44)	2.10(1.97 ; 2.23)
Calcium	2	-	2.50(1.83 ; 3.17)	2.31(1.97 ; 2.65)	2.45(2.35 ; 2.55)
Folate	1	17.50(13.63 ; 21.37)	20.00(10.50 ; 29.50)	10.30(7.44 ; 13.16)	6.00(0.93 ; 11.07)
Folate	2	25.00(16.97 ; 33.03)	30.00(15.24 ; 44.76)	-	-
Haematocrit	1	0.33(0.25 ; 0.40)	0.45(-3.45 ; 4.35)†	0.36(0.33 ; 0.38)	0.34(0.32 ; 0.36)
Haematocrit	2	0.42(0.30 ; 0.54)	-	0.41(0.39 ; 0.44)	0.36(0.33 ; 0.39)
LDH	1	327.81(289.75 ; 365.87)	365.61(119.02 ; 612.20)	446.00(372.13 ; 519.87)	333.75(60.71 ; 606.79)
LDH	2	607.04(494.76 ; 719.33)	510.26(332.32 ; 688.20)	-	751.50(645.87 ; 857.13)
Lymphocytes	1	0.72(0.48 ; 0.97)	1.11(0.05 ; 2.17)	0.90(0.64 ; 1.16)	2.59(2.36 ; 2.82)
Lymphocytes	2	3.46(3.11 ; 3.81)	-	1.58(1.27 ; 1.89)	-
Magnesium	1	0.75(0.70 ; 0.80)	0.72(0.63 ; 0.82)	0.73(0.50 ; 0.96)	0.70(0.59 ; 0.81)
Magnesium	2	0.88(0.84 ; 0.91)	1.04(0.93 ; 1.15)	0.91(0.40 ; 1.42)	-
Magnesium	3	0.94(0.89 ; 1.00)	-	-	-
MCHC	1	32.53(31.72 ; 33.34)	35.25(34.73 ; 35.76)	31.00(23.54 ; 38.46)	31.48(30.04 ; 32.93)
MCHC	2	32.71(32.27 ; 33.15)	-	35.00(31.01 ; 38.99)	-
MCV	1	90.35(82.73 ; 97.97)	69.97(65.42 ; 74.52)	77.50(67.49 ; 87.51)	86.90(68.40 ; 105.40)
MCV	2	-	91.84(79.14 ; 104.54)	104.73(98.75 ; 110.71)	92.60(84.89 ; 100.31)
Monocytes	1	0.44(-1.27 ; 2.15)†	0.16(0.09 ; 0.23)	0.34(0.27 ; 0.41)	0.47(-2.85 ; 3.79)†
Monocytes	2	-	0.53(0.40 ; 0.66)	0.46(0.36 ; 0.56)	-
Platelet	1	287.42(150.06 ; 424.79)	181.59(160.09 ; 203.10)	457.47(379.27 ; 535.67)	331.00(161.51 ; 500.49)
Platelet	2	487.10(389.74 ; 584.46)	217.88(199.03 ; 236.74)	-	-
Platelet	3	-	455.93(368.54 ; 543.32)	-	-
Potassium	1	3.08(1.95 ; 4.21)	0.89(0.69 ; 1.09)	4.00(1.06 ; 6.94)	2.00(-5.71 ; 9.71)†
Potassium	2	3.59(2.89 ; 4.28)	3.00(2.29 ; 3.71)		5.20(-14.33 ; 24.72)†
Potassium	3	-	3.50(2.71 ; 4.29)	-	-
Total protein	1	70.00(59.27 ; 80.73)	74.73(56.72 ; 92.75)	89.60(83.16 ; 96.04)	70.00(56.76 ; 83.24)
Total protein	2	-	-	-	80.00(65.19 ; 94.81)
Total protein	3	-	-	-	90.00(83.64 ; 96.36)
Red blood cells	1	4.55(4.39 ; 4.71)	5.15(4.66 ; 5.64)	3.18(3.01 ; 3.35)	4.15(4.00 ; 4.31)
Red blood cells	2	4.77(4.63 ; 4.92)	-	5.49(5.00 ; 5.98)	4.37(4.18 ; 4.55)
Sodium	1	133.00(123.88 ; 142.12)	133.51(129.51 ; 137.50)	132.80(128.29 ; 137.31)	134.93(130.32 ; 139.54)
Sodium	2	140.00(137.62 ; 142.38)	140.65(133.48 ; 147.81)	140.70(134.91 ; 146.49)	140.72(138.68 ; 142.77)

† Insignificant at 5% level

A few of the detected breakpoints were statistically proven to be insignificant and herein marked (†) for convenience. The statistically insignificant breakpoints suggested that the rate of $CD4^+$ count change in response to that particular covariate remained the same over the entire range of the repeated measurements. This was observed in basophils (early and established phases), haematocrit (early phase), monocytes (acute and ART phases) and potassium (ART phase). Where at least two breakpoints were detected, there were either too close or far apart indicating sudden or gradual changes in the rate of the $CD4^+$ count respectively. For example, during the acute phase, the two breakpoints of the MCHC were 32.529 (31.722; 33.336) g/dL and 32.714 (32.274; 33.154) g/dL with a difference of barely 0.185g/dL whereas the two

breakpoints for the LDH were 327.809 (289.752; 365.866) U/L and 607.043 (494.757; 719.329) U/L being 279.237U/L apart. Our data showed that the maximum of three breakpoints and consequently the existence of four segmented linear relationships between the $CD4^+$ count and clinical covariates were detected in magnesium, platelet count, potassium and total protein. Although two breakpoints were common, they were not detected in every infection phase for almost all the studied covariates. Whether a single or at least two breakpoints were detected, almost all were also not consistently the same values from one infection phase to the next. The results revealed that only sodium consistently showed the same two breakpoints in all the infection phases at approximately 132mEq/L and 140mEq/L.

6.3.3 The results of the desirable and undesirable covariate subintervals

Table 6.3 shows the influential covariates in our data and their subintervals within which they showed either desirable or undesirable effects on the $CD4^+$ count outcome.

Table 6.3: The desirable and undesirable covariate subintervals

Covariate	Unit	2-Acute	3-Early	4-Est	5-ART
Lymphocytes	$x10^9$/L	0.725 - 3.459	>1.112	0.900 - 1.578	<2.591
Lymphocytes	$x10^9$/L			>1.578[a]	
Albumin	g/L	>36.500	<36.576	<40.900	>28.130
Albumin	g/L		>39.491[a]	>40.900[a]	
Platelet	$x10^9$/L	287.424 - 487.099	<181.595[a]	<457.471	
Platelet	$x10^9$/L		>217.884		
Basophils	$x10^9$/L			0.000 - 0.090[b]	<0.035
Basophils	$x10^9$/L				>0.050[a]
Calcium	mmol/L	<2.518			
Red blood cells	$x10^6$ cells/mm^3			<3.180	
ALP	IU/L			<116.393	
LDH	U/L	<327.809*		<446.000	
LDH	U/L	327.809 - 607.043		>446.000*	
MCHC	g/dL	<32.529*	>35.246		
Folate	nmol/L	<17.500*	<20.000*	<10.300*	>6.000*
Total protein	g/L	>70.000*	>74.731*	<89.600*	
Sodium	mEq/L	133.000 - 140.000*		132.800 - 140.700*	
Monocytes	$x10^9$/L		>0.532*		0.070 - 1.010*

[a] *Subinterval with higher rate of $CD4^+$ count increase due to the covariate within the given phase*

[b] *No breakpoint. The entire range (min - max) of the observed values showed desirable effects*

* *Subinterval with undesirable effects on the $CD4^+$ count*

The approximate rates of the $CD4^+$ change within the desirable segments is shown in Table 6.4.

Table 6.4: The segmented slopes and their confidence intervals

Covariate k	Slope m	2-Acute $\beta_{mk}(95\%CI)$	3-Early $\beta_{mk}(95\%CI)$	4-Est $\beta_{mk}(95\%CI)$	5-ART $\beta_{mk}(95\%CI)$
Albumin	0	9.57(-1.93 ; 21.08)	8.39(1.08 ; 15.69)*	5.49(1.54 ; 9.43)*	-37.80(-124.30 ; 48.69)
Albumin	1	6.27(0.60 ; 11.93)*	-4.07(-30.98 ; 22.84)	21.52(7.99 ; 35.05)*	3.14(0.12 ; 6.16)*
Albumin	2	-	14.86(6.69 ; 23.03)*	-	-
ALP	0	-0.47(-1.46 ; 0.52)	1.16(-0.06 ; 2.37)	1.19(0.53 ; 1.84)*	0.66(-1.06 ; 2.38)
ALP	1	5.74(-11.64 ; 23.13)	-0.06(-1.63 ; 1.52)	-2.27(-8.56 ; 4.02)	27.50(-237.51 ; 292.51)
ALP	2	-3.17(-8.68 ; 2.34)	-	-	0.14(-0.66 ; 0.95)
Basophils	0	-849.27(-2862.60 ; 1164.10)	-296.59(-1538.50 ; 945.38)	1338.50(80.34 ; 2596.70)*	2117.10(557.80 ; 3676.50)*
Basophils	1	1209.80(-992.67 ; 3412.20)	935.47(-1548.60 ; 3419.60)	-1020.20(-6876.20 ; 4835.80)	-8476.70(-17139.00 ; 185.67)
Basophils	2	-14736.00(-35348.00 ; 5875.20)	-	-	8416.50(222.64 ; 16610.00)*
Calcium	0	157.80(2.32 ; 313.28)*	655.48(-494.95 ; 1805.90)	-18.12(-349.52 ; 313.28)	-670.48(-2401.70 ; 1060.70)
Calcium	1	-976.21(-2212.90 ; 260.48)	-58.71(-217.55 ; 100.12)	-254.48(-1145.90 ; 636.90)	-57.53(-213.72 ; 98.67)
Calcium	2	-	97.00(-1176.70 ; 1370.70)	-112.97(-431.18 ; 205.24)	1094.60(-885.10 ; 3074.40)
Folate	0	-9.52(-15.11 ; -3.92)*	-5.08(-8.63 ; -1.54)*	-15.06(-24.44 ; -5.69)*	5.97(-21.97 ; 33.92)
Folate	1	4.69(-5.98 ; 15.36)	0.13(-7.73 ; 7.99)	-1.98(-4.47 ; 0.51)	-4.98(-7.10 ; -2.86)*
Folate	2	-3.63(-8.78 ; 1.51)	-4.72(-12.51 ; 3.07)	-	-
Haematocrit	0	1368.40(-1567.00 ; 4303.70)	106.16(-1842.60 ; 2054.90)	323.97(-1629.70 ; 2277.70)	-1378.90(-4218.00 ; 1460.10)
Haematocrit	1	456.49(-1857.70 ; 2770.70)	145.25(-9745.90 ; 10036.00)	-1577.70(-3533.10 ; 377.67)	2094.60(-4190.60 ; 8379.80)
Haematocrit	2	1244.20(-2974.50 ; 5463.00)	-	1259.60(-1223.80 ; 3742.90)	-657.46(-2753.40 ; 1438.50)
LDH	0	-0.98(-1.39 ; -0.57)*	0.02(-0.32 ; 0.36)	0.36(0.11 ; 0.62)*	0.22(-0.57 ; 1.02)
LDH	1	0.42(0.11 ; 0.73)*	0.30(-0.36 ; 0.95)	-0.16(-0.30 ; -0.02)*	-0.08(-0.20 ; 0.04)
LDH	2	-0.34(-0.96 ; 0.28)	-0.06(-0.34 ; 0.22)	-	0.66(-0.11 ; 1.43)
Lymphocytes	0	-771.17(-2165.00 ; 622.62)	129.61(-51.34 ; 310.55)	-10.24(-284.30 ; 263.81)	276.76(250.55 ; 302.97)*
Lymphocytes	1	172.77(147.86 ; 197.68)*	183.99(163.26 ; 204.73)*	275.74(167.02 ; 384.45)*	18.16(-59.14 ; 95.46)
Lymphocytes	2	-43.59(-190.29 ; 103.12)	-	140.43(113.65 ; 167.22)*	-
Magnesium	0	-1022.60(-2681.90 ; 636.69)	-591.57(-2206.60 ; 1023.50)	97.33(-708.40 ; 903.06)	613.27(-733.06 ; 1959.60)
Magnesium	1	449.49(-131.36 ; 1030.30)	77.08(-117.85 ; 272.01)	-98.37(-368.61 ; 171.86)	66.44(-128.04 ; 260.91)
Magnesium	2	-1056.90(-2382.20 ; 268.27)	2904.70(-5742.50 ; 11552.00)	41.77(-999.34 ; 1082.90)	-
Magnesium	3	370.32(-613.51 ; 1354.10)	-	-	-
MCHC	0	-38.55(-70.10 ; -7.01)*	-5.65(-15.96 ; 4.67)	-16.55(-93.90 ; 60.80)	32.48(-27.91 ; 92.87)
MCHC	1	171.06(-967.92 ; 1310.00)	169.34(41.19 ; 297.48)*	-7.06(-21.18 ; 7.06)	-9.01(-21.02 ; 3.01)
MCHC	2	-14.37(-35.12 ; 6.38)	-	-24.67(-102.20 ; 52.87)	-
MCV	0	3.01(-7.61 ; 13.62)	38.56(-10.19 ; 87.31)	-0.10(-14.15 ; 13.95)	5.71(-4.77 ; 16.20)
MCV	1	9.03(-3.26 ; 21.31)	3.49(-5.44 ; 12.42)	7.04(-0.53 ; 14.60)	8.20(-6.04 ; 22.44)
MCV	2	-	7.74(-2.85 ; 18.34)	73.11(-228.04 ; 374.25)	2.07(-8.35 ; 12.49)
Monocytes	0	-101.02(-349.04 ; 147.00)	-4321.90(-17123.00 ; 8479.10)	147.92(-191.47 ; 487.32)	-187.93(-340.65 ; -35.22)*
Monocytes	1	-125.96(-280.12 ; 28.21)	21.21(-123.99 ; 166.42)	-541.12(-1095.50 ; 13.24)	-175.51(-372.43 ; 21.41)
Monocytes	2	-	-332.17(-587.47 ; -76.87)*	-77.40(-235.60 ; 80.80)	-
Platelet	0	0.31(-0.18 ; 0.80)	2.01(0.31 ; 3.71)*	0.44(0.28 ; 0.60)*	0.07(-0.21 ; 0.36)
Platelet	1	0.63(0.25 ; 1.01)*	-3.20(-7.06 ; 0.65)	-0.93(-2.96 ; 1.10)	0.30(-0.09 ; 0.68)
Platelet	2	-1.20(-4.31 ; 1.92)	0.55(0.32 ; 0.77)*	-	-
Platelet	3	-	-0.65(-2.31 ; 1.01)	-	-
Potassium	0	33.63(-38.55 ; 105.81)	-1576.50(-3707.80 ; 554.82)	6.87(-16.59 ; 30.34)	-21.91(-213.13 ; 169.30)
Potassium	1	-68.67(-366.84 ; 229.49)	47.13(-13.89 ; 108.16)	31.14(-80.74 ; 143.03)	0.16(-13.85 ; 14.17)
Potassium	2	30.64(-16.70 ; 77.99)	-70.86(-278.97 ; 137.24)	-	-9.74(-438.42 ; 418.94)
Potassium	3	-	3.84(-39.09 ; 46.77)	-	-
Total protein	0	3.37(-25.03 ; 31.76)	-10.11(-23.56 ; 3.35)	-7.58(-10.10 ; -5.06)*	-13.44(-48.08 ; 21.20)
Total protein	1	-5.13(-7.50 ; -2.76)*	-7.03(-9.03 ; -5.04)*	-1.06(-6.66 ; 4.55)	-1.93(-9.25 ; 5.39)
Total protein	2	-	-	-	-5.44(-11.61 ; 0.72)
Total protein	3	-	-	-	2.67(-4.01 ; 9.35)
Red blood cells	0	37.68(-166.41 ; 241.78)	-6.20(-173.55 ; 161.14)	873.48(58.47 ; 1688.50)*	-45.56(-259.15 ; 168.03)
Red blood cells	1	-424.26(-1011.30 ; 162.80)	159.57(-237.81 ; 556.95)	52.25(-92.68 ; 197.19)	309.59(-140.39 ; 759.56)
Red blood cells	2	186.07(-64.73 ; 436.86)	-	478.48(-4584.90 ; 5541.90)	-13.44(-202.64 ; 175.76)
Sodium	0	-2.09(-107.01 ; 102.83)	-26.36(-74.78 ; 22.07)	10.77(-31.12 ; 52.66)	8.97(-9.62 ; 27.56)
Sodium	1	-11.97(-20.98 ; -2.97)*	-7.20(-15.07 ; 0.68)	-7.11(-13.82 ; -0.39)*	-0.78(-9.63 ; 8.08)
Sodium	2	23.96(-15.69 ; 63.61)	2.25(-46.74 ; 51.25)	8.42(-73.54 ; 90.38)	-50.08(-113.48 ; 13.31)

*Significant at 5% level

The behavioural patterns of each covariate's segmented linear effect on the $CD4^+$ count at each phase are shown in Figures 6.2 – 6.6.

Desirable effects (in at least two phases)

Lymphocytes and albumin had at least one significant segment or subinterval (see Table 6.3 and Figure 6.2).

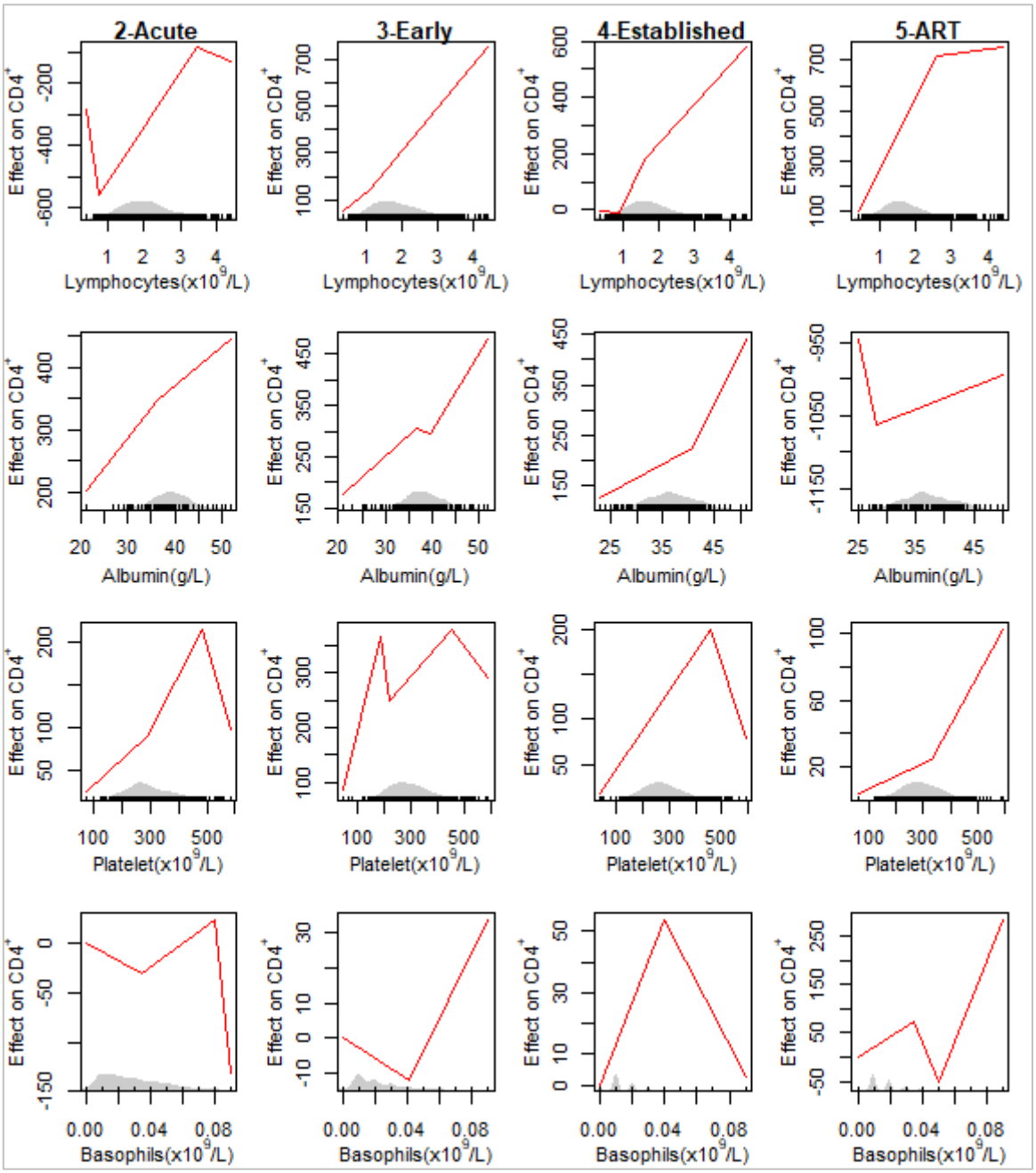


Figure 6.2: Desirable effects (in at least two phases).
The density plots of the observed records are shown on the horizontal scale.

The significant segments show a positive linear relationship with the $CD4^+$ count at each and every phase of the HIV disease progression. These segments corresponded to the densely populated region of the observed values The highest rate of the $CD4^+$ count increase of 275cells/mm^3 in response to $1x10^9$/L increase in the total lymphocytes was observed during the established and ART phases for the total lymphocyte ranges of 0.900 - $1.578x10^9$/L and <$2.591x10^9$/L respectively. The acute and early phases had similar rates of the $CD4^+$ count change of around 175cells/mm^3 per $1x10^9$/L increase in the total lymphocytes in the ranges 0.725 - $3.459x10^9$/L and >$1.112x10^9$/L respectively. Albumin: $CD4^+$ count increase per unit

change in albumin was higher during the established phase where it was 21.520(7.992;35.048) cells/mm^3 for albumin >40.9g/L followed by the acute phase where 14.859(6.686;23.031) cells/mm^3 of CD4$^+$ count change was observed between 36.576 - 39.491g/L of albumin. Platelet count: our model showed that the CD4$^+$ count did not significantly increase in response to platelet count during therapy. The highest rate of CD4$^+$ count change, 2.013(0.312;3.714) cells/mm^3 in response to 1x10^9/L change in platelet count was observed during the early phase for the platelet count <181.595x10^9/L. An increase in 1x10^9/L of basophil during the entire established phase corresponded to the same rate of the CD4$^+$ count change. The significant increase in the CD4$^+$ count during ART phase was observed at basophil count below 0.035 x10^9/L and at basophil count above 0.05x10^9/L but the basophils scale was too sensitive to show realistic rate of change.

Desirable effects (in one phase only)

Figure 6.3 shows covariates with desirable effects in a single phase.

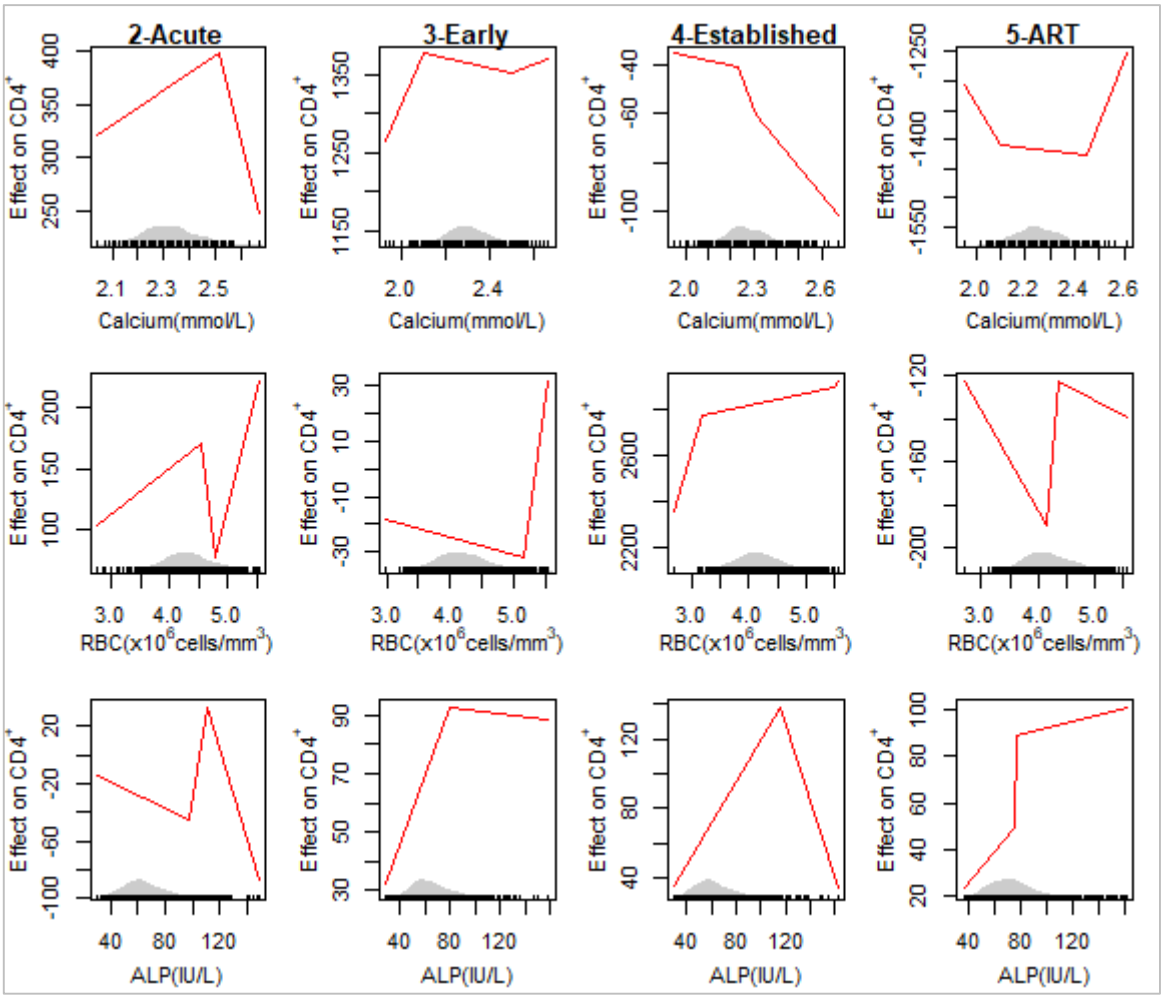


Figure 6.3: Desirable effects (in one phase only).
The density plots of the observed records are shown on the horizontal scale.

Table 6.3 shows that a unit increase in calcium <2.518mmol/L during the acute phase improved $CD4^+$ cell response by 157.800(2.318;313.280) cells/mm^3. This calcium range corresponded to the majority of the observed calcium measurements during this acute phase (see density plot in Figure 6.3). Also the positive $CD4^+$ cells response was observed during the established phase only for red blood cells <3.18x10^6cells/mm^3. ALP <116.393IU/L also showed a corresponding increase in the $CD4^+$ count by 1.187(0.531;1.843)cells/mm^3 during the established phase only.

Alternating desirable and undesirable effects

During the acute phase, for LDH <327.809U/L, the $CD4^+$ count was declining and then increasing for LDH between 327.809 - 607.043U/L. This shows that during the acute phase, higher LDH had desirable outcomes on the $CD4^+$ count (see Figure 6.4). The established phase revealed that the $CD4^+$ count increased for the LDH <446U/L and began to decline after this point. It was also notable that the 446U/L was the cut-off point for approximately the bottom 25% of the measurements recorded during the established phase. Hence, an indication that the $CD4^+$ count improved at lower LDH levels during the established phase. Although the $CD4^+$ count was generally declining as the MCHC was increasing during the acute phase, our results showed that the $CD4^+$ count significantly declined at lower MCHC (<32.529g/dL). Contrary to the acute phase, the $CD4^+$ count increased at much higher MCHC levels (>35.246g/dL) during the early phase only.

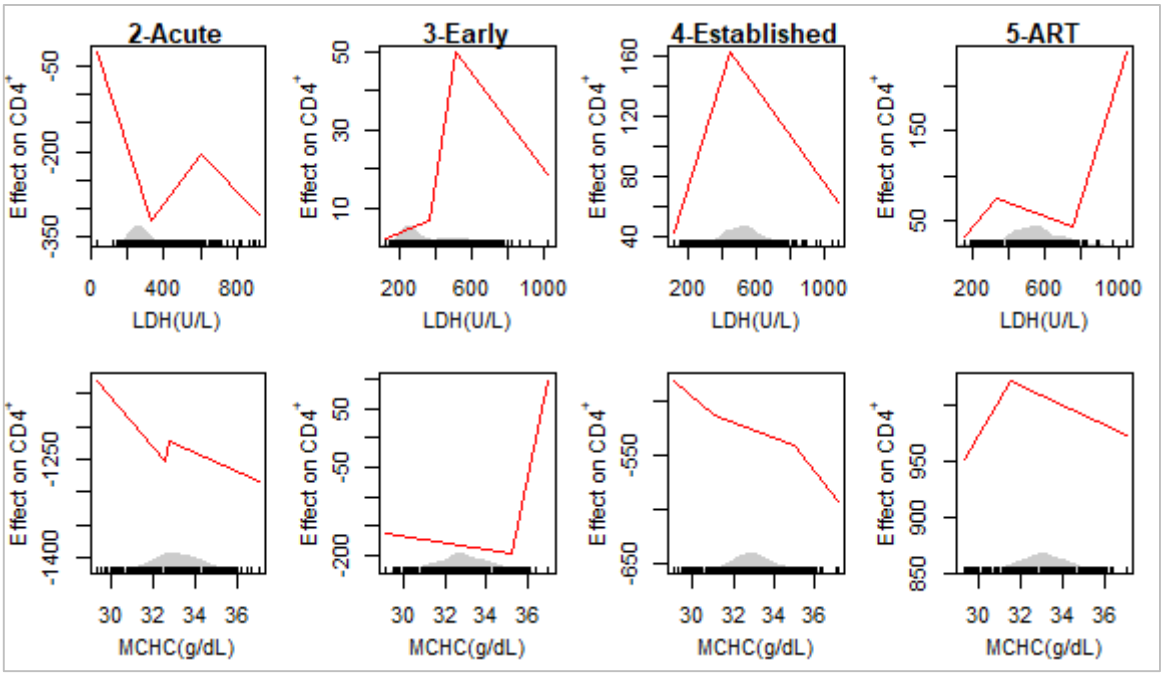


Figure 6.4: Alternating desirable and undesirable effects.

The results show two consecutive segments with either positive and then negative slope or vice versa. This also includes different slope directions from one phase to the other. The density plots of the observed records are shown on the horizontal scale.

Undesirable effects

At folate <10.30nmol/L which is about the bottom 50% of the observed values during the established phase (see Figure 6.5), the $CD4^+$ count dropped by 15.063 cells/mm^3 for a unit increase in the folate level. It has been shown to have the most detrimental effect on the $CD4^+$

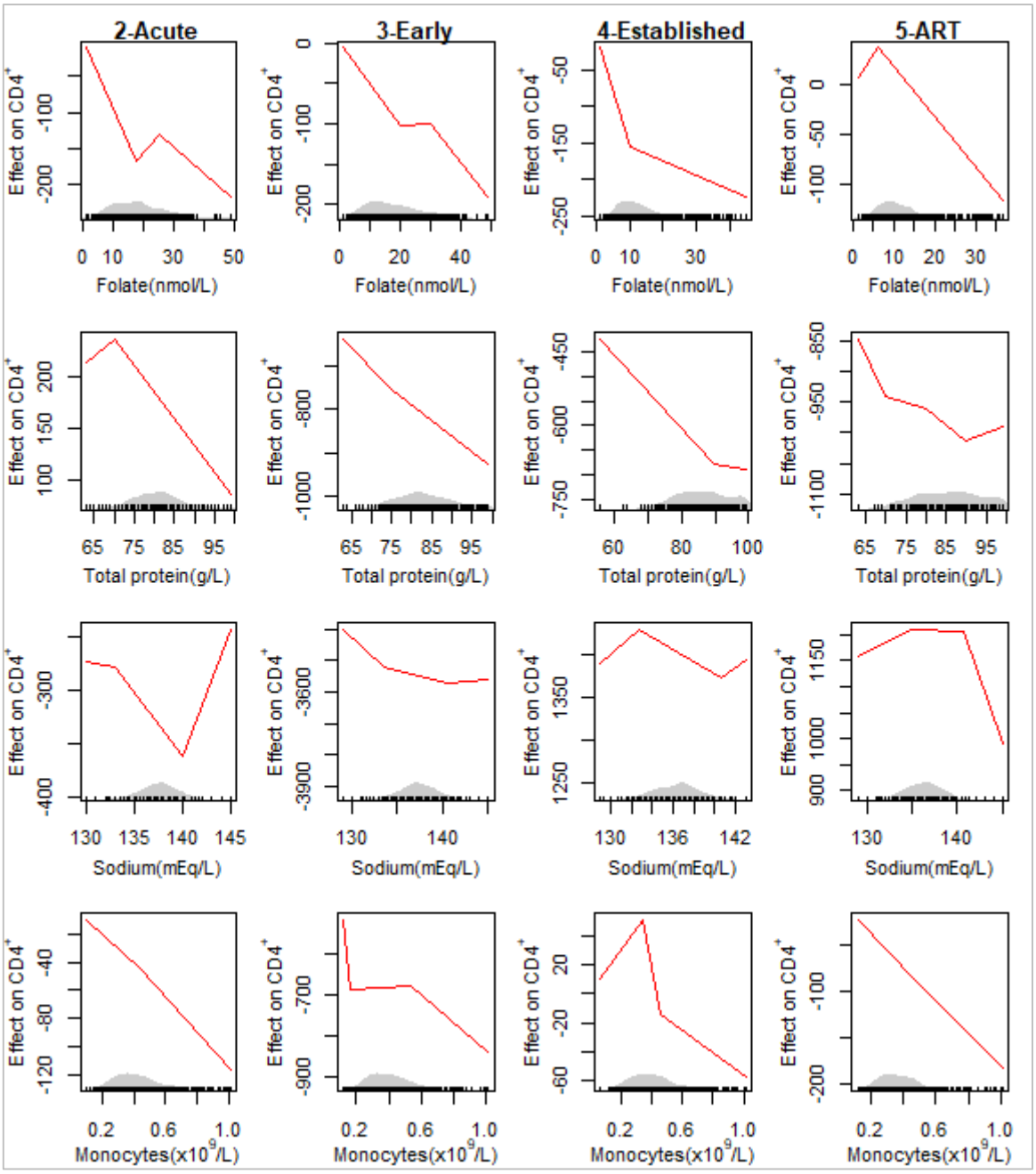


Figure 6.5: Undesirable effects.
The density plots of the observed records are shown on the horizontal scale.

count at each phase of the HIV disease progression. This was the highest rate of $CD4^+$ count decline with respect to the folate increase followed by the acute phase -9.515(-15.114; -3.917)cells/mm^3 at folate <17.50nmol/L, early phase -5.083(-8.630;-1.537)cells/mm^3 at folate <20.00nmol/L and lastly ART phase -4.982(-7.099 ; -2.864)cells/mm^3 at folate >6.00nmol/L.

The results also revealed that the $CD4^+$ count significantly dropped at higher folate levels during ART only as compared to the other infection phases. Total protein did not show effect on the $CD4^+$ count change during therapy only whereas the acute and early phases had the $CD4^+$ count dropping by around 6cells/mm^3 for total protein >70g/L and <89.6g/L during the established phase. Sodium had consistently shown two breakpoints (132mEq/L and 140 mEq/L) in the linear relationships with the $CD4^+$ count at all the infection phases but undesirable effects between these two points were found during the acute and established phases only. For a unit increase in the sodium levels, the $CD4^+$ count significantly dropped by 11.972cells/mm^3 and 7.106cells/mm^3 during the acute and established phases respectively. The $CD4^+$ count declined in response to monocytes during the early phase for the monocyte count >0.532x10^9/L and also during the entire therapy period.

No effect

Our model suggested that all the segmented linear relationships of the following covariates were not statistically significant despite some noticeable trends: haematocrit, MCV, magnesium and potassium (see Figure 6.6). All the the linear segments show no significant relationships between the $CD4^+$ count and the clinical covariates.

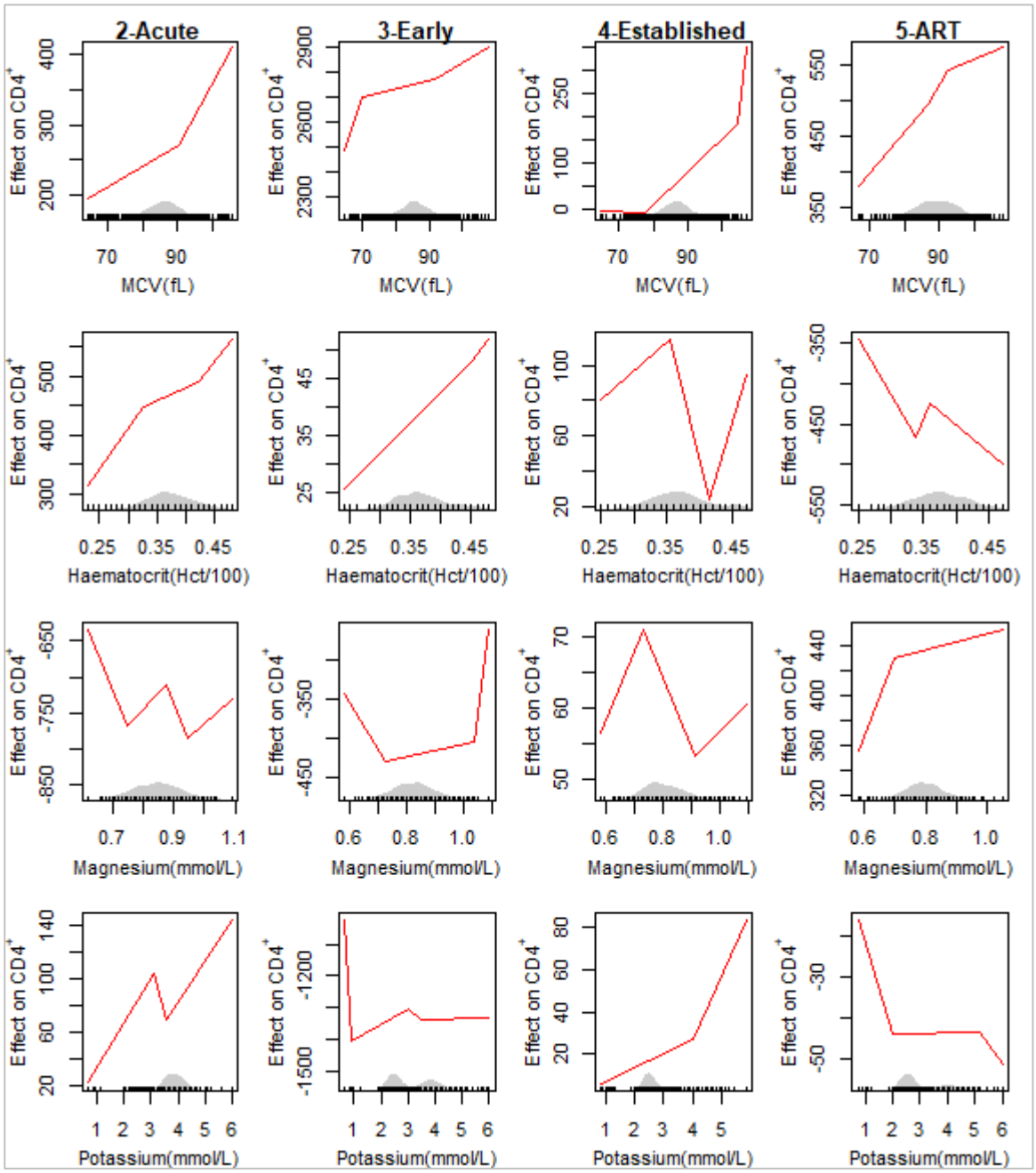


Figure 6.6: The segmented regression lines with all insignificant relationships

6.4 Clinical interpretation of the results

The desirable intervals for the total lymphocytes, albumin, platelet count, basophils, calcium, red blood cells and ALP varied depending on the infection phase. The total lymphocytes were found to have a desirable subinterval in every HIV infection phase showing that they were

closely related to the $CD4^+$ count. The total lymphocytes are referred to as B or T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012) whereas the $CD4^+$ cells are T cell type (Papagno et al. 2004), hence the positive relationship was as expected. Although the lymphocytes desirable ranges varied by phase, they were mostly within the normal references as other health females in Kenya, 1.22 - 3.24 $x10^9$/L (Omuse et al. 2018) and South Africa 0.840 - 3.256 $x10^9$/L (Lawrie et al. 2009). There were some instances where the lymphocytes desirable ranges encroached outside, below or above these normal references, for example <2.591 $x10^9$/L during the ART phase where much lower lymphocytes than usual where also considered desirable.

Albumin was found to be one of the covariates with the highest contribution to the $CD4^+$ count variation in our study following the lymphocytes and having a desirable subinterval at each infection phase too. Albumin helps with tissue nourishment (James 2017) which our results indicated to be important in enhancing the $CD4^+$ count throughout the HIV disease progression. Such positive relationship was also confirmed by (Pralhadrao et al. 2016) although this was limited to a cross sectional study and with no indication of specific desirable ranges. The suggestion to use serum albumin in pre-treatment assessment (Olawumi and Olatunji 2006, Sharma et al. 2016) also sounded to be feasible according to our results but more precisely throughout the entire established phase. However, care should be taken to consider that the rate of $CD4^+$ count change per unit increase in the albumin <40.9g/L was lower as compared to albumin >40.9g/L. The normal reference range for albumin is usually 35 - 55g/L but our results showed that lower levels of albumin were capable of influencing the $CD4^+$ cell response during the early (<36.576g/L) and established (40.900g/L) phases. Although the lower levels of albumin influenced the $CD4^+$ cell response during ART too, we found out that such albumin effect could only start from as low as 28.130g/L and then upwards. It has been reported that serum albumin concentrations increased significantly after this ART initiation (Chong et al. 2015).

The blood clotting condition is measured by platelet count (Project Inform 2007, NAM 2012, National Institutes of Health Clinical Center 2015, James 2017) and some normal female reference ranges were found to be 178 - 454$x10^9$/L in South Africa (Lawrie et al. 2009) and 152 - 443 $x10^9$/L in Kenya (Omuse et al. 2018). There was a notable increased rate of $CD4^+$ count change at lower platelet count (<181.595$x10^9$/L) and in the early phase. Such low platelet count was also found to be associated with low $CD4^+$ count (Sloand et al. 1992). It was also

during this early phase where the $CD4^+$ count increased at elevated platelet count >217.884 $x10^9$/L, an indication that the desirable platelet counts during the early phase tended to be outside of the usual reference range. The breakpoints of the desirable range in the acute phase resembled an upward shift of the normal reference range. It is known that the acute phase is the period during which the viral load is at its peak (Bellan et al. 2015, Manoto et al. 2018). According to our study design, the established phase is when the patient has been living with the HIV for at least a year and could be as long as many years before ART initiation. Our results revealed that during this phase, the desirable range for platelet count was similar to the normal reference ranges except that it could extend to even lower platelet count levels. ART is known to boost the $CD4^+$ count (Meintjes et al. 2017) and control the problem of low platelet count (Brook et al. 1997). Our segmented model showed that it was only during the ART phase where the platelet count did not have an effect on the $CD4^+$ count. Basophils are the indicators of the body tissue inflammation (Project Inform 2007) and our results revealed that this condition was positively related to the $CD4^+$ count at the later stages of the disease progression (established and ART phases). Although there was a slight change in the rate of $CD4^+$ count change during the ART, our results showed that the basophils influenced the $CD4^+$ cell response in the same known (Lawrie et al. 2009) reference range of 0 - 0.1 $x10^9$/L. However, the basophils scale was too sensitivity for the rate of $CD4^+$ count change in our model. Our results for the basophils were in contrast to (Jiang et al. 2015) whose findings indicated that regardless of $CD4^+$ depletion, the basophils' frequency remained fairly the same over the course of the disease progression.

The brain and nerve function, heart rhythm, water and blood characteristics are regulated by calcium (Project Inform 2007, James 2017, The Johns Hopkins Lupus Center 2017), which we found to increase together with the $CD4^+$ count. That relationship was observed during the acute phase only at calcium levels <2.518mmol/L. This overlapped with the normal (Kratz et al. 2004) reference range of 2.2 - 2.6mmol/L and encroached into the lower levels. This could have been attributed to the increased intracellular calcium concentration during the acute phase (Cloyd and Lynn 1991) in conjunction with the negative effects of HIV infection on the permeability of plasma membrane to calcium (Choi et al. 1998). This also shows that if phase specific behavioural patterns of the calcium and $CD4^+$ count relationships are not taken into consideration, one might conclude that serum calcium is not affected by HIV/AIDS as reported in Nigeria (Ugwuja and Eze 2007). Recalling that in our study, the established phase referred to the period during which the patient has been living with the HIV for at least a year without

medical intervention to suppress the HIV. This established phase was the only period during which the red blood cells influenced $CD4^+$ cell response at low red blood cell count (<3.18 $x10^6$ cells/mm^3). Usually the red blood cell count is referenced at 3.93 - 5.40cells/mm^3 or 4.13 – 5.67 cells/mm^3 in the local South African context (Lawrie et al. 2009). Such low count in the red blood cells is known to be associated with low $CD4^+$ count (Vanisri and Vadiraja 2016a, Lumbanraja and Siregar 2018) in HIV positive patients. Hence, our desirable results at low red blood cells count suggested that mechanisms to prevent anaemia (lower red blood cell count) during the established phase provide desirable outcomes on the $CD4^+$ count. ALP helps in detecting liver health (Patil et al. 2013) with the normal range being 30 - 120IU/L (Kratz et al. 2004) or 30 - 115IU/L (Park et al. 2013). In our study, the majority of the observed ALP values were under 116.393IU/L suggesting that our cohort seems to have no abnormal liver conditions. However, this indicates that efforts to improve the $CD4^+$ count by manipulating ALP can be directed towards the established phase because it was the only phase indicating a strong linear relationship between the ALP and the $CD4^+$ count.

Folate, total protein, sodium and monocytes had phase specific ranges showing undesirable effects to the $CD4^+$ count and prevention of such subintervals has been found to be important too. Although folate is needed for cell growth and metabolism (Arya and Kumar 2012, Dieticians of Canada 2014), our data showed that it had detrimental effects on our cohort's $CD4^+$ count which is contrary to (Adhikari et al. 2016) findings where it was reported that folate improves the $CD4^+$ count. Our results showed that the $CD4^+$ count decline in response to total protein in HIV positive patients was predictable before treatment and that relationship vanished with ART. It has been reported that the serum protein increases with highly active antiretroviral therapy (Ugwuja and Eze 2007). Because sodium helps in the functioning of vital organs such brain (Project Inform 2007), the negative effects on the $CD4^+$ count might have been attributed to the osmotic gradient between extracellular and intracellular fluid in cells created by sodium changes (Shu et al. 2018). Our data has shown that such balance was crucial since the breakpoints were consistent (132mEq/L and 140 mEq/L) across all the HIV positive phases. Monocytes, together with the $CD4^+$ cells are known to be infected by HIV (Pasupathi et al. 2008) such that both are expected to have similar changes. However, during the early and ART phases, our data showed an inverse relationship between the monocytes and the $CD4^+$ count. Subintervals of LDH and MCHC showed both desirable and undesirable effects on the $CD4^+$ count in some of the post-HIV infection phases.

6.5 Summary

This chapter aimed to identify the subintervals of the clinical covariates that had desirable influence on the $CD4^+$ cell response in the face of the HIV disease complexities. The objective was to detect the breakpoints of the subintervals of each covariate based on the $CD4^+$ count change in the linear relations. We also determined the nature of the linear relationship within the subintervals of the clinical covariates. As a point of reference for assessing the adaptive behaviour of the desirable ranges, we compared our results with some known laboratory reference ranges. We conclude that the $CD4^+$ count has segmented linear relationships with the covariates. Some segmented relationships were significant whilst others were not. Generally, the phase specific desirable ranges overlapped with the normal clinical reference ranges but some tended to encroach more outside, either below or above the normal reference ranges. The detected breakpoints may also give an insight into the feasibility of universal and biologically meaningful cut-off points in categorising the continuous $CD4^+$ count clinical covariates when categorical data is really called for in meta-analysis, say. It is important to avoid the undesirable subintervals of some of the clinical covariates in the management of the HIV disease.

The segmented regression model was effective in the breakpoint detection as well as providing the slopes, a limitation of the GAMM procedure. Ideally, the intention of the segmented model was to detect the break-points of the turning points that were discovered in smooth curves but this was a challenge. Currently there is no way to inform the segmented model in the R software to adapt to the same GAMM smooths turning points and then detect the break-points for consistence. The detected break-points are rather the locally weighted scatterplot smoothers which are not necessarily penalized as in GAMM. Hence, instead of the general linear models used, future studies are recommended to develop a segmented model with a built-in GAMM technique or some form of penalisation to allow for consistence in extracting a close set of the same information.

Together with the previous chapters, the segmented regression model has not incorporated a realistic approach of considering the relationship between the covariates themselves and the influential effects of time lag on the $CD4^+$ count during the HIV disease progression. The following chapter discusses the suitability of the structural equation models in shedding more light on the nature of the complexity in the relationships in our data.

CHAPTER 7 STRUCTURAL EQUATION MODELS (SEM)

A human body is a complex machine that usually responds to the changing internal and external environments. It is capable of automatically self-regulating its harmonious functional systems to support life with many variables at play (Kelly 2006). Hence, every piece of the captured information during the HIV patient follow up care should ideally contribute to the overall health status of an individual. In the previous chapters, the observed changes in the $CD4^+$ count in response to each covariate assumed that the other covariates were held constant, which may not necessarily be the ideal case in real life. Taking, for example, the multilevel model in Chapter 4, the standard linear models were obtained for each covariate in relation to the $CD4^+$ count. We then later compared the variations in these linear models. Similarly, with the GAMM, the technique itself is specifically developed for univariate analysis where the smooths are for each covariate at a time. The general linear model underlying the segmented regression model is also not equipped to have a covariate that simultaneously explains the response as well as any other influential effects on the other covariates. In real biological terms, at any given time, every covariate has a role to play. Accordingly, health practitioners rely on the $CD4^+$ count and the covariates to determine the course of the HIV disease progression. In essence, the $CD4^+$ count responds to the covariates whilst some covariates being responsive to the other covariates. More so, because the patients are followed up over time, the effect of the current measurements on the next ones is likely to exist, a major drawback of the previous chapters. Structural equation models (SEM) are capable of simultaneously modelling the interconnection between the covariates, the response and the time lag to resemble a more realistic system in the body.

Like any other statistical procedures where *observed* variables in a given data set are used for modelling, SEM have the same capability but in addition, they model *unobserved* variables too. It is a technique that allows specification, estimation and evaluation of models of linear relationships in a given set of *observed* variables but in a reduced set of *unobserved* variables (Shah and Goldstein 2006). A hypothesized model is set up at the beginning of the analysis and then tested as to whether it is supported by the data. The data consists mainly of the observed variables that can be measured but there is also a belief (Norm and Larry 2013) that there is a phenomenon of theoretical interest conceptually termed a construct (Edwards and Bagozzi 2000) which is related to the measured variables. The data may or may not fit well the hypothesized model in which case some modifications to the model can be applied until a

reasonably acceptable and parsimonious model that closely correspond to the data is obtained. The selection of alternative models is useful when model generation is the main objective as this allows the arrival at the best fit model to the data (Kline 2011). However, in a strictly confirmatory application, a single model is either accepted or rejected without considering alternatives (Jöreskog 1993). In this chapter, we intend to try and modify the hypothesized model until we obtain the best fit for our data. The integration of different discipline-specific advances over a long period of time created the multi-disciplinary method called SEM. Many well-known conventional statistical techniques such as regression analysis and correlation analysis can be considered as special cases of SEM. It is a combination (Newsom 2015) of two main techniques: path analysis (Hauser and Goldberger 1971) and factor analysis (Jöreskog 1969). It utilises the application of generalised least squares (Zellner 1970).

Path analysis found its roots in genetics dating back to the early 20th century (Wright 1918, 1920, 1921, 1923). The approach estimated causal relations among variables based on the correlation matrix of observed variables only with the aid of path diagrams depicting regression coefficients (Wright 1934). Basically, a path analysis model focuses on the relationships of multiple observed variables only using several regression equations simultaneously (Bian 2011). The path diagrams consist of variable labels that are connected by lines that either end with a single head or double headed arrows. The direct effects are represented by single straight line headed arrows and the correlations shown by curved line double headed arrows. The regression path coefficients are then indicated along the straight or curved lines.

Factor analysis has been introduced as a variable reduction technique that uses a correlation matrix of observed variables to obtain fewer underlying latent variables that capture much of the variation in the original variables. The development of this technique has been mostly in line with the psychology discipline (Coffman and MacCallum 2005). There are two types of factor analysis namely exploratory factor analysis and confirmatory factor analysis (CFA). Exploratory factor analysis has been in existence for more than a century and originally developed by Spearman in 1904 (Kline 2011). Its goal is to group together correlated variables and is useful in the early stage of research. On the other hand, CFA has been around since the mid-1960s (Brown 2006) and is a hypothesis-driven approach in which alternative models are tested empirically (Newsom 2015). The hypothesis testing about a factor structure involves coming up with a theory first, deriving a model and then followed by testing of the model consistency with the observed data (Bian 2011). Grouping of the variables that form a construct is the work of exploratory factor analysis (Jollife 1986, Kline 1994, Everitt 2001) and this

chapter dealt with variables that are known to already belong to certain groups of clinical platforms. Hence, the emphasis is on CFA which can assume different types of the SEM components.

7.1 Components of the general Structural Equation Model

In general, a SEM model is a network of two main forms of relationships: (i) measurement model that connects the observed and latent and (ii) structural model that shows the relationships between latent variables only. The latent variables are free from the random error that would have been estimated and removed leaving the common variance (Byrne 2001). Figure 7.1 illustrates the main components of the SEM.

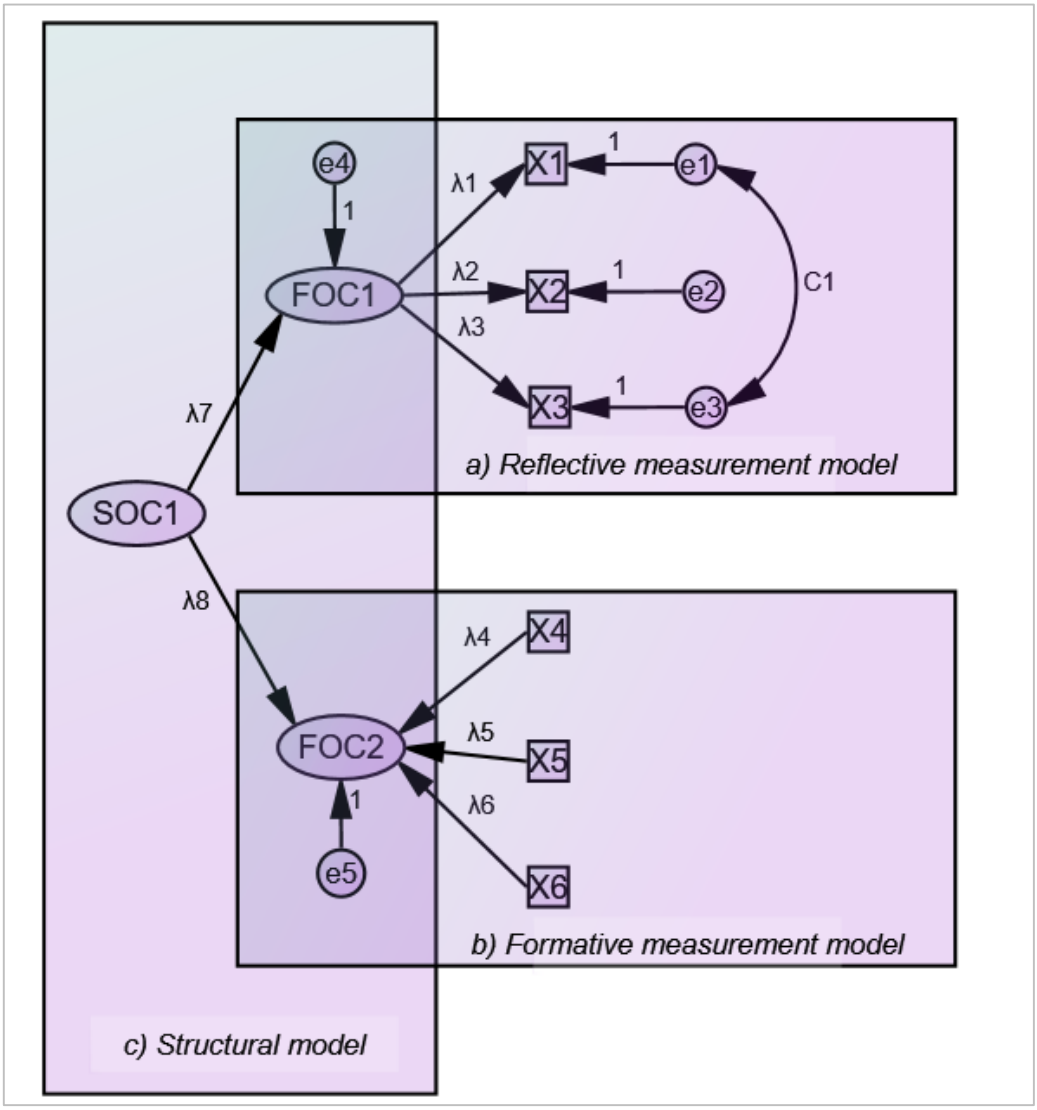


Figure 7.1: Components of the SEM and types of constructs
Source: Author

First order constructs (FOC) are directly related to the indicator variables and the next level, second order constructs (SOC) related to the FOC (Marsh and Hocevar 1985) and so on. For example, the latent constructs FOC1 and FOC2 are directly related to the indicator variables $x_1 - x_3$ and $x_4 - x_6$ respectively. The SOC1 is in turn a resultant from FOC1 and FOC2. The causal relationships are indicated by the arrows and can be in any direction depending on the theory behind the variables and researcher's hypothesis (Roy et al. 2012). The variable from which the arrow is coming from is called an exogenous variable (analogous to the independent variable). In the SEM terminology it can either be an observed exogenous (e.g. $x_4 - x_6$) or unobserved exogenous (e.g. FOC1) variable (Schreiber et al. 2006). Similarly, the variable in which the arrow is going into is analogous to the dependent variable and is called an observed endogenous (e.g. $x_1 - x_3$) or unobserved endogenous variable (FOC2). In hypothesizing a model, the exogenous variable is assumed to cause the changes in the endogenous variable which in turn is believed to be associated with a measurement error (e.g. $e_1 - e_5$) that cannot be explained by the exogenous variable. To some extent these disturbances (errors) can be considered as latent variables in their own right (Kline 2011). Any covariation or correlation between two variables is depicted by double headed arrows (C_1). The pictorial representation of the network of arrows (as in Figure 7.1) that show the causal relationships is what is basically called a path analysis diagram where the regression weights e.g. $\lambda_1 - \lambda_8$ are the strengths of the relationships often found along the single headed arrows in the path diagrams. The λ's represent the influence of at least one variable on another variable (Byrne 2006).

7.1.1 The types of measurement models

The specification of the measurement model is the basis of the SEM and has two main types: the formative or reflective measurement models (Bollen and Lennox 1991). Indicator variables are either influenced by an underlying reflective construct (Figure 7.1a) or the aggregation of these indicators resulting in a formative construct (Figure 7.1b). The choice of the appropriate measurement model is crucial as this affects the model fit parameters, path coefficients and consequently the interpretation of the interrelationships among the variables in the model (Jarvis et al. 2003, MacKenzie et al. 2005). The application of the incorrect measurement model is called misspecification; hence the need for justification of the choice of a measurement model becomes a necessity. Most researchers have been found to assume reflective models, yet prior to such decisions emphasis should have been placed on considering the nature of the

construct. That is, the causality direction between the indicator and construct including indicator characteristics (Coltman et al. 2008).

Reflective indicators (effect model)

The choice of a reflective measurement model requires that the latent construct be independent of the measures (Rossiter 2002, Borsboom et al. 2004). The indicators should have a positive and high intercorrelations to enable the assessment of composite and individual reliabilities of the indicators (Trochim 2006). Causality flows from the construct to the indicators where change in the construct effects change in the indicators (Figure 7.1a). The reflective model is a series of equations that share a common factor and the mathematical representation for Figure 7.1a is given by

$$X_1 = \lambda_1 FOC1 + e_1$$
$$X_2 = \lambda_2 FOC1 + e_2$$
$$X_3 = \lambda_3 FOC1 + e_3 .$$

The sharing of a common theme and interchangeability of the reflective indicators allows for the construct to be measured by a sample of few relevant indicators that are underlying the domain of the construct (Gilbert and Churchil 1979, Nunnally and Bernstein 1994) and this becomes useful in practical situations where the number of observed variables can be unprecedentedly huge. Also, sharing of the common theme using the factor analysis procedure helps to identify and eliminate measurement error for each indicator (Spearman 1904).

Formative indicators (causal model)

In the case of the formative measurement model, the latent construct is dependent on the indicators (Borsboom et al. 2003), for example, human development exists as an index comprising of health, education and income (UNDP 2006). Indicators do not necessarily share a common theme and have weak intercorrelation patterns. The formative model does not require highly correlated indicators as they pose a challenge in the estimation of their weights resulting in imprecise values (Coltman et al. 2008). In setting up the hypothesis, causality flows from indicators to the construct where the change in the indicators affects the construct as shown in Figure 7.1b. The formative measurement model is closely associated with the PLS where the principal components are an aggregation of the multi-dimensional scale. This achieves model parsimony but the rich, diverse and unique information within individual indicators is lost. Nevertheless, (Blunch 2008) argued that the fewer underlying factors in the

CFA explain the correlations in the data with minimum loss of information. The formative first order construct in Figure 7.1b can be formulated as $FOC2 = \lambda_4 X_4 + \lambda_5 X_5 + \lambda_6 X_6 + e_5$. Dropping an indicator affects the conceptual domain of the construct but on the other hand this can be necessary as long as the conceptual domain of interest is represented by the indicators based on the empirical prediction (Coltman et al. 2008). Hence, there is no need to call for a census of indicators (Rossiter 2002) which may prove difficult although some authors belief that a complete enumeration of the indicators is important (Bollen and Lennox 1991).

Considerations for minimising the study misspecification

Most importantly, both formative and reflective techniques have the same goal and share some common characteristics in their applications although they differ in some way in the methodologies. Both are dimension reduction techniques and represent some aspect of the covariance or correlation matrices. However, the reflective model is based on an underlying hypothetical model whereas the formative being inclined towards PLS which is based on PCA that is more of unsupervised dimension reduction technique (Maitra and Yan 2008). This makes the reflective constructs unattractive for the already grouped $CD4^+$ count clinical covariates for the problem at hand. The data exist as clinical platforms that have gone through variable reduction in which multicollinearity has been minimised. More so, the adopted variable selection, SPLS utilises the concept of formative constructs. Hence, our data can better be modelled by formative constructs in the CFA.

However, the current study is focused on the $CD4^+$ count changes over time in response to the clinical covariates such that the CFA has to be assessed in waves where at least two occasions are required to understand the shifts in the causal effects in the HIV disease progression. This calls for Cross-lagged panel models (CLPM).

7.1.2 Cross-lagged path models (CLPM)

Data that are collected at one moment in time is usually referred to as a cross-sectional data and there is no way to determine if the inferences are correct (Kearney 2017). The $CD4^+$ count and the covariates have been repeatedly measured over time such that the CFA model requires assessment at more than one cross-section calling for the incorporation of longitudinal data analysis techniques in the SEM framework. When the variables have directional influences on each other over time, the CFA is then referred to as a cross-lagged path or panel model (CLPM) which is a powerful tool for examining lagged relationships between two or more variables

(Kearney 2017). The interactions and reciprocal influences between the variables over time are longitudinally assessed including the between-person variations in the SEM perspective (Mund et al. 2018). In CLPM, it is assumed that individuals fluctuate around a common group mean whereby there is no stable between-person differences for each variable over time. Growth curve models have been designed to capture such between-variations (De Leeuw and Kreft 1986, Hedeker 2004) but they suffer from the limitation in the flexibility to answer a number of research questions (Mund et al. 2018). It is argued that if stable between-person variations exist, the CLPM will include them in the estimation of the autoregressive and cross-lagged paths (Hamaker et al. 2015, Berry and Willoughby 2017).

There are two main components of the CLPM: the "cross" and the "lagged" paths. Both show lagged effects in the sense that they examine the effect of a variable at time $j = 1$ on the other variable at time $j = 2$. However, the "lagged" strictly refers to the effect of the same variable on its next measurements and usually called the autoregressive path as shown in Figure 7.2 (blue arrows). On the other hand, when one variable at $j = 1$ influences a different variable at $j = 2$, this is referred to as a "cross" lagged path as depicted by a red arrow in Figure 7.2. Furthermore, Figure 7.2 shows that the CLPM can also involve latent variables. The estimation of CLPM effects also control for correlations within time-points as well as the autoregressive effects or stability across time. Stability refers to the influence of the previous time point. Less stability is shown by smaller autoregressive coefficients that are closer to zero indicating more variance in the construct. Similarly, larger autoregressive coefficients are an indication that there is more influence from the previous time point and are associated with little variance over time.

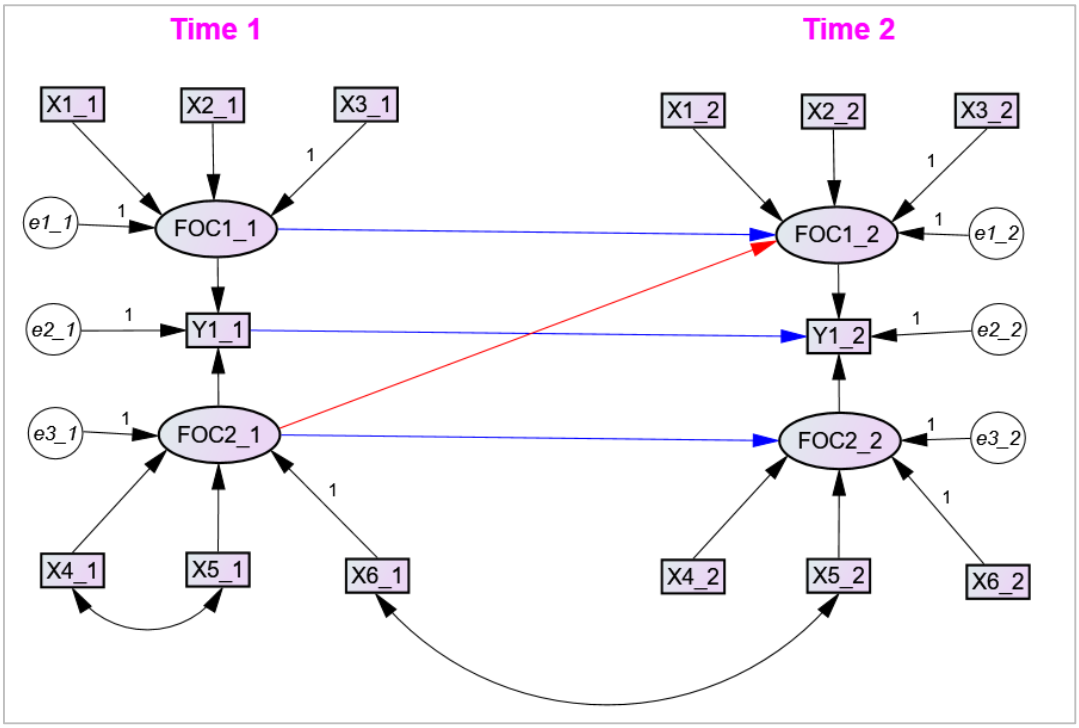


Figure 7.2: Schematic representation of a cross lagged path model
The blue arrows represent the autoregressive paths whereas the red arrow represents the cross-lagged paths and curved lines are the covariances usually standardised to correlations. The Xi_j represents the i^{th} variable with records obtained at the j^{th} time point whereas the FOCi_ j represents the i^{th} first order construct formed at the j^{th} time point. The ei_ j represents the i^{th} random error at the j^{th} time point.
Source: Author

7.2 Model specification

7.2.1 The CLPM

We hypothesized that at time $j \in (1,...,T)$, a set of covariates x_j's form a first order construct FOC_j which in turn explain the endogenous CD4$^+$ count y_{ij}, for the i^{th} individual. The unobserved values FOC_{ij} and observed y_{ij} at follow up time j, are first modelled as

$$FOC_{ij} = \mu_{(FOC)j} + FOC^{*}_{ij}$$
$$y_{ij} = \mu_{yj} + y^{*}_{ij}$$

where $\mu_{(FOC)j}$ and μ_{yj} are temporal group means at time point $j \in (1,...,T)$ whilst FOC^{*}_{ij} and y^{*}_{ij} are temporal deviations from the temporal group means for individual i (Watkins and Styck 2017, Mund et al. 2018). For $j \geq 2$, FOC_{ij} and y_{ij} are initially modelled as

$$FOC^{*}_{ij} = \beta_{(FOC)j} FOC^{*}_{i(j-1)} + \gamma_{(FOC)j} y^{*}_{i(j-1)} + \varepsilon_{(FOC)ij} \tag{7.1}$$

$$y^{*}_{ij} = \beta_{yj} y^{*}_{i(j-1)} + \gamma_{yj} FOC^{*}_{i(j-1)} + \varepsilon_{yij}$$

where at time point $j \in (1,...,T)$, $\beta_{(FOC)j}$ and β_{yj} are the autoregressive parameters. The cross-lagged regression parameters are given by $\gamma_{(FOC)j}$ and γ_{yj}, whereas $\varepsilon_{(FOC)ij}$ and ε_{yij} are the residuals that are assumed to be normally distributed and correlated. For $j=1$, the FOC_{i1} and y_{i1} are modelled as exogenous variables. The individual differences have been studied in Chapter 4 and of interest are the lagged effects of both the CD4^{+} count latent variables on their next measurements. Hence, the simultaneous equations in the SEM framework ignored the i and γ in (7.1) for the FOC to solve a system of equations with

$$FOC^{*}_{j} = \beta_{(FOC)j} FOC^{*}_{(j-1)} + \varepsilon_{(FOC)j}. \tag{7.2}$$

Also in this study, at follow up time j, the CD4^{+} count y_j is explained by both the FOC_j and the previously measured CD4^{+} count y_{j-1} such that the system of equations includes

$$y^{*}_{ij} = \beta_{yj} y^{*}_{i(j-1)} + \gamma_{yj} FOC^{*}_{ij} + \varepsilon_{yij}. \tag{7.3}$$

Since the CD4^{+} count at time j is explained by more than one latent variable that can either be FOC or SOC, (7.3) is extended to model

$$y^{*}_{ij} = \beta_{yj} y^{*}_{i(j-1)} + \sum_{k=1}^{q-a} \gamma_{kyj} FOC^{*}_{kij} + \sum_{k=a}^{q} \gamma_{kyj} SOC^{*}_{kij} + \varepsilon_{yij} \tag{7.4}$$

where q is the total number of latent exogenous variables consisting of FOC and SOC. Our hypothesized CLPM is underpinned by a system of equations in the form of (7.2) and (7.3). When the CLPM is hypothesised, testing of the model is the same as the general CFA where the parameter estimation is based on fitting a model that can closely reproduce the covariance structure of the given data (He et al. 2017, Watkins and Styck 2017, Mund et al. 2018). Discussed below is the general covariance structure of a CFA model and the associated hypothesis testing.

7.2.2 The CFA model

In population terms, the CFA model predicts the observed covariance matrix. Hence the CFA models are also referred to as the covariance structure models. Recalling that the general SEM

consists of the measurement model and the latent variable model, the system of mathematical equations is represented as follows (Bollen 1989):

(a) The latent model is given by $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\varepsilon}$ where

$\boldsymbol{\eta}_{m\times 1}$ = vector of latent endogenous random variables

$\boldsymbol{\xi}_{q\times 1}$ = vector of latent exogenous random variables

$\mathbf{B}_{m\times m}$ = coefficient matrix showing the influence of latent endogenous variables on each other

$\boldsymbol{\Gamma}_{m\times q}$ = coefficient matrix for the effects of latent exogenous on latent endogenous variables

$\boldsymbol{\varepsilon}$ = disturbance vector

(b) The measurement model is given by $\mathbf{Y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}_y$ and $\mathbf{X} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\varepsilon}_x$ where

$\mathbf{Y}_{r\times 1}$ and $\mathbf{X}_{p\times 1}$ are vectors of the observed variables

$\boldsymbol{\Lambda}_y(r\times m)$ = coefficient matrices showing the relation of observed variables to latent endogenous

$\boldsymbol{\Lambda}_x(p\times q)$ = coefficient matrices showing the relation of observed variables to latent exogenous

$\boldsymbol{\varepsilon}_y(r\times 1)$ = measurement errors for $\mathbf{Y}_{r\times 1}$

$\boldsymbol{\varepsilon}_x(p\times 1)$ = measurement errors for $\mathbf{X}_{p\times 1}$

The $\mathrm{Cov}(\mathbf{X})$ as a function of $\boldsymbol{\theta}$ is given by $\boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Psi}_{\varepsilon_y}$ where $\boldsymbol{\Phi}$ is the covariance of the factors $\boldsymbol{\eta}$, the variance of the unique factors represented by $\boldsymbol{\Psi}$ and the vector containing unknown and unconstrained parameters of the model given by $\boldsymbol{\theta}$. The

$\text{Cov}(\mathbf{XY}) = \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_y(\mathbf{1}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}'_x$ is such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{yx}(\boldsymbol{\theta}) \\ \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) \end{bmatrix}$ which contains the implied covariances (Bollen 1989). If the covariance matrix of the population data is denoted as $\boldsymbol{\Sigma}$, the hypothesis testing in CFA model tests the departure of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ from $\boldsymbol{\Sigma}$. However, in practice the observed covariance is usually represented by $\mathbf{S}$ and the corresponding model implied covariance is then denoted by $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ or simply $\hat{\boldsymbol{\Sigma}}$. Hence in CFA hypothesis testing, the desired outcome is failure to reject $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ or practically $H_0 : \mathbf{S} = \hat{\boldsymbol{\Sigma}}$. Several goodness of fit measures are considered to assess the equivalence of these two matrices ($\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$).

7.2.3 Model estimation

The estimation procedure requires the satisfaction of several characteristics such as the data requirements, model identification, parameter estimation rules and evaluation criteria.

Data requirements in the estimation of the CFA model

There are some minimum requirements desired for the estimation of the CFA model including the sample size. A ratio of at least four subjects per free parameter has been recommended for sample size determination (Tanaka 1987) but this is multifaceted with statistical power where the CFA's desired outcome is failure to reject H_0 (MacCallum et al. 1996). Insufficient statistical power leads to a failure to detect nontrivial departures of $\hat{\boldsymbol{\Sigma}}$ from $\mathbf{S}$. Also taking indices of fit into account, their asymptotic properties begin to show for a sample size of least 400 (Hu et al. 1992) and to protect against chance features of the data in re-specifying a poor model, a sample size of least 800 is needed (MacCallum et al. 1992). Hence users are encouraged to aim for sample sizes of at least 200 and preferably 400 then 800 if necessary (Hoyle 2000). Since CFA is a multivariate statistical model, the estimation procedures such as maximum likelihood and generalised least squares assume multi-normally distributed data (Bollen 1989) although the actual data in many instances are usually non-normal (Micceri 1989). For a reasonably large sample, the CFA is usually estimated by maximum likelihood under most violations of multivariate normality (Chou and Bentler 1995, West et al. 1995) and assumes that the indicators are on a continuous scale (Jöreskog 1994).

Model identification

After the CFA model has been hypothesized, its parameters are specified in a path diagram as a set of measurement equations or represented as a set of matrices (MacCallum 1995). Solving the set of equations requires the identification of the model, a procedure for determining if the available equations and the number of unknown parameters can be solved satisfactorily. There are three possible situations: (a) the number of unknown parameters in the equations being equal to the number of equations, usually referred to as just identified CFA model but this cannot be tested; (b) an under-identified model where there are more unknown parameters than the number of available equations and this cannot be solved; (c) if the number of unknown parameters is less than the number of equations, the CFA model is over-identified and this can be tested. There are set rules for model identification in standard CFA applications (Bollen 1989) but not sufficient for complex models, in which case there is need to rely on the computer software or algebraic expressions (Hoyle 2000). The goal of the factor model identification is to reproduce the observed covariance matrix with as few estimated parameters as possible. The parameter estimates from the measurement model include the factor loadings, item intercepts and error variances whereas the structural model estimates are factor variances, covariances and means.

Rules for parameter estimation

The model identification depends on the number of items p and is given as degrees of freedom (df) that are mathematically represented as $df = \frac{1}{2}r(r+1) - t$ where $\frac{1}{2}r(r+1)$ is the number of possible (non-redundant) parameters, the known values (covariances) and *t* the number of free parameters in the model. This rule is sometimes called the *t*-rule where it is necessary that $t \leq \frac{1}{2}r(r+1)$ for parameter estimation although not sufficient for model identification. Taking note that $\mathbf{Y}_{r\times 1}$ and $\mathbf{X}_{p\times 1}$ are vectors of observed variables and as a continuation from CFA, it suffices to reason that the *t*-rule for the general SEM becomes $t \leq \frac{1}{2}(r+p)(r+p+1)$. The Two-step rule is also applied where the first step involves identification of a reformulated original model as a measurement model and then followed by establishing identification of latent variable model as if latent variables were observed with no measurement error.

Parameter estimation

The most widely used fitting function (Bollen 1989, Chou and Bentler 1995) for parameter estimation is the maximum likelihood fitting function which is given by

$$F_{ML} = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + tr\left[\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right] - \log|\mathbf{S}| - (r+p).$$

It is iterative in nature where there are incremental attempts to optimize parameter estimates that recover the observed data by minimizing the value of the fitting function. Convergence problems with the iteration process indicate issues with either the starting values, data, the model or a combination of these. Once the parameters have been estimated and substituted into the equations, the observed variances and covariances are most precisely recovered. The CFA produces two types of solutions: (a) un-standardised – useful for comparing solutions across groups or time (b) standardised – which are useful when comparing items within a solution on the same scale and the slopes are in a correlation metric. The variances of the implied model and those of the observed data are always equal but the covariances do not match and the difference reflects the degree of discrepancy in recovering the data. Interpretation of residuals in this case is difficult due to the metric of the indicators reflected in the covariances. There are further techniques for model diagnostics that are used to assess the adequacy of the implied model in recovering the covariance in the observed data.

CFA Model evaluation

Ideally the value of the fitting function should be zero for a set of parameter estimates that perfectly recover the observed variances and covariances where in such a case $\mathbf{S} = \hat{\boldsymbol{\Sigma}}$. This is not practically easy since $\hat{\boldsymbol{\Sigma}}$ can only be as close as possible to $\mathbf{S}$. Instead, goodness of fit is applied where several measures are actually used to assess the degree of discrepancy because there is no single number for a definitive inference (Bollen and Long 1993, Tanaka 1993) since the SEM evaluation is connected with difficult issues (Mulaik et al. 1989, Bollen and Long 1993, Steiger 2013). The fit indices can be absolute or comparative. The absolute fit indexes measures the degree to which $\hat{\boldsymbol{\Sigma}}$ is matched to $\mathbf{S}$ and the most prominent one is the $\chi^2 = (n-1)F_{ML}$. In a perfect model fit this value becomes zero regardless of the sample size. However, the χ^2 value increases with sample size as the value of the fitting function departs from zero (Hoyle 2000). The $CMIN / DF = \chi^2 / df$ is another measure called the minimum

discrepancy where the ratio is obtained by dividing the χ^2 value with the degrees of freedom. A ratio of approximately ≤ 5 is considered as being reasonable (Wheaton et al. 1977) although not clear how far it should be from 1 (Arbuckle 2012) but other researchers have recommended a ratio between 2 and 5 (Marsh and Hocevar 1985) whilst others being more restrictive to accept ratios ≤ 3.only. Other absolute fit indexes include the goodness of fit index > 0.9, standardised root mean square residual (SRMR) ≤ 0.08 and the root mean square error of approximation $\left(RMSEA = \sqrt{\frac{F_{ML}}{df}} \right) \leq 0.05$ according to (Arbuckle 2012). However, (Browne and Cudeck 1993) indicated that for a reasonable good fit model, the RMSEA should not exceed 1. The CMIN/DF and RMSEA are among the most popular measures of model fit (Holmes-Smith 2000). In the case of competing models that seem to capture the same theoretical framework, the model with the smallest Akaike Information Criterion (AIC) is selected.

Although these model fit measures are commonly directed on the implied model, additional models are also reported for comparison purposes. For example, a model with observed variables assumed to be uncorrelated (independence model) is severely constrained and expected to provide a poor fit to any data. Whereas a saturated model will be having no constraints and is considered the most general model possible. The comparative fit indexes compare the fit of a proposed model to a baseline model (Bentler and Bonett 1980) which is also referred to as the null or independence model. A commendable measure in this category is the comparative fit index (CFI) > 0.9 (Bentler 1990). Instead of comparing fit indexes, the models themselves can be compared in terms of parsimony using the fit indices. CFA models can be miss-specified as a result of either wrong number of factors, un-modelled sub-factors or wrong loading patterns (indicators assigned to wrong factors or loading on multiple factors). The R^2, a proportion of variance attributable to the factor(s) on which each item loads can be used to implicate the un-modelled factors that contributed to the variability in the responses. To improve the performance of the indices, re-specification can be considered where some parameters can be freed or fixed (Bollen and Long 1993) although care should be taken to avoid Type I error (MacCallum et al. 1992).

7.3 Data analysis and software

The analysis of a moment structures (AMOS) software, version 25 was used for path modelling and the parameter results graphically displayed in R software, version 3.5.3 of the R Core Team. The study pertains to medical variables that exist in standard known groups with at least two indicators per construct as required (Bollen 1989). However, the indicators have weak correlations and the objective was to obtain the causal relationships between the formative latent constructs of the independent variables (the $CD4^+$ count clinical covariates) and dependent variable ($CD4^+$ count). In such causal relationships, the SEM is considered as the most appropriate application (Schumacker and Lomax 2010, Hair et al. 2014). The task was not strictly confirmatory and as such several models were tested to obtain the one that closely fits the data. The standardised direct effects as well as the correlations among the variables were estimated. The estimates in AMOS are computed by the maximum likelihood which is known to produce estimates with very desirable properties (Arbuckle 2012). The causal effect diagrams were identical across all the phases but the correlation patterns were not restricted and allowed to vary as suggested by the modification indices. The standardised regression weights and correlation results were of interest as they are independent of the units in which all the variables are measured (Arbuckle 2012). Furthermore, results interpretation was not feasible from the meshed regressions paths and the correlations across the four phases. Again exporting the results to R as csv files was the best option. The library **ggplot2** and parallel coordinate plots proved once more again to be very helpful in the visual displays. The main challenge faced with the CLPM was that the raw data for analysis was presented in the long format. Prior to data analysis in AMOS 25, the longitudinal data were reshaped into wide format using the function `dcast` in R. Reshaping the data to wide format with many variables required high versatility in a software such as R due the variable naming that tend to increase dramatically. All the R codes are presented in Appendix C7.

7.4 The results of fitting the CLPM

7.4.1 The results of the model diagnostics

The model fit diagnostics are given in Table 7.1 and the results showed that the implied models had smaller discrepancies as compared to the constrained (independence) models. All the CMIN/DF ratios were below 5 and mostly between 2 and 3. Only the phase 2 ratio was slightly above 3 with its AIC also slightly higher than the other models showing that it was slightly

harder to fit the phase 2 data than the other phases. The RMSEAs were all below 1 indicating that the models were acceptable. An attempt to fit the reflective measurement models produced higher AIC values and CMIN/DF ratios although the RMSEAs were relatively the same. Hence, the hypothesised formative measurement models were the best model fit and selected.

Table 7.1: Results of model fit diagnostics

Phase	Independence		Implied model				
	CMIN	DF	CMIN	DF	CMIN/DF	RMSEA	AIC
2-Acute	10381.457	2278	2232	2278	3.182	.096	7331.152
3-Early	7958.422	2278	2224	2278	2.636	.083	6106.523
4-Established	7604.282	2278	2222	2278	2.675	.084	6191.427
5-ART	9259.683	2278	2225	2278	2.925	.090	6750.485

7.4.2 The resultant layout of the CLPM path diagrams

In order to accommodate the variable names in the path diagrams, only the first three characters were considered. In our case, there was no similar names upon truncation but this requires careful attention. Even though the variable names were shortened, it was necessary to reduce the font size to create enough space for the path diagrams. Further, the readability of the congested path coefficients required some colour codes.

The path diagrams of the fitted (implied) models were set up as shown in Figures 7.3 and 7.4. The single headed red arrows represent causal effects and the associated red values show the regression weights. The rectangles are the observed variables whose names have been trimmed to leave the first three characters only for convenience. The trimmed variable names are suffixed with a number representing the time point at which the observation was recorded. As such, all the observations recorded at Time 1 say, appear in a block headed “Time 1”. The double headed aqua (light blue) coloured curved lines represent the correlations between some of the observed variables (also see Appendix B for enhanced correlation view). It was hypothesized that the white blood cells (WBC) were an aggregation (oval shaped) of the lymphocytes, basophils and monocytes. Similarly, the red blood cell group (RBC) being a composite variable formed by MCV, MCHC and the red blood cells. Total protein, ALP, calcium, magnesium, potassium, sodium, albumin, LDH and folate were meant to form the biochemistry (BIOCHEM) construct.

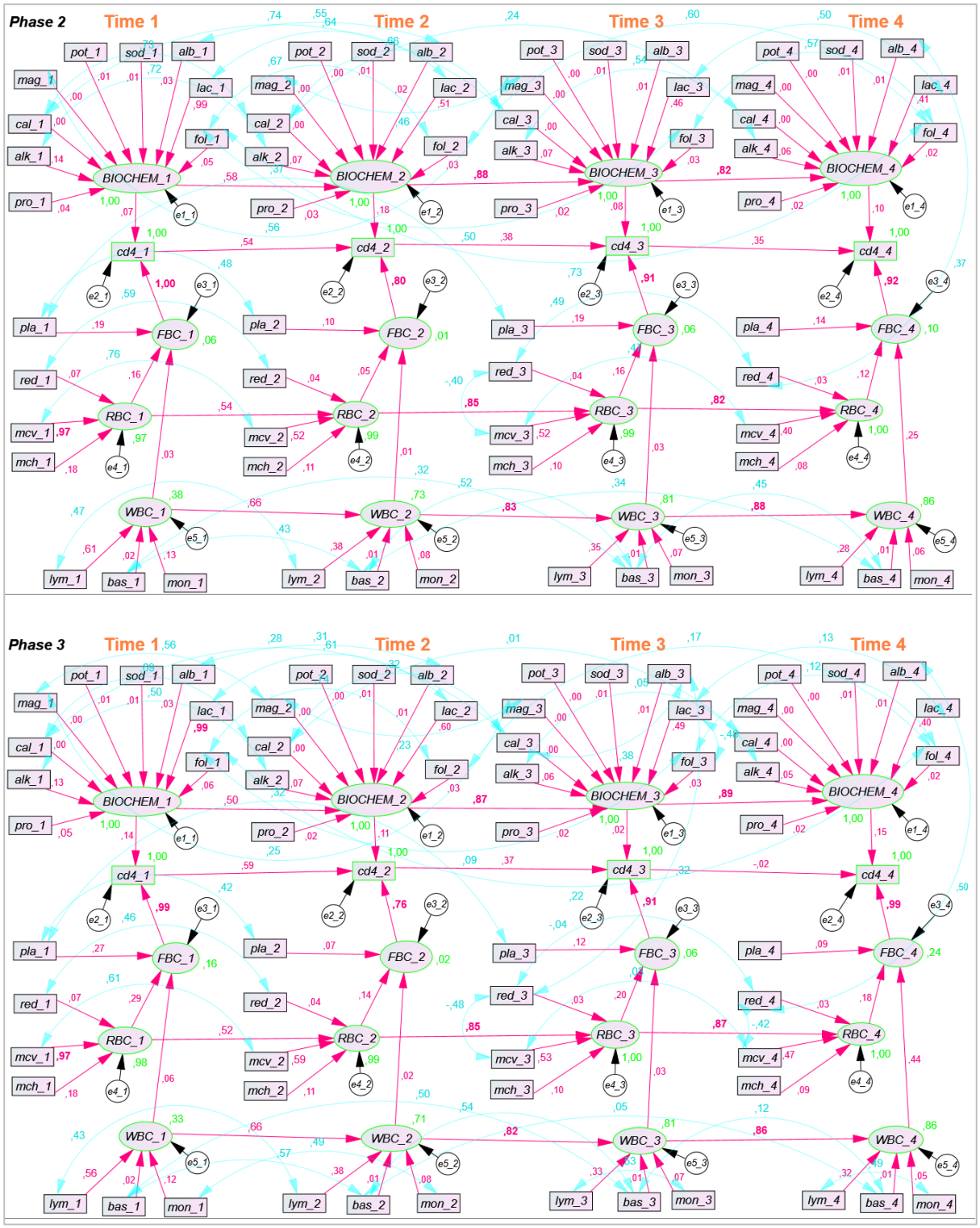


Figure 7.3: Results of path analysis diagrams of HIV infection phases 2 and 3

Abbreviations: *mon*-monocytes; *bas*-basophils; *lym*-lymphocytes; *mch*-MCHC (mean corpuscular haemoglobin concentration); *mcv*-MCV (mean corpuscular volume); *hae*-haemoglobin; *red*-red blood cells; *glu*-glucose; *pla*-platelets; *alk*-alkaline phosphatase(ALP); *cal*-calcium; *mag*-magnesium; *pot*-potassium; *sod*-sodium; *alb*-albumin; *lac*-lactate dehydrogenase(LDH); *fol*-folate; *BIOCHEM*-biochemistry; *RBC*-red blood cell group and *WBC*-white blood cell group. The number suffix represents the time point of observation.

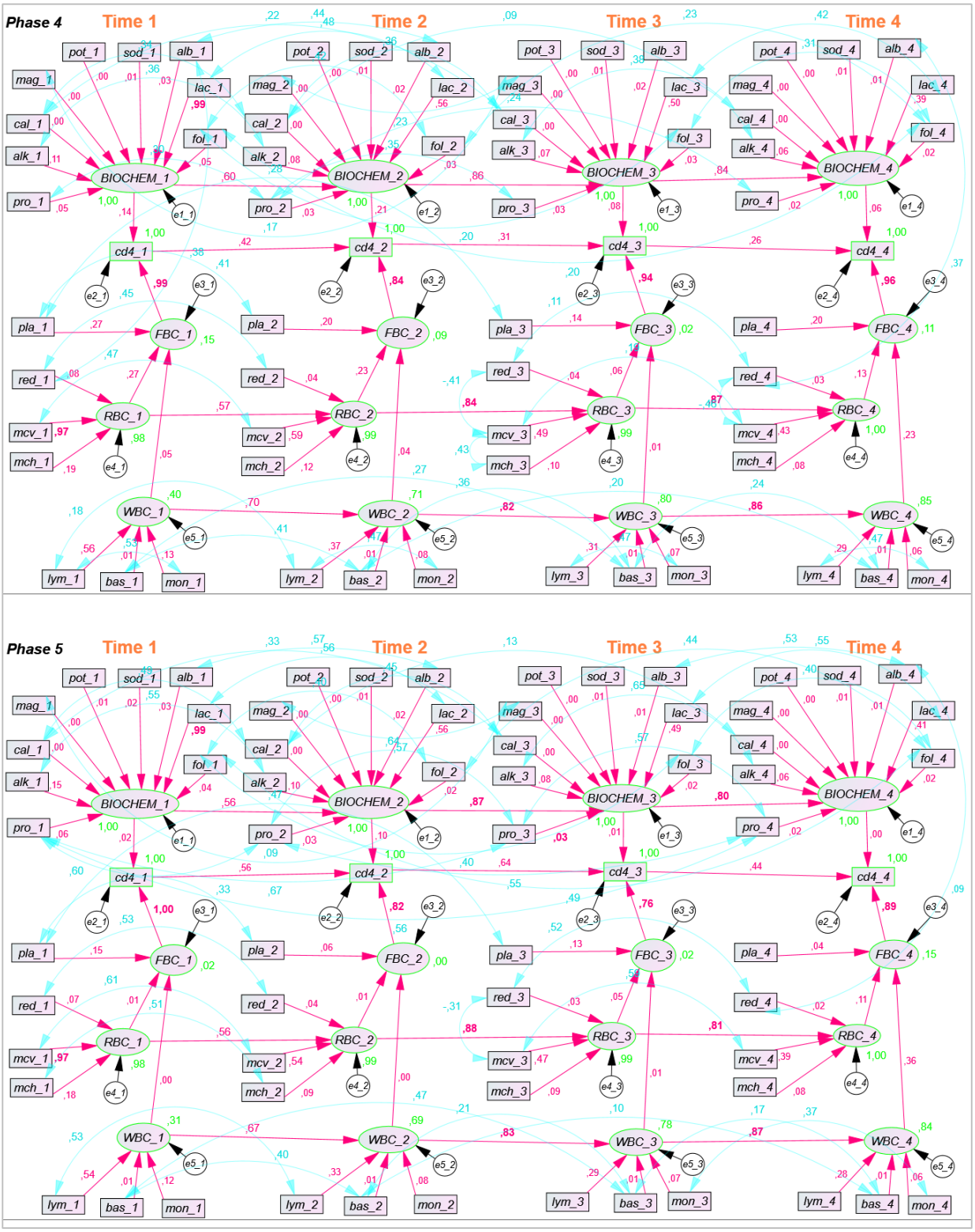


Figure 7.4: Results of path analysis diagrams of HIV infection phases 4 and 5

Abbreviations: *mon*-monocytes; *bas*-basophils; *lym*-lymphocytes; *mch*-MCHC (mean corpuscular haemoglobin concentration); *mcv*-MCV (mean corpuscular volume); *hae*-haemoglobin; *red*-red blood cells; *glu*-glucose; *pla*-platelets; *alk*-alkaline phosphatase(ALP); *cal*-calcium; *mag*-magnesium; *pot*-potassium; *sod*-sodium; *alb*-albumin; *lac*-lactate dehydrogenase(LDH); *fol*-folate; *BIOCHEM*-biochemistry; *RBC*-red blood cell group and *WBC*-white blood cell group. The number suffix represents the time point of observation.

The RBC and WBC were considered as the FOC that formed a SOC called the full blood count (FBC). In addition to the RBC and WBC, the platelet count was also considered to contribute to the formation of the FBC formative construct. The $CD4^+$ count of the previous time point was allowed to have an effect on the $CD4^+$ count of the next time point. In addition to this, it was also believed that the $CD4^+$ count at any given time point was as a result of the influence of the FBC and the BIOCHEM. An attempt to include the direct effect of glucose to the $CD4^+$ count showed a very insignificant effect and the glucose variable was dropped from the model together with haematocrit which was found to cause estimation challenges due to a high correlation with the red blood cells. The small circles represent the random error terms and the green values indicate the variation explained by the exogenous variables with 1 showing a better explanation of the endogenous variables bordered in green colour. The BIOCHEM, RBC and WBC of the previous time points were also allowed to have an effect on their respective values of the next time points. The lagged effects of the FBC were not supported by the covariance structure resulting in a poor model fit and this was not modelled. The path diagrams provided a system of relationships among the variables showing waves in the disease progression. However, the assessment of the effect strength of the large number of regression coefficients was difficult considering that we had four CFA models each with four time waves. To alleviate this challenge and that the regression weights were standardised, parallel coordinate plots were employed and augmented by multiple bar charts throughout the rest of this chapter.

7.4.3 The results of the regression weights and correlations across the phases

A detailed pattern of the comparison of the regression weights across all the infection phases is shown in parallel coordinate plots of Figures 7.5 and 7.6. However, it is important to note that the parallel coordinate plots are too sensitive to the differences in the regression weights as compared to the figures shown in the path diagrams which are shown just to 2 decimal places. The effect of the previous $CD4^+$ count on the $CD4^+$ count of next time point increased continuously during the ART whilst it was observed to be continuously dropping during the early phase. Although the FBC dominantly influenced the $CD4^+$ count, its effect on the $CD4^+$ count was also continuously declining at each time point during the ART. Contrary to this, the FBC effect on the $CD4^+$ count was increasing at each visit time during the early infection phase. Despite being overshadowed by the greater contribution of the previous BIOCHEM measurements, the LDH influence on the current BIOCHEM increased at each visit time during the ART. The other noticeable patterns were the decline in the MCV effect on the RBC over

time during the ART phase and an increasing effect at each visit time during the early phase. Despite the low regression weights in the paths shown in Figure 7.6, there was a noticeable mirror image in the pattern of weights between the established and ART phases. When the established regression weights were low, there tended to be a corresponding higher regression weights during the ART phase.

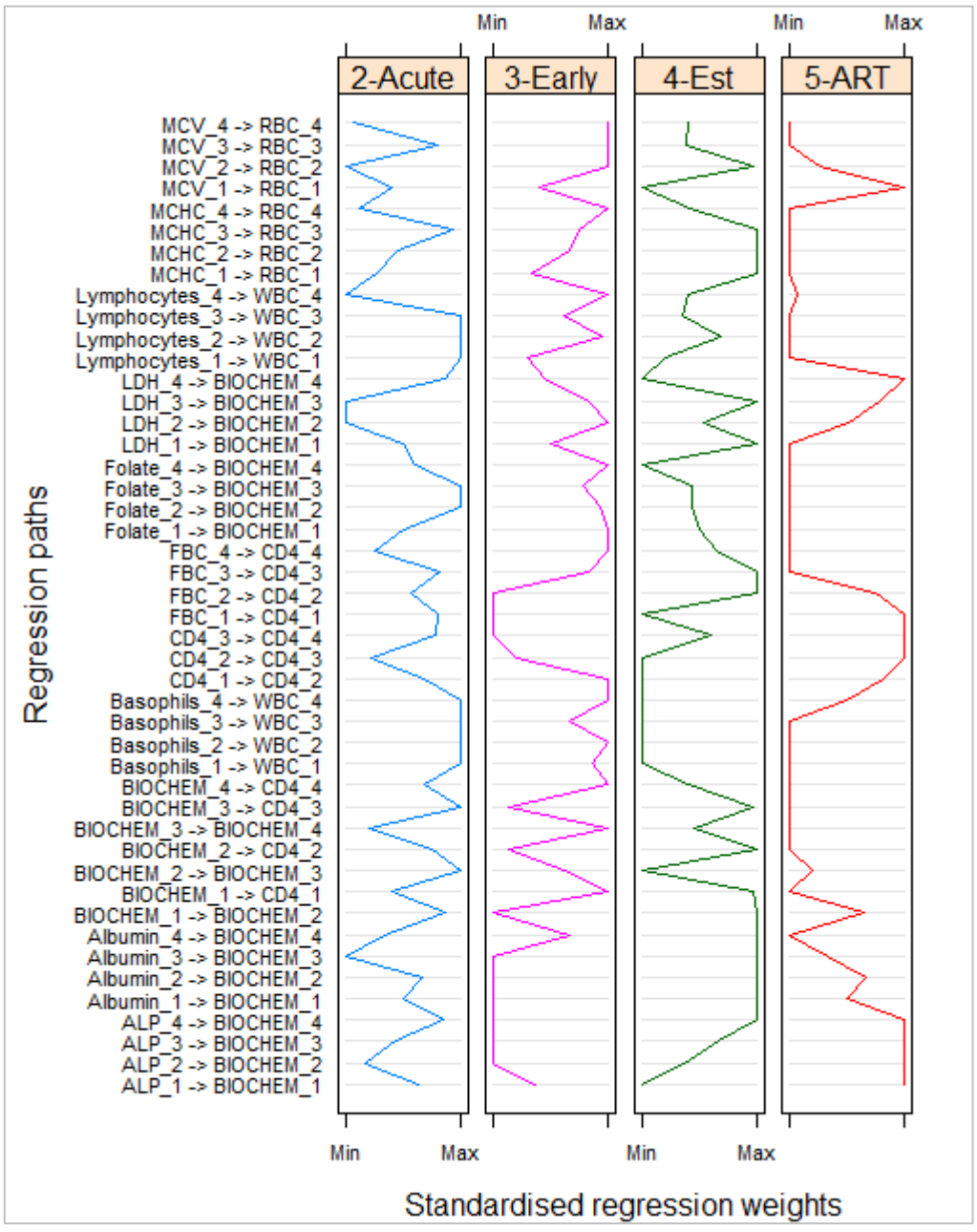


Figure 7.5: Results of the path regression weights across all the phases (a)

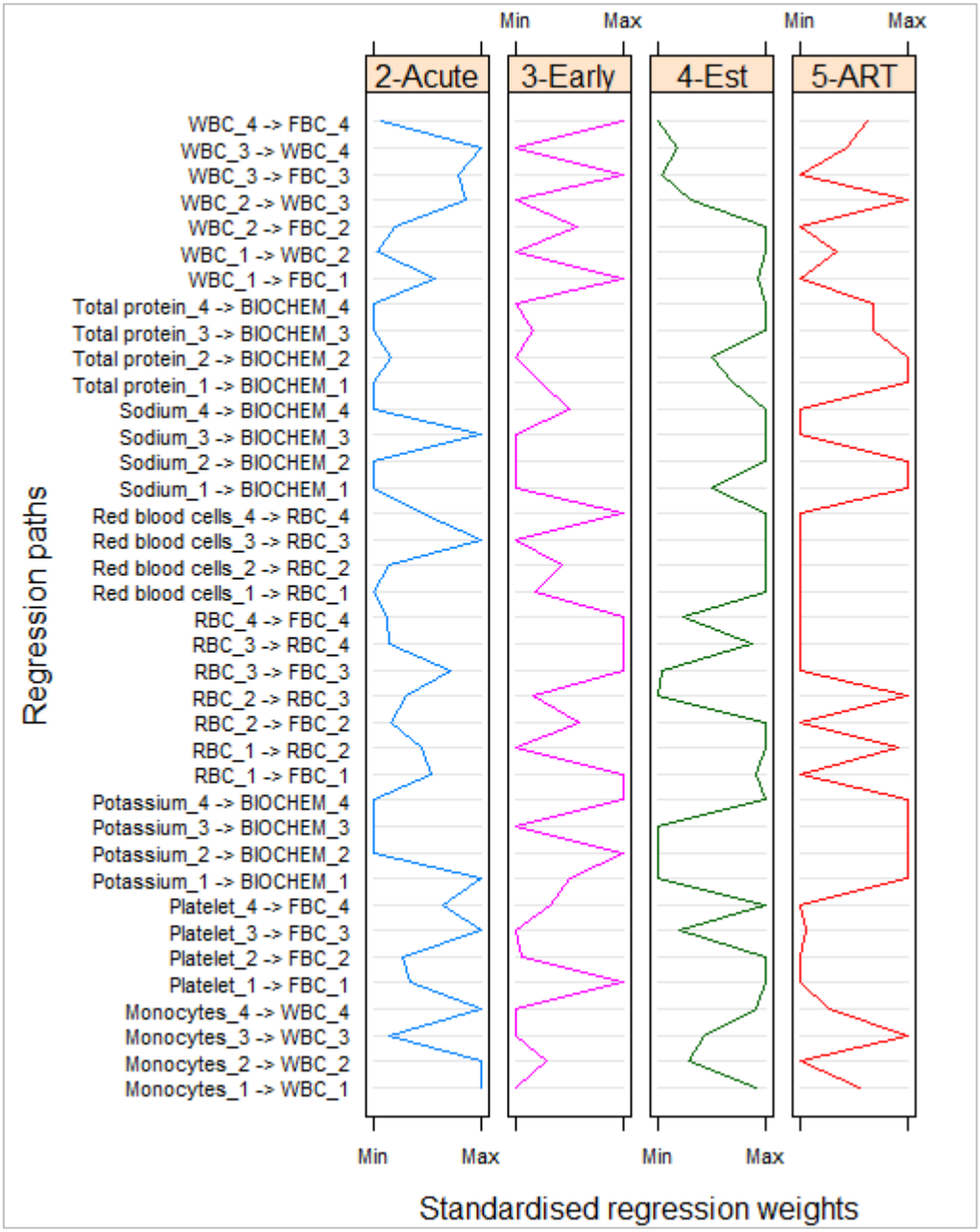


Figure 7.6: Results of the path regression weights across all the phases (b)

The overall strengths of the interrelationships between the variables is shown in Figure 7.7. We selected only those with regression weights > 0.4, although the main interest was in the weights $\lambda > 0.7$. The results showed that the first order constructs (RBC, WBC and BIOCHEM) in all the infection phases had influential effects ($\lambda > 0.75$) on their next time measurements from Time 2 → Time 3 → Time 4. The second order construct (FBC) also showed strong causal effects on the $CD4^+$ count at all next time points and within each infection phase were high, $\lambda_{All\ phases}^{FBC\_All\ time\ points} \geq 0.75$.

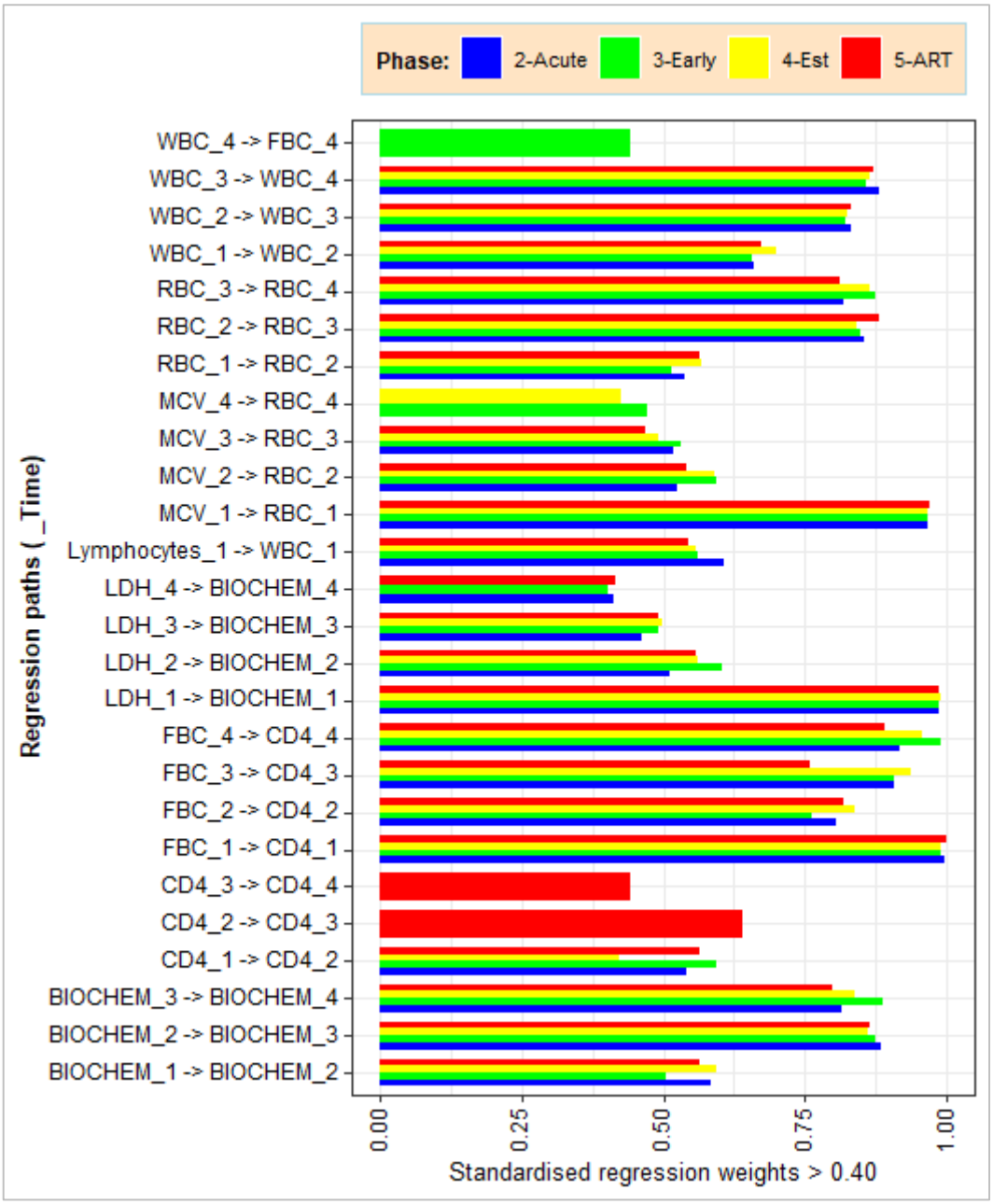


Figure 7.7: Standardised regression weights > 0.40

It is important to note that at Time 1 measurements, the effects of the previous measurements are not taken into consideration. Hence, this provided a cross sectional understanding of the weights of the observed variables on the latent variables (clinical categories) in the absence of the effects due to the previous time measurements. Accordingly, and with the aid of Figures 7.5 and 7.6, Figure 7.7 further showed that the Time 1 results indicated that the observed LDH dominantly contributed to the BIOCHEM group with standardised regression weights of $\lambda_{All\ phases}^{LDH\_1} = 0.99$. Similarly, the MCV contributed the most to the FBC where the Time 1 regression weights in all the phases were $\lambda_{All\ phases}^{MCV\_1} = 0.97$. Although the lymphocytes were

dominant among the WBC, the regression weights were not all the same across the phases where $\lambda_{Phase\_1}^{Lymphocytes\_1} = 0.61$, $\lambda_{Phase\_2}^{Lymphocytes\_1} = \lambda_{Phase\_3}^{Lymphocytes\_1} = 0.56$ and $\lambda_{Phase\_4}^{Lymphocytes\_1} = 0.54$. These results also showed that at any given time across all the infection phases, the influence of the previous $CD4^+$ count on the current $CD4^+$ count was always less than that due to the current FBC. The current BIOCHEM, RBC and WBC were mostly determined by their previous measurements rather than the LDH, MCV and lymphocytes respectively as observed during the cross sectional view at Time 1. The weights of the RBC and WBC on the FBC were very small at Time 1 and improved well at Time 4. It means that generally the autoregressive coefficients of the $CD4^+$ count, BIOCHEM, RBC and WBC had more influence from the previous time points and also showing little variance over time. Figure 7.8 shows the correlations of different variables that were recorded at different time points.

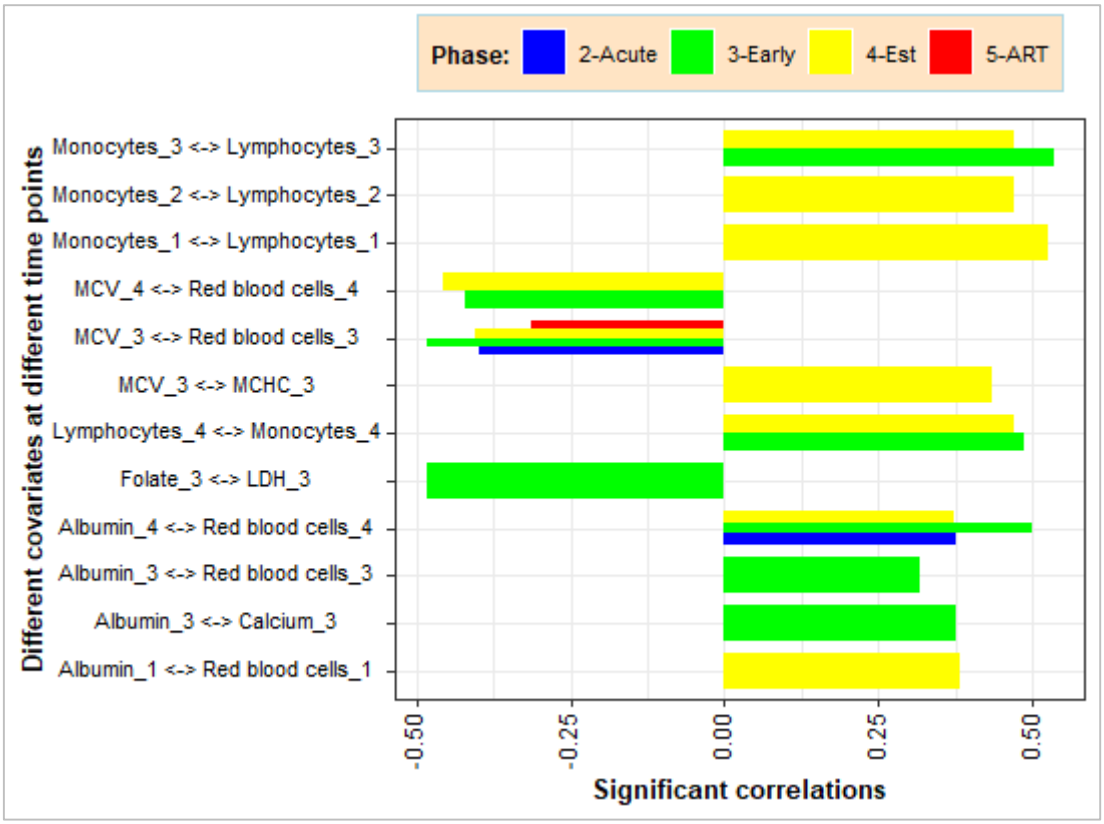


Figure 7.8: Results of the correlation patterns for different variables at different time points

Recalling Figure 2.9 that during the early and established phases the cohort's average $CD4^+$ counts were the least and below the population average at each visit time, these were the infection phases that also showed some variable correlations at the different time points. At these lower $CD4^+$ counts, the WBC, monocytes and lymphocytes were found to be highly and

positively correlated. The measurements of the MCV and red blood cells were negatively correlated at such low $CD4^+$ counts. The investigated data at hand showed that the MCV and red blood cells were negatively correlated at Time 3 of all the infection phases. The folate and LDH were correlated and negatively during the early phase only. The results also showed that the measurements of the same variable obtained at different time points were highly correlated (Figure 7.9).

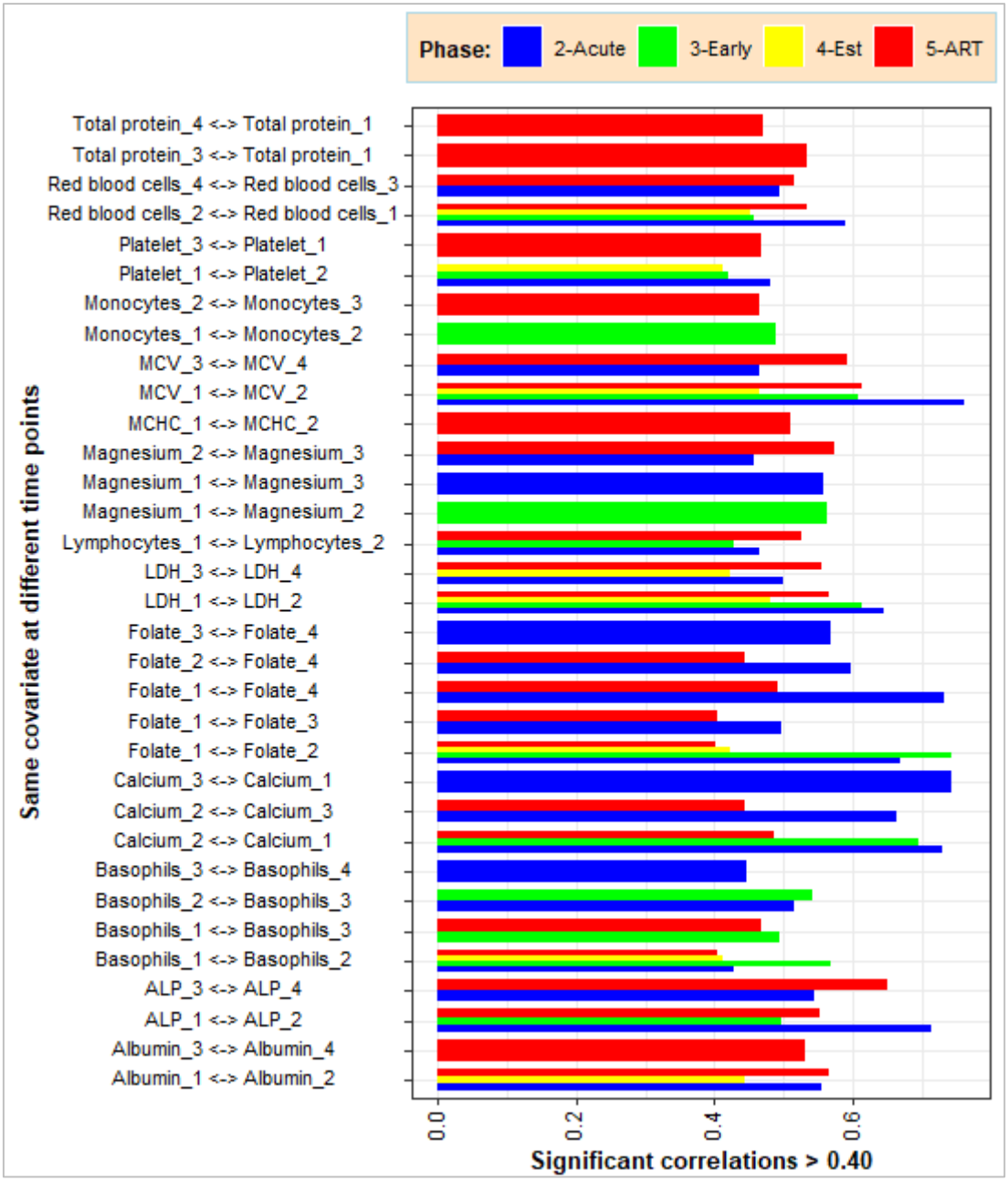


Figure 7.9: Results of the correlation patterns for the same variable at different time points

The MCV, folate, calcium and ALP had the highest correlations ($r \geq 0.70$) among their measurements recorded at different time points and these were observed during the acute phase only. Folate was the only $CD4^+$ count covariate to show high correlation ($r \geq 0.70$) between measurements at different time points and this was observed during the early phase.

7.5 Clinical interpretation of the results

The $CD4^+$ cells are a T cell type (Papagno et al. 2004) whereas the lymphocytes are either B or T cells (Shapiro et al. 1998, Project Inform 2007, Obirikorang et al. 2012). Our study showed that given a cross sectional view of the health status based on the $CD4^+$ cell count, the FBC played a major role in determining the immune system's status of HIV infected patients as compared to the biochemistry. Due to the fact that the FBC was a resultant of the WBC which was in turn an aggregate consisting of the lymphocytes, hence the FBC was the most contributing factor of the $CD4^+$ count. Excluding the effect of the previous time points, the lymphocytes dominantly influenced the WBC and this effect was high during the acute infection phase which is known to be associated with a high viral load (Bellan et al. 2015, Manoto et al. 2018). This suggests the active involvement of the lymphocytes during foreign invasion. Lymphocytes and monocytes are both defence cells and both are known to be attacked by the HIV (Pasupathi et al. 2008). Hence their likely positive correlation during the low $CD4^+$ counts in our data.

The LDH is known to catalyse the fulfilment of short-term energy requirements in the case of insufficient oxygen (Valvona et al. 2016). In our data, the enzymatic activities of the LDH were found to be the most influential among the biochemistry components. This pointed out to a lack of oxygen in generating energy during the HIV infection phases as indicated by the high glucose cell consumption to maintain the patient's $CD4^+$ cell count. In our results, the short-term insufficient of oxygen could be linked to the fact that the red blood cells effect on the composite RBC was insignificant. This supports the reports that HIV patients are anaemic (Kulkarni et al. 2015). In our data, the coincidental shortage of oxygen as depicted by the role of LDH in controlling the biochemistry, indicated a reduced aerobic endurance. Aerobic endurance refers to the functional state of the oxygen transport system and has also been reported to be a health related parameter (Baquet et al. 2003). Reduced aerobic endurance in HIV positive patients has also been reported in the Western populations (Cade et al. 2003, Oursler et al. 2006). The results of the present study on reduced aerobic endurance were similar

to a study in the same geographical setting in Malawian HIV positive patients whose aerobic endurance was also found to be lower than in the HIV negative control group (Chisati and Vasseljen 2015).

Currently it is common practice to monitor the HIV disease progression based the $CD4^+$ count because they are the main target of the HIV (Weston and Marett 2009). Our results confirmed that indeed the $CD4^+$ count had influence from the previous time points but more attention should be given to the current full blood count. Over time, the $CD4^+$ count lagged effects were decreasing whilst the full blood count current effects on the $CD4^+$ count were increasing. The fact that the $CD4^+$ count lagged effects were declining suggests less stability over time which is associated with more variance.

7.6 Summary

In considering a more complex causal relationship between the $CD4^+$ count clinical covariates, the important players in our data were the FBC as compared to the biochemistry. The biochemistry was dominantly controlled by the LDH whereas the RBC and WBC were dominated by the MCV and lymphocytes respectively. Although the variable selection methods in Chapter 3 were instrumental in alleviating the multidimensional curse, the unimportant variables were completely discarded whereas the SEM would utilize any smallest effect of such variable into the constructs as part of variable reduction. The multilevel models improved the understanding of the $CD4^+$ count by capturing the variations at individual level but nothing was provided on how the covariates were related themselves. In a way the GAMM and segmented models attempted to "listen" to the data by modelling the data patterns closely. To some extent dropped the insignificant covariates. However, unlike the multilevel models, the data dependence in GAMM was lost and again with no information on the interlinks between the covariates. The CLPM proved to be versatile in providing an overview of the causal relationships among many variables as well as giving the opportunity and ability to measure the unobserved variables that were part of the theoretical framework for the problem at hand. With the SEM, we were able to visualise the system of unobserved variables that went silently untapped in Chapters 3 to 6. Further, it was clear to realise the chained covariate effects towards the response with even fewer construct variables. The constructs acted as variable reduction in which the effects of the observed variables were collectively and indirectly accommodated. Because the monitoring of the HIV disease progression happens in a time space, we also gained

insight into developmental system of all the variable relationships over time through the lagged effects. However, like the multilevel models, the underlying generalised least squares in the SEM may be summarising some information as compared to the GAMM and the segmented models that are more of data-driven

CHAPTER 8 DISCUSSION AND CONCLUSION

The study strived to examine the adaptive effects of clinical covariates on $CD4^+$ cell influence in reaction to the body's weakened immune system as induced by HIV invasion. In particular, we looked at 46 continuous clinical covariates of the $CD4^+$ count, repeatedly measured from 237 patients' post-HIV infection phases, at the Centre for the Programme of AIDS research in South Africa (CAPRISA). The phases after seroconversion were acute, early, established and ART. Understanding the data, called for the application of raw data handling techniques for multidimensional continuous variables. The initial stage of data exploratory analysis was conducted for two objectives: data preparation and pattern discovery.

The data preparation began with an assessment of distributional properties of the covariates from an overall point of view, followed by similarly assessing within each HIV infection phase. Displaying boxplots of variables with different scales on a single graph is challenging and not common in the literature. With the multidimensional curse of our data, this was successfully achieved by scaling the values to percentiles. Ranking each scaled value of a variable, misrepresented the actual distributions. To overcome this problem, distinct values were scaled and the ranks were then matched to all the other values. Some of the limitations of the boxplots were augmented by the strengths of parallel coordinate plots, which were further instrumental in visualising the distributional patterns among the many variables in their original scale. The visualisation of the variable distributions detected both the erroneous and outlier measurements, which were then subsequently removed. This contributed to the challenge of the proportion of missingness present in the data. A further challenge was that the CAPRISA raw data were in format of an unbalanced design, where each patient did not have identical number of repeated measurements per infection phase. The interest of this study however, was to work with a balanced design as anticipated in the statistical applications. In an attempt to solve this problem, the last four records of each patient before each phase transition were selected. The corresponding missingness proportion was thus reduced. The missingness mechanism was subsequently tested, as well as being visually displayed. The conclusion was reached that the data thus obtained were missing not completely at random. Hence a multilevel data imputation method, using a multivariate imputation by chained equations, was applied. This proved to be powerful, since the distributional properties of the observed and imputed data sets were almost identical.

The exploratory data analysis for the pattern discovery began with some descriptive statistics, which were subsequently noted to give limited information based on phase specific and overall summaries. The summary results included the central tendencies and measures of dispersion. Although the magnitudes of the variable scales were informative, it was realised that pattern discovery was very limited due to the loss of information in the summarisation process. The procedure was further complemented by some common graphical displays such as individual linear trajectories and locally weighted scatterplot smoothers, for each infection phase. Some trends were observed but tended to be challenging due to the multidimensional curse. This prompted the recalling of the parallel coordinate plots, which helps to visualise the complex behavioural patterns between the covariate and $CD4^+$ count relationships, in a consolidated fashion. A correlation plot, as another form of multidimensional visualisation, was subsequently introduced. This plot revealed some redundant features, which were subsequently dropped as they are known to pose some multicollinearity problems. The dropped redundant features consisted of those variables with the highest mean absolute correlation.

The whole exploratory data analysis exercise eventually resulted in a clean and complete balanced design data set. The patterns discovered in the data structure were very insightful and indicated that indeed the covariate influence on the $CD4^+$ cell response existed, with infection phase being a factor. Although the patterns were limited in some way, they provided strong leads on the best course of action for suggesting more advanced investigative statistical modelling techniques in order to obtain a more comprehensive understanding of the covariate influence on the $CD4^+$ cell response.

Even though redundant features were dropped during the exploratory data analysis, few features were dropped, with selection merely based on direct correlation between the covariates. Of interest were the clinical covariates that capture the greatest variation in the $CD4^+$ count. Their effects were believed to be influential in the management of the patient health status during HIV disease progression. In addition, this potentially enhances resource optimisation during the HIV/AIDS prospective studies. As such, a further variable screening process was deemed necessary. Three recently developed longitudinal variable selection techniques that focus on different objectives, were evaluated. These are the time-varying coefficient models, penalised generalised estimation equations and sparse partial least squares. The sparse partial least squares approach was considered to be a more robust technique, as it takes into account the multilevel nature of the longitudinal data across all the phases, as well as simultaneously accounting for the highest variation in both covariates and responses. Since

there was only a single response variable, the $CD4^+$ count, the main task was to reduce the 46 covariates. The adopted results of the sparse partial least squares process selected only 18 covariates. These were mostly from electrolytes, proteins and red blood cells. Having successfully screened the covariates to obtain the most salient ones, this paved the way for more parsimonious model building with other statistical techniques to untangle the covariate influence on the $CD4^+$ cell count.

We then investigated the individual $CD4^+$ count linear trajectories in response to the covariates within each infection phase. These were then used to compare the $CD4^+$ count variation between patients in a bid to identify the clinical covariates that induced some sources of variation. The aim was to promote a tailored medical plan for patients if it turned out that each patient has a unique relationship between the response and the covariate. Wider variations in the $CD4^+$ counts in response to a covariate would be an indication of the uniqueness due to that particular covariate. Hence, well-informed management of the covariate influencing the $CD4^+$ count was called for. It turned out that multilevel models were appropriate for this task as they provide evidence of specific individual effects. Unlike regular regression, they allow study effects to vary by groups, which was suitable for the HIV infection phases at hand. The covariates were mean-centred, to compare the $CD4^+$ count variation, for an average patient. Although the variable reduction helped to reduce the size to 18, the selected covariates were still large for SAS. More so, there was limited literature available on multilevel model formulation for high-dimensional covariates, with a grouping factor. This work further contributed towards establishing such a mathematical representation. We systematically investigated the effect of each covariate on the $CD4^+$ count, for each patient. On average, each patient had a unique relationship of the response and every covariate during ART and majority of the covariates in the initial upsurge of viral load, acute infection phase. Some of the $CD4^+$ count variation patterns included an increased rate of $CD4^+$ count for the already high $CD4^+$ count, or the worsening of the $CD4^+$ counts, which were already lower in response to an increase in the covariates. The results further revealed that an increase in some of the covariates was associated with less variation in the $CD4^+$ counts among the patients. Although the multilevel models were used in capturing the individual effects, the relationships between the $CD4^+$ count and the covariates were forced to assume a linear dependency. This approach relied on interpreting parametric coefficients that summarises the relationships. It was then crucial to visualise the actual behavioural trends using additive models.

The additive models allow interesting discoveries to be made due to their flexibility, which is not controlled by assumptions. They provide clear visualisations in displaying the response curves, whilst relaxing the linearity assumption of the multilevel models. Their strength of this approach lies in the ability to be highly selective of the data points that are giving a true reflection of the response behaviour. The smoothing technique penalises the data points that brings in the "wiggliness" in the shapes of the curves, whilst automatically taking care of overfitting. Similar to the case of multilevel models, the random effects are incorporated but as random smooths, which we obtain by fitting a generalised additive mixed model. In our data, smoothing individual trajectories to a few data points per infection phase was not feasible since even if ignoring the phases, the number of patients was simply too large for visualisation. Consequently, specific trends were modelled. Although the additive models accommodate semi-parametric terms, the best model fit for our data demanded that we have strictly non-parametric terms. The traditional way of presenting such graphs was not easy due to the high demand for programming codes and space requirements. Hence, this work also contributed some literature on embedding the usual models into macros for productivity in modelling multidimensional data. The results showed that an overall increase in lymphocytes, haematocrit, platelets, albumin and ALP, corresponded to a general upward trend in the $CD4^+$ count, whilst total protein and sodium corresponded to general downward trends in the $CD4^+$ count. On the other hand, LDH and red blood cells showed irregular overall trends. Given these overall trends, it was also observed that all the associated phase specific trends were significantly different. For the other covariates such as MCV, basophils, monocytes and folate, their phase specific trends were not significantly different from each other, but their overall trajectories indicated a significant effect on the $CD4^+$ count. A comparison of the trends in relation to the cohort's average $CD4^+$ count provided an understanding of where the covariates influenced the $CD4^+$ count to desirable levels. The points at which the curves were above the average suggested the covariate set-points which could potentially be of use to manage towards a better $CD4^+$ cell outcome. These set-points were found to mostly differ across the HIV infection phases. Also, the smooth curves were in the form of straight lines that joined at some turning points. Their interpretation was based on approximate values extracted from the graphs. For meaningful biological references in managing the HIV disease, a more specialised technique was required to detect the break-points. In addition, the additive model transformed the response scale which could accordingly not be interpreted. As a way forward, a segmented regression approach was used for the break-point detection.

The segmented linear regression model attempts to identify the number of turning points in the curves and then fits linear equations between them. Locally weighted scatterplot smoothers can be used to provide initial estimates during the estimation procedure to minimise the chances of having the algorithm getting stuck in a localised turning point. The strength of the linear relationships within the covariate sub-intervals produced by the break-points is also assessed by confidence intervals. The mathematical representation of multiple covariates with multiple break-points was enhanced, and further, macros were written to embed the models. The results showed that the $CD4^+$ count had mostly three segments in which there was different relationships with the covariates. However, in some cases for a given covariate, the break-points were not detected in all the infection phases. Similarly, the break-points of a covariate that were identified at each infection phase, were mostly not the same value. Only sodium consistently showed that its linear relationship with the $CD4^+$ count changed in all the phases at approximately 132mEq/L and 140mEq/L. The $CD4^+$ count relationships within the covariates sub-intervals were either significantly positive, negative or insignificant. Although it was important to pin-point these covariate values as possible reference ranges in which they influence the $CD4^+$ count, the segmented model and all the previously discussed models looked at each covariate in isolation. In more realistic terms, the $CD4^+$ cells, the covariates and the HIV co-exist in a host whose system consists of complex interrelationships. We subsequently introduced the structural equation models to unravel such connections.

The structural equation models allow for the modelling of unmeasured variables. Like any other research activity where a study is usually triggered by a theoretical framework, the structural equation models proposes a system of relationships that are subject to evaluation. A confirmatory factor analysis model was considered where the original hypothesis was modified until the best fit to the data was achieved. The clinical platforms of the covariates were assumed to be the theoretical constructs that cannot be measured. Together with the observed covariates, the constructs were believed to form a system that had influential effects on the $CD4^+$ cell count. Because the $CD4^+$ count was monitored over time, the developmental changes in the confirmatory factor analysis were captured as autoregressive effects. Hence, the structural equation models variant called cross-lagged path model was the most suitable approach due to its resemblance to the homeostatic nature that exists in real life supporting systems. The reshaping of data to suit this model gives rise to an increased number of variables to analyse and consequently the output. Again there was no documented approach available to present such massive amount of cross-lagged path model results data. We also resorted to results data

wrangling and graphical displays of the standardised regression weights using parallel coordinate plots, that helped with pattern discovery too. The correlations were better displayed as multiple bar charts. The system of interrelationships showed that at any given time, the unmeasured secondary construct from full blood count, had the greatest influential effects of the $CD4^+$ count. This secondary construct was formed by other first order constructs namely red blood cell group (the dominant one) and white blood cells. In turn, the red blood cell group and white blood cells group were dominated by MCV and lymphocytes respectively. On the other hand, the biochemistry construct, which did not show significant effect on the $CD4^+$ cell count was dominated by LDH. All the biochemistry, red blood cell group and white blood cells clinical platforms, that were unmeasured variables, indicated that they were influenced by the effects of their previous time points. The cross-lagged path model further revealed that some measurements at different time points were highly correlated, either from the same variable, or from different covariates. Although there were chained effects in the complex relationships, the $CD4^+$ count indicated that it was also influenced by its own lagged effects, following the full blood count effect.

It is important to acknowledge that there were some limitations in the data. The baseline records before HIV infection were not available, limiting the opportunity to compare the covariate influence on the $CD4^+$ cell response, before and after the HIV infection. Potentially important confounders that could have been adjusted were also not available. These include dehydration, underlying infection, comorbidities and patient dietary conditions, especially their effect on the biochemistry covariates. The second short coming of this study is lack of development of influence diagnostics in the models adopted.

Evaluating the clinical covariates for patients with extremely low $CD4^+$ counts, is also recommended owing to the key driver for prophylaxis and surveillance for opportunistic infections related to $CD4^+$ count < 250cells/mm^3. Given big enough sample size for such records alone would suffice in a similar investigation. It is also important to note that the policy on ART initiation has changed recently, where medication is now started as soon as a patient is diagnosed with HIV. The strongest covariates identified in this study are based on a study design that delayed the medication process. Hence, one of the future areas of research is to identify the strongest covariates and their influential effects on the $CD4^+$ cell count, with data that take into account the context of the recent policies on ART initiation upon diagnosis.

The effectiveness of the GAMM can further be improved by incorporating the break-point detection of the segmented model. In both methods, the original scale of the covariates is preserved such that detecting the GAMM break-points can still provide meaningful biological interpretation. Instead of detecting the GAMM smooth turning points, the segmented model tended to pinpoint the break-points of the locally weighted scatterplot smoothers. Accordingly, the other future direction of research would be integrating GAMM with breakpoint analysis. The SEM approach modelled the linear relationships even though the GAMM and breakpoint analysis shows that the relationship between the covariates and $CD4^+$ count is nonlinear. This might be considered as another shortcoming of this study. Hence, one of the possible future direction of the study would be nonlinear SEM model.

References

ABBASTABAR, H., REZAIANZADEH A., RAJAEEFARD A., GHAEM H., MOTAMEDIFAR M. & KAZEROON P. A. (2016). "Determining Factors of Cd4 Cell Count in HIV Patients: In a Historical Cohort Study." *International Journal of Life Science and Pharma Research* **SP 2016**(1): 93-101.

ABDI, H. & WILLIAMS L. J. (2010). "Principal component analysis." *WIREs Comp Stat* **2**: 433-459.

ABDOLLAHI, A., SAFFAR H., SHOAR S. & JAFARI S. (2014). "Is total lymphocyte count a predictor for CD4 cell count in initiation antiretroviral therapy in HIV-infected patients? ." *Niger Med J.* **55**(4): 289–293.

ABE, M. (1999). "A Generalized Additive Model for Discrete Choice Data." *Journal of Business & Economic Statistics* **17**(3): 271-284.

ADHIKARI, P. M., CHOWTA M. N., RAMAPURAM J. T., RAO S. B., UDUPA K. & ACHARYA S. D. (2016). "Effect of Vitamin B12 and folic acid supplementation on neuropsychiatric symptoms and immune response in HIV-positive patients." *J Neurosci Rural Pract.* **7**(3): 362–367.

ADHIKARI, P. M. R., CHOWTA M. N., RAMAPURAM J. T., RAO S. B., UDUPA K. & ACHARYA S. D. (2016). "Prevalence of Vitamin B12 and folic acid deficiency in HIV-positive patients and its association with neuropsychiatric symptoms and immunological response." *Indian Journal of Sexually Transmitted Diseases and AIDS* **37**(2): 178-184.

ADRIAN, L., EDWARD J. C. & MICHELLE A. L. (2016). "Shaping Variation in the Human Immune System." *Trends in Immunology* **37**(10): 637-646.

AGGARWAL, V. & KOSIAN S. (2011). Feature Selection and Dimension Reduction Techniques in SAS. NESUG. Statistics & Analysis**:** 1-6.

AL-KINDI, S. G., KIM C. H., MORRIS S. R., FREEMAN M. L., FUNDERBURG N. T., RODRIGUEZ B., MCCOMSEY G. A., DALTON J. E., SIMON D. I., LEDERMAN M. M., LONGENECKER C. T. & ZIDAR D. A. (2017). "Elevated Red Cell Distribution Width (RDW) Identifies Elevated Cardiovascular Disease Risk in Patients with HIV infection " *J Acquir Immune Defic Syndr* **74**(3): 298–302.

ALAVI, S. M., AHMADI F. & FARHAD M. (2009). "Correlation between Total Lymphocyte Count, Hemoglobin, Hematocrit and CD4 Count in HIV/AIDS Patients." *Acta Medica Iranica* **47**(1): 1-4.

ALTMAN, D. G., LAUSEN B., SAUERBREI W. & SCHUMACHER M. (1994). "Dangers of using "Optimal" cutpoints in the evaluation of prognostic factors. ." *Journal of the National Cancer Institute* **86**: 829-835.

ALTMAN, D. G. & ROYSTON P. (2006). "The cost of dichotomising continuous variables." *BMJ* **332**: 1080.

ANDREWS, S. M. & ROWLAND-JONES S. (2017). "Recent advances in understanding HIV evolution[version 1; referees: 2 approved]." *F1000Research* **6(F1000 Faculty Rev)**(597).

ARBUCKLE, J. L. (2012). IBM®SPSS®Amos™ 21 User's Guide.

ARIKA, W., NYAMAI D., MUSILA M., NGUGI M. & NJAGI E. (2016). "Hematological Markers of In Vivo Toxicity." *Journal of Hematology & Thromboembolic Diseases* **4**(2).

ARYA, S. S. & KUMAR P. K. (2012). "Folate: sources , production and bioavailability." *Agro Food Industry Hi Tech* **23**(4): 23-27.

ATERE, A. D., AKINBO B. D., OKAFOR A. M.-J., EGBUCHULEM K. I. & AKINOLA E. A. (2016). "Evaluating correlation between total lymphocyte counts and CD4 counts in monitoring HIV patients." *Archives of Applied Science Research* **8**(3): 22-28.

BACON, D. & WATTS D. (1971). "Estimating the transition between two intersecting straight lines." *Biometrika* **58**: 525–534.

BAJPAI, S. & BAJPAI R. (2014). "Goodness of Measurement: Reliability and Validity." *International Journal of Medical Science and Public Health* **3**(1): 173-176.

BANESHI, M. R. & TALEI A. R. (2010). "Dichotomisation of Continuous Data- Revie w of Methods, Advantages and Disadvantages." *Iranian Journal of Cancer Prevention* **1**: 26-32.

BANI-SADR, F., LAPIDUS N., ROSENTHAL E., GERARD L., FOLTZER A., PERRONNE C., CACOUB P., POL S., CARRAT F. & TEAM A. H. R. S. (2009). "Gamma Glutamyl Transferase Elevation in HIV/Hepatitis C Virus–Coinfected Patients During Interferon–Ribavirin Combination Therapy." *J Acquir Immune Defic Syndr* **50**(4).

BAQUET, G., VAN PRAAGH E. & BERTHOIN S. (2003). "Endurance Training and Aerobic Fitness in Young People." *Sports Med* **33**(15): 1127-1143.

BATEMAN, C. (2013). "Drug stock-outs: Inept supply-chain management and corruption." *S Afr Med J* **103**(9): 600-602.

BATES, D. M., Ed. (2010). lme4:Mixed-Effects modeling with R, Springer.

BEARE, A., STOCKINGER H., ZOLA H. & NICHOLSON I. (2008). "The CD System of Leukocyte Surface molecules: Monoclonal Antibodies to Human Cell Surface Antigens." *Current Protocols in Immunology* **73**(80): A.4A.1-A.4A.73.

BEAUJEAN, A. A. (2012). Package 'BaylorEdPsych': R Package for Baylor University Educational Psychology Quantitative Courses. version 0.5. .

BECKMAN, R. & COOK R. (1979). "Testing for two-phase regressions." *Technometrics* **21**: 65–69.

BELLAN, S. E., DUSHOFF J., GALVANI A. P. & MEYERS L. A. (2015). "Reassessment of HIV-1 acute phase infectivity: accounting for heterogeneity and study design with simulated cohorts." *PLoS Med* **12**(3): e1001801.

BEN-GAL, I. (2005). Outlier detection. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. M. O. and R. L., Kluwer Academic Publishers.

BENTLER, P. & BONETT D. (1980). "Significance Tests and Goodness-of-Fit in the Analysis of Covariance Structures." *Psychological Bulletin* **88**: 588-606.

BENTLER, P. M. (1990). "Comparative fit indexes in structural models." *Psychological Bulletin* **107**: 238-246.

BENTWICH, Z. (2005). "CD4 Measurements in Patients with HIV: Are They Feasible for Poor Settings?" *PLoS Med* **2**(7): e214.

BERRY, D. & WILLOUGHBY M. T. (2017). "On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse." *Child Development* **88**(4): 1186–1206.

BIAN, H. (2011). "Structural Equation Modeling with AMOS II." Retrieved 09/02/2018, from https://www.slideshare.net/JordanSitorus1/sem-with-amos-ii.

BISWAS, P., MANTELLI B., SICA A., MALNATI M., PANZERI C., SACCANI A., HASSON H., VECCHI A., SANIABADI A., LUSSO P., LAZZARIN A. & BERETTA A. (2003) "Expression of CD4 on human peripheral blood neutrophils." Blood DOI: 10.1182/blood-2002-10-3056.

BLOOMFIELD, G. S., HOGAN J. W., KETER A., HOLLAND T. L., SANG E., KIMAIYO S. & VELAZQUEZ E. J. (2014). "Blood pressure level impacts risk of death among HIV seropositive adults in Kenya: a retrospective analysis of electronic health records." *BMC Infectious Diseases* **14**: 284.

BLUNCH, N. J. (2008). Introduction to Structural Equation Modelling Using SPSS and AMOS. Califonia, SAGE Publication Inc.

BOCK, R. D. (1989). Measurement of human variation: A two stage model. In R. D. Bock (Ed.), . New York, Academic Press.

BOLLEN, K. & LENNOX R. (1991). "Conventional wisdom in measurement- a structural equation perspective." *Pyschological bulletin* **110**(2): 305-314.

BOLLEN, K. A. (1989). Structural equations with latent variables. New York, Wiley.

BOLLEN, K. A. & LONG J. S. (1993). Introduction. Testing structural equation models. K. A. Bollen and J. S. Long. Newbury Park, CA, Sage.

BOLLEN, K. A. & LONG J. S., Eds. (1993). Testing structural equation models. Newbury Park, CA, Sage.

BOROVICKA, T., JIRINA JR. M., KORDIK P. & JIRINA M. (2012). Selecting Representative Data Sets. Advances in Data Mining Knowledge Discovery and Applications.

BORSBOOM, D., GIDEON J. & VAN HEERDEN J. (2003). "The Theoretical Status of Latent Variables." *Psychological Review* **110**(2): 203–219.

BORSBOOM, D., J. G. & VAN HEERDEN J. (2004). "The Concept of Validity." *Psychological Review* **111**(4): 1061–1071.

BRACONNIER, P., DELFORGE M., GARJAU M., WISSING K. M. & DE WIT S. (2017). "Hyponatremia is a marker of disease severity in HIV-infected patients: a retrospective cohort study." *Braconnier et al. BMC Infectious Diseases* **17**: 98.

BRENER, J., GALL A., HURST J., BATORSKY R., LAVANDIER N., CHEN F., EDWARDS A., BOLTON C., DSOUZA R., ALLEN T., PYBUS O. G., KELLAM P., MATTHEWS P. C. & GOULDER P. J. R. (2018). "Rapid HIV disease progression following superinfection in an HLA-B*27:05/B*57:01-positive transmission recipient." *Retrovirology* **15**(7): 1-13.

BROOK, M., AYLES H., HARRISON C., ROWNTREE C. & MILLER R. (1997). "Diagnostic utility of bone marrow sampling in HIV positive patients." *Genitourin Med* **73**: 117–121.

BROWN, T. A. (2006). Confirmatory Factor Analysis for Applied Research. . New York, The Guilford Press.

BROWNE, M. W. & CUDECK R. (1993). Alternative ways of assessing model fit. Newbury Park, CA, Sage.

BRUNO, R., SCUDERI D., LOCATELLI M., PAMPALONI A. & PINZONE M. (2017).

"Prevalence of micronutrients deficiencies in a cohort of HIV-positive individuals on ART." *Infect Dis Trop Med* **3**(4): E431.

BRYK, A. S. & RAUDENBUSH S. W. (1992). Hierarchical Linear Models. Newbury Park, CA., Sage.

BUNYASI, E. W. & COETZEE D. J. (2017). "Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa." *BMJ Open* **7**: e016232.

BURCH, L. S., SMITH C. J., ANDERSON J., SHERR L., RODGER A. J., O'CONNELL R., GERETTI A.-M., GILSON R., FISHER* M., ELFORD J., JONES M., COLLINS S., AZAD Y., PHILLIPS A. N., SPEAKMAN A., JOHNSON M. A., LAMPE F. C. F. T. A. & SEXUAL TRANSMISSION RISK AND ATTITUDES (ASTRA) STUDY GROUP (2016). "Socioeconomic status and treatment outcomes for individuals with HIV on antiretroviral treatment in the UK: cross-sectional and longitudinal analyses." *Lancet Public Health* **1**: e26–36.

BURNS, E. A., KORN K. & WHYTE IV J. (2011). Oxford American Handbook of Clinical Examination and Practical Skills. Oxford New York Oxford University Press Inc.

BUSHER, J. T. (1990). Clinical Methods: The History, Physical, and Laboratory Examinations. . Boston, Butterworths.

BUTT, A. A., MICHAELS S., GREER D., CLARK R., KISSINGER P. & MARTIN D. H. (2002). "Serum LDH Level as a Clue to the Diagnosis of Histoplasmosis." *The AIDS Read* **12**(7).

BUTT, A. A., MICHAELS S. & KISSINGER P. (2002). "The association of serum lactate dehydrogenase level with selected opportunistic infections and HIV progression." *International Journal of Infectious Diseases* **6**: 178-181.

BUZANOVSKII, V. A. (2017). "Determination of Proteins in Blood. Part 1: Determination of Total Protein and Albumin." *Review Journal of Chemistry* **7**(1): 79-124.

BYRNE, B. M. (2001). Structural equation modeling with AMOS, basic concepts, applications and programming LEA Publishers, NJ.

BYRNE, B. M. (2006). STRUCTURAL EQUATION MODELING WITH EQS Basic Concepts, Applications, and Programming.

CADE, W. T., FANTRY L. E., NABAR S. R. & KEYSER R. E. (2003). "Decreased Peak Arteriovenous Oxygen Difference During Treadmill Exercise Testing in Individuals Infected With the Human Immunodeficiency Virus." *Arch Phys Med Rehabil* **84**: 1595-1603.

CAMPBELL JR , E. W. & LYNN C. K. (1990). The Physical Examination. Clinical Methods: The History, Physical, and Laboratory Examinations**:** 37-39.

CANADIAN CANCER SOCIETY. (2017). "Blood chemistry tests." Retrieved 13/10/2017, from http://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/blood-chemistry-tests/?region=on.

CASO, J. A. A., MINGO S. C. & TENA J. G. (1999). "Effect of Highly Active Antiretroviral Therapy on Thrombocytopenia in Patients with HIV Infection." *The New England Journal of Medicine* **341**: 1239-1240.

CHEN, C. (2000). "Generalized additive mixed models." *Communications in Statistics - Theory and Methods* **29**(5-6): 1257-1271.

CHEN, C. & MU LIU L. (2015). "Joint Estimation of Model Parameters and Outlier Effects in Time Series." *Journal of the American Statistical Association, Taylor & Francis, Ltd. and American Statistical Association* **88**(421): 284-297.

CHEN, W.-T., SHIU C.-S., YANG J. P., SIMONI J. M., FREDRIKSEN-GOLDSEN K. I., SZU-HSIEN LEE T. & ZHAO H. (2013). "Side Effects of Antiretroviral Therapy (ART) Are Associated with Depression in Chinese Individuals with HIV: A Mixed Methods Study." *AIDS & Clinical Research* **4**(6).

CHHABRA, N. (2013). Structure of Hemoglobin-An Overview. Biochemistry for Medics.

CHISATI, E. M. & VASSELJEN O. (2015). "Aerobic endurance in HIV-positive young adults and HIV-negative controls in Malawi." *Malawi Medical Journal* **27**(1): 5-9.

CHO, H. & QU A. (2013). "Model selection for correlated data with diverging number of parameters." *Statistica Sinica* **23**: 901–927.

CHOI, B., GATTI P. J., HAISLIP A. M., FERMIN C. D. & GARRY R. F. (1998). "Role of Potassium in Human Immunodeficiency Virus Production and Cytopathic Effects." *VIROLOGY* **247**: 189-199.

CHONG, I.-G. & JUN C.-H. (2005). "Performance of some variable selection methods when multicollinearity is present." *Chemometrics and Intelligent Laboratory Systems* **78**: 103-112.

CHONG, J. J., FRAGASZY E., DUKES O., CASON J. & KOZLAKIDIS Z. (2015). "Serum Albumin Concentrations in a Multi-Ethnic Cohort of Patients with Human Immunodeficiency Virus Infection from

South East London." *BioResearch Open Access* **4**(1): 160-163.

CHORBA, T. L., MAURICE C., NKENGASONG J., MARAN M., ROELS T. H. & DJOMAND G. (2002). "Assessing Eosinophil Count as a Marker of Immune Activation among Human Immunodeficiency Virus–Infected Persons in Sub-Saharan Africa." *Clinical Infectious Diseases 2002;34:* **34**: 1264–1266.

CHOU, C.-P. & BENTLER P. M. (1995). Estimates and tests in structural equation modeling. Structural equation modeling: Concepts, issues, and applications R. H. Hoyle. Thousand Oaks, CA, Sage.

CHU, W., LI R. & REIMHERR M. (2016). "Feature Screening for Time-Varying Coefficient Models with Ultrahigh-Dimensional Longitudinal Data." *The Annals of Applied Statistics* **10**(2): 596–617.

CHUN, H. & KELES S. (2010). "Sparse partial least squares regression for simultaneous dimension reduction and variable selection." *Jornal of the Royal Statistical. Society. B* **72**(1): 3-25.

CLARK, M. (2014). Getting Started with Additive Models in R.

CLOYD, M. W. & LYNN W. S. (1991). "Perturbation of host-cell membrane is a primary mechanism of HIV cytopathology." *Virology* **181**: 500±511.

CMA MEDIA INC. (2005). "Inexpensive CD4 counting for the developing world." *JAMC* **175**(3): 478.

COELHO, R., INFANTE P. & SANTOS M. N. (2013). "Application of Generalized Linear Models and Generalized Estimation Equations to model at-haulback mortality of blue sharks captured in a pelagic longline fishery in the Atlantic Ocean." *Fisheries Research* **145**: 66-75.

COFFMAN, D. L. & MACCALLUM R. C. (2005). "Using Parcels to Convert Path Analysis Models Into latent Variable Models." *MULTIVARIATE BEHAVIORAL RESEARCH* **40**(2): 235–259.

COHEN, A. J. & STEIGBIGEL R. T. (1996). "Eosinophilia in Patients Infected with Human Immunodeficiency Virus." *The Journal of Infectious Diseases* **174**: 615-618.

COLLINS, A. J., PITT B., MCGAUGHEY K., REAVEN N., WILSON D., FUNK S. & BUSHINSKY D. A. (2017). "Association of Serum Potassium with All-Cause Mortality in Patients with and without Heart Failure, Chronic Kidney Disease, and/or Diabetes." *American Journal of Nephrology* **46**: 213–221.

COLTMAN, T., DEVINNEY T. M., MIDGLEY D. F. & VENIAK S. (2008). "Formative versus reflective measurement models- Two applications of formative measurement." *Journal of Business Research* **61**(12): 1250-1262.

COX, C. (1987). "Threshold dose-response models in toxicology." *Biometrics* **43**: 511–523.

COX , I. & GAUDARD M. (2013). Discovering Partial Least Squares with JMP. Cary, North Carolina, USA, SAS Institute, Inc.

CUEVAS, J. M., GELLER R., GARIJO R., LÓPEZ-ALDEGUER J. & SANJUÁN R. (2015). "Extremely High Mutation Rate of HIV-1 In Vivo." *PLoS Biology* **13**(9).

CUZZOCREA, A. & ZALL D. (2013). Parallel Coordinates Technique in Visual Data Mining-Advantages, Disadvantages, and Combinations. 2013 17th International Conference on Information Visualisation.

CYRINO, E. S., OKANO A. H., GLANER M. F., ROMANZINI M., GOBBO L. A., MAKOSKI A., BRUNA N., CORDEIRO DE MELO J. & TASSI G. N. (2003). "Impact of the use of different skinfold calipers for the analysis of the body composition " *Bras Med Esporte* **9**(3).

DAKA, D. & LOHA E. (2008). "Relationship between Total Lymphocyte count (TLC) and CD4 count among peoples living with HIV, Southern Ethiopia: a retrospective evaluation." *AIDS Research and Therapy* **5** 26.

DANNHAUSER, A., VAN STADEN A., VAN DER RYST E., NEL M., MARAIS N., ERASMUS E., ATTWOOD E., BARNARD H. & LE ROUX G. (1999). "Ntritional status of HIV-1 seropositive patients in the Free State Province of South Africa: Anthropometric and dietary profile." *European Journal of Clinical Nutrition* **53**: 165-173.

DASGUPTA, A. & KOSARA R. (2010). Pargnostics-Screen-Space Metrics for Parallel Coordinates.

DAVENPORT, T. H., AND DYCHÉ, J (2013). Big Data in Big Companies. International Institute for Analytics. Thomas H. Davenport and SAS Institute Inc.

DAVID, F. L. & JORDAN S. O. (2008). "Lymphocytes." *J ALLERGY CLIN IMMUNOL* **121**(2): S364-S369.

DE LEEUW, J. & KREFT I. (1986). "Random Coefficient Models for Multilevel Analysis." *Journal of Educational and Behavioral Statistics* **11**: 57–85.

DE SANTIS, G. C., BRUNETTA D. M., VILAR F. C., BRANDA R. A., MUNIZ R. Z. D. A., DE LIMA G. M. N., AMORELLI-CHACEL M. E., COVAS D. T. & MACHADO A. A. (2011). "Hematological abnormalities in HIV-infected patients." *International journal of infectious diseases* **15** e808–e811.

DEGENHARDT, F., SEIFERT S. & SZYMCZAK S. (2019). "Evaluation of variable selection methods for random forests and omics data sets." *Briefings in Bioinformatics* **20**(2): 492–503.

DEMCHENKO, Y., NGO, C. AND MEMBREY, P. (2013). Architecture Framework and Components for the Big Data Ecosystem Draft U. V. AMSTERDAM.

DEMPSTER, A. P., RUBIN D. B. & TSUTAKAWA R. K. (1981). "Estimation in Covariance Components Models." *Journal of the American Statistical Association* **76**(374): 341-353.

DIETICIANS OF CANADA (2014). Food Sources of Folate. Canadian Nutrient File Health Department. Canada, Dieticians of Canada.

DIMALA, C. A., KADIA B. M., KEMAH B.-L., TINDONG M. & CHOUKEM S.-P. (2018). "Association between CD4 Cell Count and Blood Pressure and Its Variation with Body Mass Index Categories in HIV-Infected Patients." *International Journal of Hypertension* **Volume 2018, Article ID 1691474**.

DIXON, W. J. (1950). "Analysis of extreme values. ." *Annals of Mathematical Statistics* **21**: 488-506.

DONG, Y. & PENG C.-Y. J. (2013). "Principled missing data methods for researchers." *SpringerPlus* **2**(222).

DOS SANTOS, A. C. O. & ALMEIDA A. M. R. (2013). "Nutritional status and CD4 cell counts in patients with HIV/AIDS receiving antiretroviral therapy." *Revista da Sociedade Brasileira de Medicina Tropical* **46**(6): 698-703.

DUSINGIZE, J. C., HOOVER D. R., SHI Q., MUTIMURA E., RUDAKEMWA E., NDACYAYISENGA V., GAKINDI L. O., MULVIHILL M., SINAYOBYE J. D. A., MUSABEYEZU E. & ANASTOS K. (2015). "Association of Abnormal Liver Function Parameters with HIV Serostatus and CD4 Count in Antiretroviral-Naive Rwandan Women." *AIDS RESEARCH AND HUMAN RETROVIRUSES* **31**(7): 723-730.

EDWARDS, J. R. & BAGOZZI R. P. (2000). "On the Nature and Direction of Relationships between Constructs and Measures." *Psychological Methods* **5**(2): 155-174.

EHOLIÉ, S. P., BADJE A., KOUAME G. M., N'TAKPE J.-B., MOH R., DANEL C. & ANGLARET X. (2016). "Antiretroviral treatment regardless of CD4 count: the universal answer to a contextual question." *AIDS Research and Therapy* **13**(27).

ELSA, Z. (2016). "Healthcare Systems in Sub-Saharan Africa: Focusing on community-based delivery (CBD) of health services and the development of local research institutes." *United Nations Peace and Progress* **3** (1): 44-49.

EMUCHAY, C., OKENIYI S. & OKENIYI J. (2014). "Correlation between total lymphocyte count, hemoglobin, hematocrit and CD4 count in HIV patients in Nigeria." *Pak J Biol Sci.* **17**(4): 570-573.

ENDERS, C. K. (2010). Applied Missing Data Analysis. , The Guilford Press, New York, NY.

ERTEL, J. & FOWLKES E. (1976). "Some algorithms for linear spline and piecewise multiple linear regression." *Journal of the American Statistical Association* **71**: 640–648.

ESPOSITO, F. M., COUTSOUDIS A., VISSER J. & KINDRA G. (2008). "Changes in Body Composition and Other Anthropometric Measures of Female Subjects on Highly Active Antiretroviral Therapy (HAART): A Pilot Study in Kwazulu-Natal, South Africa." *The Southern African Journal of HIV Medicine* **9**(4): 36-42.

EVERITT, B. S., AND DUNN, G. (2001). Applied Multivariate Data Analysis, John Wiley & Sons, Ltd.

FAN, J. & LI R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* **96**(456).

FAN, J. & LI R. (2004). "New estimation and model selection procedures for semi-parametric modelling in longitudinal data analysis." *Journal of the American Statistical Association* **99**: 710-723.

FAN, J. & LV J. (2010). "A Selective Overview of Variable Selection in High Dimensional Feature Space." *Statistica Sinica* **20**: 101-148.

FAN, J., MAITY A., WANG Y. & WU Y. (2013). "Parametrically Guided Generalized Additive Models with Application to Mergers and Acquisitions Data." *J Nonparametr Stat.* **25**(1): 109–128.

FARAWAY, J. J. (2006). Extending the Linear Model with R. Generalised Linear, Mixed Effects and Nonparametric Regression Models. , Chapman & Hall/CRC.

FASAKIN , K., OMISAKIN C., ESAN A., ADEBARA I., OWOSENI I., OMONIYI D., AJAYI O., OGUNDARE R. & MORONKEJI M. (2014). "Total and CD4+ T- lymphocyte count correlation in newly diagnosed HIV patients in resource-limited setting." *Journal of Medical Laboratory and Diagnosis* **5**(2): 22-28.

FEDER, P. (1975). "The log likelihood ratio in segmented regression." *Annals of Statistics* **3**(1): 84–97.

FEW, S. (2006). Multivariate Analysis Using Parallel Coordinates.

FISSEHA, G. G. & ONANA C. A. (2017). "High Dimensional Data Visualization: Advances and Challenges." *International Journal of Computer Applications (0975 – 8887)* **162**(10).

FLEISCHBEINA, E., O'BRIENB J., MARTELINOC R. & FENSTERSHEIBD M. (2008). "Elevated alkaline phosphatase with raltegravir in a treatment experienced HIV patient." *AIDS* **22**(17): 2401–2407.

FLORIS-MOORE, M., HOWARD A., LO Y., ARNSTEN J., SANTORO N. & SCHOENBAUM E. (2006). "Increased serum lipids are associated with higher CD4 lymphocyte count in HIV-infected women." *HIV Medicine* **7**: 421–430.

FOFANA, K. C. (2016). Correlation Between Nutritional Indicators and Low CD4 Count (<200 cells>/mm3) among HIV Positive Adults in Kapiri, Zambia Master of Public Health Thesis, Georgia State University.

GALLEGO, M. L., PÉREZ-HERNÁNDEZ I. A., PALACIOS R., RUIZ-MORALES J., NUÑO E., MÁRQUEZ M. & SANTOS J. (2012). "Red cell distribution width in patients with HIV infection." *Open Journal of Internal Medicine* **2**: 7-10.

GARTHWAI, P. H. (1994). "An lnterpretation of Partial Least Squares." *American Statistical Association* **89**(425): 122-127.

GELADI, P. & KOWALSKI B. R. (1986). "Partial Least-Squares Regression: A Tutorial." *Analytica Chimica Acta* **185**: 1-17.

GELMAN, A. & HILL J. (2007). Data Analysis Using Regression and Multi-level/Hierarchical Models. United States of America, Cambridge: Cambridge University Press. New York.

GHOSH, D. & VOGT A. (2012). Outliers: An Evaluation of Methodologies. Joint Statistical Meetings. Section on Survey Research Methods

GIBBONS, R. D., HEDEKER D., WATERNAUX C. M. & DAVIS J. M. (1988). "Random regression models: A comprehensive approach to the analysis of longitudinal psychiatric data." *Psychopharmacology Bulletin* **24**: 438–443.

GILBERT, A. & CHURCHIL J. (1979). "A paradigm for developing better measures of marketing constructs." *Journal of Marketing Research* **XVI**: 64-73.

GOLDSTEIN, H. (1995). Multilevel statistical models. New York, Halstead.

GOMO, E., NDHLOVU P., VENNERVALD B., NYAZEMA N. & FRIIS H. (2001). "Enumeration of CD4 and CD8 T-cells in HIV infection in Zimbabwe using a manual immunocytochemical method." *Cent Afr J Med.* **47**(3): 64-70.

GOSSL, C. & KUCHENHOFF H. (2001). "Bayesian analysis of logistic regression with an unknown change point and covariate measurement error. ." *Statistics in Medicine* **20**: 3109–3121.

GOUD, T. G. & RAMESH K. (2014). "Opportunistic infections among HIV patients attending Tertiary Carehospital, Karnataka, India." *Int. J. Curr. Microbiol. App. Sci* **3**(4): 824-829.

GRAHAM, J. W. (2009). "Missing Data Analysis: Making It Work in the Real World." *The Annual Review of Psychology* **60**: 549-576.

GREGG, X. (2011). "Parallel coordinates in JMP " Retrieved 07/10/2017, from https://community.jmp.com/t5/JMP-Blog/Parallel-coordinates-in-JMP/ba-p/31024.

GRIFFITHS, D. & MILLER A. (1973). "Hyperbolic regression – a model based on two-phase piecewise linear regression with a smooth transition between regimens." *Communications in Statistics* **2**: 561–569.

GRÖBER, U., SCHMIDT J. & KISTERS K. (2015). "Magnesium in Prevention and Therapy." *Nutrients* **7**: 8199-8226.

GUISAN, A., EDWARDS T. & HASTIE T. (2002). "Generalized linear and generalized additive models in studies of species distributions: Setting the scene." *Ecological Modelling* **157**: 89-100

HAIR, J. F., BLACK W. C., BABIN B. J. & ANDERSON R. E. (2014). Multivariate Data Analysis, Pearson Education Limited.

HAMAKER, E. L., KUIPER R. M. & GRASMAN R. P. P. P. (2015). "A critique of the cross-lagged panel model." *Psychological Methods* **20**: 102–116.

HARDIN, J. W. & HILBE J. M. (2003). Generalized Estimating Equations., Wiley Online Library.

HASTIE, T. & TIBSHIRANI R. (1986). "Generalized additive models." *Statistical Science* **1**(3): 297-318.

HASTIE, T. & TIBSHIRANI R. (1990). Generalized Additive Models. . London, Chapman & Hall.

HAUSER, R. M. & GOLDBERGER A. S. (1971). "The treatment of unobservable variables in path analysis. ." *Sociological Methodology* **3**: 81-117. .

HAWKINS, D. (1976). "Point estimation of the parameters of piecewise regression models." *Applied Statistics* **25**: 51–57.

HAWKINS, D. M. (2004). "The Problem of Overfitting." *Journal of Chemical Information Computer Science* **44**(1): 1-12.

HE, F., TEIXEIRA-PINTO A. & HAREZLAK J. (2017). "Autoregressive and Cross-lagged Model for Bivariate Non-commensurate Outcomes." *Stat. Med.* **36**(19): 3110–3120.

HEART FOUNDATION (2016). Blood pressure. National Heart Foundation of Australia. Australia.

HEDEKER, D. (2004). An introduction to growth modeling. The SAGE Handbook of Quantitative Methodology for the Social Sciences. D. Kaplan. Thousand Oaks CA, Sage.

HEDEKER, D. & GIBBONS R. D. (2006). Longitudinal Data Analysis Wiley.

HEINRICH, J. (2013). Visualization Techniques for Parallel Coordinates.

HEINRICH, J. & WEISKOPF D. (2013). "State of the Art of Parallel Coordinates."

HERSH, W. R. (2014). Healthcare Data Analytics. Health Informatics: Practical Guide for Healthcare and Information Technology Professionals. R. E. Hoyt, and Yoshihashi, A. , Pensacola, FL, Lulu.com.

HOENIGL, M., GREEN N., CAMACHO M., GIANELLA S., MEHTA S. R., SMITH D. M. & LITTLE S. J. (2016). "Signs or Symptoms of Acute HIV Infection in a Cohort Undergoing Community-Based Screening." *Emerging Infectious Diseases* **22**(3): 532-534.

HOLMES-SMITH, P. (2000). Introduction to Structural Equation Modelling. ACSPRI 2000.

HORN, P. & PESCE A. (2003). "Reference intervals: an update." *Clinica Chimica Acta* **334**(1-2): 5-23.

HOWARD, M. & HAMILTON P. (2002). Haematology. Blood cells

HOWELL, D. C. (2008). The Treatment of Missing Data. The analysis of missing data, Handbook of Social Science Methodology. London: Sage.

HOYLE, R. H. (2000). Confirmatory Facator Analysis. Handbook of Applied Multivariate Statistics and Mathematical Modeling, Academic Press.

HSUE, P. Y., HUNT P. W., HO J. E., FARAH H. H., SCHNELL A., HOH R., MARTIN J. N., DEEKS S. G. & BOLGER A. F. (2010). "Impact of HIV Infection on Diastolic Function and Left Ventricular Mass." *Circ Heart Fail.* **3**(1): 132–139.

HU, L.-T., BENTLER P. M. & KANO Y. (1992). "Can test statistics in covariance structure analysis be trusted? ." *Psychological Bulletin* **112**: 351-362.

HUI, S. L. & BERGER J. O. (1983). "Empirical Bayes estimation of rates in longitudinal studies." *Journal of the American Statistical Association* **78**: 753–759.

HUNT, P. W., DEEKS S. G., RODRIGUEZ B., VALDEZ H., SHADE S. B., ABRAMS D. I., KITAHATA M. M., KRONE M., NEILANDS T. B., BRAND R. J., LEDERMAN M. M. & MARTIN J. N. (2003). "Continued CD4 cell count increases in HIV-infected adults experiencing 4 years of viral suppression on antiretroviral therapy." *AIDS* **17**: 1907–1915.

HUTA, V. (2014). "When to Use Hierarchical Linear Modeling." *The Quantitative Methods for Psychology* **10**(1): 13-28.

IBEH, B. O., OMODAMIRO O. D., IBEH U. & HABU J. B. (2013). "Biochemical and haematological changes in HIV subjects receiving winnicure antiretroviral drug in Nigeria." *Journal of Biomedical Science* **20**(1): 73.

IFFEN, T. S., EFOBI H., USORO C. A. O. & UDONWA N. E. (2010). "Lipid Profile of HIV-Positive Patients Attending University of Calabar Teaching Hospital, Calabar - Nigeria." *World Journal of Medical Sciences* **5**(4): 89-93.

INAN, G. & WANG L. (2017). "PGEE: An R Package for Analysis of Longitudinal Data with High-Dimensional Covariates." *The R Journal* **9**(1): 393-402.

INSELBERG, A. (1997). Mutidimensional detective.

JAMES, A. G. (2017). Understanding Blood Tests. The James. Cancer Hospital and Richard J. Solove Research Institute.

JARVIS, C. B., MACKENZIE S. B. & PODSAKOFF P. M. (2003). "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research." *Journal of Consumer Research* **30**(2): 199-218.

JEFFERY, A. (2016). The NHI Proposal Risking Lives For No Good Reason. South African Institute of Race Relations. South Africa, South African Institute of Race Relations. **6**.

JEMIKALAJAH, J. & ADU M. (2015). "Assessment Of Serum Proteins In Human Immunodeficiency Virus Patients In Auchi, Nigeria." *African Journal of Cellular Pathology* **5**: 14-17.

JENSEN, F. B., FAGO A. & WEBER R. E. (1998). "Hemoglobin structure and function." *Fish Physiology* **17**: 1-40.

JIANG, A.-P., JIANG J.-F., GUO M.-G., JIN Y.-M., LI Y.-Y. & WANGA J.-H. (2015). "Human Blood-Circulating Basophils Capture HIV-1 and Mediate Viral trans-Infection of CD4+ T Cells." *J Virol* **89**: 8050-8062.

JOHNSON, B. A., LIN D. & ZENG D. (2008). "Penalized estimating functions and variable selection in semipara-metric regression

models." *Journal of the American Statistical Association* **103**(482): 672–680.
JOHNSON, R. A. & WICHERM D. W. (2007). Applied Multivariate Statistical Analysis, Pearson Education Inc.
JOLLIFE, I. T. (1986). Principal Component Analysis, Springer-Verlag New York Inc.
JÖRESKOG, K. G. (1969). "A general approach to confirmatory maximum likelihood factor analysis. ." *Psychometrika* **34**(2): 183-202.
JÖRESKOG, K. G. (1993). Testing structural equation models. Newbury Park, CA: Sage.
JÖRESKOG, K. G. (1994). "On the estimation of polychoric correlations and their asymptotic covariance matrix." *Psychometrika* **59**: 381-389.
JULIOUS, S. (2001). "Inference and estimation in a changepoint regression problem." *Statistician* **50**: 51–61.
JUN, C.-H., LEE S.-H., PARK H.-S. & LEE J.-H. (2009). "Use of Partial Least Squares Regression for Variable Selection and Quality Prediction." *International Conference on Computers & Industrial Engineering*: 1302--1307.
JUNG, T. & WICKRAMA K. A. S. (2008). "An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling." *Social and Personality Psychology Compass* **2**(1): 302–317.
JUNQUEIRA, L., CARNEIRO J. & KELLEY R. (2006). Basic Histology. A Lange Medical Book, Appleton and Lange.
KEARNEY, M. W. (2017). Cross Lagged Panel Analysis. (in press). The SAGE Encyclopedia of Communication Research Methods. M. R. Allen. Thousand Oaks, CA, Sage.
KELLY, A. U., MCSORLEY S. T., PATEL P. & TALWAR D. (2017). "Interpreting iron studies." *BMJ* **357**: j2513.
KELLY, G. (2006). "Body Temperature Variability (Part 1): A Review of the History of Body Temperature and its Variability Due to Site Selection, Biological Rhythms, Fitness, and Aging." *Alternative Medicine Review* **11**(4): 278-293.
KHAIDUKOV, S. V. & LITVINOV I. S. (2005). "Calcium Homeostasis Change in CD4+ T Lymphocytes from Human Peripheral Blood during Differentiation in vivo." *Biochemistry (Moscow)* **70**(6): 692-702.
KIMBERGER, O. & QUAST S. (2013). The significance of core temperature – Pathophysiology and measurement methods. Germany, Drägerwerk AG & Co. KGaA.
KLEIN, S. & ZION M. (2015). "The characteristics of homeostasis: a new perspective on teaching a fundamental principle in biology." *SSR* **97**(358): 85-93.
KLINE, P. (1994). An Easy Guide to Factor Analysis, Routledge.
KLINE, R. B. (2011). Principles and Practice of Structural Equation Modeling.
KRATZ, A., FERRARO M., SLUSS P. M. & LEWANDROWSKI K. B. (2004). "Laboratory Reference Values." *The New England Journal of Medicine* **351**: 1548-1563.
KUCHENHOFF, H. (1997). "An exact algorithm for estimating breakpoints in segmented generalized linear models." *Computational Statistics* **12**: 235–247.
KUCHENHOFF, H. & ULM K. (1997). "Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology." *Computational Statistics* **12**(2): 249–264.
KULKARNI, M. B., BHALERAO M. M., MUNGAL S. U. & DUBE S. P. (2015). "Anemia in People Living With HIV/AIDS: A Cross Sectional Study from India." *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)* **14**(2): 04-08.
KUNST, A., LOOMAN C. & MACKENBACH J. (1993). "Outdoor air temperature and mortality in the Netherlands: a time series analysis." *American Journal of Epidemiology* **137** 331–341.
KUUPIEL, D., BAWONTUO V. & MASHAMBA-THOMPSON T. P. (2017). "Improving the Accessibility and Efficiency of Point-of-Care Diagnostics Services in Low-and Middle-Income Countries: Lean and Agile Supply Chain Management." *Diagnostics* **7**(58).
KWANTWI, L. B., TUNU B. K., BOATENG D. & QUANSAH D. Y. (2017). "Body Mass Index, Haemoglobin, and Total Lymphocyte Count as a Surrogate for CD4 Count in Resource Limited Settings." *Journal of Biomarkers* **Volume 2017, Article ID 7907352**.
LAB CHIP (2017). "Rapid, label-free CD4 testing using a smartphone compatible device." *The Royal Society of Chemistry* **17**: 2910–2919
LAIRD, N. M. (1988). "Missing data in longitudinal studies." *Statistics in Medicine* **7**: 305–315.
LAIRD, N. M. & WARE J. H. (1982). "Random-Effects Models for Longitudinal Data." *Biometrics* **38**(4): 963-974.
LANDS, L. (2006). A Practical Guide to HIV Drug Side Effects for People Living with HIV/AIDS. R. Pustil, The Canadian AIDS Treatment Information Exchange (CATIE).
LAWRIE, D., COETZEE L. M., BECKER P., MAHLANGU J., STEVENS W. & GLENCROSS D. K. (2009). "Local reference ranges for full blood count and CD4

lymphocyte count testing." *S Afr Med J* **99**: 243-248.

LÊ CAO, K. A., ROSSOUW D., ROBERT-GRANIE C. & BESSE P. (2008). "A sparse PLS for variable selection when integrating Omics data." *Statistical Applications in Genetics and Molecular Biology* **7**(1): 35.

LETICIA, O. I., UGOCHUKWU A., IFEANYI O. E., ANDREW A. & IFEOMA U. E. (2014). "The Correlation of Values of CD4 Count, Platelet, Pt, Aptt, Fibrinogen and Factor VIII Concentrations among HIV Positive Patients in FMC Owerri." *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)* **13**(9 Ver II): 94-101.

LEUNG, N.-H. Z., CHEN A., YADAV P. & GALLIEN J. (2016). "The Impact of Inventory Management on Stock-Outs of Essential Drugs in Sub-Saharan Africa: Secondary Analysis of a Field Experiment in Zambia." *PLoS One* **11**(15): e0156026.

LIANG, K.-Y. & ZEGER S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika* **73**(1): 13–22.

LIN, X. & ZHANG D. (1999). "Inference in generalized additive mixed modelsby using smoothing splines." *J. R. Statist. Soc. B* **61**(2): 381-400.

LIPPI, G. & PLEBANI M. (2014). "Red blood cell distribution width (RDW) and human pathology. One size fits all." *Clinical Chemistry and Laboratory Medicine* **52**(9): 1247–1249.

LIQUET, B., LÊ CAO K.-A., HOCINI H. & THIEBAUT R. (2012). "A novel approach for biomarker selection and the integration of repeated measures experiments from two platforms." *BMC Bioinformatics* **13**: 325.

LIQUET, B. T., DE MICHEAUX P. L., HEJBLUM B. P. & THIE´BAUT R. (2016). "Group and sparse group partial least square approaches applied in genomics context." *Bioinformatics* **32**(1): 35-42.

LITTLE, R. J. A. (1988). "A Test of Missing Completely at Random for Multivariate Data With Missing Values." *Joumal of the American Statistical Association*, **83**(404): 1198-1202.

LIU, Y. (2012). Performing response surface analysis using the SAS RSREG procedure. MidWest SAS® Users Group (MWSUG), Minneapolis, MN.

LONG, T. V. (2009). Visualizing High-density Clusters in Multidimensional Data.

LONGFORD, N. T. (1987). "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced MixedModels with Nested Random Effects." *Biometrika* **74**(4): 817-827.

LONGFORG, N. (1993). Random coefficient models Clarendon Press, Oxford.

LORENTE, N., FERNÀNDEZ-LÓPEZ L., FUERTES R., CASTRO D. R., PICHON F., CIGAN B., CHANOS S., MEIRELES P., LUCAS R., MOREL S., KAYE P. S., AGUSTÍ C., KLAVS I., PLATTEAU T., CASABONA J. & THE EURO HIV EDAT STUDY GROUP (2016). "COBA-Cohort: a prospective cohort of HIV-negative men who have sex with men, attending community-based HIV testing services in five European countries (a study protocol)." *BMJ Open* **6**: e011314.

LUMBANRAJA, S. & SIREGAR D. (2018). "Association between red blood cell indices and CD4 count in HIV-positive reproductive women." *IOP Conference Series: Earth and Environmental Science* **125**(012027).

MACCALLUM, R. C. (1995). Model specification: Procedures, strategies, and related issues. . Structural equation modeling: Concepts, issues, and applications R. H. Hoyle. Thousand Oaks, CA, Sage**:** 16-36.

MACCALLUM, R. C., BROWNE M. W. & SUGAWARA H. M. (1996). "Power analysis and determination of sample size for covariance structure modeling " *Psychological Methods* **1**: 130-149.

MACCALLUM, R. C., ROZNOWSKI M. & NECOWITZ L. B. (1992). "Model modifications in covariance structure analysis: The problem of capitalization on chance." *Psychological Bulletin* **111**: 490-504.

MACKENZIE, S. B., PODSAKOFF P. M. & JARVIS C. B. (2005). "The Problem of Measurement Model Misspecification in Behavioral and Organisational Research and Some Recommended Solutions." *Journal of Applied Psychology* **90**(4): 710-730.

MAGANGA, E., SMART L. R., KALLUVYA S., KATARAIHYA J. B., SALEH A. M., OBEID L., DOWNS J. A., FITZGERALD D. W. & PECK R. N. (2015). "Glucose Metabolism Disorders, HIV and Antiretroviral Therapy among Tanzanian Adults." *PLoS ONE* **10**(8:e0134410).

MAITRA, S. & YAN J. (2008). Principal component analysis and partial least squares: two dimension deduction techniques for regression. Casualty Actuarial Society. Discussion Paper Program 79-90.

MANABE, Y. C., WANG Y., ELBIREER A., AUERBACH B. & CASTELNUOVO B. (2012). "Evaluation of Portable Point-of-Care CD4 Counter with High Sensitivity for Detecting Patients Eligible for Antiretroviral " *PLoS ONE* **7**(4).

MANNER, I. W., TRØSEID M., OEKTEDALEN O., BAEKKEN M. & OS I. (2013). "Low Nadir CD4 Cell Count Predicts Sustained Hypertension in HIV-Infected Individuals." *The Journal of Clinical Hypertension* **15**(2): 101-106.

MANOTO, S. L., LUGONGOLO M., GOVENDER U. & MTHUNZI-KUFA P. (2018). "Point of Care Diagnostics for HIV in Resource Limited Settings: An Overview." *MDPI* **54**(3).

MARONE, G., VARRICCHI G., GALDIERO M. R., LOFFREDO S., RIVELLESE F. & DE PAULIS A. (2016). "Are Basophils and Mast Cells Masters in HIV Infection?" *International Archives of Allergy and Immunology* **171**: 158–165.

MARSH, H. W. & HOCEVAR D. (1985). "Application of confirmatory factor analysis to the study of self-concept: First- and higher-order factor models and their invariance across groups." *Psychological Bulletin* **97**: 562-582.

MCCULLAGH, P. & NELDER J. (1989). Generalized Linear Models London, Chapman & Hall.

MCKNIGHT, P. E., MCKNIGHT K. M., SIDANI S. & FIGUEREDO A. J. (2007). Missing data: A gentle introduction, Guilford Press.

MCKNIGHT, T. R., YOSHIHARA H. A. I., SITOLE L. J., MARTIN J. N., STEFFENSD F. & MEYER D. (2014). "A combined chemometric and quantitative NMR analysis of HIV/AIDS serum discloses metabolic alterations associated with disease status." *Mol. BioSyst.* **10**: 2889-2897.

MEINTJES, G., MOORHOUSE M. A., CARMONA S., DAVIES N., DLAMINI S., VAN VUUREN C., MANZINI T., MATHE M., MOOSA Y., NASH J., NEL J., PAKADE Y., WOODS J., VAN ZYL G., CONRADIE F. & VENTER F. (2017). "Adult antiretroviral therapy guidelines 2017." *Southern African Journal of HIV Medicine* **18**(1): a776.

MEVIK, B.-H. & CEDERKVIST H. R. (2004). "Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR)." *JOURNAL OF CHEMOMETRICS* **18**: 422–429.

MGOGWE, J., SEMVUA H., MSANGI R., MATARO C., KAJEGUKA D. & CHILONGOLA J. (2012). "The evolution of haematological and biochemical indices in HIV patients during a six-month treatment period." *African Health Sciences* **12**(1): 2-7.

MICCERI, T. (1989). "The unicorn, the normal curve, and other improbable creatures." *Psychological Bulletin* **105**: 156-166.

MIN, B., BROWN M. A. & LEGROS G. (2011). "Understanding the roles of basophils: breaking dawn." *The Journal of cells, molecules, systems and technologies* **135**: 192–197.

MISRA, R., CHANDRA P., RIECHMAN S. E., LONG D. M., SHINDE S., POWNALL H. J., CORAZA I., LEWIS D. E., SEKHAR R. V. & BALASUBRAMANYAM A. (2013). "Relationship of ethnicity and CD4 Count with glucose metabolism among HIV patients on Highly-Active Antiretroviral Therapy (HAART)." *BMC Endocrine Disorders* **13**(13).

MOHAJAN, H. K. (2017). "Two Criteria for Good Measurements in Research: Validity and Reliability." *Annals of Spiru Haret University* **17**(3): 58-82.

MOLINARI, N., DAURES J.-P. & DURAND J.-F. (2001). "Regression splines for threshold selection in survivaldata analysis." *Statist. Med.* **20**: 237–247.

MONTARROYOS, U. R., MIRANDA-FILHO D. B., CÉSAR C. C., SOUZA W. V., LACERDA H. R., ALBUQUERQUE M. D. F. P. M., AGUIAR M. F. & XIMENES R. A. D. A. (2014). "Factors Related to Changes in CD4+ T-Cell Counts over Time in Patients Living with HIV/AIDS: A Multilevel Analysis." *PLoS One* **9**(2): e84276.

MONTESSORI, V., PRESS N., HARRIS M., AKAGI L. & MONTANER J. S. G. (2004). "Adverse effects of antiretroviral therapy for HIV infection." *CMAJ* **170**(2): 229–238.

MOOLLA, Y., MOOLLA Z., REDDY T. & MAGULA N. (2015). "The use of readily available biomarkers to predict CD4 cell counts in HIV-infected individuals." *South African Family Practice* **57**(5): 293-296.

MOORE, D. S., MCCABE B. A. & CRAIG G. P. (2009). Looking at Data: Distributions, W. H. Freeman and Company.

MORINET, F. D. R., PARENT M., BERGERON C., PILLET S. & CAPRON C. (2015). "Oxygen and viruses: a breathing story." *Journal of General Virology* **96**: 1979–1982.

MUGGEO, V. M. R. (2003). "Estimating regression models with unknown break-points." *Statistics in Medicine* **22**: 3055–3071

MUGGEO, V. M. R. (2017). Package 'segmented' : Regression Models with Break-Points / Change-Points Estimation**:** Given a regression model, segmented `updates' the model by adding one or more segmented (i.e., piece-wise linear) relationships. Several variables with multiple breakpoints are allowed.

MULAIK, S. A., JAMES L. R., VAN ALSTINE J., BENNET N., LIND S. & STILWELL C. D. (1989). "Evaluation of Goodness-of-Fit

Indices for Structural Equation Models." *Psychological Bulletin* **105**(3): 430-445.

MUND, M., NESTLER S. & (IN PRESS) (2018). "Beyond the cross-lagged panel model: Next-generation tools for analyzing interdependencies across the life course." *Advances in Life Course Research*.

MYERS, J. L. & WELL A. D. (2003). Research design and statistical analysis, Lawrence Erlbaum Associates.

NAKAGAWA, S. (2015). Missing data: mechanisms, methods and messages. Ecological Statistics: Contemporary Theory and Application., Oxford University Press**:** 81-105.

NAM (2012). CD4, viral load & other tests. UK.

NATIONAL CLINICAL GUIDELINE CENTRE (2012). Tailoring healthcare services for each patient. Patient Experience in Adult NHS Services: Improving the Experience of Care for People Using Adult NHS Services: Patient Experience in Generic Terms. London, Royal College of Physicians (UK). **(NICE Clinical Guidelines, No. 138.)**.

NATIONAL INSTITUTES OF HEALTH (2016). Vitamin B12 Fact Sheet for Consumers. Department of Health & Human Sciences. USA.

NATIONAL INSTITUTES OF HEALTH CLINICAL CENTER (2015). Understanding your complete blood count (CBC) and common blood deficiencies. National Institutes of Health Clinical Center. Bethesda.

NDITUNZE, L., MAKUZA S., AMOROSO C. L., ODHIAMBO J., NTAKIRUTIMANA E., CEDRO L., MUSHINZIMANA J. & HEDT-GAUTHIER B. (2015). "Assessment of Essential Medicines Stock-Outs at Health Centers in Burera District in Northern Rwanda." *Rwanda Journal Series F: Medicine and Health Sciences* **2**(1).

NEWSOM, J. T. (2015). Longitudinal Structural Equation Modeling. A Comprehensive Introduction, Routledge.

NGATIA, M. (2012). "Outliers: The good and the bad." Retrieved 09/10/2017, from http://www.water.ca.gov/environmentalservices/docs/qaqc/outliers_workgroup_06-20-12.pdf.

NIOS (2012). Kidney Function Test. BIOCHEMISTRY**:** 219-229.

NORM, O. R. & LARRY H. (2013). A Step-by-Step Approach to Using SAS ® for Factor Analysis and Structural Equation Modeling. Cary, North Carolina, USA., SAS Institute Inc.

NUNNALLY, J. C. & BERNSTEIN I. H. (1994). Psychometric Theory. New York, McGraw-Hill.

NZOU, C., KAMBARAMI R. A., ONYANGO F. E., NDHLOVU C. E. & CHIKWASHA V. (2010). "Clinical predictors of low CD4 count among HIV-infected pulmonary tuberculosis clients: A health facility-based survey." *S Afr Med J* **100**: 602-605.

OBIRIKORANG, C., QUAYE L. & ACHEAMPONG I. (2012). "Total lymphocyte count as a surrogate marker for CD4 count in resource-limited settings." *BMC Infectious Diseases* **12**: 128.

OBIRIKORANG, C. & YEBOAH F. A. (2009). "Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting." *Journal of Biomedical Science* **16**(102).

OKA, F., NAITO T., OIKE M., IMAI R., SAITA M., INUI A., MITSUHASHI K., ISONUMA H. & SHIMBO T. (2012). "Correlation between HIV disease and lipid metabolism in antiretroviral-naı¨ve HIV-infected patients in Japan." *J Infect Chemother* **18**: 17–21.

OLAWUMI, H. & OLATUNJI P. (2006). "The value of serum albumin in pretreatment assessment and monitoring of therapy in HIV/AIDS patients." *HIV Medicine* **7**: 351–355.

OLSON, H. M., WEISSFELD L., BOUDREAU R., AIZENSTEIN H., NEWMAN A., SIMONSICK E., VAN DOMELEN D., THOMAS F., YAFFE K. & ROSANO C. (2014). "A variant of sparse partial least squares for variable selection and data exploration." *Front. Neuroinform* **8**(18).

OMOREGIE, R., OSAKUE S., IHEMEJE V., OMOKARO E. & OGEFERE H. (2009). "Correlation of CD4 Count with Platelet Count, Prothrombin Time and Activated Partial Thromboplastin Time among HIV Patients in Benin City, Nigeria." *The West Indian medical journal* **58**(5): 437-440.

OMUSE, G., MAINA D., MWANGI J., WAMBUA C., RADIA K., KANYUA A., KAGOTHO E., HOFFMAN M., OJWANG P., PREMJI Z., ICHIHARA K. & ERASMUS R. (2018). "Complete blood count reference intervals from a healthy adult urban population in Kenya." *PLoS One* **13**(6): e0198444.

OPIYO, W. O., NG'WENA A. G. M. & OFULLA A. V. O. (2013). "Liver Function Markers And Associated Serum Electrolytes Changes In Hiv Patients Attending Patient Support Centre Of Jaramogi Oginga Odinga Teaching And Referral Hospital, Kisumu County, Kenya." *EAST AFRICAN MEDICAL JOURNAL* **90**(9): 276-287.

OPIYO, W. O., NG'WENA A. G. M. & OFULLA A. V. O. (2015). "Kidney Function Predictors and Associated Serum Electrolytes Changes

in HIV out Patients Attending Jaramogi Oginga Odinga Teaching And Referral Hospital, Kisumu County, Kenya." *East African medical journal* **92**(12): 442-453.

OSBORNE, J. W. & OVERBAY A. (2004). "The power of outliers (and why researchers should ALWAYS check for them)." *Practical Assessment, Research & Evaluation* **9**(6).

OURSLER, K., SORKIN J., SMITH B. & KATZEL L. (2006). "Reduced aerobic capacity and physical functioning in older HIV-infected men. ." *AIDS Res Hum Retroviruses.* **22**(11): 1113–1121.

PALACIOS, R., SANTOS J., GARCı´A A., CASTELLS E., GONZA´LEZ M., RUIZ J. & MA´RQUEZ M. (2006). "Impact of highly active antiretroviral therapy on blood pressure in HIV-infected patients. A prospective study in a cohort of naive patients." *HIV Medicine* **7**: 10–15.

PALERMO, G., PIRAINO P. & ZUCHT H.-D. (2009). "Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data." *Advances and Applications in Bioinformatics and Chemistry* **2**: 57–70.

PAPAGNO, L., SPINA C. A., MARCHANT A., SALIO M., RUFER N., LITTLE S., DONG T., CHESNEY G., WATERS A., EASTERBROOK P., DUNBAR P. R., SHEPHERD D., CERUNDOLO V., EMERY V., GRIFFITHS P., CONLON C., MCMICHAEL A. J., RICHMAN D. D., ROWLAND-JONES S. L. & APPAY V. (2004). "Immune activation and CD8+ T-cell differentiation towards senescence in HIV-1 infection. ." *PLoS BIOLOGY* **2**(2): 0173-0185.

PARK, J.-B., KANG D.-Y., YANG H.-M., CHO H.-J., PARK K., LEE H.-Y., KANG H.-J., KOO B.-K. & KIM H.-S. (2013). "Serum alkaline phosphatase is a predictor of mortality, myocardial infarction, or stent thrombosis after implantation of coronary drug-eluting stent." *European Heart Journal* **34**: 920–931.

PASTOR, R. & GUALLAR E. (1998). "Use of two-segmented logistic regression to estimate change-points in epidemiologic studies." *American Journal of Epidemiology* **148**: 631–642.

PASUPATHI, P., BAKTHAVATHSALAM G., SARAVANAN G. & DEVARAJ A. (2008). "Changes in CD4+ cell count, lipid profile and liver enzymes in HIV infection and AIDS patients." *Journal of APPLIED BIOMEDICINE* **6**: 139–145.

PATERSON, L. & GOLDSTEIN H. (1991). "New Statistical Methods for Analysing Social Structures: An Introduction to Multilevel Models." *British Educational Research Journal* **17**(4): 387-393.

PATIENTS AGAINST LYMPHOMA. (2004). "Peripheral Blood: Reference Ranges." Retrieved 20/10/2017, from http://www.lymphomation.org/peripheral-blood-ref.pdf.

PATIL, R., KAMBLE P. & RAGHUWANSHI U. (2013). "Serum ALP & GGT Levels in HIV Positive Patients." *International Journal of Recent Trends in Science And Technology* **5**(3): 155-157.

PEDERSEN, E., MILLER D., SIMPSON G. & ROSS N. (2019). "Hierarchical generalized additive models in ecology: an introduction with mgcv." *PeerJ* **7**: e6876

PHAIR, J. P. (2009). "Variations in the Natural History of HIV Infection." *AIDS Research and Human Retroviruses* **10**(8).

PILLINGER, R. (2018). "Random slope models." Retrieved 05/08/2018, 2018, from http://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html.

PINHEIRO, J. C. & BATES D. M. (2000). Mixe-Effects Models in S and S-Plus, Springer.

PRALHADRAO, H. S., KANT C., PHEPALE K., MALI M. K. & RAGHUNATH (2016). "Role of serum albumin level compared to CD4+ cell count as a marker of immunosuppression in HIV infection." *Indian Journal of Basic and Applied Medical Research* **5**(3): 495-502.

PRAVINA, P., SAYAJI D. & AVINASH M. (2013). "Calcium and its Role in Human Body." *International Journal of Research in Pharmaceutical and Biomedical Sciences* **4**(2): 659-668.

PROJECT INFORM (2007). Monitoring HIV Blood Work: A Complete Guide for Monitoring HIV. New York State Department of Health. SAN FRANCISCO, CA 94103 2621, Project Inform.

QUENE, H. & VAN DEN BERGH H. (2004). "On multi-level modeling of data from repeated measures designs: a tutorial." *Speech Communication* **43**: 103–121.

RAMSAY, J. O. & SILVERMAN B. W. (2002). Applied functional data analysis: Methods and case studies. New York, Springer.

RAMSAY, J. O. & SILVERMAN B. W. (2005). Functional data analysis. New York, Springer.

RAUDENBUSH, S. & BRYK A. S. (1986). "A Hierarchical Model for Studying School Effects." *Sociology of Education* **59**: 1-17.

RAUDENBUSH, S. W. & BRYK A. S. (2002). Hierarchical Linear Models: Applications and data analysis methods. Newbury Park, CA, Sage.

RIGBY, R. & STASINOPOULOS D. (1992). "Detecting break points in the hazard function

in survival analysis." *Statistical Modelling*: 303–311.

ROBBINS, K. R., SAXTON A. M. & SOUTHERN L. L. (2006). "Estimation of nutrient requirements using broken-line regression analysis." *J. Anim. Sci.* **84(E. Suppl.)**: E155–E165.

ROHART, F., GAUTIER B. T., SINGH A. & LÊ CAO K.-A. (2017). "mixOmics: An R package for 'omics feature selection and multiple data integration." *PLoS Comput Biol* **13**(11): e1005752.

ROSSITER, J. R. (2002). "The C-OAR-SE procedure for scale development in marketing." *International Journal of Research in Marketing* **19**(4): 305-335.

ROY, S., TARAFDAR M., RAGU-NATHAN T. S. & MARSILLAC E. (2012). "The Effect of Misspecification of Reflective and Formative Constructs in Operations and Manufacturing Management Research." *The Electronic Journal of Business Research Methods* **10**(1): 34-52.

ROYSTON, P., ALTMAN D. G. & SAUERBREI W. (2006). "Dichotomizing continuous predictors in multiple regression: A bad idea. ." *Statistics in Medicine* **25**(1): 127-141.

SARRO, Y., TOUNKARA A., TANGARA E., GUINDO O., WHITE H., CHAMOT E. & KRISTENSEN S. (2010). "Serum Protein Electrophoresis: Any role in monitoring for Antiretroviral Therapy?" *African Health Sciences* **10**(2): 138 - 143.

SAS INSTITUTE INC (2009). SAS.STAT 9.2 User's Guide, Second Edition. Cary, NC: SAS Institute Inc.

SAS INSTITUTE INC (2014). SAS/STAT(R) 13.2 User's Guide: The PLS Procedure. Cary, NC, USA, SAS Institute Inc,.

SAS INSTITUTE INC (2014). SAS® Visual Analytics 7.1. User's Guide. . Cary, NC, SAS Institute Inc.

SAWATSKY, M. L., CLYDE M. & MEEK F. (2015). "Partial Least Squares Regression in the Social Sciences." *The Quantitative Methods for Psychology* **11**(2): 52-62.

SCHREIBER, J. B., STAGE F. K. & KING J. (2006). "Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review." *The Journal of Educational Research* **99**(6): 323-337.

SCHUMACKER, R. E. & LOMAX R. G. (2010). A Beginner's Guide to Structural Equation Modeling, Routledge.

SEBER, G. & WILD C. (1989). Nonlinear Regression. New York, Wiley.

SECKO, D. (2005). "Inexpensive CD4 counting for the developing world." *CMAJ* **173**(5): 478.

SEGOLODI, T. M., HENDERSON F. L., ROSE C. E., TURNER K. T., ZEH C., FONJUNGO P. N., NISKA R., HART C. & PAXTON L. A. (2014). "Normal Laboratory Reference Intervals among HealthyAdults Screened for a HIV Pre-Exposure ProphylaxisClinical Trial in Botswana." *PloS ONE* **9**(4): e93034.

SEMEERE, A. S., NAKANJAKO D., DDUNGU H., KAMBUGU A., MANABE Y. C. & COLEBUNDERS R. (2012). "Sub-Optimal Vitamin B-12 Levels among ART-Naı¨ve HIV-Positive Individuals in an Urban Cohort in Uganda." *PLoS One* **7**(7): e40072.

SEN., L. C. S., VYAS A., SANGHI L. C. S., SHANMUGANANDAN C. K., GUPTA C. R., KAPILA B. K., PRAHARAJ S. C. A., KUMAR C. S. & BATRA C. R. (2011). "Correlation of CD4+ T cell Count with Total Lymphocyte Count, Haemoglobin and Erythrocyte Sedimentation Rate Levels in Human Immunodeficiency Virus Type-l Disease." *MJAFI* **67**(1): 15-20.

SERPA, J., HAQUE D., VALAYAM J., BREAUX K. & RODRIGUEZ-BARRADAS M. C. (2010). "Effect of combination antiretroviral treatment on total protein and calculated globulin levels among HIV-infected patients." *International Journal of Infectious Diseases* **14S**: e41–e44.

SEYOUM, A., NDLOVU P. & ZEWOTIR T. (2017). "Joint longitudinal data analysis in detecting determinants of CD4 cell count change and adherence to highly active antiretroviral therapy at Felege Hiwot Teaching and Specialized Hospital, North-west Ethiopia (Amhara Region)." *AIDS Research and Therapy* **14**(14).

SHAFER, R. W., RHEE S.-Y., PILLAY D., MILLER V., SANDSTROM P., SCHAPIRO J. M., KURITZKES D. R. & BENNETT D. (2007). "HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance." *AIDS* **21**: 215–223.

SHAH , R. & GOLDSTEIN S. M. (2006). "Use of structural equation modeling in operations management research: Looking back and forward." *Journal of Operations Management* **24**: 148-169.

SHAPIRO, N., KARRAS D. J., LEECH S. H. & HEILPERN K. L. (1998). "Absolute lymphocyte count as a predictor of CD4 count." *Annals of Emergency Medicine* **32**(3): 323-328.

SHARMA, S. S., JAMRA Y., HAWALDAR S. & MESHRAM A. (2016). "Study of serum albumin as surrogate marker of immune suppression in patients living with HIV and AIDS." *International Journal of Advances in Medicine* **3**(2): 152-156.

SHIEH, Y.-Y. & FOULADI R. T. (2003). "The Effect of Multicollinearity On Multilevel Modeling Parameter Estimates and Standard Errors." *Educational and Psychological Measurement* **63**(6): 951-985.

SHIFERAW, M. B., TULU K. T., ZEGEYE A. M. & WUBANTE A. A. (2016). "Liver Enzymes Abnormalities among Highly Active Antiretroviral Therapy Experienced and HAART Na\ve HIV-1 Infected Patients at Debre Tabor Hospital, NorthWest Ethiopia: A Comparative Cross-Sectional Study." *AIDS Research and Treatment* **Volume 2016, Article ID 1985452**.

SHU, Z., TIAN Z., CHEN J., MA J., ABUDUREYIMU A., QIAN Q. & ZHUO L. (2018). "HIV/AIDS-related hyponatremia: an old but still serious problem." *RENAL FAILURE* **40**(1): 68-74.

SILVA, I. D. S. (1999). Cancer Epidemiology: Principles and Methods. France, International Agency for Research on Cancer.

SINGER, J. D. (1998). "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics* **23**(4): 323-355.

SIRACUSA, M. C., KIM B. S., SPERGEL J. M. & ARTIS D. (2013). "Basophils and allergic inflammation." *J Allergy Clin Immunol* **132**(4): 789–788.

SIVARAM, M., WHITE A. & RADCLIFFE K. (2012). "Eosinophilia: clinical significance in HIV-infected individuals." *Int J STD AIDS.* **23**(9): 635-638.

SLOAND, E., KLEIN H., BANKS S., VARELDZIS B., MERRITT S. & PIERCE P. (1992). "Epidemiology of thrombocytopenia in HIV infection." *Eur J Haematol* **48**(3): 168-172.

SMITH, C. J., SABIN C. A., YOULE M. S., KINLOCH-DE LOES S., LAMPE F. C., MADGE S., CROPLEY I., JOHNSON M. A. & PHILLIPS A. N. (2004). "Factors Influencing Increases in CD4 Cell Counts of HIV-Positive Persons Receiving Long-Term Highly Active Antiretroviral Therapy." *The Journal of Infectious Diseases* **190**: 1860–1868.

SNIJDERS, T. (1996). "Analysis of longitudinal data using the hierarchical linear." *Quality & Quantity* **30**: 405-426.

SPEARMAN, C. (1904). "General Intelligence, Objectively Determined and Measured." *The American Journal of Psychology* **15**(2): 201-292.

SPECTRUM (2007). Total protein: Biuret Reagent, Egyptian Company for Biotechnology (S.A.E).

STASINOPOULOS, D. & RIGBY R. (1992). "Detecting break points in generalised linear models." *Computational Statistics and Data Analysis* **13** (461–471).

STEIGER, J. H. (2013). "Confirmatory Factor Analysis with R."

STRAZZULLO, P. (2014). "Sodium." *Adv Nutr.* **5**(2): 188–190.

STRENIO, J. F., WEISBERG H. I. & BRYK A. S. (1983). "Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates." *Biometrics* **39**: 71–86.

SUDFELD, C. R., ISANAKA S., ABOUD S., MUGUSI F. M., WANG M., CHALAMILLA G. E. & FAWZI W. W. (2013). "Association of Serum Albumin Concentration With Mortality, Morbidity, CD4 T-cell Reconstitution Among Tanzanians Initiating Antiretroviral Therapy." *The Journal of Infectious Diseases* **207**: 1370–1378.

SUN, J., AND REDDY, C.K. (2013). Big Data Analytics for Healthcare. SIAM International Conference on Data Mining. Austin, TX.

SUNITHA , L., BALRAJU M., SASIKIRAN J. & RAMANA E. V. (2014). "Automatic Outlier Identification in Data Mining Unisng IQR in Real-Time Data." *International Journal of Advanced Research in Computer and Communication Engineering* **3**(6): 7255-7257.

TABACHNICK, B. G. & FIDELL L. S. (2012). Using multivariate statistics,, Allyn & Bacon, Needham Heights, MA.

TANAKA, J. S. (1987). "How big is big enough? Sample size and goodness-of-fit in structural equation models with latent variables." *Child Development* **58**: 134-146.

TANAKA, J. S. (1993). Multifaceted conceptions of fit in structural equation models. Testing structural equation models K. A. Bollen and J. S. Long. Newbury Park, CA, Sage**:** 10-39.

THAIRU, L., KATZENSTEIN D. & ISRAELSKI D. (2011). "Operational challenges in delivering CD4 diagnostics in sub-Saharan Africa." *AIDS Care* **23**(7): 814-821.

THE JOHNS HOPKINS LUPUS CENTER. (2017). "Blood Chemistry Panel." Retrieved 13/10/2017, from https://www.hopkinslupus.org/lupus-tests/screening-laboratory-tests/blood-chemistry-panel/.

TIBCO-SPOTFIRE®. (2014). "What is a Parallel Coordinate Plot?" Retrieved 17/09/2017, from https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/para/para_what_is_a_parallel_coordinate_plot.htm.

TIBSHIRANI, R. (1996). "Regression shrinkage and selection via the lasso." *J. Roy. Statist. Soc. Ser. B* **58**: 58 267–288.

TIBSHIRANI, R., SAUNDERS M., ROSSET S., ZHU J. & KNIGHT K. (2005). "Sparsity and smoothness via the fused lasso." *J. R. Stat. Soc. Ser. B* **67**: 91–108.

TISHLER, A. & ZANG I. (1981). "A maximum likelihood method for piecewise regression models with a continuous dependent variable." *Applied Statistics* **30**: 116–124.

TISHLER, A. & ZANG I. (1981). "A new maximum likelihood algorithm for piecewise regression." *Applied Statistics* **76**: 980–987.

TOBIAS, R. D. (1995). "An Introduction to Partial Least Squares Regression." *SAS Institute, Inc*.

TOMITA, A., GARRETT N., WERNER L., BURNS J. K., MPANZA L., MLISANA K., VAN LOGGERENBERG F. & ABDOOL KARIM S. S. (2014). "Health-related quality of life dynamics of HIV-positive South African women up to ART initiation: evidence from the CAPRISA 002 acute infection cohort study." *AIDS Behav* **18**(6): 1114-1123.

TROCHIM, W. M. K. (2006, 20/10/2006). "Types of Reliability." Retrieved 11/02/2018, from http://www.socialresearchmethods.net/kb/reltypes.php.

U.S. DEPARTMENT OF VETERANS AFFAIRS. (2016). "Blood chemistry tests for HIV - HIV/AIDS." Retrieved 14/10/2017, from https://www.hiv.va.gov/patient/diagnosis/labs-blood-chemistry.asp.

UGWUJA, E. I. & EZE N. A. (2007). "A Comparative Study of Serum Electrolytes, Total Protein, Calcium and Phosphate Among Diabetic and HIV/AIDS Patients in Abakaliki, Southeastern, Nigeria." *The Internet Journal of Laboratory Medicine* **2**(1).

ULM, K. A. (1991). "Statistical methods for assessing a threshold in epidemiological studies." *Statistics in Medicine* **10**: 341–349.

UNDP (2006). Human Development Report. New York: Palgrave.

VALVONA, C. J., FILLMORE H. L., NUNN P. B. & PILKINGTON G. J. (2016). "The Regulation and Function of Lactate Dehydrogenase A: Therapeutic Potential in Brain Tumor." *Brain Pathology* **26**: 3-17.

VAN BUUREN, S. (2012). " Flexible Imputation of Missing Data. Boca Raton, FL: Chapman & Hall/CRC. ."

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN K. (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* **45**(3): 01-67.

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN K. (2017). "Package 'mice'." Retrieved 05/11/2017, from https://cran.r-project.org/web/packages/mice/mice.pdf.

VAN DEN BROECK, J., CUNNINGHAM S. A., EECKELS R. & HERBST K. (2005). "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities." *PLoS Medicine* **2**(10).

VAN DER VOET, H. (1994). "Comparing the Predictive Accuracy of Models Using a Simple Randomization Test " *Chemometrics and Intelligent Laboratory Systems* **25**: 313–323.

VAN LOGGERENBERG, F., MLISANA K., WILLIAMSON C., AULD S. C., MORRIS L., GRAY C. M., ABDOOL KARIM Q., GROBLER A., BARNABAS N., IRIOGBE I., ABDOOL KARIM S. S. & CAPRISA ACUTE INFECTION STUDY TEAM (2008). "Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study." *PLoS One* **3**(4): e1954.

VAN RIJ, J. (2016). "Checking for and handling autocorrelation." Retrieved 11/11/2018, 2018, from https://cran.r-project.org/web/packages/itsadug/vignettes/acf.html.

VANDEKERCKHOVE, J., MATZKE D. & WAGENMAKERS E.-J. (2014). Model Comparison and the Principle of Parsimony. The Oxford Handbook of Computational and Mathematical Psychology. J. R. Busemeyer, Z. Wang, J. T. Townsend and A. Eidels.

VANISRI, H. & VADIRAJA N. (2016a). "Relationship between Red blood cell parameters and immune status in HIV infected females." *Indian Journal of Pathology and Oncology* **3**(2): 255-259.

VANISRI, H. & VADIRAJA N. (2016b). "Association between Red blood cell parameters and immune status in HIV infected males." *Indian Journal of Pathology and Oncology* **3**(4): 684-689.

VENTER, E., GERICKE G. & BEKKER P. (2009). "Nutritional status, quality of life and CD4 cell count of adults living with HIV/AIDS in the Ga-Rankuwa area (South Africa)." *South African Journal of Clinical Nutrition* **22**(3): 124-129.

VERMONT, J., BOSSON J., FRANCOIS P., RUE R. & DEMONGEOT J. (1991). "Strategies for graphical threshold determination." *Computer Methods and Programs in Biomedicine* **35**: 141–150.

VINCENT, J.-L., DUBOIS M.-J., NAVICKIS R. J. & WILKES M. M. (2003). "Hypoalbuminemia in Acute Illness: Is There aRationale for Intervention?" *ANNALS OF SURGERY* **237**(3): 319–334.

VOLBERDING, P. A., LEVINE A. M., DIETERICH D., DONNA MILDVAN,

MITSUYASU R. & SAAG M. (2004). "Anemia in HIV Infection: Clinical Impact and Evidence-Based Management Strategies." *Clinical Infectious Diseases* **38**: 1454–1463.

VOSS, T. G., FERMIN C. D., LEVY J. A., VIGH S., CHOI B. & GARRY R. F. (1996). "Alteration of Intracellular Potassium and Sodium Concentrations Correlates with Induction of Cytopathic Effects by Human Immunodeficiency Virus." *JOURNAL OF VIROLOGY* **70**(8): 5447–5454.

WANG, L., ZHOU J. & QU A. (2012). "Penalized generalized estimating equations for high-dimensional longitudinal data analysis." *Biometrics* **68**(2): 353–360.

WANG, M. & BODNER T. E. (2007). "Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. ." *Organizational Research Methods* **10**: 635-656. .

WANG, Q., DING H., XU J., GENG W., LIU J., GUO X., KANG J., LI X., JIANG Y. & SHANG H. (2016). "Lipids profile among ART-naïve HIV infected patients and men who have sex with men in China: a case control study." *Lipids in Health and Disease* **15**: 149.

WANG, S., YANG Y., LIN C., JIH-SHENG & FANG-PANG (2013). "Using Penalized Regression with Parallel Coordinates for Visualization of Significance in High Dimensional Data." *(IJACSA) International Journal of Advanced Computer Science and Applications* **4**(10).

WATKINS, M. W. & STYCK K. M. (2017). "A Cross-Lagged Panel Analysis of Psychometric Intelligence and Achievement in Reading and Math." *J. Intell.* **5**(31).

WEGMAN, E. J. (2001). "On some mathematics for visualizing high dimensional data " *Sankhya : The Indian Journal of Statistics San Antonio Conference: selected articles* **64**: 429-452.

WEST, S. G., FINCH J. F. & CURRAN P. J. (1995). Structural equation models with nonnormal variables. . Structural equation modeling: Concepts, issues, and applications R. H. Hoyle. Thousand Oaks, CA, Sage**:** 56-75.

WESTAT INC. (1993). National Health and Nutrition Examination Survey III Cycle 2: Pulse and Blood Pressure Procedures For Household Interviewers.

WESTON, R. & MARETT B. (2009). "HIV infection pathology and disease progression " *Clinical Pharmacist* **1**: 387.

WHEATON, B., MUTHEN B., ALWIN D. F. & SUMMERS G. (1977). "Assessing Reliability and Stability in Panel Models." *Sociological Methodology* **8**(1): 84-136.

WHITFIELD, J. (2001). "Gamma glutamyl transferase." *Crit Rev Clin Lab Sci. 2001 Aug;38(4):* **38**(4): 263-355.

WINGFIELD, T. & WILKINS E. (2010). "Opportunistic infections in HIV disease." *British Journal of Nursing* **19**(10).

WINTROBE, M. & GREER J. (2009). Wintrobe's Clinical Hematology. Philadelphia., Lippincott Williams & Wilkins, .

WOLD, H. (1966). Multivariate Analysis. New York, Wiley, Academic Press.

WOLfiNGER, R. D. (1993). "Covariance structure selection in general mixed models. ." *Communications in Statistics, Simulation and Computation* **22**: 1079–1106.

WOOD, S. N. (2000). "Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. ." *J. R. Statist. Soc. B* **62**(2): 413-428.

WOOD, S. N. (2001). "Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting. ." *Biometrics* **57**: 240–244.

WOOD, S. N. (2010). "Generalized Additive Models." Retrieved 03/06/2019, 2019, from https://people.maths.bris.ac.uk/~sw15190/mgcv/tampere/gam.pdf.

WOOD, S. N. (2010). "More advanced use ofmgcv." Retrieved 03/06/2019, 2019, from https://people.maths.bris.ac.uk/~sw15190/mgcv/tampere/mgcv-advanced.pdf.

WOOD, S. N. (2016). Package 'mgcv'. Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation.

WOOD, S. N. (2017). Generalized Additive Models: an introduction with R.

WOOD, S. N. (2019, 12/03/2019). "Factor smooth interactions in GAMs." Retrieved 11/11/2018, 2018, from https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/smooth.construct.fs.smooth.spec.html.

WRIGHT, S. (1918). "On the nature of size factors." *Genetics* **3**(4): 367-374.

WRIGHT, S. (1920). "The relative importance of heredity and environment in determining the piebald pattern of guinea pigs " *Proceedings of the National Academy of Sciences* **6**: 320-332.

WRIGHT, S. (1921). "Correlation and causation. Part I- Methods of path coefficients." *Journal of Agricultural Research* **7**: 557-585.

WRIGHT, S. (1923). "The theory of path coefficients. A replay to Niles's criticism." *Genetics* **8**(3): 239-255.

WRIGHT, S. (1934). "The Method of Path Coefficients." *The Annals of Mathematical Statistics* **5**(3): 161-215.

WU, C. & MA S. (2015). "A selective review of robust variable selection with applications in

bioinformatics." *Brief. Bioinform* **16**: 873–883.

WU, C., ZHOU F., REN J., LI X., JIANG Y. & MA S. (2019). "A Selective Review of Multi-Level Omics Data Intergration using Variable selection." *High-Throughput* **8**(4): 01-25.

XIANG, D. H. (2001). P256-26 Fitting Generalized Additive Models with the GAM Procedure. Statistics, Data Analysis, and Data Mining.

YAKUBU, T., DEDU V. K. & BAMPOH P. O. (2016). "Factors Affecting CD4 Count Response in HIV Patients within 12 Months of Treatment: A Case Study of Tamale Teaching Hospital." *American Journal of Medical and Biological Research* **4**(4): 78-83.

YALLAMRAJU, S. R., MEHROTRA R., SINHA A., GATTUMEEDHI S. R., GUPTA A. & KHADSE S. V. (2014). "Use of mid upper arm circumference for evaluation of nutritional status of OSMF patients." *Journal of Internatioanl Society of Prevention & Community Dentistry* **4**(2): 122-125.

YEE, T. W. & MITCHELL N. D. (1991). "Generalized additive models in plant ecology." *J. Vegetation. Science* **2**: 587- 602.

YERLY, S. & HIRSCHEL B. (2012). "Diagnosing acute HIV infection." *Expert Rev. Anti Infect. Ther.* **10**(1): 31-41.

ZEGER, S. L. & DIGGLE P. J. (1994). "Semiparametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters." *Biometrics* **50**: 689–699.

ZELLNER, A. (1970). "Estimation of regression relationships containing unobservable variables. ." *International Economic Review* **11**: 441-454. .

ZHANG, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *Ann. Appl. Stat.* **38**(2): 894–942.

ZHENG, Y., CASTRO D., GAY D. & CAI D. (2015). "Mean Corpuscular Hemoglobin Concentration in Hemoglobin CC, SC, and AC." *North American Journal of Medicine and Science* **8**(1): 1-4.

ZHOU, H., CUI W., QU H., WU Y., YUAN X. & ZHUO W. (2009). "Splatting the Lines in Parallel Coordinates." *Eurographics/ IEEE-VGTC Symposium on Visualization 2009* **28**(3).

ZOU, H. (2006). "The adaptive lasso and its oracle properties." *J. Am. Stat. Assoc.* **101**: 1418–1429.

ZOU, H., HASTIE T. & TIBSHIRANI R. (2006). "Sparse principal component analysis." *J. Comput. Graph. Stat.* **15**(2): 265–286.

ZOU, H. H., T. (2005). "Regularization and variable selection via the elastic net." *J. R. Stat. Soc. Ser. B* **67**: 301–320.

ZUUR, A. F., IENO E. N. & SMITH G. M. (2007). Analysing Ecological Data, Springer.

Appendices

Appendix A: Full research papers I and II

Infect Dis Ther (2019) 8:269–284
https://doi.org/10.1007/s40121-019-0235-4

ORIGINAL RESEARCH

An Evaluation to Determine the Strongest CD4 Count Covariates during HIV Disease Progression in Women in South Africa

Partson Tinarwo · Temesgen Zewotir · Nonhlanhla Yende-Zuma · Nigel J. Garrett · Delia North

Received: November 23, 2018 / Published online: February 12, 2019

ABSTRACT

Introduction: Past endeavours to deal with the obstacle of expensive Cluster of Difference 4 ($CD4^+$) count diagnostics in resource-limited settings have left a long trail of suggested continuous $CD4^+$ count clinical covariates that turned out to be a potentially important integral part of the human immunodeficiency virus (HIV) treatment process during disease progression. However, an evaluation to determine the strongest candidates among these $CD4^+$ count covariates has not been well documented.

Methods: The Centre for the AIDS Programme of Research in South Africa (CAPRISA) initially enrolled HIV-negative (phase 1) patients into different study cohorts. The patients who seroconverted (237) during follow-up care were enrolled again into a post-HIV infection cohort where they were further followed up with weekly to fortnightly visits up to 3 months (phase 2: acute infection), monthly visits from 3–12 months (phase 3: early infection) and quarterly visits thereafter (phase 4: established infection) until antiretroviral therapy (ART) initiation (phase 5). The $CD4^+$ count and 46 covariates were repeatedly measured at each phase of the HIV disease progression. A multilevel partial least squares approach was applied as a variable reduction technique to determine the strongest $CD4^+$ count covariates.

Results: Only 18 of the 46 investigated clinical attributes were the strongest $CD4^+$ count covariates and the top 8 were positively and independently associated with the $CD4^+$ count. Besides the confirmatory *lymphocytes*, these were *basophils*, *albumin*, *haematocrit*, *alkaline phosphatase (ALP)*, *mean corpuscular volume (MCV)*, *platelets*, *potassium* and *monocytes*. Overall, electrolytes, proteins and red blood cells were the dominant categories for the strongest covariates.

Conclusion: Only a few of the many previously suggested continuous $CD4^+$ count clinical covariates showed the potential to become an important integral part of the treatment process. Prolonging the pre-treatment period of the HIV disease progression by effectively

Enhanced digital features To view enhanced digital features for this article go to https://doi.org/10.6084/m9.figshare.7637267

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40121-019-0235-4) contains supplementary material, which is available to authorized users.

P. Tinarwo (✉) · T. Zewotir · D. North
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa
e-mail: partsont@gmail.com

N. Yende-Zuma · N. J. Garrett
Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa

incorporating and managing the covariates for long-term influence on the $CD4^+$ cell response has the potential to delay challenges associated with ART side effects.

Keywords: $CD4^+$ count; Continuous clinical covariates; HIV disease progression; Multilevel partial least squares; Prospective cohort studies; Variable reduction

INTRODUCTION

The Cluster of Difference 4 ($CD4^+$) count is the most common indicator of health status and immune function of patients infected with the human immunodeficiency virus (HIV) [1]. Several $CD4^+$ count covariates from different clinical platforms have been investigated in HIV-positive patients. The quest for understanding the behavioural patterns of the $CD4^+$ count covariates has been due to different reasons ranging from their potential use as either cost-effective $CD4^+$ count surrogates [2–4] or predictors [5–7] to pre-treatment assessment and monitoring of therapy in HIV-positive patients [8]. Such endeavours to keep abreast of the health status of HIV-positive patients in the absence of the $CD4^+$ count were triggered by high costs of the $CD4^+$ count diagnostic devices in the past [9, 10], making them not easily accessible to resource-limited settings in the developing world [11] where the health facilities are usually overburdened [12]. The challenge was exacerbated by operational and logistical issues [13, 14] in the supply of essential medicine for the patients [15–17] including frequent instrument breakdown and poor manufacturer maintenance of $CD4^+$ count diagnostics [18]. Recently, obtaining the $CD4^+$ count has become extraordinarily inexpensive [19–21] and, in the contemporary era of antiretroviral therapy (ART), the monitoring and restoration of patient's $CD4^+$ count to acceptable levels are now relatively easy [22] and have led to improved patient survival periods [23]. Despite the breakthrough in ART, recommendations have been made to suggest other factors that influence long-term $CD4^+$ cell response in conjunction with the therapy [24]. As such, previous studies have been inclined more towards social, demographic and other categorical factors [25–27], which suffer from information loss due to their grouping nature [28, 29]. On the other hand, the "richer" continuous clinical covariates are more sensitive to sources of variation [30] in the $CD4^+$ count and better capable of capturing and explaining realistic behavioural patterns of the $CD4^+$ cell response in the face of the rapidly mutating [31] HIV that is known to attack the $CD4^+$ cells [32]. The suggestion of the $CD4^+$ count surrogates, predictors and pre-treatment assessment options in the past turned out to have the potential to be manipulated as drivers for influencing long-term $CD4^+$ cell response in HIV-positive patients. For example, a close follow-up on sodium has been reported to improve outcomes [33] as it positively influences the $CD4^+$ count and its early management was found to be a contributing factor to the survival rates of HIV-positive patients [34]. Among other $CD4^+$ count clinical covariates, sodium and calcium levels are affected by dietary conditions [35, 36], which can become a potentially important integral part of the HIV treatment process during disease progression. Other blood chemistry components have also been suggested [5, 7, 33, 37–52] including $CD4^+$ count covariates from other clinical platforms such as the full blood count [3, 5, 53–57], lipids [58–60], sugar [61–63] and clinical examination measurements [2, 6, 64–71]. It then stands to reason that endeavours to deal with the obstacle of expensive $CD4^+$ count diagnostic devices in the past left a long trail of suggested continuous $CD4^+$ count clinical covariates that have potential to be an important integral part of the treatment process during HIV disease progression. The list of such potentially manageable continuous $CD4^+$ count clinical covariates has also grown over the past few years owing to the tremendously high volume of patient electronic health records that are being stored at a faster pace and relatively cheaper than in the past [72]. However, an evaluation to determine the strongest candidates of these continuous $CD4^+$ count clinical covariates during HIV disease progression has not been well documented.

This bioinformatics study aimed to pool and evaluate the previously and independently suggested continuous $CD4^+$ count clinical covariates to give an insight on the strongest drivers of the long-term $CD4^+$ cell response in HIV-positive patients during the disease progression. Our goal was to shed more light on the possibilities of integrating and managing the continuous $CD4^+$ count clinical covariates in the HIV treatment process. For example, ART is a major milestone in HIV treatment [73] but it is associated with side effects that can lead patients into challenging situations [74, 75]. Hence, the realisation of managing this continuous clinical covariate influence on the $CD4^+$ cell response would potentially prolong the pre-treatment period and increase the likelihood of delaying patients in experiencing ART-related issues at an early stage of the disease progression. Some of the statistical tools previously used to assess the $CD4^+$ count and covariate associations either were limited or suffered from information loss, for example, analysis of variance (ANOVA) [51, 76, 77], confidence intervals [64], t-tests [58–60], non-parametric tests [33, 38], chi-square tests [61, 78], linear regression [65, 79], sensitivity, specificity and positive prediction [2, 8] and correlation analysis [63, 66, 80]. As such, we also sought to pave the way for other areas such as predictive modelling with streamlined influential clinical covariates that are richer in information preserved in their continuous nature to explain the $CD4^+$ count variation. We evaluated available measurements of the continuous $CD4^+$ count clinical covariates routinely collected at the Centre for the AIDS Programme of Research in South Africa (CAPRISA).

METHODS

The Study Design

The CAPRISA 002 enrolled 245 HIV-negative (phase 1: pre- HIV infection) female sex workers into an Acute Infection study. The establishment of the acute infection study, cohort screening and seroconverts; routine evaluation procedures; CAPRISA-participant interaction and data management have been previously documented [81]. The study protocol and informed consent documents were reviewed and approved by the local ethics committees of the University of KwaZulu-Natal, the University of Cape Town, the University of the Witwatersrand in Johannesburg and the Prevention Sciences Review Committee (PSRC) of the Division of AIDS (DAIDS, National Institutes of Health, USA). The study was also performed in accordance with the Helsinki Declaration of 1964 and its later amendments. The consent forms were translated into vernacular language, isi-Zulu, and written informed consent was obtained at each stage of the study. All minors under the age of 18 years were excluded from the study as part of the screening procedure. The HIV-negative cohort was followed up and upon HIV infection they were further followed up with weekly to fortnightly visits up to 3 months (phase 2: acute infection), monthly visits from 3–12 months (phase 3: early infection) and quarterly visits thereafter (phase 4: established infection) until ART initiation (phase 5). Eventually 27 seroconversions were recorded. In addition to the 27 seroconverts, 210 more patients who seroconverted from other CAPRISA studies were also enrolled and similarly followed up post infection from the acute to ART phase. Figure 1 summarises how the total sample size of 237 seroconverts for this study was obtained.

Data

Four time points prior to each phase transition were selected, which resulted in a total of 16 repeated measurements being investigated for each patient. The baseline (Phase 1) repeated measurements were scarce; hence, this study focused on phases 2 to 5 only. The $CD4^+$ count covariates include: full blood count, lipids, sugar, blood chemistry and clinical examination. Several of these variables have been studied as potential covariates for the $CD4^+$ count but mostly contested in isolation or within a small group of barely just under five variables confined within their respective clinical

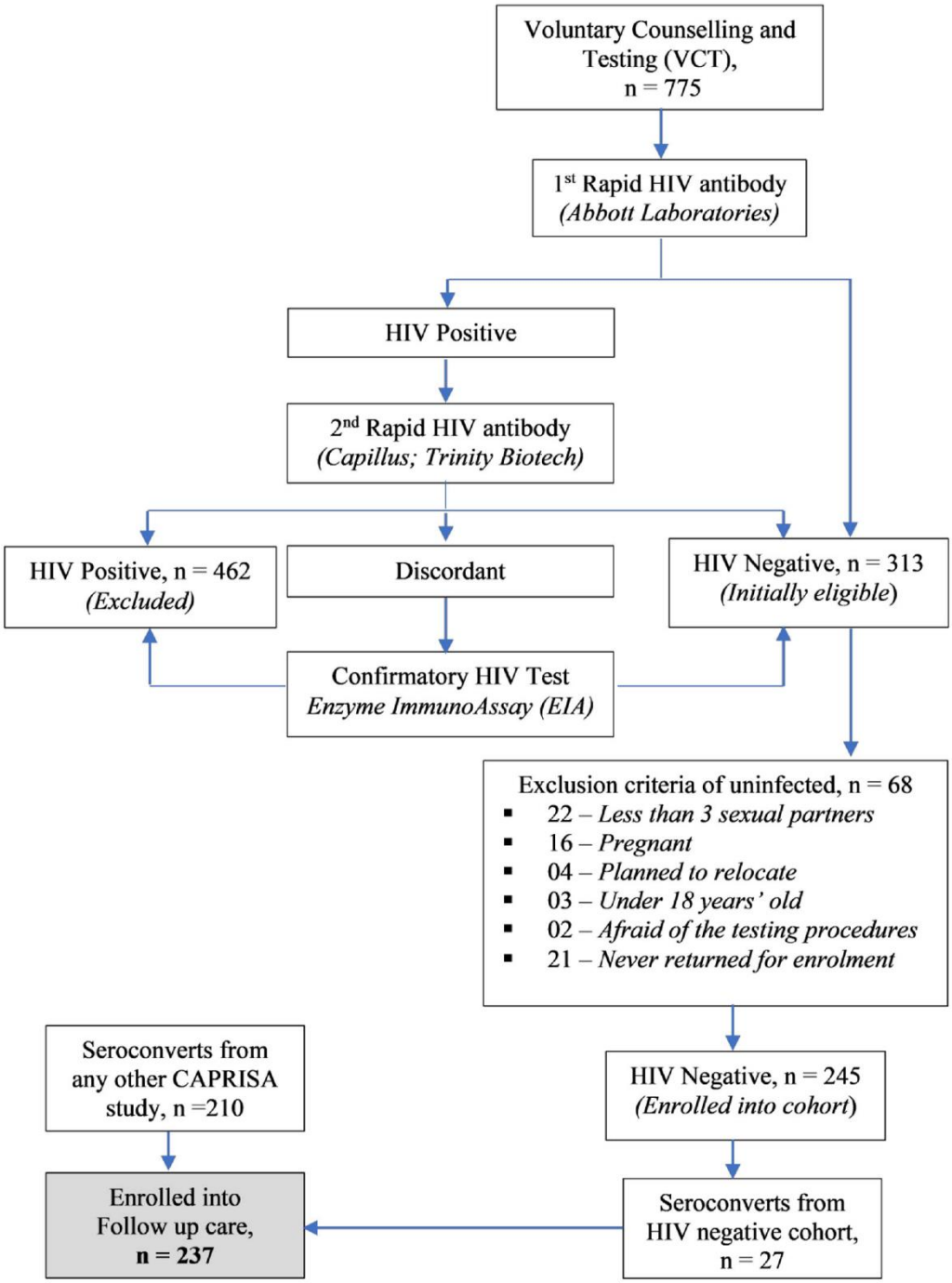


Fig. 1 Study design. The HIV-negative cohort screening involved 775 voluntary potential candidates of which 462 were already HIV positive and 313 initially eligible. Of the 313 HIV-negative patients, only 245 were enrolled and the rest excluded for various reasons according to the eligibility criteria. Eventually 27 out of the 245 seroconverted were enrolled into follow-up care. Seroconverts from other CAPRISA studies (210) were also included into the follow-up care that resulted in a total of 237 patients for this study

platforms. All the data sets from the different clinical platforms were pooled into a single data set.

Statistical Analysis

All the analysis was performed in the open-source R software, version 3.5.0. Firstly, a descriptive summary of the repeated measurements was provided using the function stat.desc in the **pastecs** library. Secondly, redundant features among the covariates were investigated using correlation analysis that dropped off the covariates with the highest mean absolute correlation using the findCorrelation function. Thirdly and last, this was then followed by the partial least squares (PLS) approach to model building with the application of the spls function in the **mixOmics** library, which is capable of handling the complex structure of repeated measurements. The package incorporates a design matrix to account for variation in the multilevel structure of the longitudinal data. PLS handles multicollinearity and a very large number of variables in longitudinal data. It ranks the covariates from strongest to weakest allowing variable selection and consequently dimension reduction. Since the PLS is a multi-dimensional analysis technique, graphical displays of the results were vital to comprehensively visualise the variable selection process with the aid of the instrumental R libraries: the **ggplot2** and **ggrepel**.

RESULTS

Descriptive Statistics

Table 1 shows that throughout the follow-up care, the minimum and maximum $CD4^+$ counts recorded were 45 and 1395 cells/mm^3, respectively. During the follow-up period, at least 50% of the $CD4^+$ count repeated measurements were above 539 cells/mm^3 and averaging 571.14 ± 238.45 cells/mm^3 with an overall variation of 41.75% around the cohort average. The greatest variation in the covariates was observed in *eosinophils* (101.11%), *basophils* (75.74%) and *gamma glutamyl transferase* (64.20%).

Redundant Feature Selection

Table 2 shows that haemoglobin (Hb), mean corpuscular haemoglobin (MCH), leucocytes, cholesterol, hip circumference, weight (kg) and body mass index (BMI) were highly correlated with the other covariates. The anthropometric measurements were the most highly correlated among themselves but the BMI, although marked as a redundant feature, was intuitively included in the second stage of variable reduction using the PLS.

Variable Selection

The optimal principal component (Fig S1) explained 68.95% of the variance in the response ($CD4^+$ count) and the variable selection simultaneously considered both the variable importance in projection (VIP) and regression coefficients (see Fig S2 for details). We presented all three VIP cut-off points where a cut-off point of 1.5 can be considered as a strict selection, 1.0 as moderate and 0.8 as lenient. A stricter variable selection process selected two covariates, the moderate (13) and lenient (18), of the 40 non-redundant features available for our study. We developed an interest in all the 18 strongest covariates as selected by the lenient cut-off point.

Figure 2 provides a list of all 40 covariates from the strongest to the weakest significance as well as their behavioural patterns in the predictive power (coefficients), component construction (loadings) and independent association (correlation) with the $CD4^+$ count together with the associated p values. The covariate loadings and regression coefficients indicated more or less the same effects in component construction and predictive power, respectively. Among the significant covariates, folate, magnesium, calcium and sodium had the highest reducing effect on the $CD4^+$ count, whereas alkaline phosphatase (ALP), mean corpuscular volume (MCV) and lactate dehydrogenase (LDH) corresponded to an increased

Table 1 A descriptive summary of the investigated variables

		Variable	Min	Median	Mean	Max	SD	CV (%)
Full blood count	Red blood cells	CD4 count (cells/mm^3)	45.00	539.00	571.14	1395.00	238.45	41.75
		Red blood cell count	2.69	4.23	4.28	5.55	0.49	11.48
		Haemoglobin(Hb)	7.20	12.20	12.28	16.20	1.59	12.96
		Haematocrit	0.23	0.37	0.37	0.48	0.04	11.18
		MCV	64.20	86.90	87.11	108.00	6.99	8.02
		MCH	19.40	28.70	28.78	37.20	2.88	10.02
		MCHC	29.10	33.00	33.02	37.10	1.34	4.05
		RDW	11.10	14.30	14.68	22.40	1.87	12.71
		Platelet	43.00	287.00	295.13	591.00	81.01	27.45
	White blood cells	Leucocytes	1.85	4.98	5.31	11.90	1.73	32.62
		Neutrophils	0.59	2.43	2.78	7.79	1.33	47.84
		Lymphocytes	0.37	1.79	1.90	4.42	0.74	38.78
		Monocytes	0.07	0.39	0.41	1.01	0.15	37.13
		Eosinophils	0.00	0.10	0.15	0.80	0.15	101.11
		Basophils	0.00	0.02	0.02	0.09	0.02	75.74
Lipids		Cholesterol	1.10	3.80	3.84	7.00	0.87	22.73
		LDL	0.60	2.30	2.31	4.80	0.71	30.73
		Triglycerides	0.00	0.90	0.97	2.80	0.46	47.20
		Glucose	1.00	2.00	1.62	3.00	0.67	40.96
Blood chemistry	Liver function	ALT(GPT)	4.00	19.00	21.22	64.00	9.43	44.43
		AST(GOT)	12.50	25.00	27.30	72.00	10.06	36.86
		Bilirubin	0.00	7.00	7.32	20.00	3.35	45.71
		Alkaline phosphatase	28.00	66.00	70.52	162.00	21.52	30.51
		Gamma glutamyl transferase	4.00	19.00	23.58	94.00	15.14	64.20
	Electrolytes	Calcium	1.93	2.28	2.29	2.67	0.11	4.67
		Chloride	95.00	105.00	105.14	115.00	2.83	2.69
		Magnesium	0.58	0.82	0.82	1.09	0.08	9.63
		Potassium	0.65	2.92	3.19	6.00	0.82	25.73
		Sodium	129.00	137.00	136.90	145.00	2.35	1.72
	Protein	Protein	56.00	83.00	84.16	99.00	7.84	9.32
		Albumin	21.00	38.00	37.90	52.00	4.51	11.89
		Lactate dehydrogenase	27.00	440.00	438.31	1076.00	169.58	38.69
	Iron and vitamins	Iron (Fe) (mcg/dl)	1.00	11.00	11.59	34.45	6.09	52.52
		Folate (nmol/l)	1.15	13.65	15.69	49.28	8.47	53.99
		Vitamin B12 (ng/ml)	0.00	274.00	296.56	829.75	126.54	42.67
		Urea (mmol/l)	1.10	3.30	3.42	6.80	0.95	27.87

Table 1 continued

		Variable	Min	Median	Mean	Max	SD	CV (%)
Clinical examination	Physical	BP(systolic) (mmHg)	74.00	118.00	118.64	168.00	13.81	11.64
		BP(diastolic) (mmHg)	46.00	74.50	75.14	109.00	9.67	12.87
		Pulse (bpm)	48.00	80.00	81.03	118.00	10.08	12.44
		Axillary temperature (°C)	34.30	36.25	36.24	37.90	0.49	1.36
	Anthropometric	Waist circumference (cm)	31.00	84.00	86.99	145.00	16.25	18.69
		Hip circumference(cm)	64.12	106.00	107.99	160.00	14.58	13.50
		Arm(right) circumference (cm)	13.50	29.00	29.87	47.00	5.35	17.92
		Triceps skin fold (mm)	5.00	25.00	27.29	61.00	10.53	38.60
		Height (m)	1.31	1.57	1.58	1.81	0.08	5.14
		Weight (kg)	38.50	69.00	73.66	150.00	20.96	28.46
		Body mass index (BMI) (kg/m^2)	16.00	27.53	29.52	61.10	8.10	27.45

Table 2 Redundant features: highly correlated ($r > 0.75$) covariates of the CD4 count

	Hb[a]	MCH[a]	Leucocytes[a]	Cholesterol[a]	Hip circumference[a]	Weight (kg) [a]	BMI[b]
Haematocrit	0.9499						
MCV		0.9211					
Neutrophils			0.8464				
LDL				0.8636			
Waist circumference					0.7580	0.8158	0.7810
Hip circumference						0.8884	0.8414
Arm(right) circumference					0.7867	0.8006	0.7925
Weight(kg)[a]							0.9149

Hb, haemoglobin, *MCH*, mean corpuscular haemoglobin, *MCV*, mean corpuscular volume, *LDL*, low-density lipoproteins, *BMI*, body mass index

[a] Dropped redundant feature

[b] Intuitively included redundant feature in the PLS variable reduction

$CD4^+$ count. In this study, the lymphocytes had the highest direct independent positive correlation with the $CD4^+$ ($r = 0.5421$, $p < 0.0001$) followed by haematocrit ($r = 0.2337$, $p < 0.0001$). On the other hand, protein had the highest negative correlation ($r = -0.1740$, $p < 0.0001$) with the $CD4^+$ count followed by folate ($r = -0.1530$, $p < 0.0001$). The results showed that the top 8 of the 18 selected covariates were positively and independently associated with the $CD4^+$ count. Of all the investigated 40 non-redundant covariates, red blood cell distribution width (RDW), pulse, urea, alanine aminotransferase (glutamate pyruvate transaminase) ALT(GPT) and axillary temperature were the least important.

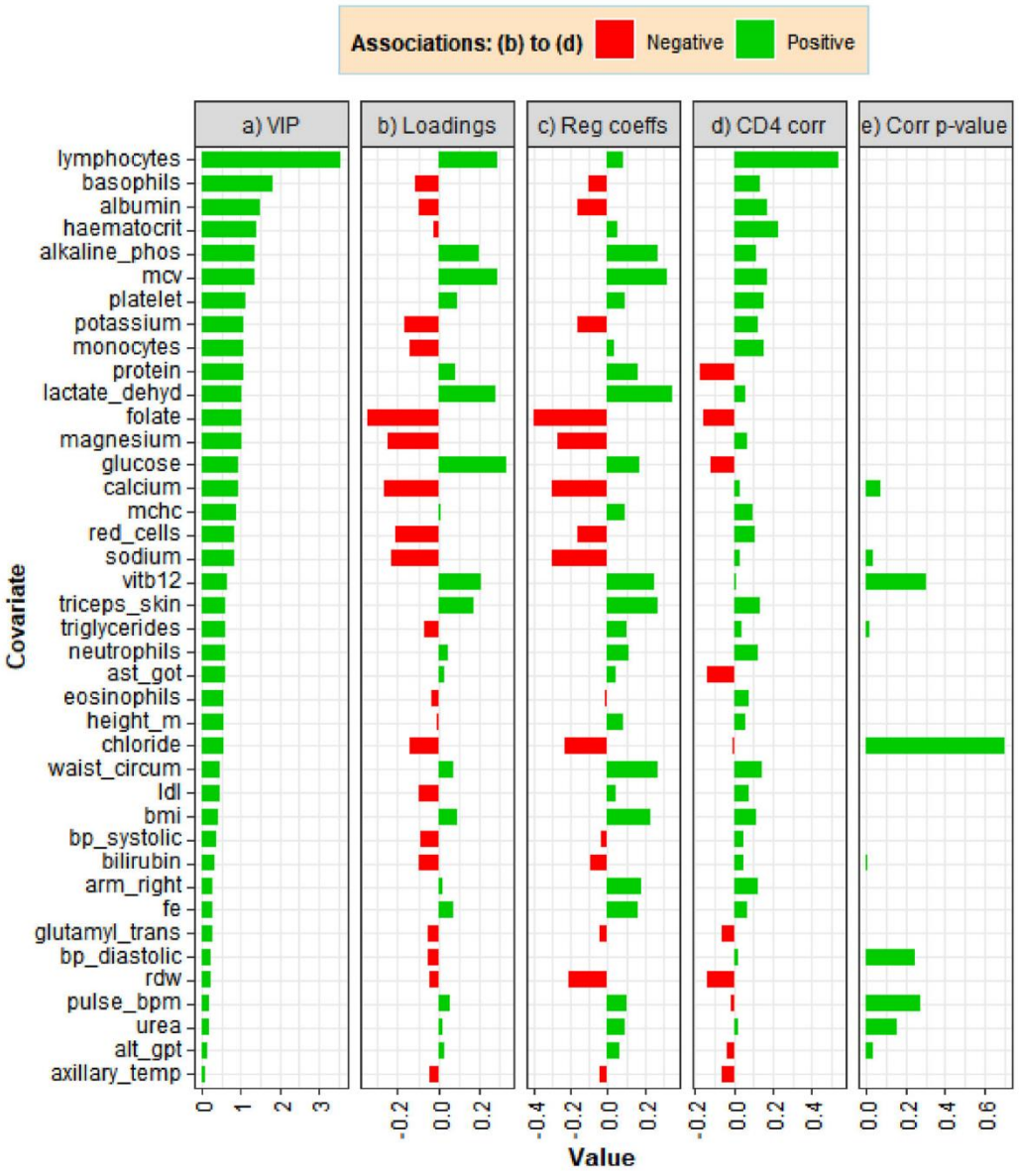


Fig. 2 Variable importance. Also shown are the related loadings, standardised regression coefficients and correlations of each covariate with the response variable (CD4 count)

A look at the significant variables by clinical category (Fig. 3) revealed that there was no significant variable selected from lipids, physical examination and anthropometric measurements. Folate was the only significant variable in its category and similarly alkaline phosphatase only among the liver function indicators. The PLS suggested chloride and RDW as the only insignificant $CD4^+$ count covariates among the electrolytes and red blood cells, respectively. Given the lymphocytes, basophils and monocytes, the significant covariates within the white blood cells group, the lymphocytes were dominantly significant. Generally, most of the significant $CD4^+$ count covariates were selected from electrolytes, proteins and red blood cells. The data for all the variable selection plots are given in File S1.

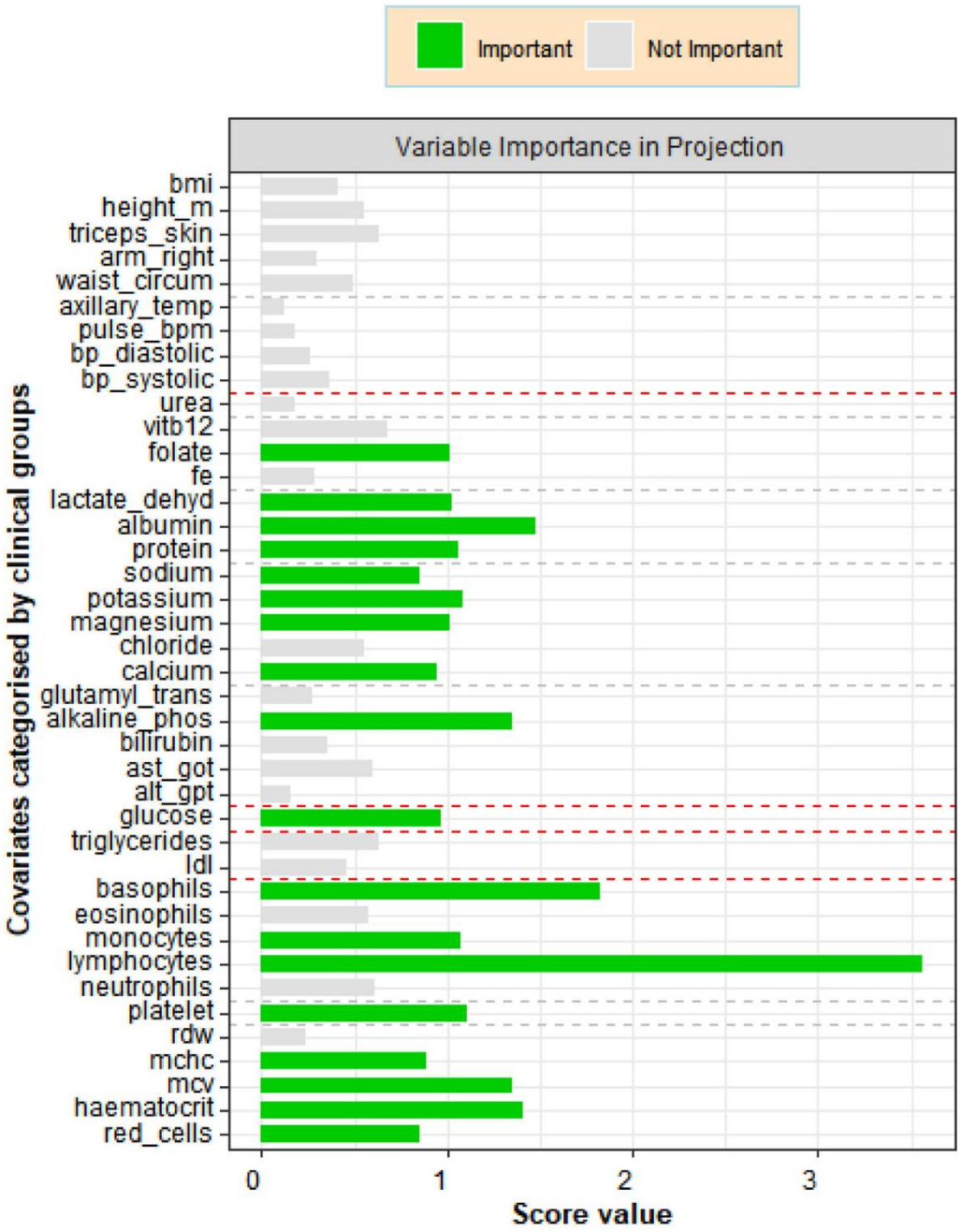


Fig. 3 Variable importance by clinical category. The broken red horizontal lines divide the major groups. From the top, the groups are clinical examination, blood chemistry, sugar, lipids and full blood count. The horizontal broken grey lines divide the subgroups within the major groups

DISCUSSION

In the present study, we evaluated a list of continuous $CD4^+$ count clinical covariates that were available at CAPRISA to determine the strongest candidates that can potentially become an important integral part of the HIV treatment process. The HIV targets and kills $CD4^+$ cells resulting in the $CD4^+$ count being an important outcome indicator for the

patient's health status. ART is known to supress the viral load and consequently an increased number of $CD4^+$ cells are spared giving rise to an improved immune system [73]. Hence, during the HIV treatment phase, ART is a major determinant of the $CD4^+$ count distribution. The intention of this study was to select the continuous clinical covariates that contributed to the greatest variation in the $CD4^+$ count from an overall perspective throughout the post-HIV period including ART. We used the PLS approach to achieve this and variable reduction is possible [82] given the long list of covariates under study. The PLS also handles the variation in the multilevel structure of the data. The evaluated covariates were already known to be associated with the $CD4^+$ count based on other statistical methods that were limited in some way or suffered from information loss due to grouping and details given in the introduction section. The predictive nature of the selected continuous covariates was beyond the scope of this work as our focus was on variable selection yet paving the way for such areas as predictive modelling with streamlined and richer continuous $CD4^+$ count clinical covariates. In this discussion we provided a brief summary of the functions of the selected and strongest 18 (out of 46) covariates according to our PLS model to point out the direction for future studies on the feasibility of incorporating them in the HIV treatment process to influence long-term $CD4^+$ cell response especially in an attempt to prolong the pre-treatment period and hence the likelihood of delaying the patients from experiencing the ART side effects, although the covariates can still be influential in the long-term $CD4^+$ cell response during therapy as previously reported [33, 34]. On our list of selected continuous clinical covariates, the lymphocytes were the strongest, as expected, because the $CD4^+$ cells are a T cell type [83] whereas the lymphocytes are either B or T cells [4, 56, 84]. Our results also showed the lymphocytes to have the highest independent positive correlation with the $CD4^+$ count ($r = 0.5421$, $p < 0.0001$). Hence, efforts to improve the $CD4^+$ cell response seem to be similar to those for the lymphocytes and the results obtained hereby serve to give an assurance of the effectiveness of our statistical methodology. In light of the other selected variables, our results showed the need to pay much attention to the white blood cells (basophils and monocytes) and platelet count. Basophils and monocytes control damage to body tissues and inflammation and fight pathogens, respectively [84]. Platelet count measures the blood clotting condition [84–87]. Although they are the least abundant leucocytes [88], our study has found basophils to explain the greatest variation in the $CD4^+$ count following the lymphocytes. However, the direct contact between human basophils and $CD4^+$ T cells is known to mediate viral transinfection of T cells through the formation of viral synapses [89, 90]. Also, the presence of basophils and other white blood cells in the blood is affected by underlying infection [91]. Areas of potential consideration in the blood chemistry group included potassium, sodium, calcium, magnesium, ALP and folate. Potassium regulates the acid-base chemistry and water balance [92], nerve impulses and heart muscle [84, 85]. Potassium's effect on the $CD4^+$ count is affected by underlying comorbidities [93]. Sodium and calcium regulate the water balance, blood pressure, blood volume, heart rhythm and most importantly the brain and nerve function [84, 85, 92]. Changes in the sodium concentration are known to create an osmotic gradient between the extra- and intracellular fluid in cells [94] suggesting that a proper balance is essential. Magnesium is involved in muscle contractions and protein processing [84], ALP in detecting liver health [1, 95, 96] and folate for cell growth and metabolism [97, 98]. Red blood cells indices [haematocrit, MCV, mean corpuscular haemoglobin concentration (MCHC) and red blood cells] are related to haemoglobin [99], which binds oxygen for transport to tissues and binds tissue carbon dioxide to transport it back for exhalation [100, 101]. The indices indicate the volume, concentration and proportions of red blood cells [101, 102]. Because volume contributes to the haematocrit, dehydration becomes a confounder of the $CD4^+$ count relationship. Details on patient dehydration were not available and this has not been taken into consideration in

this study. In line with the red blood cell indices, our results revealed that LDH also needs attention. LDH is a cytosolic enzyme that enables the fulfilment of short-term energy requirements in the absence of sufficient oxygen at the expense of a greater consumption of glucose cells [103]. Proteins (total protein, albumin and LDH) were included in the selected list for the maintenance of normal water distribution between the tissues and blood as well as acid-base balance [104].

It is important to acknowledge that there were some limitations to this study. Several variables that influence the clinical covariates may not have been included, for example, dehydration, underlying infection, comorbidities and patient dietary conditions, especially their effect on the biochemistry covariates. These are potentially important confounders that could have been adjusted. Furthermore, the study findings were limited to adult females. We recommend future studies to consider the effect of gender and age on the strongest $CD4^+$ count covariates during HIV disease progression. Given a large enough sample size, evaluating the clinical covariates for subjects with $CD4^+$ count < 250 cells/mm^3 is also recommended owing to the key driver for prophylaxis and surveillance for opportunistic infections related to $CD4^+$ count < 250 cells/mm^3.

CONCLUSION

Only a few of the many clinical attributes routinely collected during the HIV disease progression were found to be strong $CD4^+$ count covariates and mostly from electrolytes, proteins and red blood cells. Prolonging the pretreatment period of the HIV disease progression by effectively incorporating and managing the covariates for the long-term influence on the $CD4^+$ cell response has the potential to delay the challenges associated with ART side effects. Damage to body tissues and inflammation as indicated by basophils was found to be the strongest $CD4^+$ count covariate to effectively incorporate and manage for long-term influence on the $CD4^+$ cell response. Lipids, physical examination and anthropometric measurements are not worth considering as important drivers of the $CD4^+$ count when monitoring the health status of HIV-infected women during disease progression. There is a possibility of resource optimisation by streamlining the amount of routinely collected information when monitoring the health status of HIV-infected patients during the disease progression using just a few of the clinical attributes that strongly co-vary with the $CD4^+$ count.

ACKNOWLEDGMENTS

We thank the CAPRISA study teams and all participants for their important personal contribution to the availability of the data for the HIV research through their support and participation in the project.

Funding. No funding or sponsorship was received for this study or publication of this article. The article processing charges were funded by the authors.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole and have given their approval for this version to be published.

Disclosures. Partson Tinarwo, Temesgen Zewotir, Nonhlanhla Yende-Zuma, Nigel J. Garrett and Delia North have nothing to disclose.

Compliance with Ethics Guidelines. All procedures performed in studies involving human participants were in accordance with the local ethics committees of the University of KwaZulu-Natal, the University of Cape Town, the University of the Witwatersrand in Johannesburg, and the Prevention Sciences Review Committee (PSRC) of the Division of AIDS (DAIDS, National Institutes of Health, USA). The study was also performed in accordance with the Helsinki Declaration of 1964 and its

later amendments. The consent forms were translated into vernacular language, isiZulu, and written informed consent was obtained at each stage of the study. All the minors under the age of 18 years were excluded from the study as part of the screening procedure.

Data Availability. All data generated or analysed during this study are included in this published article as supplementary information files.

REFERENCES

1. Beare A, Stockinger H, Zola H, Nicholson I. The CD system of leukocyte surface molecules: monoclonal antibodies to human cell surface antigens. *Current Protocols in Immunology*. 2008;73(80):A.4A.1-A.4A.73.

2. Kwantwi LB, Tunu BK, Boateng D, Quansah DY. Body mass index, haemoglobin, and total lymphocyte count as a surrogate for CD4 count in resource limited settings. *Journal of Biomarkers*. 2017;Volume 2017, Article ID 790735

3. Alavi SM, Ahmadi F, Farhad M. Correlation between total lymphocyte count, hemoglobin, hematocrit and CD4 count in HIV/AIDS patients. Acta Medica Iranica. 2009;47(1):1–4.

4. Obirikorang C, Quaye L, Acheampong I. Total lymphocyte count as a surrogate marker for CD4 count in resource-limited settings. BMC Infect Dis. 2012;12:128.

5. Obirikorang C, Yeboah FA. Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting. J Biomed Sci. 2009;16(102).

6. Nzou C, Kambarami RA, Onyango FE, Ndhlovu CE, Chikwasha V. Clinical predictors of low CD4 count among HIV-infected pulmonary tuberculosis clients: A health facility-based survey. S Afr Med J. 2010;100:602–5.

7. Moolla Y, Moolla Z, Reddy T, Magula N. The use of readily available biomarkers to predict CD4 cell counts in HIV-infected individuals. South Afr Fam Pract. 2015;57(5):293–6.

8. Olawumi H, Olatunji P. The value of serum albumin in pretreatment assessment and monitoring of therapy in HIV/AIDS patients. HIV Med. 2006;7:351–5.

9. Secko D. Inexpensive CD4 counting for the developing world. CMA J. 2005;173(5):478.

10. Bentwich Z. CD4 Measurements in patients with HIV: are they feasible for poor settings? PLoS Med. 2005;2(7):e214.

11. Manoto SL, Lugongolo M, Govender U, Mthunzi-Kufa P. Point of care diagnostics for HIV in resource limited settings: An Overview. *MDPI* 2018;54(3).

12. Elsa Z. Healthcare systems in Sub-Saharan Africa: focusing on community-based delivery (CBD) of health services and the development of local research institutes. United Nations Peace and Progress. 2016;3(1):44–9.

13. Kuupiel D, Bawontuo V, Mashamba-Thompson TP. Improving the accessibility and efficiency of point-of-care diagnostics services in low-and middle-Income countries: Lean and Agile Supply Chain Management. *Diagnostics* 2017;7(58).

14. Leung N-HZ, Chen A, Yadav P, Gallien J. The impact of inventory management on stock-outs of essential drugs in sub-Saharan Africa: secondary analysis of a field experiment in Zambia. *PLoS One*. 2016;11(15):e0156026.

15. Jeffery A. The NHI proposal risking lives for no good reason. In: South African Institute of Race Relations, editor. South Africa: South African Institute of Race Relations; 2016.

16. Bateman C. Drug stock-outs: inept supply-chain management and corruption. S Afr Med J. 2013;103(9):600–2.

17. Nditunze L, Makuza S, Amoroso CL, Odhiambo J, Ntakirutimana E, Cedro L, et al. Assessment of essential medicines stock-outs at health centers in Burera District in Northern Rwanda. *Rwanda Journal Series F: Medicine and Health Sciences* 2015;2(1).

18. Thairu L, Katzenstein D, Israelski D. Operational challenges in delivering CD4 diagnostics in sub-Saharan Africa. AIDS Care. 2011;23(7):814–21.

19. Chip Lab. Rapid, label-free CD4 testing using a smartphone compatible device. The Royal Society of Chemistry. 2017;17:2910–9.

20. CMA Media Inc. Inexpensive CD4 counting for the developing world. *JAMC*. 2005;175(3):478.

21. Manabe YC, Wang Y, Elbireer A, Auerbach B, Castelnuovo B. Evaluation of portable point-of-care CD4 counter with high sensitivity for detecting patients eligible for antiretroviral therapy. *PLoS ONE* 2012;7(4).

22. Hunt PW, Deeks SG, Rodriguez B, Valdez H, Shade SB, Abrams DI, et al. Continued CD4 cell count increases in HIV-infected adults experiencing 4 years of viral suppression on antiretroviral therapy. AIDS. 2003;17:1907–15.

23. Egger M, Hirschel B, Francioli P, Sudre P, Wirz M, Flepp M, et al. Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study. BMJ. 1997;315:1194.

24. Smith CJ, Sabin CA, Youle MS, Kinloch-de Loes S, Lampe FC, Madge S, et al. Factors influencing increases in CD4 cell counts of HIV-positive persons receiving long-term highly active antiretroviral therapy. J Infect Dis. 2004;190:1860–8.

25. Yakubu T, Dedu VK, Bampoh PO. Factors affecting CD4 count response in HIV patients within 12 months of treatment: a case study of Tamale Teaching Hospital. American Journal of Medical and Biological Research. 2016;4(4):78–83.

26. Burch LS, Smith CJ, Anderson J, Sherr L, Rodger AJ, O'Connell R, et al. Socioeconomic status and treatment outcomes for individuals with HIV on antiretroviral treatment in the UK: cross-sectional and longitudinal analyses. Lancet Public Health. 2016;1:e26–36.

27. Bunyasi EW, Coetzee DJ. Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa. BMJ Open. 2017;7:e016232.

28. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "Optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829–35.

29. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. Stat Med. 2006;25(1):127–41.

30. Lawrence Erlbaum Associates. Continuous and discrete variables. Journal of Consumer Psychology. 2001;10(1&2):37–533.

31. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biology*. 2015;13(9).

32. Weston R, Marett B. HIV infection pathology and disease progression. Clinical Pharmacist. 2009;1:387.

33. Braconnier P, Delforge M, Garjau M, Wissing KM, De Wit S. Hyponatremia is a marker of disease severity in HIV-infected patients: a retrospective cohort study. *BMC Infectious Diseases* 2017;17:98.

34. Xu L, Ye H, Huang F, Yang Z, Zhu B, Xu Y, et al. Moderate/severe hyponatremia increases the risk of death among hospitalized Chinese human immunodeficiency virus/acquired immunodeficiency syndrome patients. PLoS ONE. 2014;9(10):1–9.

35. Gie Liem D, Miremadi F, Keast RSJ. Reducing sodium in foods: the effect on flavor. Nutrients. 2011;3:694–711.

36. Pravina P, Sayaji D, Avinash M. Calcium and its role in human body. International Journal of Research in Pharmaceutical and Biomedical Sciences. 2013;4(2):659–68.

37. Adhikari PM, Chowta MN, Ramapuram JT, Rao SB, Udupa K, Acharya SD. Effect of vitamin B12 and folic acid supplementation on neuropsychiatric symptoms and immune response in HIV-positive patients. J Neurosci Rural Pract. 2016;7(3):362–7.

38. Semeere AS, Nakanjako D, Ddungu H, Kambugu A, Manabe YC, Colebunders R. Sub-optimal vitamin B-12 levels among ART-naïve HIV-positive individuals in an urban cohort in Uganda. PLoS ONE. 2012;7(7):e40072.

39. Volberding PA, Levine AM, Dieterich D, Donna Mildvan, Mitsuyasu R, Saag M. Anemia in HIV infection: clinical impact and evidence-based management strategies. *Clinical Infectious Diseases* 2004;38:1454–63.

40. Butt AA, Michaels S, Greer D, Clark R, Kissinger P, Martin DH. Serum LDH level as a clue to the diagnosis of histoplasmosis. *The AIDS Read*. 2002;12(7).

41. Butt AA, Michaels S, Kissinger P. The association of serum lactate dehydrogenase level with selected opportunistic infections and HIV progression. International Journal of Infectious Diseases. 2002;6:178–81.

42. Sudfeld CR, Isanaka S, Aboud S, Mugusi FM, Wang M, Chalamilla GE, et al. Association of serum albumin concentration with mortality, morbidity, CD4 T-cell reconstitution among Tanzanians

initiating antiretroviral therapy. J Infect Dis. 2013;207:1370–8.

43. dos Santos ACO, Almeida AMR. Nutritional status and CD4 cell counts in patients with HIV/AIDS receiving antiretroviral therapy. Rev Soc Bras Med Trop. 2013;46(6):698–703.

44. Pralhadrao HS, Kant C, Phepale K, Mali MK, Raghunath. Role of serum albumin level compared to CD4+ cell count as a marker of immunosuppression in HIV infection. *Indian Journal of Basic and Applied Medical Research*. 2016;5(3):495-502.

45. Voss TG, Fermin CD, Levy JA, Vigh S, Choi B, Garry RF. Alteration of intracellular potassium and sodium concentrations correlates with induction of cytopathic effects by human immunodeficiency virus. J Virol. 1996;70(8):5447–544.

46. Choi B, Gatti PJ, Haislip AM, Fermin CD, Garry RF. Role of potassium in human immunodeficiency virus production and cytopathic effects. Virology. 1998;247:189–99.

47. Khaidukov SV, Litvinov IS. Calcium homeostasis change in CD4+ T lymphocytes from human peripheral blood during differentiation in vivo. Biochemistry (Moscow). 2005;70(6):692–702.

48. Bani-Sadr F, Lapidus N, Rosenthal E, Gerard L, Foltzer A, Perronne C, et al. Gamma glutamyl transferase elevation in HIV/hepatitis C virus-coinfected patients during interferon-ribavirin combination therapy. *J Acquir Immune Defic Syndr* 2009;50(4).

49. Fleischbeina E, O'Brienb J, Martelinoc R, Fenstersheibd M. Elevated alkaline phosphatase with raltegravir in a treatment experienced HIV patient. AIDS. 2008;22(17):2401–7.

50. Gomo E, Ndhlovu P, Vennervald B, Nyazema N, Friis H. Enumeration of CD4 and CD8 T-cells in HIV infection in Zimbabwe using a manual immunocytochemical method. Cent Afr J Med. 2001;47(3):64–70.

51. Dusingize JC, Hoover DR, Shi Q, Mutimura E, Rudakemwa E, Ndacyayisenga V, et al. Association of abnormal liver function parameters with HIV serostatus and CD4 count in antiretroviral-naive Rwandan women. Aids Research and Human Retroviruses. 2015;31(7):723–30.

52. Shiferaw MB, Tulu KT, Zegeye AM, Wubante AA. Liver enzymes abnormalities among highly active antiretroviral therapy experienced and HAART naïve HIV-1 infected patients at Debre Tabor Hospital, Northwest Ethiopia: a comparative cross-sectional study. *AIDS Research and Treatment*. 2016;Volume 2016, Article ID 19854

53. Vanisri H, Vadiraja N. Association between Red blood cell parameters and immune status in HIV infected males. Indian Journal of Pathology and Oncology. 2016;3(4):684–9.

54. Vanisri H, Vadiraja N. Relationship between red blood cell parameters and immune status in HIV infected females. Indian Journal of Pathology and Oncology. 2016;3(2):255–9.

55. Leticia OI, Ugochukwu A, Ifeanyi OE, Andrew A, Ifeoma UE. The correlation of values of CD4 count, platelet, PT, APTT, fibrinogen and factor VIII concentrations among HIV positive patients in FMC Owerri. *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*. 2014;13(9 Ver II):94-101.

56. Shapiro N, Karras DJ, Leech SH, Heilpern KL. Absolute lymphocyte count as a predictor of CD4 count. Ann Emerg Med. 1998;32(3):323–8.

57. Sivaram M, White A, Radcliffe K. Eosinophilia: clinical significance in HIV-infected individuals. Int J STD AIDS. 2012;23(9):635–8.

58. Iffen TS, Efobi H, Usoro CAO, Udonwa NE. Lipid profile of HIV-positive patients attending university of Calabar Teaching Hospital, Calabar - Nigeria. World Journal of Medical Sciences. 2010;5(4):89–93.

59. Oka F, Naito T, Oike M, Imai R, Saita M, Inui A, et al. Correlation between HIV disease and lipid metabolism in antiretroviral-naïve HIV-infected patients in Japan. J Infect Chemother. 2012;18:17–211.

60. Floris-Moore M, Howard A, Lo Y, Arnsten J, Santoro N, Schoenbaum E. Increased serum lipids are associated with higher CD4 lymphocyte count in HIV-infected women. HIV Medicine. 2006;7:421–30.

61. Misra R, Chandra P, Riechman SE, Long DM, Shinde S, Pownall HJ, et al. Relationship of ethnicity and CD4 Count with glucose metabolism among HIV patients on highly-active antiretroviral therapy (HAART). *BMC Endocrine Disorders* 2013;13(13).

62. Maganga E, Smart LR, Kalluvya S, Kataraihya JB, Saleh AM, Obeid L, et al. Glucose metabolism disorders, HIV and antiretroviral therapy among Tanzanian adults. *PLoS ONE* 2015;10(8:e0134410).

63. McKnight TR, Yoshihara HAI, Sitole LJ, Martin JN, Steffensd F, Meyer D. A combined chemometric and quantitative NMR analysis of HIV/AIDS serum discloses metabolic alterations associated with disease status. Mol BioSyst. 2014;10:2889–977.

64. Dannhauser A, van Staden A, van der Ryst E, Nel M, Marais N, Erasmus E, et al. Nutritional status of HIV-1 seropositive patients in the Free State Province of

South Africa: anthropometric and dietary profile. Eur J Clin Nutr. 1999;53:165–73.

65. Dimala CA, Kadia BM, Kemah B-L, Tindong M, Choukem S-P. Association between CD4 cell count and blood pressure and its variation with body mass index categories in HIV-infected patients. International Journal of Hypertension. 2018; Volume 2018, Article ID 1691474.

66. Esposito FM, Coutsoudis A, Visser J, Kindra G. Changes in body composition and other anthropometric measures of female subjects on highly active antiretroviral therapy (HAART): a pilot study in Kwazulu-Natal, South Africa. The Southern African Journal of HIV Medicine. 2008:36–42.

67. Fofana KC. Correlation between nutritional indicators and low CD4 count (<200 cells>/mm3) among HIV positive adults in Kapiri, Zambia [Thesis]: Georgia State University; 2016.

68. Venter E, Gericke G, Bekker P. Nutritional status, quality of life and CD4 cell count of adults living with HIV/AIDS in the Ga-Rankuwa area (South Africa). South African Journal of Clinical Nutrition. 2009;22(3):124–9.

69. Manner IW, Trøseid M, Oektedalen O, Baekken M, Os I. Low nadir CD4 cell count predicts sustained hypertension in HIV-infected individuals. The Journal of Clinical Hypertension. 2013;15(2):101–6.

70. Hsue PY, Hunt PW, Ho JE, Farah HH, Schnell A, Hoh R, et al. Impact of HIV infection on diastolic function and left ventricular mass. Circ Heart Fail. 2010;3(1):132–9.

71. Palacios R, Santos J, Garcı´a A, Castells E, González M, Ruiz J, et al. Impact of highly active antiretroviral therapy on blood pressure in HIV-infected patients. A prospective study in a cohort of naive patients. *HIV Medicine* 2006;7:10–5.

72. Sun J, and Reddy, C.K. Big Data Analytics for Healthcare. SIAM International Conference on Data Mining; Austin, TX.2013.

73. Meintjes G, Moorhouse MA, Carmona S, Davies N, Dlamini S, van Vuuren C, et al. Adult antiretroviral therapy guidelines 2017. Southern African Journal of HIV Medicine. 2017;18(1):a776.

74. Chen W-T, Shiu C-S, Yang JP, Simoni JM, Fredriksen-Goldsen kI, Szu-Hsien Lee T, et al. Side effects of antiretroviral therapy (ART) are associated with depression in Chinese individuals with HIV: a mixed methods study. *AIDS & Clinical Research*. 2013;4(6).

75. Lands L. A Practical Guide to HIV Drug Side Effects for People Living with HIV/AIDS. In: Pustil R, editor: The Canadian AIDS Treatment Information Exchange (CATIE); 2006.

76. Opiyo WO, Ng'wena AGM, Ofulla AVO. Liver function markers and associated serum electrolytes changes in HIV patients attending Patient Support Centre Of Jaramogi Oginga Odinga Teaching And Referral Hospital, Kisumu County, Kenya. *East African Medical Journal*. 2013;90(9):276-87.

77. Adhikari PMR, Chowta MN, Ramapuram JT, Rao SB, Udupa K, Acharya SD. Prevalence of vitamin B12 and folic acid deficiency in HIV-positive patients and its association with neuropsychiatric symptoms and immunological response. Indian Journal of Sexually Transmitted Diseases and AIDS. 2016;37(2):178–84.

78. Cohen AJ, Steigbigel RT. Eosinophilia in patients infected with human immunodeficiency virus. J Infect Dis. 1996;174:615–8.

79. Fasakin K, Omisakin C, Esan A, Adebara I, Owoseni I, Omoniyi D, et al. Total and CD4+ T-lymphocyte count correlation in newly diagnosed HIV patients in resource-limited setting. Journal of Medical Laboratory and Diagnosis. 2014;5(2):22–8.

80. Sen. LCS, Vyas A, Sanghi LCS, Shanmuganandan CK, Gupta CR, Kapila BK, et al. Correlation of CD4+ T cell count with total lymphocyte count, haemoglobin and erythrocyte sedimentation rate levels in human immunodeficiency virus type-l disease. *MJAFI* 2011;67(1):15-20.

81. van Loggerenberg F, Mlisana K, Williamson C, Auld SC, Morris L, Gray CM, et al. Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study. PLoS ONE. 2008;3(4):e1954.

82. Maitra S, Yan J. Principal component analysis and partial least squares: two dimension deduction techniques for regression. Casualty Actuarial Society Discussion Paper Program 2008. p. 79-90.

83. Papagno L, Spina CA, Marchant A, Salio M, Rufer N, Little S, et al. Immune activation and CD8+ T-cell differentiation towards senescence in HIV-1 infection. PLoS Biol. 2004;2(2):0173–185.

84. Project Inform. Monitoring HIV Blood Work: A Complete Guide for Monitoring HIV. In: New York State Department of Health, editor. San Francisco, CA 94103 26212007.

85. James AG. Understanding Blood Tests. In: The James, editor. Cancer Hospital and Richard J. Solove Research Institute.2017.

86. National Institutes of Health Clinical Center. Understanding your complete blood count (CBC)

and common blood deficiencies. In: National Institutes of Health Clinical Center, editor. Bethesda2015.

87. NAM. CD4, viral load & other tests. UK2012.

88. Min B, Brown MA, LeGros G. Understanding the roles of basophils: breaking dawn. The Journal of Cells, Molecules, Systems and Technologies. 2011;135:192–7.

89. Marone G, Varricchi G, Galdiero MR, Loffredo S, Rivellese F, de Paulis A. Are basophils and mast cells masters in HIV infection? Int Arch Allergy Immunol. 2016;171:158–65.

90. Jiang A-P, Jiang J-F, Guo M-G, Jin Y-M, Li Y-Y, Wanga J-H. Human blood-circulating basophils capture HIV-1 and mediate viral trans-infection of CD4+ T cells. J Virol. 2015;89:8050–62.

91. Siracusa MC, Kim BS, Spergel JM, Artis D. Basophils and allergic inflammation. J Allergy Clin Immunol. 2013;132(4):789–98.

92. The Johns Hopkins Lupus Center. Blood Chemistry Panel America2017 [Available from: https://www.hopkinslupus.org/lupus-tests/screening-laboratory-tests/blood-chemistry-panel/.

93. Collins AJ, Pitt B, McGaughey K, Reaven N, Wilson D, Funk S, et al. Association of serum potassium with all-cause mortality in patients with and without heart failure, chronic kidney disease, and/or diabetes. Am J Nephrol. 2017;46:213–21.

94. Shu Z, Tian Z, Chen J, Ma J, Abudureyimu A, Qian Q, et al. HIV/AIDS-related hyponatremia: an old but still serious problem. Ren Fail. 2018;40(1):68–74.

95. Whitfield J. Gamma glutamyl transferase. *Crit Rev Clin Lab Sci 2001 Aug;38(4):*. 2001;38(4):263-355.

96. Patil R, Kamble P, Raghuwanshi U. Serum ALP & GGT levels in HIV positive latients. International Journal of Recent Trends in Science And Technology. 2013;5(3):155–7.

97. Arya SS, Kumar PK. Folate: sources, production and bioavailability. Agro Food Industry Hi Tech. 2012;23(4):23–7.

98. Dieticians of Canada. Food Sources of Folate. In: Canadian Nutrient File Health Department, editor. Canada: Dieticians of Canada; 2014.

99. Junqueira L, Carneiro J, Kelley R. Basic Histology. A Lange Medical Book. 7th ed: Appleton and Lange; 2006.

100. Jensen FB, Fago A, Weber RE. Hemoglobin structure and function. Fish Physiology. 1998;17:1–40.

101. Wintrobe M, Greer J. Wintrobe's Clinical Hematology. Philadelphia.: Lippincott Williams & Wilkins, ; 2009.

102. Arika W, Nyamai D, Musila M, Ngugi M, Njagi E. Hematological markers of in vivo toxicity. *Journal of Hematology & Thromboembolic Diseases*. 2016;4(2).

103. Valvona CJ, Fillmore HL, Nunn PB, Pilkington GJ. The regulation and function of lactate dehydrogenase A: therapeutic potential in brain tumor. Brain Pathol. 2016;26:3–17.

104. Spectrum. Total protein: Biuret Reagent: Egyptian Company for Biotechnology (S.A.E); 2007.

HIV/AIDS - Research and Palliative Care

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Covariate random effects on the CD4 count variation during HIV disease progression in women

This article was published in the following Dove Press journal:
HIV/AIDS - Research and Palliative Care

Partson Tinarwo
Temesgen Zewotir
Delia North

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Purpose: To investigate the variation in CD4 count between HIV positive patients due to clinical covariates at each phase of the HIV disease progression.

Patients and methods: The Centre for the AIDS Programme of Research in South Africa (CAPRISA) conducted different studies in which female patients were initially enrolled in HIV negative cohorts (phase 1). Seroconverts were further followed-up weekly to fortnightly visits up to 3 months (phase 2: acute infection), monthly visits from 3 to 12 months (phase 3: early infection), quarterly visits thereafter (phase 4: established infection) until antiretroviral therapy (ART) initiation (phase 5).

Results: Eighteen out of the 46 CD4 count covariates investigated were significant. Low average CD4 counts at acute and early phase entry improved at a faster rate than entries at higher average CD4 count. During therapy, all the 18 covariates induced significantly different patients' average CD4 counts. The rate of change of CD4 count greatly varied in response to lactate dehydrogenase during the acute phase. Red blood cells increase resulted in the patients' CD4 counts approaching a common higher level during the early phase. During therapy, the already high CD4 counts improved faster than lower ones in response to the red blood cells increase. As the monocytes increased, patients with lower average CD4 counts became worse than those with higher average CD4 counts.

Conclusion: Changes in the covariates measurements either induced no variation effects in certain phases or improved the CD4 count at a faster rate for those patients whose average CD4 was already high or worsen the CD4 level which was already low or caused the patients' CD4 counts to approach the same level – higher or lower than the general cohort. The studied covariates induced wide variations in the CD4 count between HIV positive patients during the ART phase.

Keywords: parallel plot, redundant features, partial least squares, mixOmics, mixed models, between variation

Introduction

A human body is a complex machine that usually responds automatically to the changing internal and outside environments.[1] Although there are measurement reliabilities[2,3] in recording patient information, by nature, the repeated measurements from the same individual are bound to vary. Inasmuch as this variation exists within an individual and so does between any two given individuals. Regardless of this inter-individual variation, the health care fraternity generally administers an average dose of medication to patients irrespective of their differences in either the

Correspondence: Partson Tinarwo
Statistics, University of KwaZulu-Natal, Private Bag X 54001, Durban 4000, South Africa
Tel +27 31 260 3011
Fax +27 31 260 1009
Email partsont@gmail.com

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/HIV.S193652

body tolerances, specific needs, or preferred medical treatment. However, there ought to be some medical measurements that are likely to remain fairly the same across patients whilst others greatly fluctuating to bring about the individual or time uniqueness. There is a need to understand these components that vary widely among individuals to streamline the focus areas in providing specific treatment needs during patients' care.

Cohort studies,[4,5,6] especially in the context of HIV/AIDS, commonly record the CD4 cell count, the prime target of HIV,[7] for monitoring the HIV disease progression,[8] and hence the CD4 count being regarded as a health indicator.[8] Alongside the CD4 count, many other covariates have also been recorded and these include the full blood count,[9–15] lipids,[16–18] sugar,[19–21] blood chemistry,[10,22–39] and clinical examination.[40–49] However, an evaluation to determine the clinical covariates that bring the variation in the CD4 count between HIV patients during the disease progression has not been well documented. This gives an insight on the potential to manipulate and incorporate these influential CD4 count covariates to streamline the pathway to tailored medical attention for HIV-infected individuals at a specific HIV infection phase.

Previously, the associations of these covariates with the CD4 count have been analyzed with statistical methods that ranged from Pearson or Spearmen correlation analysis,[50,51] sensitivity, specificity and positive prediction,[52,53] linear regression[54,55] multivariate regression,[18] logistic regression,[26,45] Chi-Square tests,[28,29,56] non-parametric tests,[34,39] independent student *t*-tests,[57,58] confidence intervals,[40] the analysis of variance[59,60] to generalized estimating equations.[27] Their limitations include the inability to give the covariates an opportunity to compete in a single multidimensional model to identify the most influential ones and consequently assessing their effects on the CD4 count variation.

This study aimed to pool the covariates from five clinical platforms in order to identify the ones that bring the CD4 count variation between HIV positive patients at each phase of the HIV disease progression. Our first objective was to minimize multicollinearity among the covariates by using correlation analysis and the application of partial least squares as a multidimensional analysis approach to obtain the most salient CD4 covariates. A mixed model approach was then applied as the second objective to investigate the CD4 count variation between HIV positive patients in response to the covariate induced random effects using the data from CAPRISA studies.

Materials and methods

The study design

The CAPRISA 002 enrolled 245 HIV negative (Phase I: pre-HIV infection) female sex workers into an Acute Infection study. The establishment of the acute infection study, cohort screening and seroconverts; routine evaluation procedures, CAPRISA–participant interaction, and data management have been previously documented,[61] and was conducted in accordance with the Declaration of Helsinki. The study protocol and informed consent documents were reviewed and approved by the local ethics committees of the University of KwaZulu-Natal, the University of Cape Town, the University of the Witwatersrand in Johannesburg, and by the Prevention Sciences Review Committee (PSRC) of the Division of AIDS (DAIDS, National Institutes of Health, USA). The consent forms were translated into vernacular language, isiZulu and written informed consent obtained at each stage of the study. All the minors, under the age of 18 years were excluded from the study as part of the screening procedure. The HIV negative cohort was followed up and upon HIV infection, they were further followed-up weekly to fortnightly visits up to 3 months (Phase II: acute infection), monthly visits from 3 to 12 months (Phase III: early infection), quarterly visits thereafter (Phase IV: established infection) until antiretroviral therapy (ART) initiation (Phase V). Eventually, 27 seroconversions were recorded at the end of the study of an average period of 4.5 years. In addition to the 27 seroconverts, 210 more patients who seroconverted from other CAPRISA studies were also enrolled and similarly followed up postinfection from acute to ART phase. Figure 1 summarizes how the total sample size of 237 seroconverts for this study was obtained.

The data

Table 1 shows the studied repeated number of measurements per individual at each phase. Four-time points prior to each phase transition were selected and that resulted in a total of 16 repeated measurements being investigated for each patient. The baseline, pre-HIV (Phase I) repeated measurements were scarce and hence, this study focused on Phases II–V only. The CD4 count is the response variable and the routinely collected information on the covariates (c1–c46) consists of full blood count, biochemistry, sugar, lipids, physical examination, and anthropometric measurements. The raw data for the study are available as Supplementary material (File S1).

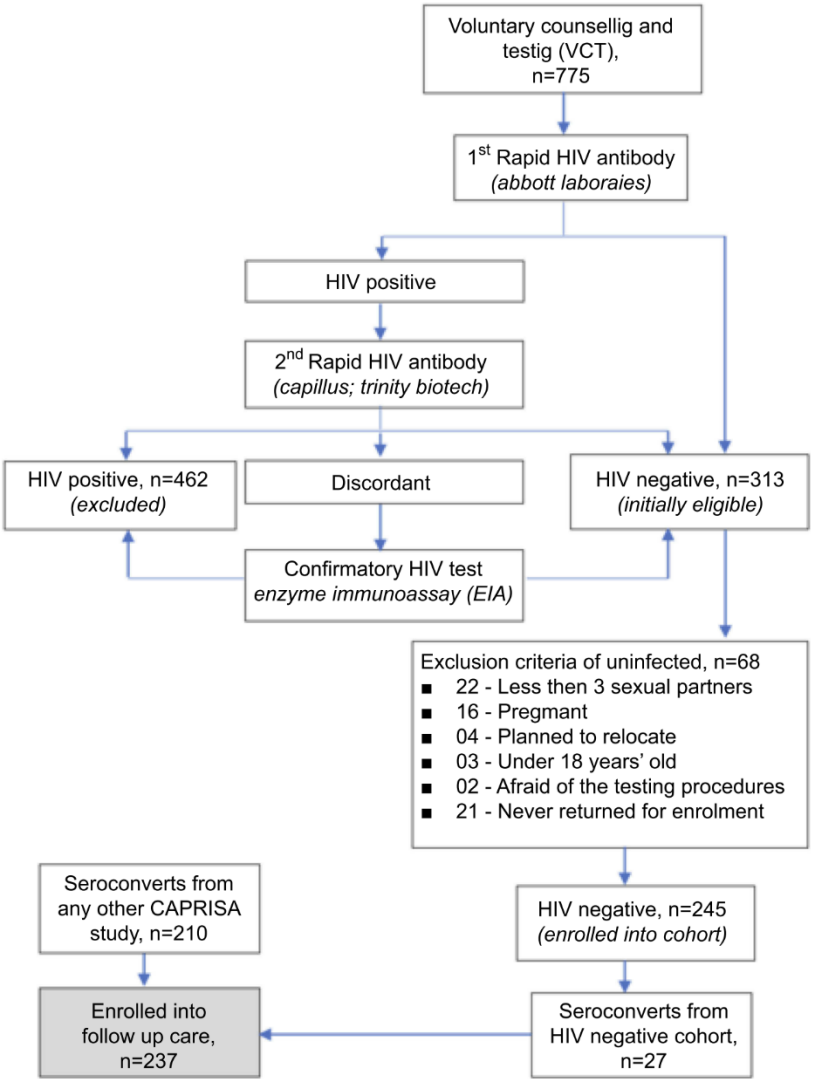


Figure 1 The recruitment of the 237 study participants. The HIV negative cohort screening involved 775 voluntary potential candidates of which 462 were already HIV positive and 313 initially eligible. Of the 313 HIV negative, only 245 were enrolled and the rest excluded for various reasons according to the eligibility criteria. Eventually, 27 out of the 245 seroconverted and enrolled into follow-up care. Seroconverts from other CAPRISA studies (210) were also included in the follow-up care that resulted in a total of 237 patients for this study.

Statistical analysis

The statistical analysis first considered a parallel plot overview of the variations in the repeated measurements around their respective means within each phase followed by dropping off of redundant features. This involved dropping off the variables with the highest mean absolute correlation using the find correlation function at the initial variable reduction stage and then another further second variable reduction stage involving the selection of the important variables using the PLS with the application of the split function in the library mixOmics. The library mixOmics is capable of handling the complex structure of repeated measurements and incorporates a design matrix to account for variation in the multilevel structure of the longitudinal data. Both the variable reduction functions were used in open source R software, version 3.5.0. Finally, a mixed model was applied using SAS 9.4 PROC HPMIXED and PROC MIXED to a reduced set of the CD4 count covariates. The model with an unstructured variance was appropriate to estimate the intercept-slope covariance and the repeated measurements took an autoregressive moving average correlation structure of [ARMA (1,1)]. Each covariate in the reduced set of the covariates was mean centered to obtain the intercepts for the average patient and scaled for estimates comparison.

Table 1 The studied number of repeated measurements per individual

	Phase:	**2-Acute**				**3-Early**				**4-Est**				**5-ART**			
	Time:	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n	T_{n-3}	T_{n-2}	T_{n-1}	T_n
ID	Variable																
01	CD4	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
01	c01	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
01	c02	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
01	C46	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
02	CD4	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
02	c01	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
02	c02	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
02	c46	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
237	CD4	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
237	c01	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
237	c02	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
237	c46	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I

Abbreviations: Est, established; ART, antiretroviral therapy; T, time; c, covariate

Results

The variations in the cohort's repeated measurements

Around the mean within each phase were presented in a parallel plot for phase comparison (Figure 2). The greatest variation in the CD4 count was observed during the established phase and the lowest during therapy. The higher CD4 count variation was associated with the highest variation in all the white blood cells. However, when the CD4 count varied the least during the ART phase, there was a corresponding low variation in all the red blood cell count components. Our data seem to show complex relationships and variations in the CD4 count and its covariates during the different phases of the HIV disease progression.

Variable reduction

The results showed that of the 46 covariates that were available for investigation,18 were found to be the strongest and none of these were from lipids, physical examination nor anthropometric category (Table S1 and Figure S1). The 18 significant CD4 count covariates were further used to fit the mixed models in which each patient was allowed to have own CD4 count trajectory in response to each of the covariates.

General trends within the phase

CD4 count general trends against each covariate within phase. Table 2 shows the results of the mixed model in which the marginal (fixed) effects indicate the cohort's general CD4 count trajectories in response to the covariates within each phase. All the significant trends are in bold. Lymphocyte increase was associated with an improved CD4 count throughout the phases of the HIV

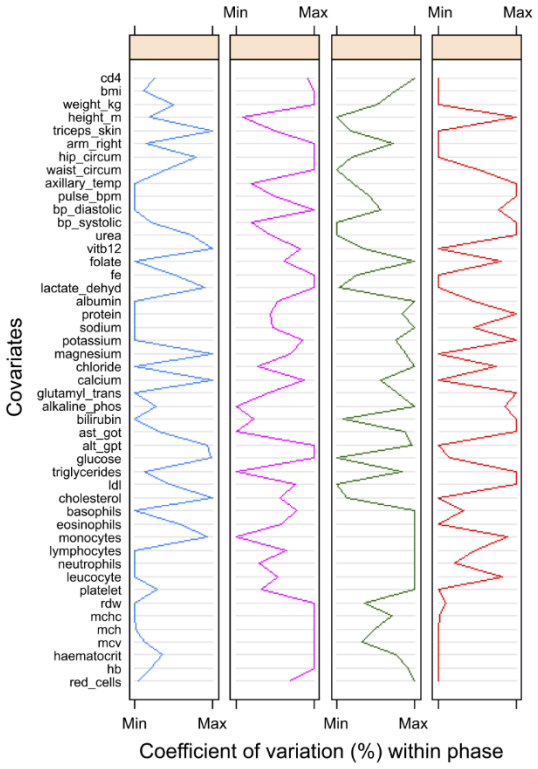


Figure 2 The coefficients of variation (CV). The CVs give information about the spread of the repeated measurements around the mean. The colour codes represent Phase II (blue), Phase III (pink), Phase IV (green) and Phase V (red). **Abbreviations:** BMI, body mass indixex; bp, blood pressure; ALT_GPT, Alamine Aminotransferase_Glutamate Pyruvate Transaminase; AST_GOT, Aspartate Aminotransferase_Glutamate Oxaloactate Transaminase; LDL, Low density lipoprotein; RDW, red blood cell distribution width; MCHC, mean corpuscular haemoglobin concentration; MCH, mean corpuscular haemoglobin; MCV, mean corpscular volume; Hb, haemoglobin.

Table 2 Fixed effects – the cohort's general trajectories within each phase

	2-Acute	3-Early	4-Est	5-ART
Effect	**Estimate(Pr > |t|)**	**Estimate(Pr > |t|)**	**Estimate(Pr > |t|)**	**Estimate(Pr > |t|)**
Phase	−52.6627(**0.0053**)	−53.3344(**0.0013**)	−34.3791(**0.0339**)	(ref)616.0300(<0.**0001**)
Time*phase	8.9971(**0.0376**)	0.6577(0.8770)	−12.1440(**0.0008**)	5.1856(0.1572)
Albumin*phase	25.9973(**0.0044**)	29.4461(<0.**0001**)	29.6052(<0.**0001**)	11.7354(0.0725)
Alkaline phosphatase*phase	−2.9375(0.7416)	10.6827(0.1402)	25.1476(**0.0002**)	14.3366(**0.0044**)
Basophils*phase	−0.4112(0.9439)	0.0747(0.9905)	7.2633(0.3959)	16.8130(**0.0454**)
Calcium*phase	8.1087(0.4206)	−6.3913(0.3409)	−9.4794(0.1370)	−5.9208(0.3670)
Folate*phase	−43.7866(<0.**0001**)	−20.1953(**0.0075**)	−16.2681(**0.0214**)	−46.6532(<0.**0001**)
Glucose*phase	2.7933(0.7398)	3.4673(0.5911)	−1.0686(0.8407)	4.8128(0.4096)
Haematocrit*phase	−20.6596(0.6311)	−12.1918(0.7407)	−2.3535(0.9300)	3.1107(0.9308)
LDH*phase	−9.6857(0.3913)	10.6586(0.1616)	−0.2509(0.9706)	−1.7578(0.8041)
Lymphocytes*phase	102.5100(<0.**0001**)	127.1300(<0.**0001**)	128.1800(<0.**0001**)	165.7000(<0.**0001**)
Magnesium*phase	4.0267(0.6886)	3.5717(0.6022)	−9.2187(0.1098)	13.8217(0.0507)
MCHC*phase	−13.5170(0.0728)	12.5703(**0.0409**)	−5.2952(0.3453)	16.1918(**0.0095**)
MCV*phase	52.9572(0.1077)	56.5794(**0.0388**)	30.7462(0.1373)	11.3666(0.6600)
Monocytes*phase	−3.2578(0.6058)	−10.1268(0.1212)	−18.5394(**0.0018**)	−18.7442(**0.0016**)
Platelet*phase	28.4224(**0.0002**)	12.7530(0.0773)	36.6385(<0.**0001**)	16.1257(**0.0291**)
Potassium*phase	−1.8457(0.8039)	−3.2011(0.4461)	7.0560(0.3034)	1.6780(0.6404)
Protein*phase	−30.8203(**0.0015**)	−39.5359(<0.**0001**)	−29.0654(<0.**0001**)	−13.3719(**0.0394**)
Red blood cells*phase	38.2319(0.3902)	18.2958(0.6231)	14.2265(0.6069)	−1.9031(0.9609)
Sodium*phase	−19.2748(**0.0148**)	−14.6409(**0.0177**)	−7.2344(0.1810)	6.0560(0.2789)

Notes: *The interaction between the clinical covariate and the HIV infection phase. Bold *p*-value indicates significant change in the $CD4^+$ count due to the covariate increase.

disease progression whereas folate and protein increase resulted in a decline of the CD4 count at each phase. Before treatment, an increase in albumin improved the CD4 count by almost the same magnitude, whereas basophils increase could only have a significant positive effect on the CD4 count during therapy. The CD4 count improved with an increase in alkaline phosphatase (ALP) during the established and ART phases with more improvement at the established phase. Contrary to ALP behavior, it was during the established and ART phases where the monocytes indicated a negative impact on the CD4 count. The platelet count showed positive effects on the CD4 count in all the stages except the early phase. Our results also showed that it was in this early phase only where the mean corpuscular volume increase significantly improved the CD4 count. The mean corpuscular hemoglobin concentration (MCHC) also indicated a positive association with the CD4 count in the early phase and then during the ART as well. The results revealed that an increase in sodium content soon after HIV infection (acute and early phases) was associated with a CD4 decline. Our data show that over time within the acute phase, the CD4 count increased by 8.9971 cells/mm^3 (*p*-value =0.0376) at each visit and dropped by 12.1440 cells/mm^3 (*p*-value =0.0008) at each visit during the established phase. Our mixed model estimated that the ART phase records were on average of 616.03 cells/mm^3 of CD4 count and those from the acute phase being 52.6627 cells/mm^3 below that of the ART average. Table 3 shows that the ART phase was at least 45 cells/mm^3 of CD4 count above that of any other investigated phase. All the average CD4 counts from the other phases before therapy (acute to established) were found not to be significantly different from each other.

Random effects due to each covariate

We further investigated the random effects due to each covariate by allowing each patient to have own CD4 count trajectory with intercept and slope. This improved the Akaike Information Criterion in the modeling of the CD4 count.

Time within each phase was also considered as a covariate. The variations in the intercepts (intr) and slopes of individual patient's CD4 counts against *time* are presented in Table 4. Also shown are the relationships between the patients' intercepts and slopes within each phase. The variations were then expressed as percentages of the total variation captured by the model. For the *time* covariate, the results show that there was greater variation among patients' average CD4 counts upon entering the acute phase (17.9147%, *p*-value =0.0012) followed by the variations in the CD4 counts recorded at the beginning of the ART phase (15.8941%, *p*-value =0.0008). The intercepts and the slopes were negatively related at the acute and early phases in which the CD4 counts had upward

Table 3 Least squares means and differences

Least squares means								
Effect	**Phase**			**Estimate**	**Standard error**	**DF**	**t Value**	**Pr > |t|**
Phase	5-ART			623.81	11.5931	3,712	53.81	<0.0001
Phase	4-Est			563.43	9.1055	3,712	61.88	<0.0001
Phase	3-Early			563.68	8.3091	3,712	67.84	<0.0001
Phase	2-Acute			576.86	12.9588	3,712	44.52	<0.0001
Least squares means differences								
Effect	Phase	_phase		Estimate	Standard error	DF	*t* Value	Pr > |t|
Phase	5-ART	4-Est		60.3735	14.7414	3,712	4.1	<0.0001
Phase	5-ART	3-Early		60.1262	14.2633	3,712	4.22	<0.0001
Phase	5-ART	2-Acute		46.9455	17.3876	3,712	2.7	0.0070
Phase	4-Est	3-Early		−0.2473	12.3269	3,712	−0.02	0.9840
Phase	4-Est	2-Acute		−13.428	15.8379	3,712	−0.85	0.3966
Phase	3-Early	2-Acute		−13.1807	15.3939	3,712	−0.86	0.3919

Table 4 Covariance parameter test of time effect and the proportions

Subject	**Phase**	**Covariance parameter**	**Estimate**	**Estimate (%)**	**Standard error**	**Z Value**	***p* Value**
Patient	2-Acute	Intr	8,011.37	17.9147	2,637.70	3.04	**0.0012**
Patient	2-Acute	Intr-Slope	−2,565.41	−5.7367	991.25	−2.59	**0.0097**
Patient	2-Acute	Slope	1,864.63	4.1696	518.92	3.59	**0.0002**
Patient	3-Early	Intr	0.0000	0.0000	–	–	–
Patient	3-Early	Intr-Slope	−1,384.33	−3.0956	451.95	−3.06	**0.0022**
Patient	3-Early	Slope	1,925.76	4.3063	441.40	4.36	**0.0001**
Patient	4-Est	Intr	0.0000	0.0000	–	–	–
Patient	4-Est	Intr-Slope	78.6057	0.1758	443.99	0.18	0.8595
Patient	4-Est	Slope	682.38	1.5259	381.74	1.79	**0.0369**
Patient	5-ART	Intr	7107.74	15.8941	2,244.07	3.17	**0.0008**
Patient	5-ART	Intr-Slope	40.3178	0.0902	555.11	0.07	0.9421
Patient	5-ART	Slope	0.0000	0.0000	–	–	–
Patient	AR	Rho	0.9439	0.0021	0.01113	84.79	**0.0001**
Patient	MA	Gamma	0.4378	0.0010	0.02703	16.20	**0.0001**
–	–	Residual	28.957	64.7526	1,340.19	21.61	**0.0001**
				100.0000			

Note: Bold *p*-value indicates the significant variation between patients.
Abbreviations: AR, Autoregressive; MA, Moving average.

trends of 11.6459 and 3.3582, respectively. This suggests that over time all the patients' CD4 count trajectories during the acute and early phases approached a higher focal level. This phenomenon indicates that the patients who entered the acute and early phases at lower CD4 count had their counts increasing at a faster rate than those who entered with a higher CD4 count already. Eventually, all the patients' CD4 counts approached the same higher CD4 count level. Similar estimate proportions of the intercepts and slope relationships for the other covariates are presented in Figure 3 where the intercepts represent the average CD4 counts at the mean covariate value (mean centered). The trajectory slopes are the rates of CD4 count change as the values of the covariates measurements increase.

Variations in the average CD4 counts as induced by each covariate. Figure 3 (intercept variation) shows that given the average values of folate, LDH, lymphocytes, and magnesium at the acute phase, there was no significant difference in the CD4 counts for all the 237 patients under study. The same phenomenon was also

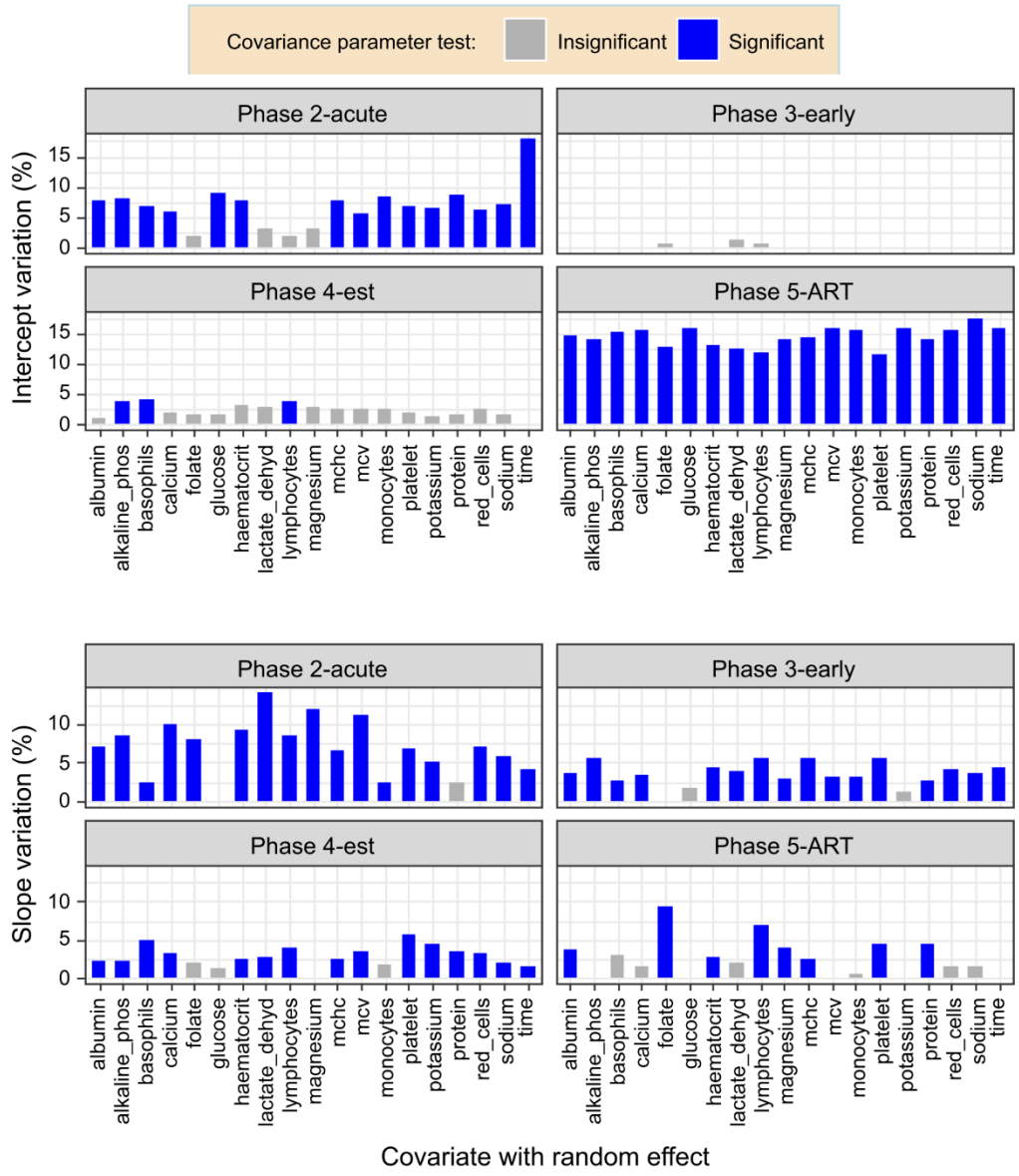


Figure 3 Proportion of variation in intercepts and slopes. The fixed effects parameters are identical, and each covariate at a time was allowed to have a random effect. Different variance parameter estimates were obtained for each phase (group) and these were expressed as a percentage of the total variation including the ARMA (1,1) and residuals.

observed during the early phase where there was no significant difference in the average CD4 counts for all the patients in response to each of the studied covariates. This was almost the same situation at the established phase except for the significant differences in the patients' average CD4 counts at the mean values of ALP, basophils, and lymphocytes. The results also show that upon taking medication, all the patients' average CD4 counts were significantly different from each other. Generally, the patients' average CD4 counts did not vary too much in response to the covariates during the early and established phases. Wide variations in the average CD4 counts were observed during the acute and ART phases.

Variations in the rate of CD4 count change as induced by each covariate. We further explored the variations in the rates of the CD4 count change in response to the increase in the values of each covariate. Figure 3 (slope variation) shows that the rate of CD4 count change in response to each of the covariates varied among the patients mostly from the acute to the established phase. The acute phase was characterized by no significant difference in the CD4 count rate of change in response to the increase in glucose and protein. Similarly, in the early phase, folate, glucose, and potassium did not induce any differences in the rate of change of CD4 count among the patients. During the established phase, an increase in the folate, glucose, magnesium, and monocytes resulted in no significant difference in the CD4 count rate of change among

all the patients. However, upon taking medication, more than half of the covariates were associated with similar rates of CD4 count change among the patients. The greatest variation in the rate of CD4 count change was observed soon after infection (acute phase) in which an increase in the LDH induced the widest variations in the CD4 count rate of change between the patients. This was followed by folate during the ART phase.

Correlation between random intercepts and slopes of CD4 count trajectories. Throughout the post-HIV infection follow up period, there was a positive relationship (r >0.80) between the intercepts and slopes of the CD4 count trajectories against lymphocytes (Figure 4 and Table 5). This indicates that at each phase of the HIV disease progression, an increase in lymphocytes resulted in the patients whose average CD4 counts that were already high to increase at a faster rate than those whose average CD4 counts were lower. The CD4 count trajectories against red blood cells (ART phase), LDH (early phase), basophils (established and ART phases), platelets (established phase), and sodium (ART phase) showed an upward trend with positive intercept and slope correlations. This means that, as these covariates increase within the indicated phases, the patients with higher average CD4 counts had their CD4 counts increasing at a faster rate than those who had lower CD4 counts. The cohort's CD4 count trajectory against monocytes was heading downwards during the ART phase with positive intercept–slope relationships. This indicated that as the monocytes increased, patients with lower average CD4 counts became worse than those with higher average CD4 counts. On the other hand, there was a negative relationship (covtest, p-value =0.0297, Figure 4) between the average CD4 counts and their rate of change with red blood cells during the early phase. This early phase's CD4 count and red blood cells trajectories followed a general upward trend suggesting that as the red blood cells increase, all the patients' CD4 counts approached a common higher CD4 count level than the cohort's average. That is, red blood cell increase during the uptake of medication, resulted in the patients whose CD4 count that was higher to increase even faster than those whose count was lower.

Discussion

The investigated data from the CAPRISA studies showed complex relationships and variations in the CD4 count and its covariates during the different phases of the HIV disease progression. The cohort's repeated measurements for the CD4 count varied widely around their mean within the established phase and narrowly during the ART phase. All

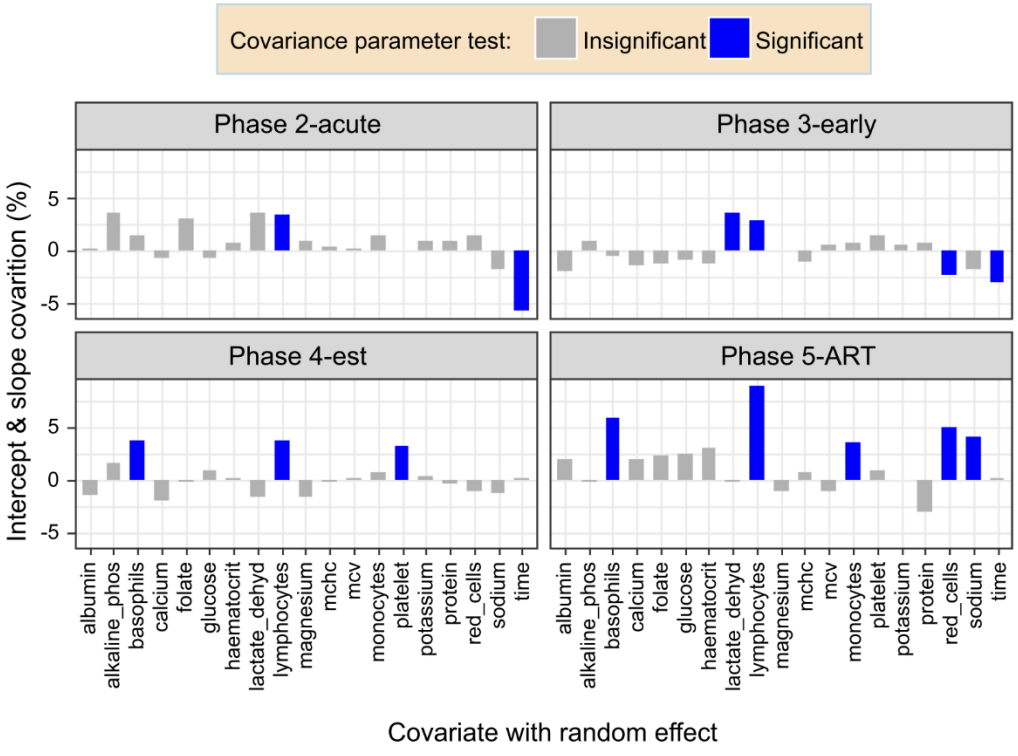


Figure 4 Proportion of variation in intercept and slope covariations. The fixed effects parameters are identical, and each covariate at a time was allowed to have a random effect. Different covariance parameter estimates were obtained for each phase (group) and these were expressed as a percentage of the total variation including the autoregressive of order 1 and moving average of order 1 (ARMA (1,1)) and residuals.

Table 5 Correlation between intercept and slope

	2-Acute	3-Early	4-Est	5-ART
Covariate	Corr(covtest)	Corr(covtest)	Corr(covtest)	Corr(covtest)
Time	−0.6638(**0.0097**)	0.0000(**0.0022**)†	0.0000(0.8595)	0.0000(0.9421)
Albumin	0.0222(0.9354)	0.0000(0.0589)	−0.9695(0.1951)	0.2702(0.2788)
Alkaline phosphatase	0.4280(0.0930)	0.0000(0.4897)	0.5497(0.2165)	0.0000(0.8407)
Basophils	0.3238(0.3120)	0.0000(0.5522)	0.8030(**0.0422**)	0.8775(**0.0104**)
Calcium	−0.1085(0.7203)	0.0000(0.2056)	−0.8544(0.0905)	0.4093(0.3003)
Folate	0.8445(0.0840)	0.0000(0.1124)	−0.1260(0.8364)	0.2060(0.4763)
Glucose	0.0000(0.6623)	0.0000(0.3932)	0.5831(0.4188)	0.0000(0.2206)
Haematocrit	0.0846(0.7095)	0.0000(0.2133)	0.0680(0.8709)	0.5053(0.0776)
LDH	0.5580(0.1187)	1.0000(**0.0050**)	−0.5951(0.2352)	−0.0552(0.8932)
Lymphocytes	0.8498(**0.0060**)	1.0000(**0.0033**)	0.9767(**0.0005**)	0.9803(**0.0001**)
Magnesium	0.1499(0.6507)	0.0000(0.9944)	−1.0000(0.1406)	−0.1519(0.5945)
MCHC	0.0528(0.8257)	0.0000(0.3265)	−0.0588(0.8901)	0.1009(0.7069)
MCV	0.0189(0.9448)	0.0000(0.6571)	0.0730(0.8570)	0.0000(0.4749)
Monocytes	0.2954(0.3582)	0.0000(0.6132)	0.3425(0.5015)	1.0000(**0.0247**)
Platelet	−0.0176(0.9416)	1.0000(0.2117)	0.9625(**0.0097**)	0.1234(0.5933)
Potassium	0.1456(0.7031)	0.0000(0.6158)	0.1396(0.8302)	0.0000(0.9515)
Protein	0.1995(0.6021)	0.0000(0.5694)	−0.1578(0.7586)	−0.3955(0.1034)
Red blood cells	0.2022(0.4729)	-	−0.3736(0.3356)	1.0000(**0.0059**)
Sodium	−0.2855(0.3131)	0.0000(0.0879)	−0.7696(0.2077)	0.8205(**0.0203**)

Notes: † The intercept variation in Table 4 was zero but covariance significant, hence the intercept and slope correlation zero. Bold *p*-value indicates the significant correlation between the intercept and slope.

the red blood cell count components were also found to narrowly vary during the ART phase as compared to the other phases. Only 18 of the 46 CD4 count covariates that were available for investigation were significant and consequently considered for further investigation.

There was great variation in the patients' average CD4 counts upon entering the acute and ART phases explaining the patients' immune responses to viral invasion and treatment, respectively. This is likely to be attributed to the high level of inter-individual diversity of the human system which is also affected by different factors.[62] An increase in the measurements or quantities of the covariates was found to change the CD4 count either for the better or worse in certain patients and in some cases causing the patients' CD4 counts to approach a common level which was higher or lower than that of the cohort. The random effects due to the covariates were either widely varying or showed no significant difference in the CD4 counts in the phases of the HIV disease progression. During the acute phase, the mean values of folate, lactate LDH, lymphocytes, and magnesium corresponded to similar CD4 count levels for all the patients. These results revealed that on average the patients' CD4 counts were not affected by the demand for cell growth and metabolism (folate[63,64]), glucose conversion (LDH[65]), and muscle contractions and protein processing (magnesium[66]). The CD4 cells are T cells[67] which are also part of the lymphocytes, and the results showed that on average the CD4 count did not significantly differ between patients during the acute and the early phases, given the average lymphocytes count. However, during the established and ART phases, our data showed that the average lymphocytes count (total B and T cells[14,66,68]), were associated with significantly different average CD4 levels among the patients. With the exception of the early phase, the indicators of damage to body tissues and inflammation (basophil[66]) and liver health (ALP[8,69,70]) were on average significantly inducing CD4 count variations among the patients. Our data show that it is not only during therapy where treatment interferes with the biochemical properties among HIV patients as found by[71] during a six-month treatment period study. We observed that liver damage was one of the most common biochemical associated with significant CD4 count variation throughout the HIV disease progression except during the early phase. Hence, it then turns out that on average, tissue damage indicators were associated with the CD4 count variation in most of the phases. However, our data further revealed that upon taking medication which

significantly improved the CD4 count than any other phase, all the 18 covariates induced wide variations in the patients' average CD4 counts. HIV treatment is known to affect the clinical attributes[72] which could consequently be the attributing factor to the CD4 count variations in response to all the 18 covariates in our data during the ART phase. This is because treatment has proved to be effective but also increasingly complex due to new developing syndromes.[73]

Our results showed that all the patients' CD4 counts changed at different rates in response to each of the covariates upon taking medication. An increase in glucose and protein did not bring about variation in the rate of change of the CD4 counts between patients during the acute phase. Early phase CD4 counts also changed at the same rate when either of folate, glucose or potassium increase. Similarly, folate, glucose, magnesium, and monocytes increase at the established phase gave rise to the same rate of the CD4 count change. Most of the covariates induced wide variations in the rate of the CD4 count change during the acute phase. Our data showed that lymphocytes increase in every phase resulted in patients whose CD4 count was already higher increasing even faster than those patients with lower average CD4 counts. Similarly, patients with higher CD4 counts were found to have their count increasing at a much faster rate as the following covariates increase in certain phases: LDH (early phase), basophils (established and ART phases), platelets (established phase), and sodium (ART phase). During the early phase of the disease progression, the patients whose average CD4 count that was lower, increased at a faster rate in response to the red blood cells increase such that all the patients' CD4 counts eventually approached a common higher CD4 count. However, upon taking medication, an increase in the red blood cell count resulted in those individuals whose CD4 count which was already higher to become even much better as compared to the ones that were lower. Our data show that red blood cells that are packed with hemoglobin[74] and plays a role in the respiratory process[75,76] are associated with CD4 count improvement. Monocytes increase during medication (ART phase) resulted in CD4 counts that were lower to become much worse than for those patients whose average CD4 counts that were higher. Although monocytes are infected together with the $CD4^+$ T cells,[77] our data show that during therapy, monocytes were spared more than the $CD4^+$ T cells.

Conclusions

Of the many CD4 count covariates that have been suggested in the previous studies, only a few were found to be significantly associated with the CD4 count variation at the different phases of the HIV disease progression. These few covariates induced either wide variations in the patients' average CD4 counts in some infection phases and show no effect in the others. An increase in the measurements or quantities of the covariates was found to either improve the CD4 count at a faster rate for those patients whose average CD4 was already high or worsen the CD4 level which was already lower than that of the other patients. In some cases, the increase in the covariates values caused the patients' CD4 counts to approach a common level which was lower or higher than that of the general cohort. Tissue damage indicators were the most common covariates associated with CD4 count variation between patients. Patients who entered either the acute or early phases with lower average CD4 counts had their count increasing at a faster rate than those who entered with a higher CD4 count already resulting in the cohort approaching a common higher CD4 count. In addition to other treatment measures, the manipulation of selected CD4 count covariates for patients within a specific phase can usefully augment tailored methods for monitoring HIV patients using the CD4 count. Generally, the studied covariates induced wide variations in the CD4 count between HIV positive patients during the ART phase.

Acknowledgments

This study would not have been a success without the assistance of Nonhlanhla Yende-Zuma and Nigel J. Garrett for their assistance in making the data available. Our gratitude also goes to the teams for the different CAPRISA studies and all participants for their important personal contribution to the availability of the data for the HIV research through their support and participation in the projects.

Disclosure

The authors report no conflicts of interest in this work.

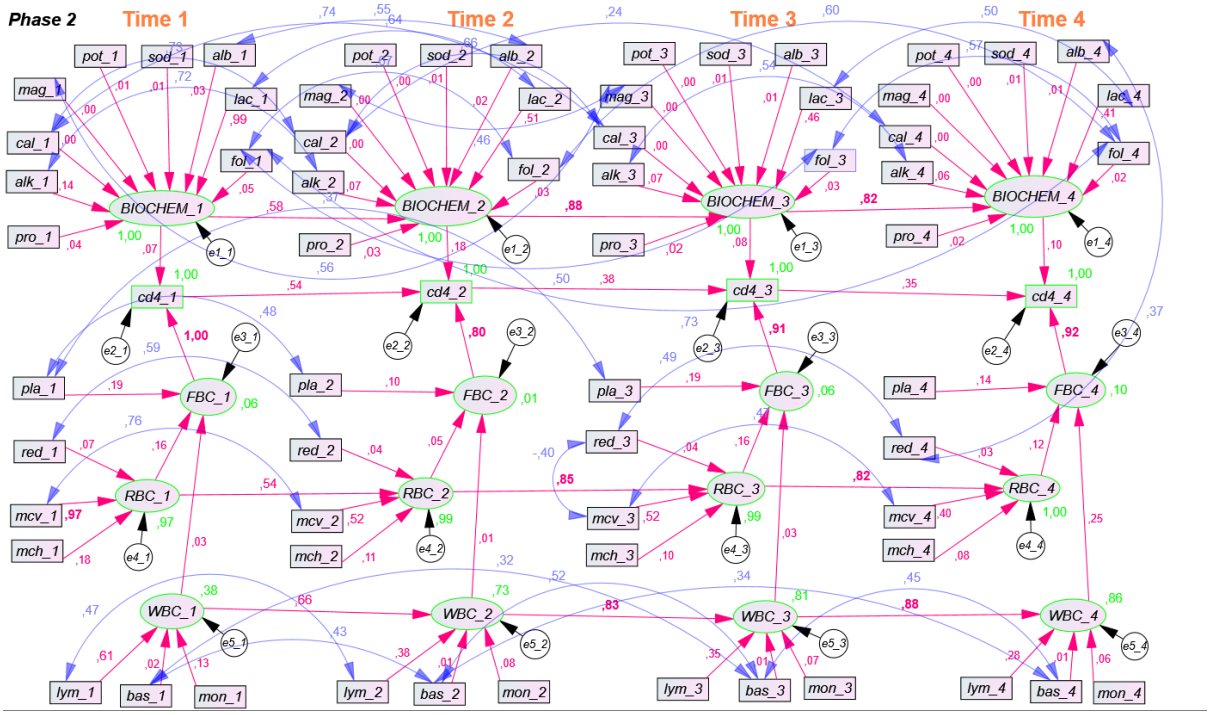
References

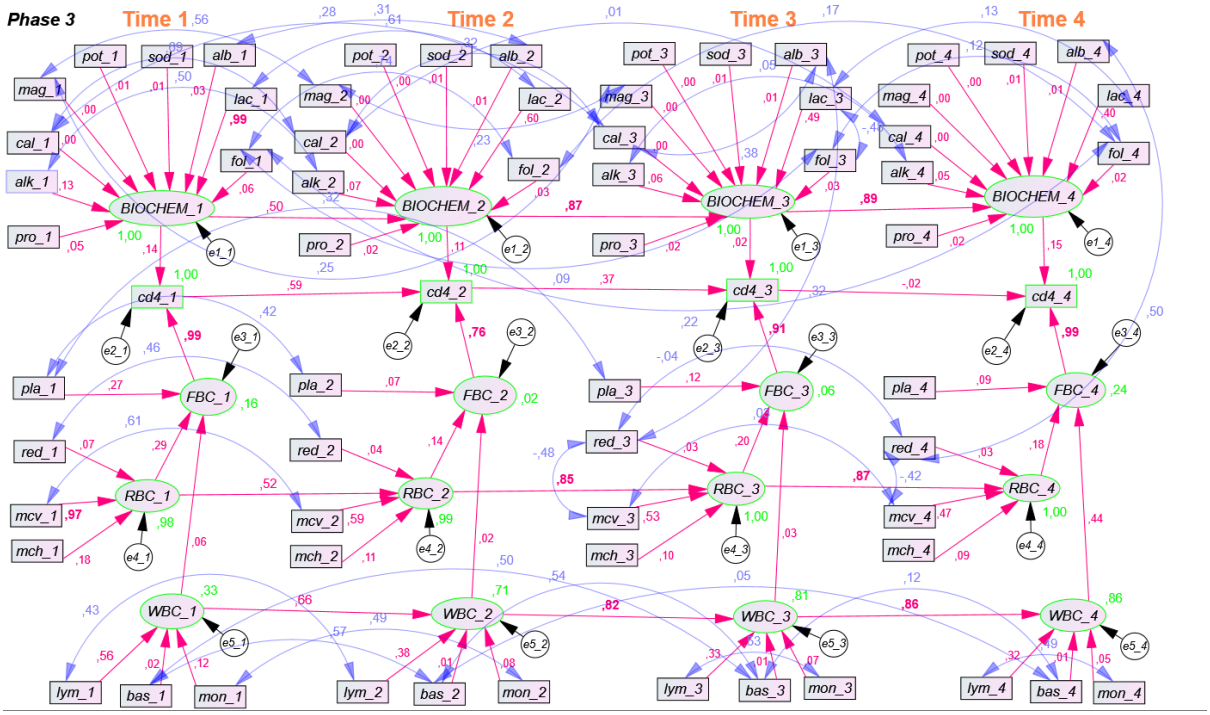
1. Kelly G. Body temperature variability (Part 1): a review of the history of body temperature and its variability due to site selection, biological rhythms, fitness, and aging. *Altern Med Rev*. 2006;11(4):278–293.
2. Mohajan HK. Two criteria for good measurements in research: validity and reliability. *Ann Spiru Haret Univ*. 2017;17(3):58–82.

3. Bajpai S, Bajpai R. Goodness of measurement: reliability and validity. *Int J Med Sci Public Health*. 2014;3(1):173–176. doi:10.5455/ijmsph.2013.191120133
4. Alexander LK, Lopes B, Ricchetti-Masterson K, Yeatts KB. Cohort Studies; 2015.
5. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. *Nephron Clin Pract*. 2009;113:c214–c217. doi:10.1159/000235241
6. Goldstein H. Longitudinal studies and the measurement of change. *J Royal Statl Soc Ser D*. 1968;18(2):93–117.
7. Weston R, Marett B. HIV infection pathology and disease progression. *Clin Pharm*. 2009;1:387.
8. Beare A, Stockinger H, Zola H, Nicholson I. The CD system of leukocyte surface molecules: monoclonal antibodies to human cell surface antigens. *Curr Protoc Immunol*. 2008;73(80):A.4A.1–A.4A.73.
9. Vanisri H, Vadiraja N. Association between red blood cell parameters and immune status in HIV infected males. *Indian J Pathol Oncol*. 2016;3(4):684–689. doi:10.5958/2394-6792.2016.00127.7
10. Obirikorang C, Yeboah FA. Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting. *J Biomed Sci*. 2009;16:102. doi:10.1186/1423-0127-16-102
11. Vanisri H, Vadiraja N. Relationship between red blood cell parameters and immune status in HIV infected females. *Indian J Pathol Oncol*. 2016;3(2):255–259. doi:10.5958/2394-6792.2016.00049.1
12. Leticia OI, Ugochukwu A, Ifeanyi OE, Andrew A, Ifeoma UE. The correlation of values of CD4 count, platelet, Pt, Aptt, fibrinogen and factor VIII concentrations among HIV positive patients in FMC owerri. *IOSR J Dent Med Sci*. 2014;13(9Ver II):94–101. doi:10.9790/0853-139294101
13. Alavi SM, Ahmadi F, Farhad M. Correlation between total lymphocyte count, hemoglobin, hematocrit and CD4 count in HIV/AIDS patients. *Acta Med Iran*. 2009;47(1):1–4.
14. Shapiro N, Karras DJ, Leech SH, Heilpern KL. Absolute lymphocyte count as a predictor of CD4 count. *Ann Emerg Med*. 1998;32(3):323–328.
15. Sivaram M, White A, Radcliffe K. Eosinophilia: clinical significance in HIV-infected individuals. *Int J STD AIDS*. 2012;23(9):635–638. doi:10.1258/ijsa.2012.011409
16. Iffen TS, Efobi H, Usoro CAO, Udonwa NE. Lipid profile of HIV-positive patients attending university of Calabar teaching hospital, Calabar - Nigeria. *World J Medl Sci*. 2010;5(4):89–93.
17. Oka F, Naito T, Oike M, et al. Correlation between HIV disease and lipid metabolism in antiretroviral-naı¨ve HIV-infected patients in Japan. *J Infect Chemother*. 2012;18:17–21. doi:10.1007/s10156-011-0275-5
18. Floris-Moore M, Howard A, Lo Y, Arnsten J, Santoro N, Schoenbaum E. Increased serum lipids are associated with higher CD4 lymphocyte count in HIV-infected women. *HIV Med*. 2006;7:421–430. doi:10.1111/hiv.2006.7.issue-7
19. Misra R, Chandra P, Riechman SE, et al. Relationship of ethnicity and CD4 count with glucose metabolism among HIV patients on highly-active antiretroviral therapy (HAART). *BMC Endocr Disord*. 2013;13:13. doi:10.1186/1472-6823-13-13
20. Maganga E, Smart LR, Kalluvya S, et al. Glucose metabolism disorders, HIV and antiretroviral therapy among tanzanian adults. *PLoS One*. 2015;10((8):e0134410. doi:10.1371/journal.pone.0134410
21. McKnight TR, Yoshihara HAI, Sitole LJ, Martin JN, Steffensd F, Meyer D. A combined chemometric and quantitative NMR analysis of HIV/AIDS serum discloses metabolic alterations associated with disease status. *Mol Biosyst*. 2014;10:2889–2897. doi:10.1039/c4mb00347k
22. Adhikari PM, Chowta MN, Ramapuram JT, Rao SB, Udupa K, Acharya SD. Effect of vitamin B12 and folic acid supplementation on neuropsychiatric symptoms and immune response in HIV-positive patients. *J Neurosci Rural Pract*. 2016;7(3):362–367. doi:10.4103/0976-3147.182774
23. Semeere AS, Nakanjako D, Ddungu H, Kambugu A, Manabe YC, Colebunders R. Sub-optimal vitamin B-12 levels among ART-Naı¨ve HIV-positive individuals in an Urban Cohort in Uganda. *PLoS One*. 2012;7(7):e40072. doi:10.1371/journal.pone.0040072
24. Volberding PA, Levine AM, Dieterich D, Donna M, Mitsuyasu R, Saag M. Anemia in HIV infection: clinical impact and evidence-based management strategies. *Clin Infect Dis*. 2004;38:1454–1463. doi:10.1086/383031
25. Butt AA, Michaels S, Greer D, Clark R, Kissinger P, Martin DH. Serum LDH level as a clue to the diagnosis of histoplasmosis. *The AIDS Read*. 2002;12:7.
26. Butt AA, Michaels S, Kissinger P. The association of serum lactate dehydrogenase level with selected opportunistic infections and HIV progression. *Int JInfect Dis*. 2002;6:178–181.
27. Sudfeld CR, Isanaka S, Aboud S, et al. Association of serum albumin concentration with mortality, morbidity, CD4 T-cell reconstitution among tanzanians initiating antiretroviral therapy. *J Infect Dis*. 2013;207:1370–1378. doi:10.1093/infdis/jit027
28. Moolla Y, Moolla Z, Reddy T, Magula N. The use of readily available biomarkers to predict CD4 cell counts in HIV-infected individuals. *South Afr Family Pract*. 2015;57(5):293–296. doi:10.1080/20786190.2015.1073895
29. Dos Santos ACO, Almeida AMR. Nutritional status and CD4 cell counts in patients with HIV/AIDS receiving antiretroviral therapy. *Rev Soc Bras Med Trop*. 2013;46(6):698–703. doi:10.1590/0037-8682-0125-2013
30. Pralhadrao HS, Kant C, Phepale K, Mali MK, Raghunath. Role of serum albumin level compared to CD4+ cell count as a marker of immunosuppression in HIV infection. *Indian J Basic Appl Med Res*. 2016;5(3):495–502.
31. Voss TG, Fermin CD, Levy JA, Vigh S, Choi B, Garry RF. Alteration of intracellular potassium and sodium concentrations correlates with induction of cytopathic effects by human immunodeficiency virus. *J VirolL*. 1996;70(8):5447–5454.
32. Choi B, Gatti PJ, Haislip AM, Fermin CD, Garry RF. Role of potassium in human immunodeficiency virus production and cytopathic effects. *Virology*. 1998;247:189–199. doi:10.1006/viro.1998.9251
33. Khaidukov SV, Litvinov IS. Calcium homeostasis change in CD4+ T lymphocytes from human peripheral blood during differentiation in vivo. *Biochemistry (Moscow)*. 2005;70(6):692–702. doi:10.1007/s10541-005-0170-8
34. Braconnier P, Delforge M, Garjau M, Wissing KM, De Wit S. Hyponatremia is a marker of disease severity in HIV-infected patients: a retrospective cohort study. *BMC Infect Dis*. 2017;17:98. doi:10.1186/s12879-017-2191-5
35. Bani-Sadr F, Lapidus N, Rosenthal E, et al. Gamma glutamyl transferase elevation in HIV/Hepatitis C virus–coinfected patients during interferon–ribavirin combination therapy. *J Acquir Immune Defic Syndr*. 2009;50:4. doi:10.1097/QAI.0b013e31819a2429
36. Fleischbeina E, O'Brienb J, Martelinoc R, Fenstersheibd M. Elevated alkaline phosphatase with raltegravir in a treatment experienced HIV patient. *Aids*. 2008;22(17):2401–2407.
37. Gomo E, Ndhlovu P, Vennervald B, Nyazema N, Friis H. Enumeration of CD4 and CD8 T-cells in HIV infection in Zimbabwe using a manual immunocytochemical method. *Cent Afr J Med*. 2001;47(3):64–70.
38. Dusingize JC, Hoover DR, Shi Q, et al. Association of abnormal liver function parameters with HIV serostatus and CD4 count in antiretroviral-naive rwandan women. *AIDS Res Hum Retroviruses*. 2015;31(7):723–730. doi:10.1089/aid.2014.0170
39. Shiferaw MB, Tulu KT, Zegeye AM, Wubante AA. Liver enzymes abnormalities among highly active antiretroviral therapy experienced and HAART Na\ve HIV-1 infected patients at Debre Tabor hospital, NorthWest Ethiopia: a comparative cross-sectional study. *AIDS Res Treat*. 2016;2016:Article ID 1985452. doi:10.1155/2016/1985452

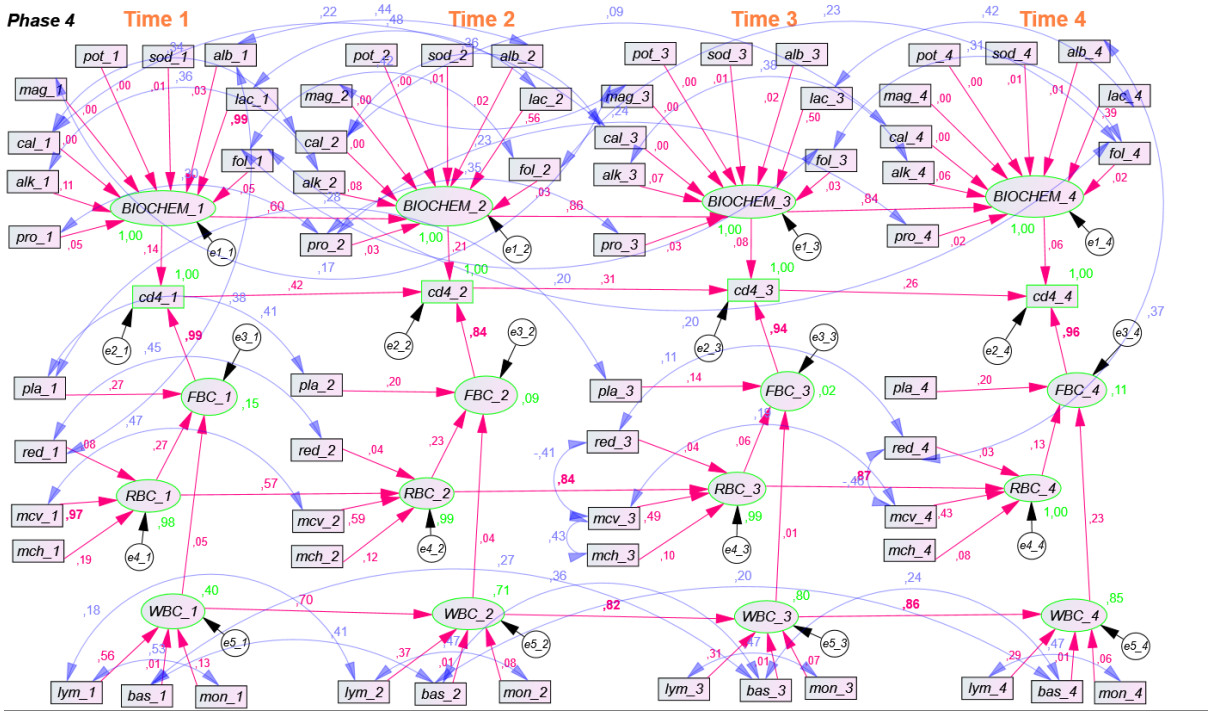
40. Dannhauser A, van Staden A, van der Ryst E, et al. Ntritional status of HIV-1 seropositive patients in the free State Province of South Africa: anthropometric and dietary profile. *Eur J Clin Nutr*. 1999;53:165–173.
41. Dimala CA, Kadia BM, Kemah B-L, Tindong M, Choukem S-P. Association between CD4 cell count and blood pressure and its variation with body mass index categories in HIV-infected patients. *Int J Hypertens*. 2018;2018:Article ID 1691474. doi:10.1155/2018/1691474
42. Nzou C, Kambarami RA, Onyango FE, Ndhlovu CE, Chikwasha V. Clinical predictors of low CD4 count among HIV-infected pulmonary tuberculosis clients: a health facility-based survey. *S Afr Med J*. 2010;100:602–605.
43. Kwantwi LB, Tunu BK, Boateng D, Quansah DY. Body mass index, haemoglobin, and total lymphocyte count as a surrogate for CD4 count in resource limited settings. *Journal of Biomarkers*. 2017;2017: Article ID 7907352. doi:10.1155/2017/7907352
44. Esposito FM, Coutsoudis A, Visser J, Kindra G. Changes in body composition and other anthropometric measures of female subjects on Highly active antiretroviral therapy (HAART): a pilot study in Kwazulu-Natal, South Africa. *Southern Afr J HIV Med*. 2008;9 (4):36–42.
45. Fofana KC. *Correlation Between Nutritional Indicators and Low CD4 Count (<200 cells>/mm3) among HIV Positive Adults in Kapiri, Zambia 2008–2009* [Thesis]. Graduate Faculty of Georgia State University, Georgia State University; 2016.
46. Venter E, Gericke G, Bekker P. Nutritional status, quality of life and CD4 cell count of adults living with HIV/AIDS in the Ga-Rankuwa area (South Africa). *South Af J Clin Nutr*. 2009;22(3):124–129. doi:10.1080/16070658.2009.11734233
47. Manner IW, Trøseid M, Oektedalen O, Baekken M, Os I. Low Nadir CD4 cell count predicts sustained hypertension in HIV-infected individuals. *J Clin Hypertens*. 2013;15(2):101–106. doi:10.1111/jch.12029
48. Hsue PY, Hunt PW, Ho JE, et al. Impact of HIV infection on diastolic function and left ventricular mass. *Circ Heart Fail*. 2010;3 (1):132–139. doi:10.1161/CIRCHEARTFAILURE.109.854943
49. Palacios R, Santos J, Garcı´A A, et al. Impact of highly active antiretroviral therapy on blood pressure in HIV-infected patients. A prospective study in a cohort of naive patients. *HIV Med*. 2006;7:10–15. doi:10.1111/j.1468-1293.2005.00333.x
50. Chorba TL, Maurice C, Nkengasong J, Maran M, Roels TH, Djomand G. Assessing eosinophil count as a marker of immune activation among human immunodeficiency virus–infected persons in sub-Saharan Africa. *Clin Infect Dis*. 2002;34:1264–1266. doi:10.1086/339940
51. Lumbanraja S, Siregar D Association between red blood cell indices and CD4 count in HIV-positive reproductive women. *IOP Conference Series: Earth and Environmental Science*; 2018:125(012027).
52. Olawumi H, Olatunji P. The value of serum albumin in pretreatment assessment and monitoring of therapy in HIV/AIDS patients. *HIV Med*. 2006;7:351–355. doi:10.1111/hiv.2006.7.issue-6
53. Sen LCS, Vyas A, Sanghi LCS, et al. Correlation of CD4+ T cell count with total lymphocyte count, haemoglobin and erythrocyte sedimentation rate levels in human immunodeficiency virus Type-l disease. *MJAFI*. 2011;67(1):15–20.
54. Daka D, Loha E. Relationship between Total lymphocyte count (TLC) and CD4 count among peoples living with HIV, Southern Ethiopia: a retrospective evaluation. *AIDS Res Ther*. 2008;5:26. doi:10.1186/1742-6405-5-26
55. Fasakin K, Omisakin C, Esan A, et al. Total and CD4+ T- lymphocyte count correlation in newly diagnosed HIV patients in resource-limited setting. *J Med Lab Diagnosis*. 2014;5(2):22–28. doi:10.5897/JMLD2014.0088
56. Cohen AJ, Steigbigel RT. Eosinophilia in patients infected with human immunodeficiency virus. *J Infect Dis*. 1996;174:615–618.
57. Atere AD, Akinbo BD, Okafor AM-J, Egbuchulem KI, Akinola EA. Evaluating correlation between total lymphocyte counts and CD4 counts in monitoring HIV patients. *Arch Appl Sci Res*. 2016;8(3):22–28.
58. Abdollahi A, Saffar H, Shoar S, Jafari S. Is total lymphocyte count a predictor for CD4 cell count in initiation antiretroviral therapy in HIV-infected patients? *Niger Med J*. 2014;55(4):289–293. doi:10.4103/0300-1652.137187
59. Opiyo WO, Ng'wena AGM, Ofulla AVO. Liver function markers and associated serum electrolytes changes in HIV patients attending patient support centre of Jaramogi Oginga Odinga teaching and referral hospital, Kisumu County, Kenya. *East Afr Med J*. 2013;90(9):276–287.
60. Adhikari PMR, Chowta MN, Ramapuram JT, Rao SB, Udupa K, Acharya SD. Prevalence of vitamin B12 and folic acid deficiency in HIV-positive patients and its association with neuropsychiatric symptoms and immunological response. *Indian J Sexually Transmitted Dis AIDS*. 2016;37(2):178–184. doi:10.4103/0253-7184.192117
61. van Loggerenberg F, Mlisana K, Williamson C, et al. Establishing a cohort at high risk of HIV infection in South Africa: challenges and experiences of the CAPRISA 002 acute infection study. *PLoS One*. 2008;3(4):e1954. doi:10.1371/journal.pone.0001954
62. Adrian L, Edward JC, Michelle AL. Shaping variation in the human immune system. *Trends Immunol*. 2016;37(10):637–646. doi:10.1016/j.it.2016.08.002
63. Arya SS, Kumar PK. Folate: sources, production and bioavailability. *Agro Food Ind Hi Tech*. 2012;23(4):23–27.
64. Dieticians of Canada. Food sources of folate. In: *Canadian Nutrient File Health Department*. ed. Canada: Dieticians of Canada; 2014.
65. Valvona CJ, Fillmore HL, Nunn PB, Pilkington GJ. The regulation and function of lactate dehydrogenase a: therapeutic potential in brain tumor. *Brain Pathol*. 2016;26:3–17. doi:10.1111/bpa.12299
66. Project Inform. Monitoring HIV blood work: a complete guide for monitoring HIV. In: *New York State Department of Health*, ed. San Francisco, CA: Project Inform, 2007; 94103 26212007.
67. Papagno L, Spina CA, Marchant A, et al. Immune activation and CD8+ T-cell differentiation towards senescence in HIV-1 infection. *PLoS Biol*. 2004;2(2):0173–0185. doi:10.1371/journal.pbio.0020020
68. Obirikorang C, Quaye L, Acheampong I. Total lymphocyte count as a surrogate marker for CD4 count in resource-limited settings. *BMC Infect Dis*. 2012;12:128. doi:10.1186/1471-2334-12-166
69. Whitfield J. Gamma glutamyl transferase. *Crit Rev Clin Lab Sci*. 2001;38(4):263–355. doi:10.1080/20014091084227
70. Patil R, Kamble P, Raghuwanshi U. Serum ALP & GGT levels in HIV positive patients. *Int J Recent Trends Sci Techy*. 2013;5(3):155–157.
71. Mgogwe J, Semvua H, Msangi R, Mataro C, Kajeguka D, Chilongola J. The evolution of haematological and biochemical indices in HIV patients during a six-month treatment period. *Afr Health Sci*. 2012;12(1):2–7.
72. Ibeh BO, Omodamiro OD, Ibeh U, Habu JB. Biochemical and haematological changes in HIV subjects receiving winnicure antiretroviral drug in Nigeria. *J Biomed Sci*. 2013;20(1):73. doi:10.1186/1423-0127-20-73
73. Montessori V, Press N, Harris M, Akagi L, Montaner JSG. Adverse effects of antiretroviral therapy for HIV infection. *Cmaj*. 2004;170 (2):229–238.
74. Junqueira L, Carneiro J, Kelley R. *Basic Histology. A Lange Medical Book*. 7th ed. Appleton and Lange; USA; 2006.
75. Jensen FB, Fago A, Weber RE. Hemoglobin structure and function. *Fish Physiol*. 1998;17:1–40.
76. Wintrobe M, Greer J. *Wintrobe's Clinical Hematology*. Philadelphia.: Lippincott Williams & Wilkins; 2009.
77. Pasupathi P, Bakthavathsalam G, Saravanan G, Devaraj A. Changes in CD4+ cell count, lipid profile and liver enzymes in HIV infection and AIDS patients. *J Appl Biomed*. 2008;6:139–145.

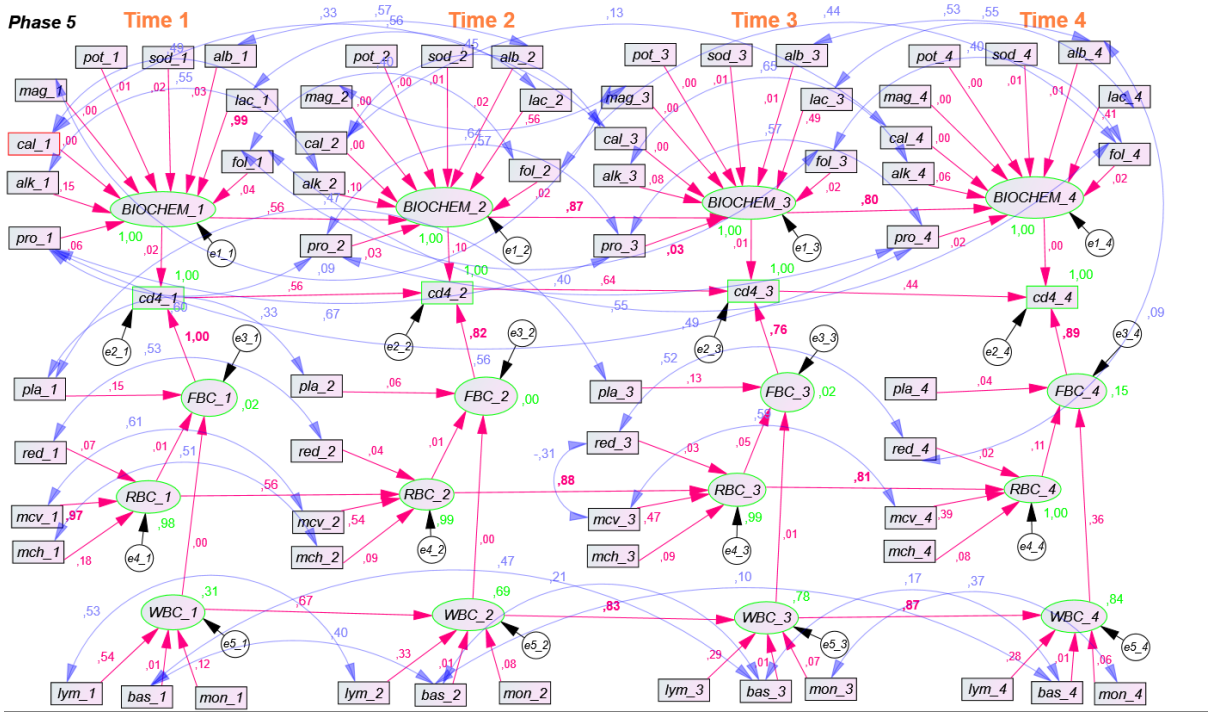
Appendix B: Additional results





The path analysis diagrams showing enhanced double headed arrows for correlation





The path analysis diagrams showing enhanced double headed arrows for correlation

Appendix C: Programming codes

C2: Chapter 2 codes

```
############################
#                          #
#      DATA EXPLORATION    #
#                          #
############################

## CLEAN UP EVERYTHING ##

rm(list= ls()[!(ls() %in% c(""))]) # clear environment and leave the selected
if(!is.null(dev.list())) dev.off() # clear all plots
cat("\014")                        # clear console
options(prompt = "R>")             # customize prompt

##################################
#Prepare  FIGURE 2.3 : BOXPLOTS  #
##################################

if(!require(ggplot2)){install.packages("ggplot2")}
load(file ="C:/Users/PARTSON/dataerror237.rda")
# Boxplot of all variables scaled
myformat_allvariables_scaled=(
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=10, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=9, face="plain",angle=0),
    axis.text.x = element_text(colour="black",size=8,angle=90,hjust = 1.0,vjust = 0.3),
    axis.title.y = element_text(color="black", size=9, face="plain"),
    axis.text.y = element_text(colour="black",size=8,angle=0),
    legend.position="top",
    legend.background = element_rect(fill="white",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(data.table)){install.packages("data.table")}
errorexplo = dataerror237
indivdistr = errorexplo
indivdistr1 =as.data.frame(indivdistr[,1])
indivdistr2 <- data.frame(lapply(indivdistr[,5:51],
    function(x) {
              x_unique <- unique(x[!is.na(x)])
              x_ranks  <- rank(x_unique)
              ifelse(is.na(x),x,
              round((x_ranks[match(x,x_unique)]-1)/(max(x_ranks)-1)*100))
                 }))
indivdistr12s = data.frame(indivdistr1, indivdistr2)
setnames(indivdistr12s, "indivdistr...1.", "patientid")
                          # Possible problem: rename this
if(!require(reshape2)){install.packages("reshape2")
indivdistr12m <- melt(indivdistr12s, id.var = "patientid")
indivdistr12m<-indivdistr12m[complete.cases(indivdistr12m$value),]

#=============================================================================#
#plot FIGURE 2.3                                                              #
#=============================================================================#

if(!require(ggplot2)){install.packages("ggplot2")}
box all =
  ggplot(data = indivdistr12m, aes(x=variable, y=value)) +
  geom_boxplot(fill = "cyan")+
  #coord_flip()+
  labs(title ="" ,x = "Clinical measurements"
       , y = "Percentile (0=min, 100=max)"
       )+
  myformat_allvariables_scaled

box_all
```

```
#Prepare FIGURE 2.4 : paired boxplots of consecutive phases

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/dataerror237.rda")
errorexplo = dataerror237

#-------Scale the data
indivdistr = errorexplo
indivdistr1 =as.data.frame(indivdistr[,c(1,3)])
indivdistr2 <- data.frame(lapply(indivdistr[,5:51],
    function(x) {
              x_unique <- unique(x[!is.na(x)])
              x_ranks  <- rank(x_unique)
              ifelse(is.na(x),x,
              round((x_ranks[match(x,x_unique)]-1)/(max(x_ranks)-1)*100))
                 }))
indivdistr = data.frame(indivdistr1, indivdistr2)
scaled =indivdistr[,c(2:49)]
#setnames(indivdistr12s, "indivdistr...1.", "patientid")
                            # Possible problem: rename this

if(!require(ggplot2)){install.packages("ggplot2")}
myformat=(
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=10, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=9, face="plain",angle=0),
    axis.text.x = element_text(color="black",size=8,angle=0),
    axis.title.y = element_text(color="black", size=9, face="plain"),
    axis.text.y = element_text(color="black",size=8,angle=0),
    legend.position="top",
    legend.background = element_rect(fill="white",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

rm(list= ls()[!(ls() %in% c("scaled","indivdistrcd4_scaled","errorexplo","myformat",
"myformat_cd4_by_infectionstage"))])

#---------boxplots by infectionstage 23
if(!require(dplyr)){install.packages("dplyr")}

datastage_23 = filter(scaled[,],
                  phase =="2-Acute" |
                  phase =="3-Early"
                  )
datastage_23s = datastage_23
if(!require(reshape2)){install.packages("reshape2")}

datastage_23m <- melt(datastage_23s, id.var = "phase")
datastage_23m<-datastage_23m[complete.cases(datastage_23m$value),]

if(!require(ggplot2)){install.packages("ggplot2")}
  ggplot(data = datastage_23m, aes(x=variable, y=value, fill=phase)) +
  geom_boxplot()+
  scale_fill_manual(values=c("green1","orange"))+
  coord_flip()+
  labs(title ="" ,x = "Clinical measurements"
       , y = "Percentile", #  (scaled: 0=min, 100=max)
       fill = "Phase:"
       )+
  myformat


rm(list= ls()[!(ls() %in% c("scaled","indivdistrcd4_scaled","no_errorexplo","myformat",
"myformat_cd4_by_phase"))])



#---------boxplots by phase_34
datastage_34 = filter(scaled[,],
                  phase =="3-Early" |
                  phase =="4-Est"
                  )
datastage_34s = datastage_34
#setnames(datastage_34s, "datastage_341", "phase" )
```

```
require(reshape2)
datastage 34m <- melt(datastage 34s, id.var = "phase")
datastage 34m<-datastage_34m[complete.cases(datastage_34m$value),]
require(ggplot2)

ggplot(data = datastage_34m, aes(x=variable, y=value, fill=phase)) +
  geom boxplot()+
  scale_fill_manual(values=c("orange","cyan"))+
  coord flip()+
  labs(title ="" ,x = "Clinical measurements"
       , y = "Percentile", #  (scaled: 0=min, 100=max)
       fill = "Phase:"
       )+
  myformat
rm(list= ls()[!(ls() %in% c("scaled","indivdistrcd4_scaled","no_errorexplo","myformat",
"myformat_cd4_by_phase"))])

#---------boxplots by phase_45
datastage_45 = filter(scaled[,],
                  phase =="4-Est" |
                  phase =="5-ART"
                  )

datastage 45s = datastage 45
#setnames(datastage_45s, "datastage_451", "phase" )

if(!require(reshape2)){install.packages("reshape2")}

datastage 45m <- melt(datastage 45s, id.var = "phase")
datastage 45m<-datastage_45m[complete.cases(datastage_45m$value),]
require(ggplot2)

ggplot(data = datastage_45m, aes(x=variable, y=value, fill=phase)) +
  geom boxplot()+
  scale_fill_manual(values=c("cyan","red"))+
  coord flip()+
  labs(title ="" ,x = "Clinical measurements"
       , y = "Percentile", #  (scaled: 0=min, 100=max)
       fill = "Phase:"
       )+
  myformat
rm(list= ls()[!(ls() %in%
c("scaled","indivdistrcd4_scaled","datax","no_errorexplo","myformat",
"myformat_cd4_by_phase"))])

# Crop and combine all graphs
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

boxplots2 3 = image_read("C:/Users/Partson/boxplots2_3.tiff")
boxplots2_3

boxplots3_4 = image_read("C:/Users/Partson/boxplots3_4.tiff")
boxplots3_4crop =image_crop(boxplots3_4, "296x670+95")
boxplots3_4crop

boxplots4 5 = image read("C:/Users/Partson/boxplots4 5.tiff")
boxplots4_5crop =image_crop(boxplots4_5, "296x670+95")
boxplots4_5crop

#=============================================================================#
#Plot FIGURE 2.4                                                              #
#=============================================================================#
boxplots2_5 = image_append(c(boxplots2_3 , boxplots3_4crop,boxplots4_5crop), stack = FALSE)
boxplots2_5scaled =image_scale(boxplots2_5,"600x670!") # ! = Resize to width and height
exactly, losing original aspect ratio.
image_write(boxplots2_5scaled,"C:/Users/Partson/boxplots2_5scaled.tiff"  , format = "tiff")

#=============================================================================#
# FIGURE 2.5: Parallel coordinate plots to detect errors                      #
#=============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/dataerror237.rda")
```

```
if(!require(lattice)){install.packages("lattice")}
parallelplot(~dataerror237[c(5:51)] | factor(phase), dataerror237,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="Value",
        ylab="Clinical attributes",
        main=""
        )
# Croping of figure above to get rid of white spaces
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

outliers yes = image read("C:/Users/Partson/outliers_yes.tiff")
outliers_yes_trim = outliers_yes %>%image_trim()
outliers_yes_trim
image_write(outliers_yes_trim,"C:/Users/Partson/outliers_yes_trim.tiff", format = "tiff")

#=============================================================================#
# FIGURE 2.6: Parallel coordinate plots to detect errors                      #
#=============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
# Replace errors with NA
if(!require(dplyr)){install.packages("dplyr")}
noerror = dataerror237 %>%
  group_by(phase)%>%
  mutate_if(is.numeric, function(x) {
      upperbound <- quantile(x,0.95, na.rm=TRUE) + (IQR(x, na.rm=TRUE) * 1.5 )
      lowerbound <- quantile(x,0.05, na.rm=TRUE) - (IQR(x, na.rm=TRUE) * 1.5 )
      ifelse(x < lowerbound |x > upperbound, NA, x)
             })

# Plot original values without errors
parallelplot(~noerror[c(5:51)] | factor(phase),
        noerror,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="Value",
        ylab="Clinical attributes",
        main=""
        )

# Croping of figure above to get rid of white spaces
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

outliers_no = image_read("C:/Users/Partson/outliers_no.tiff")
outliers_no_trim = outliers_no %>%image_trim()
outliers no trim
image_write(outliers_no_trim,"C:/Users/Partson/outliers_no_trim.tiff" , format = "tiff")

############################################################
#         PREPARE A BALANCED DATA SET                      #
############################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

#---------Create template for balanced repeated design
temp_bal_a00 = data.frame(patientid = rep(paste(c("P00"), 1:9    , sep=""),16))
temp_bal_a0  = data.frame(patientid = rep(paste(c("P0") , 10:99  , sep=""),16))
temp_bal_a   = data.frame(patientid = rep(paste(c("P")  , 100:237, sep=""),16))
temp_bal_b = rbind(
                  temp bal a00
                 ,temp_bal_a0
                 ,temp_bal_a
                  ) %>% arrange(patientid)
temp_bal_b$phase_num     = rep(2:5)
temp_bal_c = temp_bal_b %>% arrange(patientid,phase_num)
temp_bal_c$time          = rep(1:4)
temp_bal_c$patientid = as.factor(temp_bal_c$patientid)
temp_bal = temp_bal_c
save(temp_bal,file ="C:/Users/PARTSON/temp_bal.rda")
```

```
#----------Get top_4
load(file ="C:/Users/PARTSON/noerror237.rda")

table(noerror237$patientid,noerror237$phase)
if(!require(dplyr)){install.packages("dplyr")}
if(!require(tibble)){install.packages("tibble")}
top_4a = noerror237
top_4a[,"time"] =""        # create empty variable "

top_4b = top_4a[,c(1:4,52,5:51)]
top_4c =             top_4b %>%
                     tbl_df %>%
  group_by(patientid,phase) %>%
        mutate(time =1:n()) %>%
            as.data.frame()
                            # add sequential numbers from 1
                            # add sequential numbers for each patient
                            # Ideally, creates a sequential number (counter) for
                              # rows within each group of a dataframe
                            # PROBLEM: %>% tbl_df solves the ranking issue

top_4d =                    top_4c %>%
                            tbl_df %>%
        group_by(patientid,phase) %>%
                    top_n(4,time) %>%
              mutate(time =1:n()) %>%
                  as.data.frame()
table(top_4d$patientid,top_4d$phase)

top 4unbal = top 4d
save(top_4unbal,file ="C:/Users/PARTSON/top_4unbal.rda")

# ---------RECODE patientid
load(file ="C:/Users/PARTSON/top_4unbal.rda")
table(top_4unbal$patientid,top_4unbal$phase)
levels(top_4unbal$patientid)
top 4unbalcode = top 4unbal
save(top_4unbalcode,file ="C:/Users/PARTSON/top_4unbalcode.rda")


#-----------Merge "temp bal" and "top 4unbalcode"
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/temp_bal.rda")
load(file ="C:/Users/PARTSON/top_4unbalcode.rda")
bal_incomplt = full_join(temp_bal,top_4unbalcode, by = c("patientid","phase_num","time"))%>%
               arrange(patientid,phase_num,time)

table(bal_incomplt$patientid,bal_incomplt$phase_num) #test the balance - see all 4s
table(bal_incomplt$patientid,bal_incomplt$phase) #test the balance - some missing
save(bal_incomplt,file ="C:/Users/PARTSON/bal_incomplt.rda")

#----------Update "Phase" column
load(file ="C:/Users/PARTSON/bal_incomplt.rda")
bal_prep_a = bal_incomplt
bal_prep_a$phase = bal_prep_a$phase_num # prepare
if(!require(dplyr)){install.packages("dplyr")}
bal prep a$phase<-recode factor(bal_prep_a$phase,
                "2" = "2-Acute",
                "3" = "3-Early",
                "4" = "4-Est"  ,
                "5" = "5-ART"
                              )
bal2imput = bal prep a
paste(round(mean(is.na(bal2imput[,1:52]))*100,2),
                  "% missing in file: bal2imput", sep = "")
                            # missing in "bal2imput".....34.49%, (37.44% for 6:52)
dim(bal2imput[,1:52]) # 3792   42
mean(bal2imput$cd4,na.rm=T)
save(bal2imput,file ="C:/Users/PARTSON/bal2imput.rda")
#-------------Impute visitdate
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/bal2imput.rda")
#--------------------
data2imp = bal2imput
#--------------------
```

```
if(!require(dplyr)){install.packages("dplyr")}
data2imp_prep_a = transform(data2imp, visitdate = ifelse(is.na(visitdate)
                , NA, as.Date(visitdate,origin ="1970-01-01")))
                          # Convert date to numeric for imputation
if(!require(imputeTS)){install.packages("imputeTS")}
data2imp_prep_a$visitdate= as.Date(   # convert to date soon after
  na.interpolation(data2imp_prep_a$visitdate, option ="linear"), #impute
                       format="%Y-%m-%d", origin ="1970-01-01")  # required date format
                          # Impute the missing values by linear interpolation
                          # Other options: "spline" , "stine"
data2imp_prep_a$patientid = as.numeric(as.character(as.integer(data2imp_prep_a$patientid)))

data2imp_prep =   data2imp_prep_a %>%
                        tbl_df %>%
              group_by(patientid) %>%
              mutate(time =1:n()) %>%
                  as.data.frame()
save(data2imp_prep,file ="C:/Users/PARTSON/data2imp_prep.rda")
write.csv(data2imp_prep,file ="C:/Users/PARTSON/data2imp_prep.csv",na = "",row.names = FALSE)

#=============================================================================#
# FIGURE 2.7: MISSINGNESS MECHANISM                                           #
#=============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

#-----------------------------------------------------
# Test for missingness mechanism in SPSS
   # 1. Recall the Missing Value Analysis dialog box
        # a) Analyze -> Missing Value Analysis
   # 2. Click EM
   # 3. OK
# Results are at the foot notes of EM table
   # H0: Data are MCAR
   # p-value < 0.05 implies NOT MCAR
#-----------------------------------------------------

if(!require(miceadds)){install.packages("miceadds")}
if(!require(VIM)){install.packages("VIM")}

load(file ="C:/Users/PARTSON/data2imp_prep.rda")
# Plot
aggr(data2imp_prep[!names(data2imp_prep)=="glucose"], col=c("navyblue","red"),
  bars = T,
  numbers = F, prop = TRUE, combined = F, varheight = F,
  only.miss = FALSE, border = par("fg"), sortVars = TRUE,
  sortCombs = TRUE, ylabs = NULL, axes = TRUE,
  cex.lab = 1.2, cex.axis = 0.5, cex.numbers = par("cex"),
  gap = .05
       )

# Croping of figure above to get rid of white spaces
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

missingness = image_read("C:/Users/Partson/missingness.tiff")
missingness_trim = missingness %>%image_trim()
missingness trim
image_write(missingness_trim,"C:/Users/Partson/missingness_trim.tiff", format = "tiff")

############################################################
# IMPUTATION OF MISSING VALUES                             #
############################################################
load(file ="C:/Users/PARTSON/data2imp_prep.rda")
# Initialise
if(!require(mice)){install.packages("mice")}

ini <- mice(data2imp_prep
   ,maxit   = 0
   ,pred    = quickpred(data2imp prep
   ,mincor  = 0.1 # absolute corr between predictor and target
   ,method  = "pearson" # method for corr
   ,minpuc  = 0.0 # minimum proportion of usable cases
   ,include = c("patientid", "phase","visitdate","time")
   #,exclude = c("glucose")
   ))
```

```
predM  <- ini$predictorMatrix
method <- ini$meth

# adjust prediction matrix
predM[, c("patientid","phase_num")] <- -2 # class variables
predM[predM==1]=2 # replace the 1's with 2's
                  # "2" identifies both random and fixed effects
nopred = names(which(colSums(is.na(data2imp_prep))==0))
nopred            # setup cols not to predict
predM[nopred,]=0  # do not predict complete columns
# adjust prediction method
method[names(data2imp_prep)]<- "2L.pan" # predict all with "2L.pan"
method[nopred]<- "" # not predict, hence remove method
method["bmi"] <- "~I(weight_kg/(height_m)^2)"
method  # print to see
# Impute
imp <- mice(data2imp_prep, m= 5, maxit= 50, seed=123
          ,predictorMatrix  = predM
          ,imputationMethod = method
          )
#-------------
imp_bal = imp
#-------------
save(imp_bal,file ="C:/Users/PARTSON/imp_bal.rda")


# prepare FIGURE 2.8 : DENSITY PLOTS of IMPUTATIONS

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/imp_bal.rda")
#---------------
imp =imp bal
#---------------

if(!require(miceadds)){install.packages("miceadds")}

#=============================================================================#
# Plot FIGURE 2.8                                                             #
#=============================================================================#
densityplot(imp) # compare observed(blue) against m imputed(red)

# Croping of figures above to get rid of white spaces
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

densityplots = image_read("C:/Users/Partson/densityplots.tiff")
densityplots_trim = densityplots %>%image_trim()
densityplots_trim
image_write(densityplots_trim,"C:/Users/Partson/densityplots_trim.tiff", format = "tiff")

#--
plot(imp) # streams must mingle for health convergence
stacked <- complete(imp,action='long',include=TRUE)
                #  all.imp contains all M imputed data sets, stacked; the
                #  first column is the imputation number;
                #  second column is the ith imputed dataset
# Reset "time" to be within phase
if(!require(dplyr)){install.packages("dplyr")}
bal_complt_stacked_a =    stacked %>%
                           tbl_df %>%
   group_by(.imp,patientid,phase) %>% # time within phase
              mutate(time =1:n()) %>%
                  as.data.frame()

#---------------------------
bal_complt_stacked = bal_complt_stacked_a
#---------------------------
save(bal_complt_stacked,file ="C:/Users/PARTSON/bal_complt_stacked.rda")


#############################################################
# SELECTION OF A SINGLE DATA SET                            #
#############################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/bal_complt_stacked.rda")
```

```
#----------------------------
stacked = bal complt stacked
#----------------------------
if(!require(dplyr)){install.packages("dplyr")}
cvs_a = stacked[,c(1,8:54)] %>%
    group_by(.imp)%>%
    mutate_if(is.numeric, function(x) {
    round(sd(x,na.rm=TRUE)/mean(x,na.rm=TRUE)*100,2) # calculate cvs
                                                    })
cv_b = cvs_a[!duplicated(cvs_a[
                           c(1,2:48)]), ]
                           # Drop duplicates based on selected columns
if(!require(data.table)){install.packages("data.table")}
setnames(cv_b,".imp","mids")
cv c = cv b
avg_cv = cv_c %>% as.data.frame()%>%
 transmute(mids,avg=rowMeans(select(.,2:48)))
avg_cv
avg0 = avg_cv[which.min(avg_cv[,"mids"]),][[2]] # get avg for original data
avg0                          # find min in "mids" and return coresponding
                              #value in col 2
avg_cv$absdiff = abs(avg_cv$avg-avg0) # find abs diff from original avg
avg_cv_a = avg_cv[-1,]
mid selct = avg cv a[which.min(avg cv a[,"absdiff"]),][[1]] # get avg for original data
mid selct # print selected imputed data set
selctd_mid = subset(stacked[,-2],.imp==mid_selct)
paste(round(mean(is.na(selctd_mid))*100,2),
                  "% missing in file: bal_complt", sep = "")
                           # missing in "unbal_complt".....0.0%

#----------------------------
bal complt = selctd mid[,-1]
#----------------------------
names(bal_complt)
save(bal_complt,file ="C:/Users/PARTSON/bal_complt.rda")


##Prepare TABLE 2.3: Summary statistics

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/bal_complt.rda")
datatab1 = bal_complt

#--add labels
if(!require(table1)){install.packages("table1")}
label(datatab1$cd4)            <- "CD4+ Count"
label(datatab1$lymphocytes)    <- "Lymphocyte"
label(datatab1$basophils)      <- "Basophils"
label(datatab1$albumin)        <- "Albumin"
label(datatab1$haematocrit)    <- "Haematocrit"
label(datatab1$alkaline_phos)  <- "ALP"
label(datatab1$mcv)            <- "MCV"
label(datatab1$platelet)       <- "Platelet"
label(datatab1$potassium)      <- "Potassium"
label(datatab1$monocytes)      <- "Monocytes"
label(datatab1$protein)        <- "Total protein"
label(datatab1$lactate_dehyd)  <- "LDH"
label(datatab1$folate)         <- "Folate"
label(datatab1$magnesium)      <- "Magnesium"
label(datatab1$glucose)        <- "Glucose"
label(datatab1$calcium)        <- "Calcium"
label(datatab1$mchc)           <- "MCHC"
label(datatab1$red_cells)      <- "RBC"
label(datatab1$sodium)         <- "Sodium"
label(datatab1$vitb12)         <- "Vitamin B12"
label(datatab1$triceps_skin)   <- "Triceps skin fold"
label(datatab1$triglycerides)  <- "Triglycerides"
label(datatab1$neutrophils)    <- "Neutrophils"
label(datatab1$ast_got)        <- "AST(GOT)"
label(datatab1$eosinophils)    <- "Eosinophils"
label(datatab1$height_m)       <- "Height"
label(datatab1$chloride)       <- "Chloride"
label(datatab1$fe)             <- "Fe(Iron)"
label(datatab1$waist_circum)   <- "Waist circum"
label(datatab1$ldl)            <- "LDL"
label(datatab1$bmi)            <- "BMI"
label(datatab1$bp_systolic)    <- "BP(systolic)"
```

```
label(datatab1$bilirubin)      <- "Bilirubin"
label(datatab1$arm right)      <- "Arm(right)circum"
label(datatab1$glutamyl trans)<- "GGT"
label(datatab1$bp diastolic)   <- "BP(diastolic)"
label(datatab1$rdw)            <- "RDW"
label(datatab1$pulse_bpm)      <- "Pulse"
label(datatab1$urea)           <- "Urea"
label(datatab1$alt gpt)        <- "ALT(GPT)"
label(datatab1$axillary temp) <- "Ax.Temp"
label(datatab1$hb)             <- "Haemoglobin"
label(datatab1$mch)            <- "MCH"
label(datatab1$leucocyte)      <- "leucocytes"
label(datatab1$cholesterol)    <- "Cholesterol"
label(datatab1$hip circum)     <- "Hip circum"
label(datatab1$weight kg)      <- "Weight"
label(datatab1$bmi)            <- "BMI"
#--add units
units(datatab1$cd4)            <- "cells/mm^3"
units(datatab1$lymphocytes)    <- "x10^9/L"
units(datatab1$basophils)      <- "x10^9/L"
units(datatab1$albumin)        <- "g/L"
units(datatab1$haematocrit)    <- "Hct/100"
units(datatab1$alkaline_phos) <- "IU/L"
units(datatab1$mcv)            <- "fL"
units(datatab1$platelet)       <- "x10^9/L"
units(datatab1$potassium)      <- "mmol/L"
units(datatab1$monocytes)      <- "x10^9/L"
units(datatab1$protein)        <- "g/L"
units(datatab1$lactate dehyd) <- "U/L"
units(datatab1$folate)         <- "nmol/L"
units(datatab1$magnesium)      <- "mmol/L"
#units(datatab1$glucose)        <- ""
units(datatab1$calcium)        <- "mmol/L"
units(datatab1$mchc)           <- "g/dL"
units(datatab1$red cells)      <- "x10^6cells/mm^3"
units(datatab1$sodium)         <- "mEq/L"
units(datatab1$vitb12)         <- "ng/mL"
units(datatab1$triceps skin)   <- "mm"
units(datatab1$triglycerides) <- "mmol/L"
units(datatab1$neutrophils)    <- "x10^9/L"
units(datatab1$ast got)        <- "U/L"
units(datatab1$eosinophils)    <- "x10^9/L"
units(datatab1$height m)       <- "m"
units(datatab1$chloride)       <- "mEq/L"
units(datatab1$fe)             <- "mcg/dL"
units(datatab1$waist_circum)   <- "cm"
units(datatab1$ldl)            <- "mmol/L"
units(datatab1$bmi)            <- "kg/m^2"
units(datatab1$bp_systolic)    <- "mmHg"
units(datatab1$bilirubin)      <- "mmol/L"
units(datatab1$arm_right)      <- "cm"
units(datatab1$glutamyl trans)<- "U/L"
units(datatab1$bp diastolic)   <- "mmHg"
units(datatab1$rdw)            <- "%"
units(datatab1$pulse_bpm)      <- "bpm"
units(datatab1$urea)           <- "mmol/L"
units(datatab1$alt_gpt)        <- "U/L"
units(datatab1$axillary temp) <- "D.Celsius"
units(datatab1$hb)             <- "g/dL"
units(datatab1$mch)            <- "pg/cell"
units(datatab1$leucocyte)      <- "x10^9/L"
units(datatab1$cholesterol)    <- "mmol/L"
units(datatab1$hip_circum)     <- "cm"
units(datatab1$weight kg)      <- "kg"
units(datatab1$bmi)            <- "kg/m^2"

save(datatab1,file ="C:/Users/PARTSON/datatab1.rda")

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/datatab1.rda")

my.render.cont <- function(x) {
    with(stats.apply.rounding(stats.default(x), digits=2), c("                          ",
        "Mean&plusmnSD(CV%)" =sprintf("%s&plusmn;%s(%s)", MEAN,SD,CV),
        "Median(Min;Max)"    =sprintf("%s(%s;%s)",MEDIAN,MIN,MAX)
        ))
}
```

```
#=============================================================================#
#-- Table 2.3                                                                #
#=============================================================================#
if(!require(table1)){install.packages("table1")}
tab =table1(~
cd4+             red_cells+
hb+              haematocrit+     mcv+
mch+             mchc+            rdw+
platelet+        leucocyte+       neutrophils+
lymphocytes+     monocytes+       eosinophils+
basophils+       cholesterol+     ldl+
triglycerides+   glucose+         alt_gpt+
ast_got+         bilirubin+       alkaline_phos+
glutamyl_trans+ calcium+         chloride+
magnesium+       potassium+       sodium+
protein+         albumin+         lactate_dehyd+
fe+              folate+          vitb12+
urea+            bp_systolic+     bp_diastolic+
pulse_bpm+       axillary_temp+   waist_circum+
hip circum+      arm right+       triceps_skin+
height m+        weight kg+       bmi
|phase, data=datatab1, overall="Overall",
       render.continuous=my.render.cont,footnote = "");tab


# Prepare FIGURE 2.9: AVERAGE CD4 count per visit

# prepare data
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/data2model.rda")
datan = data2model
if(!require(Rmisc)){install.packages("Rmisc")}
if(!require(dplyr)){install.packages("dplyr")}

datans a <- summarySE(datan,
            measurevar="cd4",
        groupvars=c("phase","time")) %>%
                    tbl_df    %>%
        mutate(time =1:n())   %>%
            as.data.frame()

datans = datans_a[,c(1,2,4)]
datans$cd4 = round(datans$cd4,0)
save(datans,file ="C:/Users/PARTSON/datans.rda")

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/datans.rda")
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(Rcmdr)){install.packages("Rcmdr")}
if(!require(ggrepel)){install.packages("ggrepel")}

format plot mean=(
      theme bw()    +         # Background fill
      theme(                  # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element_text(size=9,angle=0,color="black", hjust=.5, vjust=0.5),
    axis.title.y = element text(color="black", size=10, face="bold"),
    axis.text.y = element text(size=9,angle=0,color="black", hjust=1,vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="white",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))
```

```
#=============================================================================#
# plot FIGURE 2.9
#=============================================================================#
dev.new(width=6, height=4, unit="in")
ggplot(datans, aes(x=time, y=  cd4)) +
    geom_line(size=.5, color="blue") + geom_point(aes(color = factor(phase)),size=5 )+
    scale_colour_manual(values=c("green1","orange","red","green4"),name="HIV infection
phase:")+
         scale_x_continuous(breaks = round(seq(min(datans$time),
                           max(datans$time), by = 1),0)) +
    labs(x="Follow up visit time)",      y=bquote("Average CD4"^"+"*"count
(cells/mm"^"3"*")"),
         title="") +
    geom_text_repel(data  = datans, aes(label=cd4),
                    size  = 2.8,box.padding = unit(0.42, "lines")
                    ,parse=TRUE,nudge_x = -.5, nudge_y = 1.5
                    ,force = 1,direction = c( "both")
                    ,arrow=arrow(length = unit(0.01, "npc"))
                    ,segment.color = 'yellow'
                    ,seed = 1)+
    theme_classic()+
    geom_hline(yintercept= 570, color ="red", lty =2)+
    format_plot_mean

# Prepare FIGURE 2.10 MEAN AND CV PARALLEL COORDINATE PLOTS

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/bal_complt.rda")
if(!require(dplyr)){install.packages("dplyr")}
if(!require(Rcmdr)){install.packages("Rcmdr")}
# Phase means
mean_bal_complt = bal_complt[c(1,3,5:52)]%>%
  group_by(phase) %>%
  mutate_if(is.numeric, function(x) { mean(x) })

if(!require(lattice)){install.packages("lattice")}
parallelplot(~mean_bal_complt[c(4:50)] | factor(phase),
        mean_bal_complt,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="Average value within phase",
        ylab="Clinical covariates",
        main=""
        )
# Phase coefficient of variation
cv_bal_complt = bal_complt[c(1,3,5:52)]%>%
  group_by(phase) %>%
  mutate_if(is.numeric, function(x) {
                      sd(x, na.rm=TRUE)/mean(x, na.rm=TRUE)*100
                                        })
#=============================================================================#
#--plot Figure 2.10                                                           #
#=============================================================================#

if(!require(lattice)){install.packages("lattice")}
parallelplot(~cv_bal_complt[c(4:50)] | factor(phase),
        cv bal complt,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="CV(%) within phase",
        ylab="Clinical covariates",
        main=""
        )

# Croping of figures above to get rid of white spaces and merge
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

parallelplot_mean = image_read("C:/Users/Partson/parallelplot_mean.tiff")
parallelplot_mean # 393x664
parallelplot_mean_trim = parallelplot_mean%>% image_trim()
parallelplot_mean_trim
```

```
parallelplot_cv = image_read("C:/Users/Partson/parallelplot_cv.tiff")
parallelplot_cv # 395x653
parallelplot_cv_trim = parallelplot_cv%>% image_trim()
parallelplot_cv_trim
parallelplot_cvcrop =image_crop(parallelplot_cv_trim, "393x664+100") #34
parallelplot_cvcrop

parallel_mean_cv = image_append(c(parallelplot_mean_trim
                                  ,parallelplot_cvcrop))%>%
                           image_trim() #emoves edges that are the background color from the
image
parallel_mean_cv
image_write(parallel_mean_cv,"C:/Users/Partson/parallel_mean_cv.tiff", format = "tiff")


#####################################################################
#                                                                   #
# Prepare FIGURE 2.11 and Figure 2.12: INDIVIDUAL PROFILES          #
#                                                                   #
# *******Create MACROS to automates tasks*********                  #
#####################################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

if(!require(gtools)){install.packages("gtools")}
# setup values to be replaced
  # if not included when running macro, default is executed
  # if not included when running macro, and default is place holder, throws error
indiv.plot= defmacro(data   = dataname  , # "dataname" is default to be replaced
                     x      = xvar      , # "xvar" is default to be replaced
                     y      = yvar      , # "yvar" is default to be replaced
                     xlab = ""      ,   # " "  default to be replaced
                     ylab = ""      ,   # ""  default to be replaced
                  legendbox = "no"    , # "no" is default condition to be replaced
# functions to be evaluated start here
    expr={
# load the required packages
if(!require(Rcmdr)){install.packages("Rcmdr")}
if(!require(ggplot2)){install.packages("ggplot2")}
# setup plot formats
format_plot_indiv=(
      theme_bw()    +         # Background fill
      theme(                  # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=8, face="plain",angle=0),
    axis.text.x = element_text(size=8,angle=0,color="black", hjust=0.5, vjust=0.5),
    axis.title.y = element_text(color="black", size=8, face="plain"),
    axis.text.y = element_text(size=8,angle=0,color="black",vjust=0.25),
    legend.position="none", # top
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))
# setup conditional functions to be executed
     # first condition always "if"
 if (legendbox=="yes"){ggplot(dataname, aes(x, y,colour = factor(patientid))) +
  geom_smooth(method = 'lm', se=FALSE, fullrange=TRUE,lwd=.3)+
  labs( x=xlab,
        y=ylab, colour="patientid")+
  format_plot_indiv
# intermediate conditions use "else if"
}else if (legendbox=="no"){
  ggplot(dataname, aes(x, y, colour = factor(patientid))) +
  geom_smooth(method = 'lm', se=FALSE, fullrange=TRUE,lwd=.3)+
  labs( x=xlab,
        y=ylab, colour="patientid")+
  format_plot_indiv

} else # last condition always "else"
  print("Nothing to plot")
})

data =load(file ="C:/Users/PARTSON/bal_complt.rda")
dataname = bal_complt

time         =indiv.plot(data=data,x=time, y=cd4,xlab="Visit time",
ylab=bquote("CD4"^"+"*"count")) ;time
```

```
red_cells    =indiv.plot(data=dataname,x=red_cells, y=cd4
,xlab=bquote("RBC(x10"^"6"*"cells/mm"^"3"*")"),ylab=bquote("CD4"^"+"*"count")) ;red_cells
hb           =indiv.plot(data=data,x=hb, y=cd4,xlab="Haemoglobin(nmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;hb
haematocrit  =indiv.plot(data=data,x=haematocrit ,y=cd4,xlab="Haematocrit(Hct/100)",
ylab=bquote("CD4"^"+"*"count")) ;haematocrit
mcv  =indiv.plot(data=data,x=mcv ,y=cd4,xlab="MCV(fL)", ylab=bquote("CD4"^"+"*"count")) ;mcv
mch  =indiv.plot(data=data,x=mch ,y=cd4,xlab="MCH(pg/cell)", ylab=bquote("CD4"^"+"*"count"))
;mch
mchc  =indiv.plot(data=data,x=mchc ,y=cd4,xlab="MCHC(g/dL)", ylab=bquote("CD4"^"+"*"count"))
;mchc
rdw  =indiv.plot(data=data,x=rdw ,y=cd4,xlab="RDW(%)", ylab=bquote("CD4"^"+"*"count")) ;rdw
platelet  =indiv.plot(data=data,x=platelet ,y=cd4,xlab=bquote("Platelet(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;platelet
leucocyte  =indiv.plot(data=data,x=leucocyte ,y=cd4,xlab=bquote("Leucocytes(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;leucocyte
neutrophils  =indiv.plot(data=data,x=neutrophils
,y=cd4,xlab=bquote("Neutrophils(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;neutrophils
lymphocytes  =indiv.plot(data=data,x=lymphocytes
,y=cd4,xlab=bquote("Lymphocytes(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;lymphocytes
monocytes  =indiv.plot(data=data,x=monocytes ,y=cd4,xlab=bquote("Monocytes(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;monocytes
eosinophils  =indiv.plot(data=data,x=eosinophils
,y=cd4,xlab=bquote("Eosinophils(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;eosinophils
basophils  =indiv.plot(data=data,x=basophils ,y=cd4,xlab=bquote("Basophils(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;basophils
cholesterol  =indiv.plot(data=data,x=cholesterol ,y=cd4,xlab="Cholesterol(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;cholesterol
ldl  =indiv.plot(data=data,x=ldl ,y=cd4,xlab="LDL(mmol/L)", ylab=bquote("CD4"^"+"*"count"))
;ldl
triglycerides  =indiv.plot(data=data,x=triglycerides ,y=cd4,xlab="Triglycerides(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;triglycerides
glucose  =indiv.plot(data=data,x=glucose ,y=cd4,xlab="Glucose",
ylab=bquote("CD4"^"+"*"count")) ;glucose
alt_gpt  =indiv.plot(data=data,x=alt_gpt ,y=cd4,xlab="ALT(GPT)(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;alt_gpt
ast_got  =indiv.plot(data=data,x=ast_got ,y=cd4,xlab="AST(GOT)(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;ast_got
bilirubin  =indiv.plot(data=data,x=bilirubin ,y=cd4,xlab="Bilirubin(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;bilirubin
alkaline_phos  =indiv.plot(data=data,x=alkaline_phos ,y=cd4,xlab="ALP(IU/L)",
ylab=bquote("CD4"^"+"*"count")) ;alkaline_phos
glutamyl_trans  =indiv.plot(data=data,x=glutamyl_trans ,y=cd4,xlab="GGT(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;glutamyl_trans
calcium  =indiv.plot(data=data,x=calcium ,y=cd4,xlab="Calcium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;calcium
chloride  =indiv.plot(data=data,x=chloride ,y=cd4,xlab="Chloride(mEq/L)",
ylab=bquote("CD4"^"+"*"count")) ;chloride
magnesium  =indiv.plot(data=data,x=magnesium ,y=cd4,xlab="Magnesium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;magnesium
potassium  =indiv.plot(data=data,x=potassium ,y=cd4,xlab="Potassium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;potassium
sodium  =indiv.plot(data=data,x=sodium ,y=cd4,xlab="Sodium(mEq/L)",
ylab=bquote("CD4"^"+"*"count")) ;sodium
protein  =indiv.plot(data=data,x=protein ,y=cd4,xlab="Protein(g/L)",
ylab=bquote("CD4"^"+"*"count")) ;protein
albumin =indiv.plot(data=data,x=albumin ,y=cd4,xlab="Albumin(g/L)",
ylab=bquote("CD4"^"+"*"count")) ;albumin
lactate_dehyd =indiv.plot(data=data,x=lactate_dehyd ,y=cd4,xlab="LDH(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;lactate_dehyd
fe =indiv.plot(data=data,x=fe ,y=cd4,xlab="Iron(Fe)(mcg/dL)", ylab=bquote("CD4"^"+"*"count"))
;fe
folate =indiv.plot(data=data,x=folate ,y=cd4,xlab="Folate(nmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;folate
vitb12 =indiv.plot(data=data,x=vitb12 ,y=cd4,xlab="Vitamin B12(ng/mL)",
ylab=bquote("CD4"^"+"*"count")) ;vitb12
urea =indiv.plot(data=data,x=urea ,y=cd4,xlab="Urea(mmol/L)", ylab=bquote("CD4"^"+"*"count"))
;urea
bp_systolic =indiv.plot(data=data,x=bp_systolic ,y=cd4,xlab="BP(systolic)(mmHg)",
ylab=bquote("CD4"^"+"*"count")) ;bp_systolic
bp_diastolic =indiv.plot(data=data,x=bp_diastolic ,y=cd4,xlab="BP(diastolic)(mmHg)",
ylab=bquote("CD4"^"+"*"count")) ;bp_diastolic
pulse_bpm =indiv.plot(data=data,x=pulse_bpm ,y=cd4,xlab="Pulse(bpm)",
ylab=bquote("CD4"^"+"*"count")) ;pulse_bpm
axillary_temp =indiv.plot(data=data,x=axillary_temp
,y=cd4,xlab=bquote("Axillary.Temp("^"o"*"C)"), ylab=bquote("CD4"^"+"*"count")) ;axillary_temp
```

```
waist_circum =indiv.plot(data=data,x=waist_circum ,y=cd4,xlab="Waist circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;waist_circum
hip_circum =indiv.plot(data=data,x=hip_circum ,y=cd4,xlab="Waist circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;hip_circum
arm_right =indiv.plot(data=data,x=arm_right ,y=cd4,xlab="Arm(right)circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;arm_right
triceps_skin =indiv.plot(data=data,x=triceps_skin ,y=cd4,xlab="Triceps skin fold(mm)",
ylab=bquote("CD4"^"+"*"count")) ;triceps_skin
height_m =indiv.plot(data=data,x=height_m ,y=cd4,xlab="Height(m)",
ylab=bquote("CD4"^"+"*"count")) ;height_m
weight_kg =indiv.plot(data=data,x=weight_kg ,y=cd4,xlab="Weight(kg)",
ylab=bquote("CD4"^"+"*"count")) ;weight_kg
bmi =indiv.plot(data=data,x=bmi ,y=cd4,xlab=bquote("BMI(kg/m"^"2"*")"),
ylab=bquote("CD4"^"+"*"count")) ;bmi


load(file ="C:/Users/PARTSON/multiplot.rda")
blankplot = plot(0,type='n',axes=FALSE,ann=FALSE)
#### indivtrends1a
multiplot(
 time        ,
 #
 red_cells   , hb           , haematocrit   , mcv       ,
 mch         , mchc         , rdw           , platelet  ,
 leucocyte   , neutrophils  , lymphocytes   , monocytes ,
 eosinophils , basophils    ,
 #
 cholesterol , ldl          , triglycerides ,
 #
 glucose     ,
 #
 alt gpt     ,
 cols=4) # save as indivtrends1a

 #### indivtrends1b
multiplot(
    #NB: repeat and crop to make it easy
 ast got     , ast got     , ast got     , ast got     , ast got     ,
 bilirubin   ,bilirubin    ,bilirubin    ,bilirubin    ,bilirubin    ,
 alkaline_phos ,alkaline_phos ,alkaline_phos ,alkaline_phos ,alkaline_phos ,
 glutamyl_trans,glutamyl_trans,glutamyl_trans,glutamyl_trans,glutamyl_trans,
 #
 cols=4) # save as indivtrends1b

#### indivtrends2a
multiplot(
 calcium     , chloride     , magnesium     , potassium   ,
 sodium      ,
 #
 protein     , albumin      , lactate_dehyd ,
 #
 fe          , folate       , vitb12        ,
 #
 urea,
 #
 bp_systolic  , bp_diastolic, pulse_bpm     ,  axillary_temp,
 #
 waist circum , hip circum  , arm_right     ,  triceps_skin,
 cols=4) # save as indivtrends2a

#### indivtrends2b
multiplot(
height_m   , height_m   ,height_m   ,height_m   ,height_m   ,
weight kg  , weight kg  ,weight kg  ,weight kg  ,weight_kg  ,
bmi      , bmi      ,bmi      ,bmi      ,bmi      ,
blankplot, blankplot, blankplot, blankplot, blankplot,
cols=4) # save as indivtrends2b

if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}
indivtrends1a = image_read("C:/Users/Partson/indivtrends1a.tiff") ;indivtrends1a
indivtrends1b = image_read("C:/Users/Partson/indivtrends1b.tiff") ;indivtrends1b
indivtrends1b_crop =image_crop(indivtrends1b, "598x652+0-535") ;indivtrends1b_crop
```

```
#=========================================================================#
#--plot Figure 2.11                                                       #
#=========================================================================#
indivtrends1 = image_append(c(indivtrends1a , indivtrends1b_crop), stack = TRUE) ;indivtrends1
image_write(indivtrends1,"C:/Users/Partson/indivtrends1.tiff", format = "tiff")

indivtrends2a = image_read("C:/Users/Partson/indivtrends2a.tiff") ;indivtrends2a
indivtrends2b = image read("C:/Users/Partson/indivtrends2b.tiff") ;indivtrends2b
indivtrends2b crop =image crop(indivtrends2b, "598x652+0-535") ;indivtrends2b_crop
#=========================================================================#
#--plot Figure 2.12                                                       #
#=========================================================================#
indivtrends2 = image_append(c(indivtrends2a , indivtrends2b_crop), stack = TRUE) ;indivtrends2
image_write(indivtrends2,"C:/Users/Partson/indivtrends2.tiff", format = "tiff")

############################################################
#                                                          #
# Prepare FIGURE 2.13 and Figure 2.14: SMOOTH CURVES       #
#                                                          #
# *******Create MACROS to automates tasks*********         #
############################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

if(!require(gtools)){install.packages("gtools")}
# setup values to be replaced
  # if not included when running macro, default is executed
  # if not included when running macro, and default is place holder, throws error
loess.plot= defmacro(data    = dataname  , # "dataname" is default to be replaced
                     x       = xvar      , # "xvar" is default to be replaced
                     y       = yvar      , # "yvar" is default to be replaced
                     xlab = ""      ,  # " "  default to be replaced
                     ylab = ""      ,  # ""  default to be replaced
                  legendbox = "no"    , # "no" is default condition to be replaced

# functions to be evaluated start here
    expr={
# load the required packages
if(!require(Rcmdr)){install.packages("Rcmdr")}
if(!require(ggplot2)){install.packages("ggplot2")}

# setup plot formats
format plot loess legend=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="white", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="white", size=8, face="bold",angle=0),
    axis.text.x = element text(size=8,angle=0,color="white", hjust=1, vjust=0.5),
    axis.title.y = element text(color="white", size=8, face="bold"),
    axis.text.y = element_text(size=8,angle=0,color="white",vjust=0.25),
    legend.position = c(0.45,0.5),
    legend.background = element_rect(fill="white",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element text(colour="black",size=8,face="bold"),
    legend.text = element text(colour="black",size=9,face="plain"),
    legend.key.size = unit(0.4,"cm")
    ,legend.key.width = unit(1.3,"cm")
     ))
format plot loess=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=8, face="plain",angle=0),
    axis.text.x = element_text(size=8,angle=0,color="black", hjust=0.5, vjust=0.5),
    axis.title.y = element text(color="black", size=8, face="plain"),
    axis.text.y = element text(size=8,angle=0,color="black",vjust=0.25),
    legend.position="none", # top
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))


# setup conditional functions to be executed
     # first condition always "if"
 if (legendbox=="yes"){ggplot(dataname, aes(x, y,colour = factor(phase))) +
  geom_smooth(method = 'loess', se=FALSE, fullrange=TRUE,lwd=.3)+
```

```
  labs( x=xlab,
        y=ylab, colour="Infection phase")+
  format plot loess legend
    # intermediate conditions use "else if"
}else if (legendbox=="no"){
 ggplot(dataname, aes(x, y,colour = factor(phase))) +
  geom_smooth(method = 'loess', se=FALSE, fullrange=TRUE,lwd=.3)+
  labs( x=xlab,
        y=ylab, colour="")+
  format plot gam
     # last condition always "else"
} else
  print("Nothing to plot")
})

data =load(file ="C:/Users/PARTSON/bal_complt.rda")
dataname = bal_complt

legend_custom=loess.plot(x=hb, y=cd4,xlab="",ylab=bquote("CD4"^"+"*"count"),legendbox="yes")
;legend custom
red_cells     =loess.plot(data=dataname,x=red_cells, y=cd4
,xlab=bquote("RBC(x10"^"6"*"cells/mm"^"3"*")"),ylab=bquote("CD4"^"+"*"count")) ;red_cells
hb            =loess.plot(data=data,x=hb, y=cd4,xlab="Haemoglobin(nmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;hb
haematocrit   =loess.plot(data=data,x=haematocrit ,y=cd4,xlab="Haematocrit(Hct/100)",
ylab=bquote("CD4"^"+"*"count")) ;haematocrit
mcv   =loess.plot(data=data,x=mcv ,y=cd4,xlab="MCV(fL)", ylab=bquote("CD4"^"+"*"count")) ;mcv
mch   =loess.plot(data=data,x=mch ,y=cd4,xlab="MCH(pg/cell)", ylab=bquote("CD4"^"+"*"count"))
;mch
mchc   =loess.plot(data=data,x=mchc ,y=cd4,xlab="MCHC(g/dL)", ylab=bquote("CD4"^"+"*"count"))
;mchc
rdw   =loess.plot(data=data,x=rdw ,y=cd4,xlab="RDW(%)", ylab=bquote("CD4"^"+"*"count")) ;rdw
platelet   =loess.plot(data=data,x=platelet ,y=cd4,xlab=bquote("Platelet(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;platelet
leucocyte   =loess.plot(data=data,x=leucocyte ,y=cd4,xlab=bquote("Leucocytes(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;leucocyte
neutrophils   =loess.plot(data=data,x=neutrophils
,y=cd4,xlab=bquote("Neutrophils(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;neutrophils
lymphocytes   =loess.plot(data=data,x=lymphocytes
,y=cd4,xlab=bquote("Lymphocytes(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;lymphocytes
monocytes   =loess.plot(data=data,x=monocytes ,y=cd4,xlab=bquote("Monocytes(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;monocytes
eosinophils   =loess.plot(data=data,x=eosinophils
,y=cd4,xlab=bquote("Eosinophils(x10"^"9"*"/L)"), ylab=bquote("CD4"^"+"*"count")) ;eosinophils
basophils   =loess.plot(data=data,x=basophils ,y=cd4,xlab=bquote("Basophils(x10"^"9"*"/L)"),
ylab=bquote("CD4"^"+"*"count")) ;basophils
cholesterol   =loess.plot(data=data,x=cholesterol ,y=cd4,xlab="Cholesterol(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;cholesterol
ldl   =loess.plot(data=data,x=ldl ,y=cd4,xlab="LDL(mmol/L)", ylab=bquote("CD4"^"+"*"count"))
;ldl
triglycerides   =loess.plot(data=data,x=triglycerides ,y=cd4,xlab="Triglycerides(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;triglycerides
glucose   =loess.plot(data=data,x=glucose ,y=cd4,xlab="Glucose",
ylab=bquote("CD4"^"+"*"count")) ;glucose
alt_gpt   =loess.plot(data=data,x=alt_gpt ,y=cd4,xlab="ALT(GPT)(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;alt_gpt
ast_got   =loess.plot(data=data,x=ast_got ,y=cd4,xlab="AST(GOT)(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;ast_got
bilirubin   =loess.plot(data=data,x=bilirubin ,y=cd4,xlab="Bilirubin(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;bilirubin
alkaline_phos   =loess.plot(data=data,x=alkaline_phos ,y=cd4,xlab="ALP(IU/L)",
ylab=bquote("CD4"^"+"*"count")) ;alkaline_phos
glutamyl_trans   =loess.plot(data=data,x=glutamyl_trans ,y=cd4,xlab="GGT(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;glutamyl_trans
calcium   =loess.plot(data=data,x=calcium ,y=cd4,xlab="Calcium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;calcium
chloride   =loess.plot(data=data,x=chloride ,y=cd4,xlab="Chloride(mEq/L)",
ylab=bquote("CD4"^"+"*"count")) ;chloride
magnesium   =loess.plot(data=data,x=magnesium ,y=cd4,xlab="Magnesium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;magnesium
potassium   =loess.plot(data=data,x=potassium ,y=cd4,xlab="Potassium(mmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;potassium
sodium   =loess.plot(data=data,x=sodium ,y=cd4,xlab="Sodium(mEq/L)",
ylab=bquote("CD4"^"+"*"count")) ;sodium
protein   =loess.plot(data=data,x=protein ,y=cd4,xlab="Protein(g/L)",
ylab=bquote("CD4"^"+"*"count")) ;protein
```

```
albumin =loess.plot(data=data,x=albumin ,y=cd4,xlab="Albumin(g/L)",
ylab=bquote("CD4"^"+"*"count")) ;albumin
lactate_dehyd =loess.plot(data=data,x=lactate_dehyd ,y=cd4,xlab="LDH(U/L)",
ylab=bquote("CD4"^"+"*"count")) ;lactate_dehyd
fe =loess.plot(data=data,x=fe ,y=cd4,xlab="Iron(Fe)(mcg/dL)", ylab=bquote("CD4"^"+"*"count"))
;fe
folate =loess.plot(data=data,x=folate ,y=cd4,xlab="Folate(nmol/L)",
ylab=bquote("CD4"^"+"*"count")) ;folate
vitb12 =loess.plot(data=data,x=vitb12 ,y=cd4,xlab="Vitamin B12(ng/mL)",
ylab=bquote("CD4"^"+"*"count")) ;vitb12
urea =loess.plot(data=data,x=urea ,y=cd4,xlab="Urea(mmol/L)", ylab=bquote("CD4"^"+"*"count"))
;urea
bp_systolic =loess.plot(data=data,x=bp_systolic ,y=cd4,xlab="BP(systolic)(mmHg)",
ylab=bquote("CD4"^"+"*"count")) ;bp_systolic
bp_diastolic =loess.plot(data=data,x=bp_diastolic ,y=cd4,xlab="BP(diastolic)(mmHg)",
ylab=bquote("CD4"^"+"*"count")) ;bp_diastolic
pulse_bpm =loess.plot(data=data,x=pulse_bpm ,y=cd4,xlab="Pulse(bpm)",
ylab=bquote("CD4"^"+"*"count")) ;pulse_bpm
axillary_temp =loess.plot(data=data,x=axillary_temp
,y=cd4,xlab=bquote("Axillary.Temp("^"o"*"C)"), ylab=bquote("CD4"^"+"*"count")) ;axillary_temp
waist_circum =loess.plot(data=data,x=waist_circum ,y=cd4,xlab="Waist circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;waist_circum
hip_circum =loess.plot(data=data,x=hip_circum ,y=cd4,xlab="Waist circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;hip_circum
arm_right =loess.plot(data=data,x=arm_right ,y=cd4,xlab="Arm(right)circum(cm)",
ylab=bquote("CD4"^"+"*"count")) ;arm_right
triceps_skin =loess.plot(data=data,x=triceps_skin ,y=cd4,xlab="Triceps skin fold(mm)",
ylab=bquote("CD4"^"+"*"count")) ;triceps_skin
height_m =loess.plot(data=data,x=height_m ,y=cd4,xlab="Height(m)",
ylab=bquote("CD4"^"+"*"count")) ;height_m
weight_kg =loess.plot(data=data,x=weight_kg ,y=cd4,xlab="Weight(kg)",
ylab=bquote("CD4"^"+"*"count")) ;weight_kg
bmi =loess.plot(data=data,x=bmi ,y=cd4,xlab=bquote("BMI(kg/m"^"2"*")"),
ylab=bquote("CD4"^"+"*"count")) ;bmi


load(file ="C:/Users/PARTSON/multiplot.rda")

#### random_smooths1a
multiplot(
 legend_custom        ,
 #
 red cells   , hb           , haematocrit   , mcv       ,
 mch         , mchc         , rdw           , platelet  ,
 leucocyte   , neutrophils  , lymphocytes   , monocytes ,
 eosinophils , basophils    ,
 #
 cholesterol , ldl          , triglycerides ,
 #
 glucose     ,
 #
 alt gpt     ,
 cols=4) # save as random_smooths1a

 #### random_smooths1b
multiplot(
    #NB: repeat and crop to make it easy
 ast got     , ast got     , ast got     , ast got     , ast got     ,
 bilirubin   ,bilirubin    ,bilirubin    ,bilirubin    ,bilirubin    ,
 alkaline phos ,alkaline phos ,alkaline phos ,alkaline phos ,alkaline phos ,
 glutamyl_trans,glutamyl_trans,glutamyl_trans,glutamyl_trans,glutamyl_trans,
 #
 cols=4) # save as random_smooths1b

#### random_smooths2a
multiplot(
 legend_custom,
 calcium     , chloride     , magnesium     , potassium   ,
 sodium      ,
 #
 protein     , albumin      , lactate_dehyd ,
 #
 fe          , folate       , vitb12        ,
 #
 urea,
 #
 bp_systolic  , bp_diastolic, pulse_bpm     ,  axillary_temp,
```

```
 #
 waist circum , hip circum  , arm right     ,
 cols=4) # save as random_smooths2a

#### random_smooths2b
multiplot(
triceps_skin, triceps_skin,triceps_skin,triceps_skin,triceps_skin,
height m   , height m   ,height m   ,height m   ,height m   ,
weight kg  , weight kg  ,weight kg  ,weight kg  ,weight_kg  ,
bmi       , bmi       ,bmi       ,bmi       ,bmi       ,
 cols=4) # save as random_smooths2b


if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}
random_smooths1a = image_read("C:/Users/Partson/random_smooths1a.tiff") ;random_smooths1a
random_smooths1b = image_read("C:/Users/Partson/random_smooths1b.tiff") ;random_smooths1b
random_smooths1b_crop =image_crop(random_smooths1b, "598x652+0-535") ;random_smooths1b_crop

#=============================================================================#
#--plot Figure 2.13                                                           #
#=============================================================================#
random_smooths1 = image_append(c(random_smooths1a , random_smooths1b_crop), stack =
TRUE);random_smooths1
image_write(random_smooths1,"C:/Users/Partson/random_smooths1.tiff" , format = "tiff")

random_smooths2a = image_read("C:/Users/Partson/random_smooths2a.tiff") ;random_smooths2a
random_smooths2b = image_read("C:/Users/Partson/random_smooths2b.tiff") ;random_smooths2b
random_smooths2b_crop =image_crop(random_smooths2b, "598x652+0-535") ;random_smooths2b_crop

#=============================================================================#
#--plot Figure 2.14                                                           #
#=============================================================================#
random_smooths2 = image_append(c(random_smooths2a , random_smooths2b_crop), stack = TRUE)
;random_smooths2
image_write(random_smooths2,"C:/Users/Partson/random_smooths2.tiff" , format = "tiff")


#=============================================================================#
# FIGURE 2.15 CORRELATION PLOT                                                #
#=============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/noerror237.rda")
load(file ="C:/Users/PARTSON/bal_complt.rda")
if(!require(corrplot)){install.packages("corrplot")}
dataxcor = bal_complt[, c(2,4,6:52)]
dataxcor = transform(dataxcor,
                visitdate = ifelse(is.na(visitdate)
                , NA, as.Date(visitdate,origin ="1970-01-01")))
                            # Convert date to numeric for imputation
datacor<- round(cor(dataxcor,use = "pairwise.complete.obs"),2)
datacor[is.na(datacor)]=0

# plot
corrplot(datacor,na.label = "NA"
         , type = "lower", order = NULL, col=c("red","blue"),
         tl.col = "black", tl.srt = 45, tl.cex = .7,mar=c(0,0,1,0)
        )

# Croping of figure above to get rid of white spaces
if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}
corrplot = image_read("C:/Users/Partson/corrplot.tiff")
corrplot_trim = corrplot %>%image_trim()
corrplot_trim
image_write(corrplot_trim,"C:/Users/Partson/corrplot_trim.tiff" , format = "tiff")
```

```
#####################################################
#                                                   #
#  REMOVE REDUNDANT FEATURES                        #
#                                                   #
#####################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/bal_complt.rda")
nocorr = c("patientid","visitdate","phase","phase_num","time","cd4")
datacor = bal complt
corrmatrix = cor(datacor[,!names(datacor)%in% nocorr])
                            # also exclude cd4 in finding correlations
save(corrmatrix,file ="C:/Users/PARTSON/corrmatrix.rda")

if(!require(caret)){install.packages("caret")}
                            # required for feature selection
set.seed(123)               # ensure the results are repeatable

hicorrindex = findCorrelation(corrmatrix, cutoff=0.75,
                              verbose = TRUE, names =FALSE, exact = FALSE)
                            # identify all correlations pairs with >= 0.75 and print(verbose)
                            # Col index with highest mean abs corr is selected for removal
hicorrindex = sort(hicorrindex)
                            # sort col indices marked for removal
datacor_a = datacor[,!names(datacor)%in%nocorr]
                            # select data used for correlations only
reducedata_a = datacor_a[,-hicorrindex]
                            # drop marked cols
dropped = datacor_a[!(datacor_a %in% reducedata_a)]
names(dropped)
save(dropped,file ="C:/Users/PARTSON/dropped.rda")

reducedata_b = datacor[!(datacor %in% dropped)]
bal_reducedata = reducedata_b
```

C3: Chapter 3 codes

```
#############################################################################################
#                                                                           #
# MODELL FITTING WITHIN PHASE AND OVERALL                                   #        VARIABLE SCREENING
#                                                                           #
#############################################################

# Overall
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
if(!require(dplyr)){install.packages("dplyr")}
data = reduced_pls[,c(1,3,5,45,6:44,46)]%>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()
#--
#data = subset(data, phase=="2-Acute")
#--
if(!require(VariableScreening)){install.packages("VariableScreening")}

Jmin <- min(table(data$patientid)) - 1
screenResults_a <- screenLD(X = data[,c(5:44)],
                            Y = data[,c(4)],
                            #z = NULL,
                            id = data[,c(1)],
                            time = scale(data[,c(2)]),
                            degree = 3, #3 for cubic splines
                            df = NULL, # NULL length(knots), =degree - intercept.
                            corstr = "AR-M",
                            M = 1
                            )
screenResults_b = as.data.frame(screenResults_a)
if(!require(data.table)){install.packages("data.table")}
screenResults_c = as.data.frame(setDT(screenResults_b, keep.rownames = TRUE))
X = data[,c(5:44)]
Xt_a = t(X)
Xt_b = as.data.frame(Xt_a)
Xt_c = as.data.frame(setDT(Xt_b, keep.rownames = TRUE))
```

```
screenResults_c$rn = Xt_c$rn
setnames(screenResults_c ,c("rn","error","rank"),c("covariate","error0","rank0"))
#--
screenResults0 = screenResults_c %>% arrange(rank0)
save(screenResults0,file ="C:/Users/PARTSON/screenResults0.rda")
#--


# Phase2
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
if(!require(dplyr)){install.packages("dplyr")}
data = reduced_pls[,c(1,3,5,45,6:44,46)]%>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()
#--
data = subset(data, phase=="2-Acute")
#--
if(!require(VariableScreening)){install.packages("VariableScreening")}

Jmin <- min(table(data$patientid)) - 1
screenResults_a <- screenLD(X = data[,c(5:44)],
                            Y = data[,c(4)],
                            #z = NULL,
                            id = data[,c(1)],
                            time = scale(data[,c(2)]),
                            degree = 3, #3 for cubic splines
                            df = NULL, # NULL length(knots), =degree - intercept.
                            corstr = "AR-M",
                            M = 1
                            )
screenResults_b = as.data.frame(screenResults_a)
if(!require(data.table)){install.packages("data.table")}
screenResults_c = as.data.frame(setDT(screenResults_b, keep.rownames = TRUE))
X = data[,c(5:44)]
Xt_a = t(X)
Xt_b = as.data.frame(Xt_a)
Xt_c = as.data.frame(setDT(Xt_b, keep.rownames = TRUE))
screenResults_c$rn = Xt_c$rn
setnames(screenResults_c ,c("rn","error","rank"),c("covariate","error2","rank2"))
#--
screenResults2 = screenResults_c %>% arrange(rank2)
save(screenResults2,file ="C:/Users/PARTSON/screenResults2.rda")
#--


# Phase3
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
if(!require(dplyr)){install.packages("dplyr")}
data = reduced_pls[,c(1,3,5,45,6:44,46)]%>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()
#--
data = subset(data, phase=="3-Early")
#--
if(!require(VariableScreening)){install.packages("VariableScreening")}

Jmin <- min(table(data$patientid)) - 1
screenResults_a <- screenLD(X = data[,c(5:44)],
                            Y = data[,c(4)],
                            #z = NULL,
                            id = data[,c(1)],
                            time = scale(data[,c(2)]),
                            degree = 3, #3 for cubic splines
                            df = NULL, # NULL length(knots), =degree - intercept.
                            corstr = "AR-M",
                            M = 1
                            )
screenResults_b = as.data.frame(screenResults_a)
if(!require(data.table)){install.packages("data.table")}
screenResults_c = as.data.frame(setDT(screenResults_b, keep.rownames = TRUE))
X = data[,c(5:44)]
Xt_a = t(X)
Xt_b = as.data.frame(Xt_a)
Xt_c = as.data.frame(setDT(Xt_b, keep.rownames = TRUE))
screenResults_c$rn = Xt_c$rn
setnames(screenResults_c ,c("rn","error","rank"),c("covariate","error3","rank3"))
```

```
#--
screenResults3 = screenResults_c %>% arrange(rank3)
save(screenResults3,file ="C:/Users/PARTSON/screenResults3.rda")
#--


# Phase4
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
if(!require(dplyr)){install.packages("dplyr")}
data = reduced_pls[,c(1,3,5,45,6:44,46)]%>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()
#--
data = subset(data, phase=="4-Est")
#--
if(!require(VariableScreening)){install.packages("VariableScreening")}

Jmin <- min(table(data$patientid)) - 1
screenResults_a <- screenLD(X = data[,c(5:44)],
                            Y = data[,c(4)],
                            #z = NULL,
                            id = data[,c(1)],
                            time = scale(data[,c(2)]),
                            degree = 3, #3 for cubic splines
                            df = NULL, # NULL length(knots), =degree - intercept.
                            corstr = "AR-M",
                            M = 1
                            )
screenResults b = as.data.frame(screenResults a)
if(!require(data.table)){install.packages("data.table")}
screenResults_c = as.data.frame(setDT(screenResults_b, keep.rownames = TRUE))
X = data[,c(5:44)]
Xt_a = t(X)
Xt_b = as.data.frame(Xt_a)
Xt_c = as.data.frame(setDT(Xt_b, keep.rownames = TRUE))
screenResults c$rn = Xt c$rn
setnames(screenResults_c ,c("rn","error","rank"),c("covariate","error4","rank4"))
#--
screenResults4 = screenResults_c %>% arrange(rank4)
save(screenResults4,file ="C:/Users/PARTSON/screenResults4.rda")
#--


# Phase5
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
if(!require(dplyr)){install.packages("dplyr")}
data = reduced_pls[,c(1,3,5,45,6:44,46)]%>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()
#--
data = subset(data, phase=="5-ART")
#--
if(!require(VariableScreening)){install.packages("VariableScreening")}

Jmin <- min(table(data$patientid)) - 1
screenResults_a <- screenLD(X = data[,c(5:44)],
                            Y = data[,c(4)],
                            #z = NULL,
                            id = data[,c(1)],
                            time = scale(data[,c(2)]),
                            degree = 3, #3 for cubic splines
                            df = NULL, # NULL length(knots), =degree - intercept.
                            corstr = "AR-M",
                            M = 1
                            )
screenResults_a@runtime;
screenResults_b = as.data.frame(screenResults_a)
if(!require(data.table)){install.packages("data.table")}
screenResults_c = as.data.frame(setDT(screenResults_b, keep.rownames = TRUE))
X = data[,c(5:44)]
Xt_a = t(X)
Xt_b = as.data.frame(Xt_a)
Xt_c = as.data.frame(setDT(Xt_b, keep.rownames = TRUE))
screenResults_c$rn = Xt_c$rn
setnames(screenResults_c ,c("rn","error","rank"),c("covariate","error5","rank5"))
```

```
#--
screenResults5 = screenResults_c %>% arrange(rank5)
save(screenResults5,file ="C:/Users/PARTSON/screenResults5.rda")
#--

############################################################
#                                                          #
# RESULTS EXTRACTION FROM THE FITTED MODELS                #
#                                                          #
############################################################
# Load results data
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/screenResults2.rda")
load(file ="C:/Users/PARTSON/screenResults3.rda")
load(file ="C:/Users/PARTSON/screenResults4.rda")
load(file ="C:/Users/PARTSON/screenResults5.rda")
load(file ="C:/Users/PARTSON/screenResults0.rda")

screenLD_ranks = Reduce(function(x,y) merge(x = x, y = y, by = "covariate",sort = FALSE),
       list(screenResults2,screenResults3, screenResults4, screenResults5,screenResults0))%>%
      arrange(-rank0)
save(screenLD_ranks,file ="C:/Users/PARTSON/screenLD_ranks.rda")
#####################################
# Prepare Table 3.2: results data    #
#####################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/screenLD_ranks.rda")
screenLD_ranks_apndx_a = screenLD_ranks%>%arrange(rank0)
screenLD ranks apndx a$covariate <-
gsub("lymphocytes","Lymphocytes",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("basophils","Basophils",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("albumin","Albumin",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("haematocrit","Haematocrit",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("alkaline_phos","ALP",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("mcv","MCV",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("platelet","Platelet",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("potassium","Potassium",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("monocytes","Monocytes",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("protein","Total
protein",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("lactate_dehyd","LDH",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("folate","Folate",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("magnesium","Magnesium",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("glucose","Glucose",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("calcium","Calcium",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("mchc","MCHC",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("red_cells","Red blood
cells",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("sodium","Sodium",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("vitb12","Vitamin
B12",screenLD ranks apndx a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("triceps_skin","Triceps skin
fold",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("triglycerides","Triglycerides",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("neutrophils","Neutrophils",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("ast_got","AST(GOT)",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("eosinophils","Eosinophils",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("height_m","Height",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("chloride","Chloride",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("fe","Fe(Iron)",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("waist_circum","Waist
circum",screenLD ranks apndx a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("ldl","LDL",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("bmi","BMI",screenLD_ranks_apndx_a$covariate)
```

```
screenLD_ranks_apndx_a$covariate <-
gsub("bp_systolic","BP(systolic)",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("bilirubin","Bilirubin",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("arm_right","Arm(right)
circum",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <-
gsub("glutamyl_trans","GGT",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("bp_diastolic","BP(diastolic)",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("rdw","RDW",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("pulse_bpm","Pulse",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("urea","Urea",screenLD_ranks_apndx_a$covariate)
screenLD ranks apndx a$covariate <-
gsub("alt_gpt","ALT(GPT)",screenLD_ranks_apndx_a$covariate)
screenLD_ranks_apndx_a$covariate <- gsub("axillary_temp","Axillary
Temp",screenLD_ranks_apndx_a$covariate)
if(!require(dplyr)){install.packages("dplyr")}
screenLD_ranks_apndx= screenLD_ranks_apndx_a %>%
  mutate_if(is.numeric, function(x) {
                                round(x,4)})
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(screenLD_ranks_apndx,file ="C:/Users/PARTSON/screenLD_ranks_apndx.xlsx"
          ,row.names = FALSE, showNA=FALSE)

#============================================================================#
# Figure 3.2: Plot errors                                                    #
#============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/screenResults0.rda")
if(!require(Rcmdr)){install.packages("Rcmdr")}
if(!require(ggplot2)){install.packages("ggplot2")}

format_plot_errors=(
      theme_bw()    +         # Background fill
      theme(                  # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(size=9,color="black",  face="plain",angle=0),
    axis.text.x = element_text(size=8,angle=90,color="black", hjust=1, vjust=0.25),
    axis.title.y = element_text(size=9,color="black",  face="plain"),
    axis.text.y = element_text(size=8,angle=0,color="black",vjust=0.25),
    legend.position="none",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))
ggplot(screenResults0,
       aes(covariate, error0, group = NULL)) +
       geom_col(aes(fill = factor(rank0)),width=.7) +
       scale_x_discrete(limits=screenResults0$covariate)+
       labs( x=bquote("Desirable <<--------------- Clinical covariates of CD4"^"+"*"count ----
----------->> Undesirable"),
             y="Sum of squared error values"
         )+
      format_plot_errors

# Reshape wide to long data
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/screenLD_ranks.rda")
screen_rank_long <- melt(screenLD_ranks,
        # ID variables - all the variables to keep but not split apart on
    id.vars=c("covariate"),
        # The source columns
    measure.vars=c("rank2", "rank3", "rank4", "rank5","rank0" ),
        # Name of the destination column that will identify the original
        # column that the measurement came from
    variable.name="Rank",
    value.name="value"
     )

if(!require(car)){install.packages("car")}
screen_rank_long$Rank<-recode(screen_rank_long$Rank,"'rank0'='Overall';
                                                    'rank2'='2-Acute';
                                                     'rank3'= '3-Early';
                                                      'rank4'='4-Est';
                                                      'rank5'='5-ART'")
```

```
#=============================================================================#
# Figure 3.3: Plot ranks                                                      #
#=============================================================================#
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

format_plot_screen_rank=(
      #theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element text(color="black", size=9, face="bold",
                              hjust = 0.5,margin = margin(t = 0, b = -0)),
    axis.title.x = element_text(size=9,color="black",  face="plain",angle=0),
    axis.text.x = element_text(size=9,angle=0,color="black", hjust=.5, vjust=1),
    axis.title.y = element text(size=9,color="black",  face="plain"),
    axis.text.y = element text(size=9,angle=0,color="black",vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid"
                                     ,colour ="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))

ggplot(screen_rank_long, aes(x=covariate, y=value)) + geom_line(color="black") +
         geom_point(aes(fill = factor(Rank)),size=4, shape=21) +
       scale_fill_manual(values=c("green3","blue","red","orange","cyan"),
                         name="Infection phase:")+
         scale_x_discrete(limits=screenLD_ranks$covariate)+
       labs( x=bquote("Clinical covariates of CD4"^"+"*"count"),
             y="1=Most - - - - - - - - Desirability rank (VariableScreening) - - - - - - - -
Least=40" ,title=""
           )+
     coord flip()+
     format_plot_screen_rank

#############################################################################################
#                                      #
# CROSS VALIDATION & MODEL FITTING     #                          PGEE
#                                      #
#######################################

# PHASE 2
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
data_a = subset(reduced_pls, phase=="2-Acute")
data_b <- data_a[,-c(2,4,5)]
if(!require(dplyr)){install.packages("dplyr")}
data_c = data_b[,c(1,42,2:41,43)] %>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()

data_d = data_c[,4:43] %>%
  mutate_if(is.numeric, function(x) {
                       (x-mean(x))/sd(x)
                                   })
data = data.frame(data_c[,1:3],data_d)
if(!require(data.table)){install.packages("data.table")}

setnames(data,c("patientid","cd4"),c("id","y"))
formula <- "y~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,5,0.2)

if(!require(PGEE)){install.packages("PGEE")}
cv <- CVfit(formula=formula,id = id, data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3)

cv$lam.opt
print(cv)

lambda <- cv$lam.opt
pgee_fit = PGEE(formula = formula, id = id, data = data, na.action = NULL,
family = family, corstr = "AR-1", Mv = NULL,
beta_int = NULL, R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = lambda, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3, silent = FALSE)

pgee_fit2 = pgee_fit
```

```
save(pgee_fit2,file ="C:/Users/PARTSON/pgee_fit2.rda")

# PHASE 3
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
data_a = subset(reduced_pls, phase=="3-Early")
data_b <- data_a[,-c(2,4,5)]
if(!require(dplyr)){install.packages("dplyr")}
data_c = data_b[,c(1,42,2:41,43)] %>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()

data_d = data_c[,4:43] %>%
  mutate_if(is.numeric, function(x) {
                      (x-mean(x))/sd(x)
                                  })
data = data.frame(data_c[,1:3],data_d)
if(!require(data.table)){install.packages("data.table")}

setnames(data,c("patientid","cd4"),c("id","y"))
formula <- "y~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,5,0.2)

if(!require(PGEE)){install.packages("PGEE")}
cv <- CVfit(formula=formula,id = id, data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3)

cv$lam.opt
print(cv)

lambda <- cv$lam.opt
pgee_fit = PGEE(formula = formula, id = id, data = data, na.action = NULL,
family = family, corstr = "AR-1", Mv = NULL,
beta_int = NULL, R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = lambda, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3, silent = FALSE)

pgee_fit3 = pgee_fit
save(pgee_fit3,file ="C:/Users/PARTSON/pgee_fit3.rda")

# PHASE 4
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
data_a = subset(reduced_pls, phase=="4-Est")
data_b <- data_a[,-c(2,4,5)]
if(!require(dplyr)){install.packages("dplyr")}
data_c = data_b[,c(1,42,2:41,43)] %>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()

data_d = data_c[,4:43] %>%
  mutate_if(is.numeric, function(x) {
                      (x-mean(x))/sd(x)
                                  })
data = data.frame(data_c[,1:3],data_d)
if(!require(data.table)){install.packages("data.table")}

setnames(data,c("patientid","cd4"),c("id","y"))
formula <- "y~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,5,0.2)

if(!require(PGEE)){install.packages("PGEE")}
cv <- CVfit(formula=formula,id = id, data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3)

cv$lam.opt
print(cv)

lambda <- cv$lam.opt
pgee_fit = PGEE(formula = formula, id = id, data = data, na.action = NULL,
family = family, corstr = "AR-1", Mv = NULL,
beta_int = NULL, R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = lambda, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3, silent = FALSE)
```

```
pgee_fit4 = pgee_fit
save(pgee_fit4,file ="C:/Users/PARTSON/pgee_fit4.rda")

# PHASE 5
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
data_a = subset(reduced_pls, phase=="5-ART")
data_b <- data_a[,-c(2,4,5)]
if(!require(dplyr)){install.packages("dplyr")}
data_c = data_b[,c(1,42,2:41,43)] %>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()

data_d = data_c[,4:43] %>%
  mutate_if(is.numeric, function(x) {
                      (x-mean(x))/sd(x)
                              })
data = data.frame(data_c[,1:3],data_d)
if(!require(data.table)){install.packages("data.table")}

setnames(data,c("patientid","cd4"),c("id","y"))
formula <- "y~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,5,0.2)

if(!require(PGEE)){install.packages("PGEE")}
cv <- CVfit(formula=formula,id = id, data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3)

cv$lam.opt
print(cv)

lambda <- cv$lam.opt
pgee_fit = PGEE(formula = formula, id = id, data = data, na.action = NULL,
family = family, corstr = "AR-1", Mv = NULL,
beta_int = NULL, R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = lambda, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3, silent = FALSE)

pgee_fit5 = pgee_fit
save(pgee_fit5,file ="C:/Users/PARTSON/pgee_fit5.rda")


# OVERALL
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
data_a = subset(reduced_pls, phase!="5-ART.")
data_b <- data_a[,-c(2,4,5)]
if(!require(dplyr)){install.packages("dplyr")}
data_c = data_b[,c(1,42,2:41,43)] %>% group_by(patientid)%>%
  mutate(time =1:n()) %>% as.data.frame()

data_d = data_c[,4:43] %>%
  mutate_if(is.numeric, function(x) {
                      (x-mean(x))/sd(x)
                              })
data = data.frame(data_c[,1:3],data_d)
if(!require(data.table)){install.packages("data.table")}

setnames(data,c("patientid","cd4"),c("id","y"))
formula <- "y~.-id"
family <- gaussian(link = "identity")
lambda.vec <- seq(0.1,5,0.2)

if(!require(PGEE)){install.packages("PGEE")}
cv <- CVfit(formula=formula,id = id, data = data, family = family,
fold = 4, lambda.vec = lambda.vec, pindex = c(1,2), eps = 10^-6, maxiter = 30,
tol = 10^-3)

cv$lam.opt
print(cv)

lambda <- cv$lam.opt
pgee_fit = PGEE(formula = formula, id = id, data = data, na.action = NULL,
family = family, corstr = "AR-1", Mv = NULL,
beta_int = NULL, R = NULL, scale.fix = TRUE,
scale.value = 1, lambda = lambda, pindex = c(1,2), eps = 10^-6, maxiter = 30,
```

```
tol = 10^-3, silent = FALSE)

pgee fit = pgee fit
save(pgee_fit,file ="C:/Users/PARTSON/pgee_fit.rda")
#####################################
#                                   #
# EXTRACTION OF RESULTS             #
#                                   #
#####################################

# load results
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/pgee_fit2.rda")
load(file ="C:/Users/PARTSON/pgee_fit3.rda")
load(file ="C:/Users/PARTSON/pgee_fit4.rda")
load(file ="C:/Users/PARTSON/pgee_fit5.rda")
load(file ="C:/Users/PARTSON/pgee_fit.rda")
#=============================================================================#
# Table 3.3                                                                   #
#=============================================================================#
summary pgee = data.frame(
        Parameter = c("Optimal lambda"   ,"Number of selected covariates"),
        phase2=
c(summary(pgee_fit2)[[10]],length(which(abs(coef(summary(pgee_fit2))[,"Estimate"]) > 10^-3))),
        phase3=
c(summary(pgee_fit3)[[10]],length(which(abs(coef(summary(pgee_fit3))[,"Estimate"]) > 10^-3))),
        phase4=
c(summary(pgee_fit4)[[10]],length(which(abs(coef(summary(pgee_fit4))[,"Estimate"]) > 10^-3))),
        phase5=
c(summary(pgee_fit5)[[10]],length(which(abs(coef(summary(pgee_fit5))[,"Estimate"]) > 10^-3))),

Overall=c(summary(pgee_fit)[[10]],length(which(abs(coef(summary(pgee_fit))[,"Estimate"]) >
10^-3))))
setnames(summary_pgee,c("phase2","phase3","phase4","phase5"),c("2-Acute","3-Early","4-Est","5-
ART"))

if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(summary_pgee,file ="C:/Users/PARTSON/summary_pgee.xlsx"
          ,row.names = FALSE, showNA=FALSE)

# Prepare Table 3.4

# Sort the estimates and add the rank column
coef2_a = summary(pgee_fit2)[[7]][-1,c(1,5)] %>%as.data.frame()
coef2_b =  as.data.frame(setDT(coef2_a, keep.rownames = TRUE))%>%
           arrange(-abs(Estimate))
pgee_coef2 = subset(coef2_b[,-3],rn!="time")%>%
          mutate(Rank =1:n()) %>% as.data.frame()
coef2_apndx = pgee_coef2
setnames(coef2_apndx,c("rn","Estimate","Rank"),c("covariate","Estimate2","Rank2"))
save(pgee_coef2, file ="C:/Users/PARTSON/pgee_coef2.rda")

coef3_a = summary(pgee_fit3)[[7]][-1,c(1,5)] %>%as.data.frame()
coef3_b =  as.data.frame(setDT(coef3_a, keep.rownames = TRUE))%>%
           arrange(-abs(Estimate))
pgee_coef3 = subset(coef3_b[,-3],rn!="time")%>%
          mutate(Rank =1:n()) %>% as.data.frame()
coef3 apndx = pgee coef3
setnames(coef3_apndx,c("rn","Estimate","Rank"),c("covariate","Estimate3","Rank3"))
save(pgee_coef3, file ="C:/Users/PARTSON/pgee_coef3.rda")

coef4_a = summary(pgee_fit4)[[7]][-1,c(1,5)] %>%as.data.frame()
coef4_b =  as.data.frame(setDT(coef4_a, keep.rownames = TRUE))%>%
           arrange(-abs(Estimate))
pgee_coef4 = subset(coef4_b[,-3],rn!="time")%>%
          mutate(Rank =1:n()) %>% as.data.frame()
coef4_apndx = pgee_coef4
setnames(coef4_apndx,c("rn","Estimate","Rank"),c("covariate","Estimate4","Rank4"))
save(pgee_coef4, file ="C:/Users/PARTSON/pgee_coef4.rda")

coef5_a = summary(pgee_fit5)[[7]][-1,c(1,5)] %>%as.data.frame()
coef5_b =  as.data.frame(setDT(coef5_a, keep.rownames = TRUE))%>%
           arrange(-abs(Estimate))
pgee_coef5 = subset(coef5_b[,-3],rn!="time")%>%
          mutate(Rank =1:n()) %>% as.data.frame()
coef5 apndx = pgee coef5
setnames(coef5_apndx,c("rn","Estimate","Rank"),c("covariate","Estimate5","Rank5"))
```

```
save(pgee_coef5, file ="C:/Users/PARTSON/pgee_coef5.rda")

coef_a = summary(pgee_fit)[[7]][-1,c(1,5)] %>%as.data.frame()
coef_b =  as.data.frame(setDT(coef_a, keep.rownames = TRUE))%>%
           arrange(-abs(Estimate))
pgee_coef0 = subset(coef_b[,-3],rn!="time")%>%
          mutate(Rank =1:n()) %>% as.data.frame()
coef apndx = pgee coef0
setnames(coef_apndx,c("rn","Estimate","Rank"),c("covariate","Estimate0","Rank0"))
save(pgee_coef0, file ="C:/Users/PARTSON/pgee_coef0.rda")
#=============================================================================#
# Table 3.4:  Combine all phase tables                                        #
#=============================================================================#
pgee_coefs = Reduce(function(x,y) merge(x = x, y = y, by = "covariate",sort = FALSE),
       list(coef2_apndx,coef3_apndx , coef4_apndx, coef5_apndx,coef_apndx))%>%
      arrange(-Rank0)
save(pgee_coefs,file ="C:/Users/PARTSON/pgee_coefs.rda")

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/pgee_coefs.rda")
pgee_coefs_apndx = pgee_coefs%>%arrange(Rank0)
pgee_coefs_apndx$covariate <- gsub("lymphocytes","Lymphocytes",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("basophils","Basophils",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("albumin","Albumin",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("haematocrit","Haematocrit",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("alkaline_phos","ALP",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("mcv","MCV",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("platelet","Platelet",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("potassium","Potassium",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("monocytes","Monocytes",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("protein","Total protein",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("lactate_dehyd","LDH",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("folate","Folate",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("magnesium","Magnesium",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("glucose","Glucose",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("calcium","Calcium",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("mchc","MCHC",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("red_cells","Red blood cells",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("sodium","Sodium",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("vitb12","Vitamin B12",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("triceps_skin","Triceps skin
fold",pgee coefs apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("triglycerides","Triglycerides",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("neutrophils","Neutrophils",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("ast_got","AST(GOT)",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("eosinophils","Eosinophils",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("height_m","Height",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("chloride","Chloride",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("fe","Fe(Iron)",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("waist_circum","Waist circum",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("ldl","LDL",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("bmi","BMI",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("bp_systolic","BP(systolic)",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("bilirubin","Bilirubin",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("arm_right","Arm(right) circum",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("glutamyl_trans","GGT",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("bp_diastolic","BP(diastolic)",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("rdw","RDW",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("pulse_bpm","Pulse",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("urea","Urea",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("alt_gpt","ALT(GPT)",pgee_coefs_apndx$covariate)
pgee_coefs_apndx$covariate <- gsub("axillary_temp","Axillary Temp",pgee_coefs_apndx$covariate)

if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(pgee_coefs_apndx,file ="C:/Users/PARTSON/pgee_coefs_apndx.xlsx"
          ,row.names = FALSE, showNA=FALSE)
#####################################
#                                   #
# PLOT RANKS                        #
#                                   #
#####################################
# Prepare Figure 3.4  #
#######################

# load the results
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/pgee_coef2.rda")
```

```
load(file ="C:/Users/PARTSON/pgee_coef3.rda")
load(file ="C:/Users/PARTSON/pgee_coef4.rda")
load(file ="C:/Users/PARTSON/pgee_coef5.rda")
load(file ="C:/Users/PARTSON/pgee_coef0.rda")

# add the effect column
pgee_coef2$effect2 = ifelse(pgee_coef2$Estimate2>0,"positive","negative")
pgee_coef3$effect3 = ifelse(pgee_coef3$Estimate3>0,"positive","negative")
pgee_coef4$effect4 = ifelse(pgee_coef4$Estimate4>0,"positive","negative")
pgee_coef5$effect5 = ifelse(pgee_coef5$Estimate5>0,"positive","negative")
pgee_coef0$effect0 = ifelse(pgee_coef0$Estimate0>0,"positive","negative")

# Plot the estimates within each phase and overall
format plot pgee=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",
                              hjust = 0.5,margin = margin(t = 10, b = -10)),
    axis.title.x = element_text(size=9,color="black",  face="plain",angle=0),
    axis.text.x = element text(size=8,angle=90,color="black", hjust=1, vjust=0.5),
    axis.title.y = element text(size=9,color="black",  face="plain"),
    axis.text.y = element_text(size=8,angle=0,color="black",vjust=0.25),
    legend.position="none",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))
pgee plot2 =ggplot(pgee coef2,
       aes(covariate, Estimate2)) +
       geom_col(aes(fill = factor(effect2)),width=.7) +
         scale_fill_manual(values=c("red","green3"),
                           name="")+
       scale_x_discrete(limits=pgee_coef2$covariate)+
       labs( x=bquote(""),
             y="Penalized estimates",title="2-Acute"
         )+
      geom hline(yintercept= 10^-3, color ="orange", lty =2)+
      geom_hline(yintercept= -10^-3, color ="orange", lty =2)+
      format_plot_pgee

pgee plot3 =ggplot(pgee coef3,
       aes(covariate, Estimate3)) +
       geom_col(aes(fill = factor(effect3)),width=.7) +
         scale_fill_manual(values=c("red","green3"),
                           name="")+
       scale x discrete(limits=pgee_coef3$covariate)+
       labs( x=bquote(""),
             y="Penalized estimates",title="3-Early"
         )+
      geom_hline(yintercept= 10^-3, color ="orange", lty =2)+
      geom hline(yintercept= -10^-3, color ="orange", lty =2)+
      format_plot_pgee

pgee_plot4 =ggplot(pgee_coef4,
       aes(covariate, Estimate4)) +
       geom_col(aes(fill = factor(effect4)),width=.7) +
         scale_fill_manual(values=c("red","green3"),
                           name="")+
       scale x discrete(limits=pgee_coef4$covariate)+
       labs( x=bquote(""),
             y="Penalized estimates",title="4-Est"
         )+
      geom hline(yintercept= 10^-3, color ="orange", lty =2)+
      geom hline(yintercept= -10^-3, color ="orange", lty =2)+
      format_plot_pgee

pgee_plot5 =ggplot(pgee_coef5,
       aes(covariate, Estimate5)) +
       geom_col(aes(fill = factor(effect5)),width=.7) +
         scale_fill_manual(values=c("red","green3"),
                           name="")+
       scale_x_discrete(limits=pgee_coef5$covariate)+
       labs( x="",
             y="Penalized estimates" ,title="5-ART"
         )+
      geom_hline(yintercept= 10^-3, color ="orange", lty =2)+
```

```
      geom_hline(yintercept= -10^-3, color ="orange", lty =2)+
      format_plot_pgee

pgee plot overall =ggplot(pgee coef0,
       aes(covariate, Estimate0)) +
       geom_col(aes(fill = factor(effect0)),width=.7) +
         scale_fill_manual(values=c("red","green3"),
                           name="")+
       scale x discrete(limits=pgee coef0$covariate)+
       labs( x=bquote("Desirable <<---------------- Clinical covariates of CD4"^"+"*"count ----
----------->> Undesirable"),
             y="Penalized estimates" ,title="Overall"
           )+
      geom hline(yintercept= 10^-3, color ="orange", lty =2)+
      geom hline(yintercept= -10^-3, color ="orange", lty =2)+
      format_plot_pgee

# Append the graphs
load(file ="C:/Users/PARTSON/multiplot.rda")
multiplot(
 pgee plot2 , pgee_plot3, pgee_plot4 ,pgee_plot5,
 cols=1)
pgee_plot_overall

if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
pgee_vip1_5 = image_read("C:/Users/Partson/pgee_vip1_5.tiff") ;pgee_vip1_5
pgee_vip_overall = image_read("C:/Users/Partson/pgee_vip_overall.tiff") ;pgee_vip_overall
#=============================================================================#
# Plot Figure 3.4: combine the estimate plots                                 #
#=============================================================================#
pgee_vip_est = image_append(c(pgee_vip1_5 , pgee_vip_overall), stack = TRUE) ;pgee_vip
image_write(pgee_vip_est,"C:/Users/Partson/pgee_vip_est.tiff", format = "tiff")


#####################################
#                                   #
# PLOT ESTIMATES                    #
#                                   #
#####################################
# Prepare Figure 3.5 #
######################

# Reshape wide to long
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/pgee_coefs.rda")
if(!require(reshape2)){install.packages("reshape2")}
coefs rank long <- melt(pgee coefs,
        # ID variables - all the variables to keep but not split apart on
    id.vars=c("covariate"),
        # The source columns
    measure.vars=c("Rank2", "Rank3", "Rank4", "Rank5","Rank0" ),
        # Name of the destination column that will identify the original
        # column that the measurement came from
    variable.name="Rank",
    value.name="value"
     )

if(!require(car)){install.packages("car")}
coefs rank long$Rank<-recode(coefs rank long$Rank,"'Rank0'='Overall';
                                                   'Rank2'='2-Acute';
                                                    'Rank3'= '3-Early';
                                                     'Rank4'='4-Est';
                                                     'Rank5'='5-ART'")
#=============================================================================#
#Plot Figure 3.5                                                              #
#=============================================================================#
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

format_plot_coefs_rank=(
      #theme_bw()    +        # Background fill
      theme(               # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",
                              hjust = 0.5,margin = margin(t = 0, b = -0)),
    axis.title.x = element text(size=9,color="black",  face="plain",angle=0),
    axis.text.x = element_text(size=9,angle=0,color="black", hjust=1, vjust=0.5),
```

```
    axis.title.y = element_text(size=9,color="black",  face="plain"),
    axis.text.y = element text(size=9,angle=0,color="black",vjust=0.25),
    legend.position="top",
    legend.background = element rect(fill="bisque",size=0.2,linetype="solid"
                                     ,colour ="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))

ggplot(coefs rank long, aes(x=covariate, y=value)) + geom line(color="black") +
        geom_point(aes(fill = factor(Rank)),size=4, shape=21) +
       scale_fill_manual(values=c("green3","blue","red","orange","cyan"),
                         name="Infection phase:")+
        scale x discrete(limits=pgee coefs$covariate)+
       labs( x=bquote("Clinical covariates of CD4"^"+"*"count"),
             y="1=Most- - - - - - - - - - - - - Desirability rank (PGEE) - - - - - - - - - - -
- - Least=40" ,title=""
           )+
     coord_flip()+
  format_plot_coefs_rank


#####################################=============================================================#
#                                    #
#               SPLS                 #
#                                    #
#####################################

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/bal_complt.rda")

#####################################################################
# CORRELATIONS OF CD4 AND PREDICTORS WITH P-VALUES                  #
#####################################################################

# Force BMI into model
load(file ="C:/Users/PARTSON/bal_reducedata.rda")
             # From DataManipFinal237
load(file ="C:/Users/PARTSON/bal_complt.rda")
             # From DataManipFinal237
reduced pls = data.frame(bal reducedata, bmi=bal complt[,"bmi"])
save(reduced_pls,file ="C:/Users/PARTSON/reduced_pls.rda")

## Summary statistics and correlation with response (CD4)
load(file ="C:/Users/PARTSON/reduced_pls.rda")
no_summary = c("patientid","visitdate","phase"
               ,"phase num","time"
               #,"bmi" # release if you wish to exclude
               )
data2summary = reduced_pls[,!names(reduced_pls)%in%no_summary]
if(!require(Hmisc)){install.packages("Hmisc")}
corr summary =rcorr(as.matrix(data2summary), type="pearson")
                           # finding correlations
rpvalues = corr summary$P
rvalues  = corr_summary$r

corr_cd4_r = rvalues[,"cd4"]  # correlation values
corr cd4 p = rpvalues[,"cd4"] # P-values for the correlations
corr cd4 = data.frame(corr cd4 r, corr cd4 p)
                              # merge r and p-values for CD4 with predictors
corr_pred_cd4_a = round(corr_cd4,4) # round of values to 4dp

if(!require(data.table)){install.packages("data.table")}
corr_pred_cd4_b = as.data.frame(setDT(corr_pred_cd4_a, keep.rownames = TRUE))
setnames(corr pred cd4 b,"rn","covariate")
corr_pred_cd4 = corr_pred_cd4_b[-which(corr_pred_cd4_b$covariate == "cd4"),]
                              # drop row with CD4
save(corr_pred_cd4,file ="C:/Users/PARTSON/corr_pred_cd4.rda")
#****************

#################################################################
# Fit model to whole data                                       #
#################################################################

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(dplyr)){install.packages("dplyr")}
if(!require(mixOmics)){install.packages("mixOmics")}
```

```
load(file ="C:/Users/PARTSON/reduced_pls.rda")
redumodel_pls = reduced_pls #

# check a multilevel design is ok:
table(redumodel_pls$patientid,redumodel_pls$phase)
design = data.frame(redumodel_pls[,c("patientid","phase","time")])
save(design,file ="C:/Users/PARTSON/design.rda")
# set up predictor and response matrices
no_X = c(
         "patientid","visitdate","phase"
        ,"phase_num","time","cd4"
        #,"bmi"    # release if you wish not to force into model
         )
X=redumodel_pls[,!names(reduced_pls)%in%no_X]
Y=redumodel_pls["cd4"]
ncomp         = round(ncol(X)) # set the number of components to number of predictors
ncomp

Xw = withinVariation(X,design)
# first, learn the model on the whole data set

spls              = spls(X,Y,
    ncomp         = ncomp, # number of components. Should be lower than
                     #the number of columns of variates
    keepX         = c(rep(ncol(X),ncomp)), # for each comp select 10 X-variables
    keepY         = c(rep(ncol(Y),ncomp)), # for each comp select 4 Y-variables
    scale         = TRUE,
    mode          = c("regression"),#"canonical","invariant","classic",
                                    # details in ?pls
    tol           = 1e-06,
    max.iter      = 500, # changed from 100
    near.zero.var = TRUE,
    logratio      = "none",
    multilevel    = design,
    all.outputs   = TRUE)
save(spls,file ="C:/Users/PARTSON/spls.rda")

perf.spls        = perf(spls,
     criterion   = c("all"),   # "MSEP", "R2", "Q2"
     validation  = c("loo"),   # "Mfold"", no need to repeat the process for "loo"
     folds       = 20,
     progressBar = TRUE,
     setseed     = 123)
                      # Throws some warning messages of non-convergence
                      # May ingore because tuned model will be error free
save(perf.spls,file ="C:/Users/PARTSON/perf.spls.rda")

#################################################################
# Prepare Table 3.5: Extract dataset of performance measures    #
#             of the fitted PLS model                           #
#################################################################
rm(list= ls()[!(ls() %in% c("datafinal","reduced_Data","spls","X","Y","design"))])
                                          # clear environment and leave the selected
#if(!is.null(dev.list())) dev.off()        # clear all plots
cat("\014")                               # clear console
load(file ="C:/Users/PARTSON/spls.rda")
load(file ="C:/Users/PARTSON/perf.spls.rda")

## Extract performance data
if(!require(data.table)){install.packages("data.table")}
MSEP = data.frame(MSEP = t(perf.spls[["MSEP"]]))
setnames(MSEP,c("Y"),c("MSEP"))

R2 = data.frame(R2 = t(perf.spls[["R2"]]))
setnames(R2,c("Y"),c("R2"))

Q2.total = perf.spls[["Q2.total"]]

PRESS = perf.spls[["PRESS"]]
Explained variance  =as.data.frame(spls$explained variance)
                  # amount of variance explained per component (note that
                  # contrary to PCA, this amount may not decrease as the
                  # aim of the method is not to maximise the variance, but
                  # the covariance between data sets
setnames(Explained_variance ,c("X","Y"),c("X-Expl.Var","Y-Expl.Var"))
```

```
#============================================================================#
# Table 3.5                                                                  #
#============================================================================#
data perf a = data.frame(MSEP,PRESS,R2,Q2.total,Explained variance)
data_perf_b = as.data.frame(setDT(data_perf_a, keep.rownames = TRUE))
setnames(data_perf_b ,"rn","Component")
if(!require(dplyr)){install.packages("dplyr")}
data_perf_c =data_perf_b %>%
  mutate(Component =1:n()) %>% as.data.frame()
data perfb4tuning = data perf c
save(data_perfb4tuning,file ="C:/Users/PARTSON/data_perfb4tuning.rda")

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/data_perfb4tuning.rda")

data_perfb4tuning_table = data_perfb4tuning
data_perfb4tuning_table$Component = as.character(data_perfb4tuning_table$Component)
data_perfb4tuning_table =data_perfb4tuning_table%>%
  mutate_if(is.numeric, function(x) {
    sprintf("%.4f",round(x,4))})
if(!require(xlsx)){install.packages("xlsx")}

write.xlsx(data_perfb4tuning_table,file ="C:/Users/PARTSON/data_perfb4tuning_table.xlsx"
          ,row.names = FALSE
          )

################################################################
# Visualise performance measures of the fitted PLS model       #
# Plot performance graphs before tuning                        #
################################################################

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/data_perfb4tuning.rda")

# Prepare Figure 3.6

# Convert data from wide to long
if(!require(reshape2)){install.packages("reshape2")}

x = colnames(data_perfb4tuning[,-1])     # Removes the first(-1) column, i.e. "ID"
W2L diagnostic b4 = melt(data perfb4tuning,id.vars = "Component",
                           # id.vars - additional variables to keep in
                             # the output
        measure.vars = x ,  # Variables to be reshaped
        variable.name="diagnostic",
                           # Name of variable used to store measured
                             # variable names
        value.name="value", # Name of variable used to store values
        na.rm = TRUE)       # The rm=Remove the "NA". If set to FALSE the
                             # NA will appear
save(W2L_diagnostic_b4,file ="C:/Users/PARTSON/W2L_diagnostic_b4.rda")

# Set up graph format
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/W2L_diagnostic_b4.rda")
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

format plot perf=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title   = element_text(color="black",size =9 , face ="bold"  ,hjust = 0.5),
    axis.title.x = element_text(color="black",size =9 , face ="bold"  ,angle =0),
    axis.text.x  = element text(color="black",size =8 ,angle=90, hjust=0.5     ,vjust =0.5),
    axis.title.y = element text(color="black",size =9 , face ="bold"),
    axis.text.y  = element_text(color="black",size =9 ,angle=0 ,vjust =0.25)   ,
    legend.position   ="top",
    legend.background = element_rect(fill    ="bisque" ,size   =0.2,
                                     linetype="solid"  ,colour ="lightblue"),
    legend.title = element text(colour="black",size=8,face="bold"),
    legend.text  = element_text(colour="black",size=8,face="plain")
     ))
```

```
#=============================================================================#
# Plot Figure 3.6                                                             #
#=============================================================================#
ggplot(W2L diagnostic b4,
       aes(Component, value, group = diagnostic)) +
       facet_grid(diagnostic~., scales="free")+
       scale_x_continuous(breaks = round(seq(min(W2L_diagnostic_b4$Component),
                          max(W2L_diagnostic_b4$Component), by = 1),0)) +
       geom line(color ="blue",lwd=0.7) +
       labs( x="SPLS components",
             y='Performance measures of the fitted SPLS model'
            ,title=" "
           )+
       #coord flip()+
       geom vline(xintercept= 2, color ="red", lty =2)+
       format_plot_perf

# Extract values of components 1 and 2
load(file ="C:/Users/PARTSON/data_perfb4tuning.rda")
data_comp1_2 = round(subset(data_perfb4tuning, Q2.total>0.09|X.Expl.Var>0.7|Y.Expl.Var>0.7),4)
                          #select components 1 and 2 for details to 4 digits
data_comp1_2
ncomp.tuned = nrow(data_comp1_2)
save(data_comp1_2,file ="C:/Users/PARTSON/data_comp1_2.rda")
save(ncomp.tuned,file ="C:/Users/PARTSON/ncomp.tuned.rda")

# Save detailed diagnostics for comp 1 and 2
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(data_comp1_2,file ="C:/Users/PARTSON/data_comp1_2.xlsx",row.names = FALSE )

#################################################################
# Re-fit the tuned model with tuned number of components        #
#################################################################
rm(list= ls()[!(ls() %in% c("datafinal","reduced_Data","spls","X","Y",
                            "design","ncomp.tuned"))])
                                        # clear environment and leave the selected
cat("\014")                             # clear console

if(!require(mixOmics)){install.packages("mixOmics")}
load(file ="C:/Users/PARTSON/design.rda")
load(file ="C:/Users/PARTSON/ncomp.tuned.rda")
load(file ="C:/Users/PARTSON/reduced_pls.rda")
redumodel_pls = reduced_pls

no_X = c(
         "patientid","visitdate","phase"
        ,"phase_num","time","cd4"
        )
X=redumodel_pls[,!names(reduced_pls)%in%no_X]
Y=redumodel_pls["cd4"]

spls2             = spls(X,Y,
    ncomp         = ncomp.tuned,
    keepX         = c(rep(ncol(X),ncomp.tuned)), # for each comp select all X-variables
    keepY         = c(rep(ncol(Y),ncomp.tuned)), # for each comp select all Y-variables
    scale         = TRUE,
    mode          = c("regression"),# "canonical", "invariant", "classic",
    tol           = 1e-06,
    max.iter      = 100, #changed from 500
    near.zero.var = TRUE,
    logratio      = "none",
    multilevel    = design,
    all.outputs   = TRUE)
                  # NB: No warnings here and sounds good
perf.spls2       = perf(spls2,
     criterion   = c("all"),   # "MSEP", "R2", "Q2"
     validation  = c("loo"), #, "loo"
     folds       = 20,  # changed from 20
     progressBar = TRUE,
     setseed     = 123)
save(spls2,file ="C:/Users/PARTSON/spls2.rda")
save(perf.spls2,file ="C:/Users/PARTSON/perf.spls2.rda")

# Confirm explained variances
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/spls2.rda")
load(file ="C:/Users/PARTSON/perf.spls2.rda")
```

```
Explained_variance2  =as.data.frame(spls2$explained_variance)
Explained_variance2

if(!require(data.table)){install.packages("data.table")}
MSEP = data.frame(MSEP = t(perf.spls2[["MSEP"]]))
setnames(MSEP,c("Y"),c("MSEP"))

R2 = data.frame(R2 = t(perf.spls2[["R2"]]))
setnames(R2,c("Y"),c("R2"))

Q2.total = perf.spls2[["Q2.total"]]

PRESS = perf.spls2[["PRESS"]]
Explained variance  =as.data.frame(spls2$explained variance)
                  # amount of variance explained per component (note that
                  # contrary to PCA, this amount may not decrease as the
                  # aim of the method is not to maximise the variance, but
                  # the covariance between data sets
setnames(Explained_variance ,c("X","Y"),c("X-Expl.Var","Y-Expl.Var"))

data perf2 a = data.frame(MSEP,PRESS,R2,Q2.total,Explained variance)
data_perf2_b = as.data.frame(setDT(data_perf2_a, keep.rownames = TRUE))
setnames(data_perf2_b ,"rn","Component")
if(!require(dplyr)){install.packages("dplyr")}
data_perf_tuned_table =data_perf2_b %>%
  mutate(Component =1:n()) %>% as.data.frame()

data_perf_tuned_table$Component = as.character(data_perf_tuned_table$Component)
data_perf_tuned_table = data_perf_tuned_table%>%
  mutate_if(is.numeric, function(x) {
    sprintf("%.4f",round(x,4))})
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(data_perfb4tuning_table,file ="C:/Users/PARTSON/data_perfb4tuning_table.xlsx"
          ,row.names = FALSE
          )

#####################################################
# Extract final results from re-fitted tuned model #
#####################################################
rm(list= ls()[!(ls() %in% c("datafinal","reduced_Data","spls","X","Y",
                            "design","ncomp.tuned","spls2"))])
                                        # clear environment and leave the selected
cat("\014")                             # clear console

if(!require(dplyr)){install.packages("dplyr")}
if(!require(mixOmics)){install.packages("mixOmics")}
load(file ="C:/Users/PARTSON/spls2.rda")

## VIP
vip_a    = as.data.frame(vip(spls2))
                         # computes the influence on the Y-responses of
                          #every predictor X in the mode
                         # VIP coefficients thus represent the importance of
                           # each X variable in fitting both the X- and Y-variates
                           # large VIP>1, are the most relevant for explaining Y

if(!require(data.table)){install.packages("data.table")}
vip_b = as.data.frame(setDT(vip_a, keep.rownames = TRUE))
setnames(vip b ,"rn","covariate")
vip = data.frame(vip_b[,c("covariate","comp 2")])
setnames(vip ,2,"VIP")

# loadings values: contribution of each variable for each component
loadings a =selectVar(spls2, comp = 2)$`X`$value #  outputs the loading value
                         # for each selected variable, the loadings are
                         # ranked according to their absolute value
loadings_a    # OR loadings = data.frame(loadings = spls2$loadings$`X`[,2])
loadings_b = as.data.frame(setDT(loadings_a, keep.rownames = TRUE))
setnames(loadings_b,c("rn","value.var"),c("covariate","Loadings"))
loadings =loadings_b

# Regression coefficients
coef_a = data.frame(Coefficients =spls2$mat.c[,2]) # matrix of coefficients from the
regression of X
coef a
coef_b = as.data.frame(setDT(coef_a, keep.rownames = TRUE))
setnames(coef_b ,"rn","covariate")
```

```
coef = coef_b

vip_load      = full_join(vip     , loadings, by = c("covariate"))
vip_load_coef = full_join(vip_load, coef    , by = c("covariate"))
save(vip_load_coef,file ="C:/Users/PARTSON/vip_load_coef.rda")

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/vip_load_coef.rda")
load(file ="C:/Users/PARTSON/corr_pred_cd4.rda")
results pls = full join(
                   vip_load_coef
                  ,corr_pred_cd4
                  ,by = c("covariate")
                       )
save(results_pls,file ="C:/Users/PARTSON/results_pls.rda")

# Prepare Table 3.6
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/results_pls.rda")
results pls a= results pls
results_pls_a$covariate <- gsub("lymphocytes","Lymphocytes",results_pls_a$covariate)
results_pls_a$covariate <- gsub("basophils","Basophils",results_pls_a$covariate)
results_pls_a$covariate <- gsub("albumin","Albumin",results_pls_a$covariate)
results_pls_a$covariate <- gsub("haematocrit","Haematocrit",results_pls_a$covariate)
results_pls_a$covariate <- gsub("alkaline_phos","ALP",results_pls_a$covariate)
results_pls_a$covariate <- gsub("mcv","MCV",results_pls_a$covariate)
results_pls_a$covariate <- gsub("platelet","Platelet",results_pls_a$covariate)
results_pls_a$covariate <- gsub("potassium","Potassium",results_pls_a$covariate)
results_pls_a$covariate <- gsub("monocytes","Monocytes",results_pls_a$covariate)
results_pls_a$covariate <- gsub("protein","Total protein",results_pls_a$covariate)
results_pls_a$covariate <- gsub("lactate_dehyd","LDH",results_pls_a$covariate)
results_pls_a$covariate <- gsub("folate","Folate",results_pls_a$covariate)
results_pls_a$covariate <- gsub("magnesium","Magnesium",results_pls_a$covariate)
results_pls_a$covariate <- gsub("glucose","Glucose",results_pls_a$covariate)
results_pls_a$covariate <- gsub("calcium","Calcium",results_pls_a$covariate)
results_pls_a$covariate <- gsub("mchc","MCHC",results_pls_a$covariate)
results_pls_a$covariate <- gsub("red_cells","Red blood cells",results_pls_a$covariate)
results_pls_a$covariate <- gsub("sodium","Sodium",results_pls_a$covariate)
results_pls_a$covariate <- gsub("vitb12","Vitamin B12",results_pls_a$covariate)
results_pls_a$covariate <- gsub("triceps_skin","Triceps skin fold",results_pls_a$covariate)
results_pls_a$covariate <- gsub("triglycerides","Triglycerides",results_pls_a$covariate)
results_pls_a$covariate <- gsub("neutrophils","Neutrophils",results_pls_a$covariate)
results_pls_a$covariate <- gsub("ast_got","AST(GOT)",results_pls_a$covariate)
results_pls_a$covariate <- gsub("eosinophils","Eosinophils",results_pls_a$covariate)
results_pls_a$covariate <- gsub("height_m","Height",results_pls_a$covariate)
results_pls_a$covariate <- gsub("chloride","Chloride",results_pls_a$covariate)
results_pls_a$covariate <- gsub("fe","Fe(Iron)",results_pls_a$covariate)
results_pls_a$covariate <- gsub("waist_circum","Waist circum",results_pls_a$covariate)
results_pls_a$covariate <- gsub("ldl","LDL",results_pls_a$covariate)
results_pls_a$covariate <- gsub("bmi","BMI",results_pls_a$covariate)
results_pls_a$covariate <- gsub("bp_systolic","BP(systolic)",results_pls_a$covariate)
results_pls_a$covariate <- gsub("bilirubin","Bilirubin",results_pls_a$covariate)
results_pls_a$covariate <- gsub("arm_right","Arm(right) circum",results_pls_a$covariate)
results_pls_a$covariate <- gsub("glutamyl_trans","GGT",results_pls_a$covariate)
results_pls_a$covariate <- gsub("bp_diastolic","BP(diastolic)",results_pls_a$covariate)
results_pls_a$covariate <- gsub("rdw","RDW",results_pls_a$covariate)
results_pls_a$covariate <- gsub("pulse_bpm","Pulse",results_pls_a$covariate)
results_pls_a$covariate <- gsub("urea","Urea",results_pls_a$covariate)
results_pls_a$covariate <- gsub("alt_gpt","ALT(GPT)",results_pls_a$covariate)
results_pls_a$covariate <- gsub("axillary_temp","Axillary Temp",results_pls_a$covariate)

#=============================================================================#
# Table 3.6: Results of variable selection based on the second component   #
#=============================================================================#
if(!require(dplyr)){install.packages("dplyr")}
results_pls_table = results_pls_a %>% arrange(-VIP) %>%
  mutate_if(is.numeric, function(x) {
    sprintf("%.4f",round(x,4))})
if(!require(data.table)){install.packages("data.table")}
setnames(results_pls_table,c("corr_cd4_r","corr_cd4_p"),c("Correlation","p-value"))
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(results_pls_table,file ="C:/Users/PARTSON/results_pls_table.xlsx",row.names = FALSE
)
```

```
##################################################
# Visualize the results of the selected variables #
##################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/results_pls.rda")
results_pls4vip = results_pls
save(results_pls4vip,file ="C:/Users/PARTSON/results_pls4vip.rda")
######################
# Prepare Figure 3.7 #
######################
# Convert data from wide to long
if(!require(reshape2)){install.packages("reshape2")}
x = colnames(results_pls4vip[,-1])      # Removes the first(-1) column, i.e. "ID"
W2L results pls4vip a = melt(results pls4vip,id.vars = "covariate",
                             # id.vars - additional variables to keep in
                               # the output
        measure.vars = x ,  # Variables to be reshaped
        variable.name="result",
                             # Name of variable used to store measured
                               # variable names
        value.name="value", # Name of variable used to store values
        na.rm = TRUE)        # The rm=Remove the "NA". If set to FALSE the
                               # NA will appear
W2L results pls4vip = W2L results pls4vip_a
W2L results pls4vip$value cutoff =
            ifelse(W2L_results_pls4vip$value< 0 ,"Negative", "Positive")

# Recode and control order in graph
W2L_results_pls4vip$result <- gsub("VIP"          ,"a) VIP"
,W2L results pls4vip$result)
W2L_results_pls4vip$result <- gsub("Loadings"     ,"b) Loadings"
,W2L results pls4vip$result)
W2L_results_pls4vip$result <- gsub("Coefficients","c) Reg coeffs"
,W2L_results_pls4vip$result)
W2L_results_pls4vip$result <- gsub("corr_cd4_r"   ,"d) CD4 corr"
,W2L results pls4vip$result)
W2L_results_pls4vip$result <- gsub("corr_cd4_p"   ,"e) Corr p-
value",W2L_results_pls4vip$result)

save(W2L_results_pls4vip,file ="C:/Users/PARTSON/W2L_results_pls4vip.rda")


load(file ="C:/Users/PARTSON/W2L_results_pls4vip.rda")
if(!require(dplyr)){install.packages("dplyr")}
pls.vars.vip.order <-results_pls4vip%>%arrange(-VIP) # descending order
W2L_vip_pls4vip = subset(W2L_results_pls4vip, result=="a) VIP")
W2L vip pls4vip$value cutoff =
  ifelse(W2L_vip_pls4vip$value<0.8,"Weak", "Strong")
W2L_vip_pls4vip$result <- gsub("a) VIP","Variable Importance in Projection(VIP)"
,W2L_vip_pls4vip$result)

if(!require(ggplot2)){install.packages("ggplot2")}
#=============================================================================#
# Plot Figure 3.7 : The strongest CD4+ count covariates based on VIP only  #
#=============================================================================#
format_plot_vip4vip=(
      theme_bw()   +        # Background fill
      theme(               # Format all the items on the graph
    plot.title = element text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element text(color="black", size=8, face="bold",angle=0),
    axis.text.x = element_text(size=8,angle=90,color="black", hjust=1, vjust=0.5),
    axis.title.y = element_text(color="black", size=8, face="bold"),
    axis.text.y = element_text(size=8,angle=0,color="black",vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=8,face="bold"),
    legend.text = element_text(colour="black",size=7,face="plain")
     ))

ggplot(W2L_vip_pls4vip,
       aes(covariate, value, group = result)) +
       facet_grid(~result, scales="free")+
       geom_col(aes(fill = factor(value_cutoff)),width=.7) +
       scale_fill_manual(values=c("green3","red"),
                         name="Covariate strength:")+
       scale_x_discrete(limits=pls.vars.vip.order$covariate)+
```

```
       labs( x="CD4+ count clinical covariates",
             y='Score'
           )+
    geom_hline(yintercept = 0.8  , color = "orange"      , lty =2)+
      format_plot_vip4vip

#####################################################################
# Visualize VIP against loadings and regression coefficients        #
#####################################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/results_pls.rda")
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(ggrepel)){install.packages("ggrepel")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

# Set up graph format
format_varselect=(
      theme_bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title   = element_text(color="black", size=9  , face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10 , face="bold",angle = 0.0),
    axis.text.x  = element_text(size =9      , angle=0),
    axis.title.y = element_text(color="black", size=9  , face="bold"),
    axis.text.y  = element_text(size =10     ,angle=0) ,
    legend.position="right",
    legend.background = element_rect(fill="white",size=0.2,linetype="solid"
                                     ,colour ="lightblue"),
    legend.title = element_text(colour="black"   ,size=8.0,face="bold"),
    legend.text  = element_text(colour="black"   ,size=8.0,face="plain")
     ))
#============================================================================#
# Figure 3.8: The scatter of VIP against the loadings                        #
#============================================================================#
ggplot(results_pls,
            aes(x=Loadings, y=VIP))+
            labs( x='Loadings',
                  y="Variable Importance in the Projection (VIP)",
              title="") +
       scale_y_continuous(breaks = round(seq(min(results_pls$VIP),
                          max(results_pls$VIP), by = .2),1)) +
  geom_point(colour = "blue", size = 1)+
  geom_hline(yintercept = 1.5  , color = "green"    , lty =2)+
  geom_hline(yintercept = 1.0  , color = "orange"   , lty =2)+
  geom_hline(yintercept = 0.8  , color = "red"      , lty =2)+
  geom_vline(xintercept =-0.025, color = "grey"     , lty =2)+
  geom_vline(xintercept = 0.025, color = "grey"     , lty =2)+
  geom_text_repel(data  =results_pls, aes(label=covariate),
                  size=3.5,box.padding = unit(0.75, "lines"))+
  format_varselect

#============================================================================#
# Figure 3.9: Plot VIP vs regression coefficients                            #
#============================================================================#

ggplot(results_pls,
            aes(x=Coefficients, y=VIP))+
            labs( x='Regression coefficients',
                  y="Variable Importance in the Projection (VIP)",
              title="") +
       scale_y_continuous(breaks = round(seq(min(results_pls$VIP),
                          max(results_pls$VIP), by = .2),1)) +
  geom_point(colour = "blue", size = 1)+
  geom_hline(yintercept = 1.5  , color = "green"    , lty =2)+
  geom_hline(yintercept = 1.0  , color = "orange"   , lty =2)+
  geom_hline(yintercept = 0.8  , color = "red"      , lty =2)+
  geom_vline(xintercept =-0.025, color = "grey"     , lty =2)+
  geom_vline(xintercept = 0.025, color = "grey"     , lty =2)+
  geom_text_repel(data  = results_pls, aes(label=covariate),
                  size  = 3.5,box.padding = unit(0.42, "lines")
                  ,parse=TRUE,nudge_x = 0, nudge_y = 0.
                  ,force = 1,direction = c( "both")
                  ,arrow=arrow(length = unit(0.01, "npc")))+
  format_varselect
```

```
#############################################################
#  Prepare Figure 3.10: COMBINED PLOT OF ALL THE RESULTS  #
#############################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

if(!require(Rcmdr)){install.packages("Rcmdr")} # brings enlarged graphics window
load(file ="C:/Users/PARTSON/results_pls.rda")

# Convert data from wide to long
if(!require(reshape2)){install.packages("reshape2")}
x = colnames(results_pls[,-1])     # Removes the first(-1) column, i.e. "ID"
W2L_results_pls_a = melt(results_pls,id.vars = "covariate",
                           # id.vars - additional variables to keep in
                             # the output
        measure.vars = x ,  # Variables to be reshaped
        variable.name="result",
                           # Name of variable used to store measured
                             # variable names
        value.name="value", # Name of variable used to store values
        na.rm = TRUE)       # The rm=Remove the "NA". If set to FALSE the
                             # NA will appear
W2L_results_pls = W2L_results_pls_a
W2L_results_pls$value_cutoff =
            ifelse(W2L_results_pls$value< 0 ,"Negative", "Positive")

# Recode and control order in graph
W2L_results_pls$result <- gsub("VIP"         ,"a) VIP"                ,W2L_results_pls$result)
W2L_results_pls$result <- gsub("Loadings"    ,"b) Loadings"           ,W2L_results_pls$result)
W2L_results_pls$result <- gsub("Coefficients","c) Reg coeffs"       ,W2L_results_pls$result)
W2L_results_pls$result <- gsub("corr_cd4_r"  ,"d) CD4 corr"    ,W2L_results_pls$result)
W2L_results_pls$result <- gsub("corr_cd4_p"  ,"e) Corr p-value"       ,W2L_results_pls$result)
save(W2L_results_pls,file ="C:/Users/PARTSON/W2L_results_pls.rda")

if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(dplyr)){install.packages("dplyr")}
pls.vars.vip.order <-results_pls%>%arrange(VIP) # descending order
#=========================================================================#
# plot Figure 3.10                                                        #
#=========================================================================#
format_plot_results=(
      theme bw()    +        # Background fill
      theme(                 # Format all the items on the graph
    plot.title = element text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element_text(size=9,angle=90,color="black", hjust=1, vjust=0.5),
    axis.title.y = element_text(color="black", size=10, face="bold"),
    axis.text.y = element text(size=9,angle=0,color="black",vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

ggplot(W2L_results_pls,
       aes(covariate, value, group = result)) +
       facet_grid(~result, scales="free")+
       geom_col(aes(fill = factor(value_cutoff)),width=.7) +
       scale_fill_manual(values=c("red","green3"),
                         name="Associations: (b) to (d)")+
       coord_flip()+
       scale_x_discrete(limits=pls.vars.vip.order$covariate)+
       labs( x="Covariate",
             y='Value'
           #,title="Variable importance per phase "
           )+
      format_plot_results


#############################################################
# Prepare FIGURE 3.11 VARIABLE IMPORTANCE BY CATEGORY       #
#############################################################

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
## Prepare data
load(file ="C:/Users/PARTSON/vars_order.rda")
                           # Borrow this file for its ordered clinical attributes
```

```
vars_dropped = c( "time"         ,"weight_kg"#,"bmi"
                 ,"hb"           ,"mch"       ,"leucocyte"
                 ,"cholesterol","hip circum" )
if(!require(dplyr)){install.packages("dplyr")}
vars_order_pls <-(filter(vars_order,!covariates%in%vars_dropped))

load(file ="C:/Users/PARTSON/W2L_results_pls.rda")
W2L_results_pls_grp = subset(W2L_results_pls, result =="a) VIP")
W2L results pls grp$value cutoff =
  ifelse(W2L_results_pls_grp$value<0.8
,"Weak"          ,
  ifelse(W2L_results_pls_grp$value>0.8 & W2L_results_pls_grp$covariate =="neutrophils."
,"Important/Not preditive",
  ifelse(W2L_results_pls_grp$value>0.8 & W2L_results_pls_grp$covariate =="protein."
,"Important/Not preditive",
  ifelse(W2L_results_pls_grp$value>0.8 & W2L_results_pls_grp$covariate
=="alkaline_phos.","Important/Not preditive",
  ifelse(W2L_results_pls_grp$value>0.8 & W2L_results_pls_grp$covariate =="arm_right."
,"Important/Not preditive",
  ifelse(W2L_results_pls_grp$value>0.8 & W2L_results_pls_grp$covariate =="mchc."
,"Important/Not preditive"
                                                  , "Strong"
                                          ))))))
#-----------------
W2L_results_pls_grp$result <- gsub("a) VIP" ,"Variable Importance in
Projection",W2L results pls grp$result)
#=========================================================================#
# Plot Figure 3.11                                                        #
#=========================================================================#
format plot grp=(
      theme bw()   +         # Background fill
      theme(                 # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element text(size=9,angle=0,color="black", hjust=1, vjust=0.5),
    axis.title.y = element text(color="black", size=10, face="bold"),
    axis.text.y = element text(size=9,angle=0,color="black",vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

ggplot(W2L_results_pls_grp,
       aes(covariate, value, group = result)) +
       facet grid(~result, scales="free")+
       geom_col(aes(fill = factor(value_cutoff)),width=.7) +
       scale_fill_manual(values=c("green3","grey88","grey100"),
                         name="")+
              geom_vline(xintercept=35.5, color ="grey",lwd=0,lty=2)+
              geom vline(xintercept=31.5, color ="red",lwd=0,lty=2)+
              geom vline(xintercept=30.5, color ="grey",lwd=0,lty=2)+
              geom vline(xintercept=27.5, color ="grey",lwd=0,lty=2)+
              geom_vline(xintercept=24.5, color ="grey",lwd=0,lty=2)+
              geom_vline(xintercept=19.5, color ="grey",lwd=0,lty=2)+
              geom_vline(xintercept=14.5, color ="red",lwd=0,lty=2)+
              geom vline(xintercept=13.5, color ="red",lwd=0,lty=2)+
              geom vline(xintercept=11.5, color ="red",lwd=0,lty=2)+
              geom vline(xintercept=06.5, color ="grey",lwd=0,lty=2)+
              geom_vline(xintercept=05.5, color ="grey",lwd=0,lty=2)+
       coord_flip()+
       scale_x_discrete(limits=vars_order_pls$covariates)+
       labs( x="Covariates categorised by clinical groups",
             y='Score value'
           #,title="Variable importance per phase "
           )+
      format_plot_grp


##=======================================================================
# GET SELECTED VARIABLES
##=======================================================================
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/results_pls.rda")
```

```
selected_vars = (results_pls %>% filter(VIP>=0.8&Coefficients>0.025|
                        VIP>=0.8&Coefficients< -0.025))$covariate
selected_vars

# Create data for selected variables
load(file ="C:/Users/PARTSON/bal_reducedata.rda")
if(!require(dplyr)){install.packages("dplyr")}
data2model_a = bal_reducedata[,c(
               "patientid","visitdate","phase","phase_num","time","cd4"
                                )]
data2model_a$timeg = data2model_a$time-1
data2model_a = data2model_a[,c(1:5,7,6)]
data2model_b = round(bal_reducedata[,names(bal_reducedata)%in%selected_vars],2)
data2model_c = data.frame(data2model_a, data2model_b)
data2model_d = data2model_c %>%
                      tbl_df %>%
        group_by(patientid) %>%
        mutate(time =1:n()) %>%
            as.data.frame()

# Time points to start from zero
data2model_e = data2model_d
data2model_e$time = data2model_e$time-1

data2model = data2model_e
save(data2model,file ="C:/Users/PARTSON/data2model.rda")
write.csv(data2model,file ="C:/Users/PARTSON/data2model.csv",na = "",row.names = FALSE)
```

C4: Chapter 4 codes

########################## SAS Sample codes ###################################

```
/*Import data from Excel*/
PROC IMPORT OUT             = WORK.std_mxd
            DATAFILE= "C:/Users/PARTSON/std_mxd.csv"
            DBMS       = CSV REPLACE;
               GETNAMES= YES;
               DATAROW = 2;
RUN;
/*PROC STANDARD DATA=std_mxd1 MEAN=0 STD=1 OUT=std_mxd; /*standardise variable 'timeg'*/*/
/* VAR timeg ;*/
/*RUN;

 Libname mysas 'C:/Users/PARTSON/Modelling'; /*creating a libray call 'mysas' to and path to
permanent folder */

data mysas.std_mxd; /*take dataset in 'mysas'*/
    Set std_mxd;  /*and save it to permanent folder set with path linked to 'mysas'*/
 Run;
/*Set all the predictor variables to global xvars for convenience*/
/*All the variables are centred*/

%let xvars1 =  timeg
               red_cells       haematocrit     mcv             mchc
               platelet        lymphocytes     monocytes       basophils
               glucose         alkaline_phos   calcium         magnesium
               potassium       sodium          protein         albumin
               lactate_dehyd   folate ;

%let xvars2 =   timeg|phase      red_cells|phase       haematocrit|phase     mcv|phase
               mchc|phase        platelet|phase        lymphocytes|phase     monocytes|phase
               basophils|phase   glucose|phase         alkaline_phos|phase   calcium|phase
               magnesium|phase   potassium|phase       sodium|phase          protein|phase
               albumin|phase     lactate_dehyd|phase   folate|phase ;

PROC IMPORT OUT             = WORK.std_hpmxd
            DATAFILE= "C:/Users/PARTSON/std_hpmxd.csv"
            DBMS       = CSV REPLACE;
               GETNAMES= YES;
               DATAROW = 2;
RUN;
```

```
####################################
# Prepare Table 4.3 and Table 4.4   #
####################################
ods graphics on;
proc hpmixed   data=std_hpmxd  method=reml ;
                     class  patientid  phase time ;     /*categorical variables*/
                     model  cd4 = phase  &xvars2 / s ;            /*define the response and
fixed part*/
                              random int &xvars1 /s type =un sub = patientid  group =phase;
/*int and slopes per phase i.e group*/
                          TEST phase;
                          LSMEANS phase/DIFF;
                          Title "HPMIXED-UN";
                          ods output  ParameterEstimates =fixef_hpmx CovParms=covs_hpmx
                            SolutionR = ranef_hpmx LSMeans = lsmeans_hpmx
                            Diffs = diffs_hpmx;
run;
ods graphics off;
/***Save csv files***/
 %ds2csv (
   data=fixef_hpmx,
   runmode=b,
   csvfile= C:\Users\PARTSON\fixef_hpmx.csv    (Table 4.3)
 );
  %ds2csv (
   data=covs_hpmx,
   runmode=b,
   csvfile= C:\Users\PARTSON\covs_hpmx.csv
 );

   %ds2csv (
   data=lsmeans_hpmx,
   runmode=b,
   csvfile= C:\Users\PARTSON\lsmeans_hpmx.csv  (Table 4.4)
 );

   %ds2csv (
   data=diffs_hpmx,
   runmode=b,
   csvfile= C:\Users\PARTSON\diffs_hpmx.csv     (Table 4.4)
 );

#####################
# Prepare Table 4.5 #
#####################
/****timeg*****/
ods graphics on;
proc mixed     data=std_mxd  method=reml  COVTEST plots=studentpanel(marginal conditional);
                     class  patientid  phase(ref="5-ART") time ;     /*categorical
variables*/
                     model  cd4 = phase &xvars2 / s ;             /*define the response and
fixed part*/
                              random int timeg /s type =un sub = patientid Gcorr group =phase;
/*int and slopes per phase i.e group*/
                Repeated time/ type = ARMA(1,1) sub = patientid  ; /*Account for corr between
repeated measurements*/
                     Title "UN:timeg ranef";
                     ods output  SolutionF =fixef_timeg CovParms=covs_timeg GCorr =
gcorr_timeg
                       FitStatistics = fitstats_timeg;
run;
ods graphics off;
/***Save csv files***/
 %ds2csv (
   data=fixef_timeg,
   runmode=b,
   csvfile= C:\Users\PARTSON\fixef_timeg.csv
 );
 %ds2csv (
   data=covs_timeg,
   runmode=b,
   csvfile= C:\Users\PARTSON\covs_timeg.csv
    );
```

```
 %ds2csv (
   data=gcorr_timeg,
   runmode=b,
   csvfile= C:\Users\PARTSON\gcorr_timeg.csv
 );

  %ds2csv (
   data=fitstats_timeg,
   runmode=b,
   csvfile= C:\Users\PARTSON\fitstats_timeg.csv
 );
#============================================================================#
Figure 4.1: From SAS output                                                  #
#============================================================================#

#-------------------------------------------------------------
# REPEAT ALL THE TABLE 4.5 STEPS ABOVE FOR THE OTHER COVARIATES
#-------------------------------------------------------------

############################ R ###################################


#############################################################################################
#                                     #
# MIXED MODELS                        #
#                                     #
#######################################
#-----------------------------------------
# Load fit statistics from SAS out put
#   * Merge
#   * Prepare Table 4.2
#-----------------------------------------
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(readr)){install.packages("readr")}
if(!require(dplyr)){install.packages("dplyr")}
fitstats_fixed <- read_csv("C:/Users/Partson/fitstats_fixed.csv")
fitstats_timeg <- read_csv("C:/Users/Partson/fitstats_timeg.csv")
fitstats_red_cells <- read_csv("C:/Users/Partson/fitstats_red_cells.csv")
fitstats_haematocrit <- read_csv("C:/Users/Partson/fitstats_haematocrit.csv")
fitstats_mcv <- read_csv("C:/Users/Partson/fitstats_mcv.csv")
fitstats_mchc <- read_csv("C:/Users/Partson/fitstats_mchc.csv")
fitstats_platelet <- read_csv("C:/Users/Partson/fitstats_platelet.csv")
fitstats_lymphocytes <- read_csv("C:/Users/Partson/fitstats_lymphocytes.csv")
fitstats_monocytes <- read_csv("C:/Users/Partson/fitstats_monocytes.csv")
fitstats_basophils <- read_csv("C:/Users/Partson/fitstats_basophils.csv")
fitstats_glucose <- read_csv("C:/Users/Partson/fitstats_glucose.csv")
fitstats_alkaline_phos <- read_csv("C:/Users/Partson/fitstats_alkaline_phos.csv")
fitstats_calcium <- read_csv("C:/Users/Partson/fitstats_calcium.csv")
fitstats_magnesium <- read_csv("C:/Users/Partson/fitstats_magnesium.csv")
fitstats_potassium <- read_csv("C:/Users/Partson/fitstats_potassium.csv")
fitstats_sodium <- read_csv("C:/Users/Partson/fitstats_sodium.csv")
fitstats_protein <- read_csv("C:/Users/Partson/fitstats_protein.csv")
fitstats_albumin <- read_csv("C:/Users/Partson/fitstats_albumin.csv")
fitstats_lactate_dehyd <- read_csv("C:/Users/Partson/fitstats_lactate_dehyd.csv")
fitstats_folate <- read_csv("C:/Users/Partson/fitstats_folate.csv")
   ## Add covariate column
fitstats_fixed$covariate = rep("fixed")
fitstats_timeg$covariate = rep("timeg")
fitstats_red_cells$covariate = rep("red_cells")
fitstats_haematocrit$covariate = rep("haematocrit")
fitstats_mcv$covariate = rep("mcv")
fitstats_mchc$covariate = rep("mchc")
fitstats_platelet$covariate = rep("platelet")
fitstats_lymphocytes$covariate = rep("lymphocytes")
fitstats_monocytes$covariate = rep("monocytes")
fitstats_basophils$covariate = rep("basophils")
fitstats_glucose$covariate = rep("glucose")
fitstats_alkaline_phos$covariate = rep("alkaline_phos")
fitstats_calcium$covariate = rep("calcium")
fitstats_magnesium$covariate = rep("magnesium")
fitstats_potassium$covariate = rep("potassium")
fitstats_sodium$covariate = rep("sodium")
fitstats_protein$covariate = rep("protein")
fitstats_albumin$covariate = rep("albumin")
fitstats_lactate_dehyd$covariate = rep("lactate_dehyd")
fitstats_folate$covariate = rep("folate")
```

```
fitstats_a = as.data.frame(rbind(
              fitstats fixed
             ,fitstats timeg
             ,fitstats red cells
             ,fitstats_haematocrit
             ,fitstats_mcv
             ,fitstats_mchc
             ,fitstats platelet
             ,fitstats lymphocytes
             ,fitstats monocytes
             ,fitstats_basophils
             ,fitstats_glucose
             ,fitstats_alkaline_phos
             ,fitstats calcium
             ,fitstats magnesium
             ,fitstats_potassium
             ,fitstats_sodium
             ,fitstats_protein
             ,fitstats_albumin
             ,fitstats lactate_dehyd
             ,fitstats_folate
             ))
save(fitstats_a,file ="C:/Users/PARTSON/fitstats_a.rda")

# FitStatitics Long to wide
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
options(prompt = "R>") # customize prompt

load(file ="C:/Users/PARTSON/fitstats_a.rda")
long = fitstats a
if(!require(reshape2)){install.packages("reshape2")}
if(!require(dplyr)){install.packages("dplyr")}
#============================================================================#
# Table 4.2: The fit statistics for the covariate effect models             #
#============================================================================#
L2W_a = dcast(long,covariate~Description,value.var = "Value") # long to wide

if(!require(data.table)){install.packages("data.table")}
setnames(L2W_a,names(L2W_a),c("covariate","-2 Res Log Likelihood","AIC","AICC","BIC"))
L2W_b =L2W_a %>% arrange(AIC)

L2W_b$info_loss = exp((min(L2W_b$AIC)-L2W_b$AIC)/2)
L2W c = L2W b
L2W_c$covariate <- gsub("lymphocytes","Lymphocytes",L2W_c$covariate)
L2W_c$covariate <- gsub("basophils","Basophils",L2W_c$covariate)
L2W_c$covariate <- gsub("albumin","Albumin",L2W_c$covariate)
L2W_c$covariate <- gsub("haematocrit","Haematocrit",L2W_c$covariate)
L2W_c$covariate <- gsub("alkaline_phos","ALP",L2W_c$covariate)
L2W_c$covariate <- gsub("mcv","MCV",L2W_c$covariate)
L2W_c$covariate <- gsub("platelet","Platelet",L2W_c$covariate)
L2W_c$covariate <- gsub("potassium","Potassium",L2W_c$covariate)
L2W_c$covariate <- gsub("monocytes","Monocytes",L2W_c$covariate)
L2W_c$covariate <- gsub("protein","Total protein",L2W_c$covariate)
L2W_c$covariate <- gsub("lactate_dehyd","LDH",L2W_c$covariate)
L2W_c$covariate <- gsub("folate","Folate",L2W_c$covariate)
L2W_c$covariate <- gsub("magnesium","Magnesium",L2W_c$covariate)
L2W_c$covariate <- gsub("glucose","Glucose",L2W_c$covariate)
L2W_c$covariate <- gsub("calcium","Calcium",L2W_c$covariate)
L2W_c$covariate <- gsub("mchc","MCHC",L2W_c$covariate)
L2W_c$covariate <- gsub("red_cells","Red blood cells",L2W_c$covariate)
L2W_c$covariate <- gsub("sodium","Sodium",L2W_c$covariate)
L2W_c$covariate <- gsub("fixed","Fixed",L2W_c$covariate)
L2W_c$covariate <- gsub("timeg","Time",L2W_c$covariate)

setnames(L2W_c,c("covariate","info_loss"),c("Covariate","Info loss"))
fitstats_results = L2W_c

save(fitstats_results,file ="C:/Users/PARTSON/fitstats_results.rda")
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(fitstats_results,file ="C:/Users/PARTSON/fitstats_results.xlsx",row.names = FALSE )
       ####--------->>Then go to the Excel spreadsheet
#----------------------------------------
# Load fixed only from SAS output
#   * Merge
#   * Prepare Table 4.3 (see SAS code)
#----------------------------------------
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
```

```
if(!require(dplyr)){install.packages("dplyr")}
if(!require(readr)){install.packages("readr")}

#=============================================================================#
# Table 4.3 : Fixed effects - The cohort's phase general trajectories       #
#=============================================================================#
fixefhpmx_a <- read_csv("C:/Users/Partson/fixef_hpmx.csv")%>%
               as.data.frame()
fixefhpmx_a$Effect <- gsub('\\*phase','',fixefhpmx_a$Effect) # the \\ forces the * to go
fixefhpmx_a$Effect <- gsub('timeg',"time",fixefhpmx_a$Effect)
fixefhpmx_a$phase = ifelse(fixefhpmx_a$Effect=="Intercept"&
is.na(fixefhpmx_a$phase),"1ART",fixefhpmx_a$phase)
fixefhpmx_a$Effect <- gsub("Intercept","phase",fixefhpmx_a$Effect)
fixefhpmx_a$phase <- gsub("1ART"  ,"5-ART"   ,fixefhpmx_a$phase)
fixefhpmx_a$phase <- gsub("2Est"  ,"4-Est"   ,fixefhpmx_a$phase)
fixefhpmx_a$phase <- gsub("3Early","3-Early" ,fixefhpmx_a$phase)
fixefhpmx_a$phase <- gsub("4Acute","2-Acute" ,fixefhpmx_a$phase)

fixefhpmx = fixefhpmx_a[-2,]
save(fixefhpmx,file="C:/Users/Partson/fixefhpmx.rda")


# long to wide
load(file="C:/Users/Partson/fixefhpmx.rda")
fixefhpmx b = fixefhpmx
if(!require(dplyr)){install.packages("dplyr")}
fixefhpmx_b$Estimate = sprintf("%.4f",round(fixefhpmx_b$Estimate,4)) # round off to 4 with
trailing zeros

fixefhpmx b$"Pr > |t|" = paste0("(",fixefhpmx b$"Pr > |t|",")") # bracket the values
fixefhpmx b$Estimate = as.character(fixefhpmx b$Estimate)
fixefhpmx_b$est_pvalue =paste(fixefhpmx_b$Estimate,fixefhpmx_b$"Pr > |t|", sep="") # combine
columns
long = fixefhpmx_b[,c(1:2,8)] %>% na.omit()

if(!require(reshape2)){install.packages("reshape2")}
L2W a = dcast(long,Effect~phase,value.var = "est_pvalue") # long to wide
L2W_b = L2W_a[c(14,20,1:13,15:19),]

if(!require(data.table)){install.packages("data.table")}
#setnames(L2W b,"5-ART","5-ART(ref)")
fixed_final_results = L2W_b

save(fixed_final_results,file ="C:/Users/PARTSON/fixed_final_results.rda")
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(fixed_final_results,file ="C:/Users/PARTSON/fixed_final_results.xlsx"
,row.names = FALSE
          )
load(file ="C:/Users/PARTSON/fixed_final_results.rda")
       ####--------->>Then go to the Excel spreadsheet

##########################################
# Load LMEANS & DIFFS from SAS out put     #
#   * Results wrangling                    #
#   * Prepare Table 4.4 (see SAS code)     #
##########################################

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(dplyr)){install.packages("dplyr")}
if(!require(readr)){install.packages("readr")}
#=============================================================================#
# Table 4.4: Least squares means and differences                            #
#=============================================================================#
#lsmeans and diffs
lsmeans_hpmx<- read_csv("C:/Users/Partson/lsmeans_hpmx.csv")%>%
               as.data.frame()
lsmeans_hpmx$phase <- gsub("1ART"  ,"5-ART"   ,lsmeans_hpmx$phase)
lsmeans_hpmx$phase <- gsub("2Est"  ,"4-Est"   ,lsmeans_hpmx$phase)
lsmeans_hpmx$phase <- gsub("3Early","3-Early" ,lsmeans_hpmx$phase)
lsmeans_hpmx$phase <- gsub("4Acute","2-Acute" ,lsmeans_hpmx$phase)

diffs_hpmx<- read_csv("C:/Users/Partson/diffs_hpmx.csv")%>%
             as.data.frame()
diffs_hpmx$phase <- gsub("1ART"  ,"5-ART"   ,diffs_hpmx$phase)
diffs_hpmx$phase <- gsub("2Est"  ,"4-Est"   ,diffs_hpmx$phase)
diffs_hpmx$phase <- gsub("3Early","3-Early" ,diffs_hpmx$phase)
diffs_hpmx$phase <- gsub("4Acute","2-Acute" ,diffs_hpmx$phase)
```

```
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(diffs_hpmx,file ="C:/Users/PARTSON/diffs_hpmx.xlsx",row.names = FALSE  )
write.xlsx(lsmeans_hpmx,file ="C:/Users/PARTSON/lsmeans_hpmx.xlsx",row.names = FALSE  )
       ####--------->>Then go to the Excel spreadsheet


#########################################################################
#  PREPARING SAMPLE RESULTS FOR COVARIATIONS                            #
#########################################################################
# Load covariance parameters from SAS out put                           #
#   * Merge                                                             #
#   * Prepare Table 4.5 and Table 4.6: covariate data                   #
#########################################################################

covs_timeg <- read_csv("C:/Users/Partson/covs_timeg.csv")
covs_red_cells <- read_csv("C:/Users/Partson/covs_red_cells.csv")
covs_haematocrit <- read_csv("C:/Users/Partson/covs_haematocrit.csv")
covs_mcv <- read_csv("C:/Users/Partson/covs_mcv.csv")
covs mchc <- read csv("C:/Users/Partson/covs mchc.csv")
covs platelet <- read csv("C:/Users/Partson/covs platelet.csv")
covs_lymphocytes <- read_csv("C:/Users/Partson/covs_lymphocytes.csv")
covs_monocytes <- read_csv("C:/Users/Partson/covs_monocytes.csv")
covs basophils <- read csv("C:/Users/Partson/covs basophils.csv")
covs glucose <- read csv("C:/Users/Partson/covs glucose.csv")
covs alkaline phos <- read csv("C:/Users/Partson/covs alkaline_phos.csv")
covs_calcium <- read_csv("C:/Users/Partson/covs_calcium.csv")
covs_magnesium <- read_csv("C:/Users/Partson/covs_magnesium.csv")
covs potassium <- read csv("C:/Users/Partson/covs potassium.csv")
covs sodium <- read csv("C:/Users/Partson/covs sodium.csv")
covs protein <- read csv("C:/Users/Partson/covs protein.csv")
covs albumin <- read csv("C:/Users/Partson/covs albumin.csv")
covs_lactate_dehyd <- read_csv("C:/Users/Partson/covs_lactate_dehyd.csv")
covs_folate <- read_csv("C:/Users/Partson/covs_folate.csv")

# Add covariate column
covs_timeg$covariate = rep("timeg")
covs_red_cells$covariate = rep("red_cells")
covs_haematocrit$covariate = rep("haematocrit")
covs_mcv$covariate = rep("mcv")
covs_mchc$covariate = rep("mchc")
covs_platelet$covariate = rep("platelet")
covs_lymphocytes$covariate = rep("lymphocytes")
covs_monocytes$covariate = rep("monocytes")
covs_basophils$covariate = rep("basophils")
covs_glucose$covariate = rep("glucose")
covs_alkaline_phos$covariate = rep("alkaline_phos")
covs_calcium$covariate = rep("calcium")
covs_magnesium$covariate = rep("magnesium")
covs_potassium$covariate = rep("potassium")
covs_sodium$covariate = rep("sodium")
covs_protein$covariate = rep("protein")
covs_albumin$covariate = rep("albumin")
covs_lactate_dehyd$covariate = rep("lactate_dehyd")
covs_folate$covariate = rep("folate")

# Add "estimate" as % of total proportion
covs_timeg$prop_est = covs_timeg$Estimate/sum(covs_timeg[,"Estimate"])*100
covs_red_cells$prop_est = covs_red_cells$Estimate/sum(covs_red_cells[,"Estimate"])*100
covs_haematocrit$prop_est = covs_haematocrit$Estimate/sum(covs_haematocrit[,"Estimate"])*100
covs_mcv$prop_est = covs_mcv$Estimate/sum(covs_mcv[,"Estimate"])*100
covs_mchc$prop_est = covs_mchc$Estimate/sum(covs_mchc[,"Estimate"])*100
covs_platelet$prop_est = covs_platelet$Estimate/sum(covs_platelet[,"Estimate"])*100
covs_lymphocytes$prop_est = covs_lymphocytes$Estimate/sum(covs_lymphocytes[,"Estimate"])*100
covs_monocytes$prop_est = covs_monocytes$Estimate/sum(covs_monocytes[,"Estimate"])*100
covs_basophils$prop_est = covs_basophils$Estimate/sum(covs_basophils[,"Estimate"])*100
covs_glucose$prop_est = covs_glucose$Estimate/sum(covs_glucose[,"Estimate"])*100
covs_alkaline_phos$prop_est =
covs_alkaline_phos$Estimate/sum(covs_alkaline_phos[,"Estimate"])*100
covs_calcium$prop_est = covs_calcium$Estimate/sum(covs_calcium[,"Estimate"])*100
covs_magnesium$prop_est = covs_magnesium$Estimate/sum(covs_magnesium[,"Estimate"])*100
covs_potassium$prop_est = covs_potassium$Estimate/sum(covs_potassium[,"Estimate"])*100
covs_sodium$prop_est = covs_sodium$Estimate/sum(covs_sodium[,"Estimate"])*100
covs_protein$prop_est = covs_protein$Estimate/sum(covs_protein[,"Estimate"])*100
covs_albumin$prop_est = covs_albumin$Estimate/sum(covs_albumin[,"Estimate"])*100
covs lactate dehyd$prop est =
covs_lactate_dehyd$Estimate/sum(covs_lactate_dehyd[,"Estimate"])*100
```

```
covs_folate$prop_est = covs_folate$Estimate/sum(covs_folate[,"Estimate"])*100

covs_a = as.data.frame(rbind(
              covs timeg
             ,covs_red_cells
             ,covs_haematocrit
             ,covs_mcv
             ,covs mchc
             ,covs platelet
             ,covs lymphocytes
             ,covs_monocytes
             ,covs_basophils
             ,covs_glucose
             ,covs alkaline_phos
             ,covs calcium
             ,covs_magnesium
             ,covs_potassium
             ,covs_sodium
             ,covs_protein
             ,covs albumin
             ,covs lactate_dehyd
             ,covs_folate
             ))
save(covs_a,file ="C:/Users/PARTSON/covs_a.rda")

#----------------
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/covs_a.rda")

if(!require(data.table)){install.packages("data.table")}
if(!require(dplyr)){install.packages("dplyr")}
setnames(covs_a,c("Cov Parm","Group","Pr Z"),c("covparm","phase","pvalue"))
covs_a$covariate <- gsub("timeg","time",covs_a$covariate)
covs_a$pvalue <- gsub("<","0",covs_a$pvalue)
covs_a$pvalue = as.numeric(covs_a$pvalue)

covs a$sig =
          ifelse(covs_a$pvalue <0.05 ,"Significant","Insignificant")
covs_a$sig[is.na(covs_a$sig)] = "Insignificant"
covs = covs_a
save(covs,file ="C:/Users/PARTSON/covs.rda")
write.csv(covs,file ="C:/Users/PARTSON/covs.csv",na = "",row.names = FALSE)

#============================================================================#
# Table 4.5                                                                  #
#============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/covs.rda")
covs_time_a = subset(covs, covariate=="time")
covs_time_a$prop_est = sprintf("%.4f",round(covs_time_a$prop_est,4)) # round off to 4 with
trailing zeros
covs_time_b = covs_time_a[,-c(8,10)]
covs_time_c = covs_time_b[,c(2,3,1,4,8,5:7)]
covs_time_c$phase <- gsub("phase ","",covs_time_c$phase) # the \\ forces the * to go
covs_time_c$Subject <- gsub("patientid","patient",covs_time_c$Subject) # the \\ forces the *
to go
covs_time_c$covparm <- gsub("UN\\(1,1\\)","Intr",covs_time_c$covparm) # the \\ forces the * to
go
covs_time_c$covparm <- gsub("UN\\(2,1\\)","Intr-Slope",covs_time_c$covparm) # the \\ forces
the * to go
covs_time_c$covparm <- gsub("UN\\(2,2\\)","Slope",covs_time_c$covparm) # the \\ forces the *
to go

if(!require(data.table)){install.packages("data.table")}
setnames(covs_time_c,c("prop_est"), c("Estimate(%)"))
covs_time_results = covs_time_c

if(!require(xlsx)){install.packages("xlsx")}
# Table 4.5: Sample covariance parameter tests of time effect and the proportions
write.xlsx(covs_time_results,file ="C:/Users/PARTSON/covs_time_results.xlsx"
          ,row.names = FALSE, showNA=FALSE
          )
       ####--------->>The go to the Excel spreadsheet
```

```
#=============================================================================#
# Table 4.6                                                                   #
#=============================================================================#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/covs.rda")
covs_red_cells_a = subset(covs, covariate=="red_cells")
covs_red_cells_a$prop_est = sprintf("%.4f",round(covs_red_cells_a$prop_est,4)) # round off to
4 with trailing zeros
covs_red_cells_b = covs_red_cells_a[,-c(8,10)]
covs_red_cells_c = covs_red_cells_b[,c(2,3,1,4,8,5:7)]
covs_red_cells_c$phase <- gsub("phase ","",covs_red_cells_c$phase) # the \\ forces the * to go
covs_red_cells_c$Subject <- gsub("patientid","patient",covs_red_cells_c$Subject) # the \\
forces the * to go
covs_red_cells_c$covparm <- gsub("UN\\(1,1\\)","Intr",covs_red_cells_c$covparm) # the \\
forces the * to go
covs_red_cells_c$covparm <- gsub("UN\\(2,1\\)","Intr-Slope",covs_red_cells_c$covparm) # the \\
forces the * to go
covs_red_cells_c$covparm <- gsub("UN\\(2,2\\)","Slope",covs_red_cells_c$covparm) # the \\
forces the * to go

if(!require(data.table)){install.packages("data.table")}
setnames(covs_red_cells_c,c("prop_est"), c("Estimate(%)"))
covs_red_cells_results = covs_red_cells_c
if(!require(xlsx)){install.packages("xlsx")}
# Table 4.6: Sample covariance parameter tests of red blood cells effect and the proportions
write.xlsx(covs_red_cells_results,file ="C:/Users/PARTSON/covs_red_cells_results.xlsx"
          ,row.names = FALSE, showNA=FALSE
          )
       ####--------->>The go to the Excel spreadsheet


##########################
# Prepare Figure 4.2     #
##########################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(scales)){install.packages("scales")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

format_plot=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element_text(size=9,angle=90,color="black", hjust=1, vjust=0.25),
    axis.title.y = element_text(color="black", size=10, face="bold"),
    axis.text.y = element text(size=9,angle=0,color="black", hjust=1,vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))
save(format_plot,file ="C:/Users/PARTSON/format_plot.rda")

load(file ="C:/Users/PARTSON/covs.rda")

# UN(1,1)
intr =ggplot(subset(covs,covparm=="UN(1,1)"& !is.na(phase)),
  aes(x=covariate, y=prop est))+
  geom_col(aes(fill = factor(sig)),width=.6)+
          scale_fill_manual(values=c("grey70","blue"),name="Covariance parameter test:")+
   facet_wrap(~phase)+
         labs( x="",
               y="Intercept variation(%)"
             )+ format_plot
intr # plot it

# UN(2,2)
slope =ggplot(subset(covs,covparm=="UN(2,2)"& !is.na(phase)),
  aes(x=covariate, y=prop_est))+
  geom_col(aes(fill = factor(sig)),width=.6)+
          scale_fill_manual(values=c("grey70","blue"),name="Covariance parameter test:")+
  facet_wrap(~phase)+
         labs( x="Covariate with random effect",
               y="Slope variation(%)"
             )+ format_plot
```

```
slope # plot it


# Combine multiple plots
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(ggpubr)){install.packages("ggpubr")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

#=========================================================================#
# Figure 4.2: Proportion of variation in intercepts and slopes.           #
#=========================================================================#

theme_set(theme_pubr())
ggarrange(
  intr, slope, ncol = 1,nrow =2,#labels = c("Intercept", "Slope"),
  common.legend = TRUE, legend = "top"
  )


########################################
# Load G correlations from SAS out put    #
#   and merge                          #
########################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(readr)){install.packages("readr")}
gcorr timeg <- read csv("C:/Users/Partson/gcorr timeg.csv")
gcorr_red_cells <- read_csv("C:/Users/Partson/gcorr_red_cells.csv")
gcorr_haematocrit <- read_csv("C:/Users/Partson/gcorr_haematocrit.csv")
gcorr mcv <- read csv("C:/Users/Partson/gcorr mcv.csv")
gcorr mchc <- read csv("C:/Users/Partson/gcorr mchc.csv")
gcorr platelet <- read csv("C:/Users/Partson/gcorr platelet.csv")
gcorr lymphocytes <- read csv("C:/Users/Partson/gcorr lymphocytes.csv")
gcorr_monocytes <- read_csv("C:/Users/Partson/gcorr_monocytes.csv")
gcorr_basophils <- read_csv("C:/Users/Partson/gcorr_basophils.csv")
gcorr glucose <- read csv("C:/Users/Partson/gcorr glucose.csv")
gcorr alkaline phos <- read csv("C:/Users/Partson/gcorr alkaline_phos.csv")
gcorr calcium <- read csv("C:/Users/Partson/gcorr calcium.csv")
gcorr magnesium <- read csv("C:/Users/Partson/gcorr magnesium.csv")
gcorr_potassium <- read_csv("C:/Users/Partson/gcorr_potassium.csv")
gcorr_sodium <- read_csv("C:/Users/Partson/gcorr_sodium.csv")
gcorr protein <- read csv("C:/Users/Partson/gcorr protein.csv")
gcorr albumin <- read csv("C:/Users/Partson/gcorr albumin.csv")
gcorr lactate dehyd <- read csv("C:/Users/Partson/gcorr lactate_dehyd.csv")
gcorr_folate <- read_csv("C:/Users/Partson/gcorr_folate.csv")


# Extract and prepare Intercept and Slope correlations
if(!require(dplyr)){install.packages("dplyr")}
if(!require(data.table)){install.packages("data.table")}

#---
gcorr_timeg_a = gcorr_timeg[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is corr=rowSums(select(.,5:8)))
                # NOTE: restart computer if there are any issues
gcorr_timeg_cor = data.frame(gcorr_timeg[-c(1,3,5,7),-c(3,5:12)],is_cor=gcorr_timeg_a[,2])
setnames(gcorr_timeg_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_timeg_cor$covparm= rep("UN(2,1)")
#---
gcorr_red_cells_a = gcorr_red_cells[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_red_cells_cor = data.frame(gcorr_red_cells[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_red_cells_a[,2])
setnames(gcorr_red_cells_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_red_cells_cor$covparm= rep("UN(2,1)")
#---
gcorr_haematocrit_a = gcorr_haematocrit[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_haematocrit_cor = data.frame(gcorr_haematocrit[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_haematocrit_a[,2])
setnames(gcorr_haematocrit_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_haematocrit_cor$covparm= rep("UN(2,1)")
#---
gcorr_mcv_a = gcorr_mcv[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_mcv_cor = data.frame(gcorr_mcv[-c(1,3,5,7),-c(3,5:12)],is_cor=gcorr_mcv_a[,2])
setnames(gcorr_mcv_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_mcv_cor$covparm= rep("UN(2,1)")
```

```
#---
gcorr_mchc_a = gcorr_mchc[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_mchc_cor = data.frame(gcorr_mchc[-c(1,3,5,7),-c(3,5:12)],is_cor=gcorr_mchc_a[,2])
setnames(gcorr_mchc_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_mchc_cor$covparm= rep("UN(2,1)")
#---
gcorr_platelet_a = gcorr_platelet[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_platelet_cor = data.frame(gcorr_platelet[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_platelet_a[,2])
setnames(gcorr_platelet_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_platelet_cor$covparm= rep("UN(2,1)")
#---
gcorr_lymphocytes_a = gcorr_lymphocytes[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_lymphocytes_cor = data.frame(gcorr_lymphocytes[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_lymphocytes_a[,2])
setnames(gcorr_lymphocytes_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_lymphocytes_cor$covparm= rep("UN(2,1)")
#---
gcorr_monocytes_a = gcorr_monocytes[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_monocytes_cor = data.frame(gcorr_monocytes[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_monocytes_a[,2])
setnames(gcorr_monocytes_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_monocytes_cor$covparm= rep("UN(2,1)")
#---
gcorr_basophils_a = gcorr_basophils[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_basophils_cor = data.frame(gcorr_basophils[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_basophils_a[,2])
setnames(gcorr_basophils_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_basophils_cor$covparm= rep("UN(2,1)")
#---
gcorr_glucose_a = gcorr_glucose[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_glucose_cor = data.frame(gcorr_glucose[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_glucose_a[,2])
setnames(gcorr_glucose_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_glucose_cor$covparm= rep("UN(2,1)")
#---
gcorr_alkaline_phos_a = gcorr_alkaline_phos[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_alkaline_phos_cor = data.frame(gcorr_alkaline_phos[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_alkaline_phos_a[,2])
setnames(gcorr_alkaline_phos_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_alkaline_phos_cor$covparm= rep("UN(2,1)")
#---
gcorr_calcium_a = gcorr_calcium[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_calcium_cor = data.frame(gcorr_calcium[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_calcium_a[,2])
setnames(gcorr_calcium_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_calcium_cor$covparm= rep("UN(2,1)")
#---
gcorr_magnesium_a = gcorr_magnesium[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_magnesium_cor = data.frame(gcorr_magnesium[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_magnesium_a[,2])
setnames(gcorr_magnesium_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_magnesium_cor$covparm= rep("UN(2,1)")
#---
gcorr_potassium_a = gcorr_potassium[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_potassium_cor = data.frame(gcorr_potassium[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_potassium_a[,2])
setnames(gcorr_potassium_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_potassium_cor$covparm= rep("UN(2,1)")
#---
gcorr_sodium_a = gcorr_sodium[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
               transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_sodium_cor = data.frame(gcorr_sodium[-c(1,3,5,7),-c(3,5:12)],is_cor=gcorr_sodium_a[,2])
setnames(gcorr_sodium_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_sodium_cor$covparm= rep("UN(2,1)")
#---
gcorr_protein_a = gcorr_protein[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
```

```
              transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_protein_cor = data.frame(gcorr_protein[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_protein_a[,2])
setnames(gcorr_protein_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_protein_cor$covparm= rep("UN(2,1)")
#---
gcorr_albumin_a = gcorr_albumin[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_albumin_cor = data.frame(gcorr_albumin[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_albumin_a[,2])
setnames(gcorr_albumin_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_albumin_cor$covparm= rep("UN(2,1)")
#---
gcorr_lactate_dehyd_a = gcorr_lactate_dehyd[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is corr=rowSums(select(.,5:8)))
gcorr_lactate_dehyd_cor = data.frame(gcorr_lactate_dehyd[-c(1,3,5,7),-
c(3,5:12)],is_cor=gcorr_lactate_dehyd_a[,2])
setnames(gcorr_lactate_dehyd_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_lactate_dehyd_cor$covparm= rep("UN(2,1)")
#---
gcorr_folate_a = gcorr_folate[-c(1,3,5,7),-c(6,8,10,12)]%>% as.data.frame()%>%
              transmute(phase,is_corr=rowSums(select(.,5:8)))
gcorr_folate_cor = data.frame(gcorr_folate[-c(1,3,5,7),-c(3,5:12)],is_cor=gcorr_folate_a[,2])
setnames(gcorr_folate_cor,c("Row","Effect"),c("covparm","covariate"))
gcorr_folate_cor$covparm= rep("UN(2,1)")
#---

# Merge the data sets
gcorrs_a = as.data.frame(rbind(
              gcorr timeg cor
             ,gcorr red cells cor
             ,gcorr haematocrit_cor
             ,gcorr_mcv_cor
             ,gcorr_mchc_cor
             ,gcorr platelet cor
             ,gcorr lymphocytes cor
             ,gcorr monocytes cor
             ,gcorr basophils cor
             ,gcorr_glucose_cor
             ,gcorr_alkaline_phos_cor
             ,gcorr calcium cor
             ,gcorr magnesium cor
             ,gcorr potassium cor
             ,gcorr_sodium_cor
             ,gcorr_protein_cor
             ,gcorr_albumin_cor
             ,gcorr lactate dehyd_cor
             ,gcorr_folate_cor
             ))
save(gcorrs_a,file ="C:/Users/PARTSON/gcorrs_a.rda")
gcorrs_a$covariate <- gsub("timeg","time",gcorrs_a$covariate)
gcorrs = gcorrs a
save(gcorrs,file ="C:/Users/PARTSON/gcorrs.rda")
write.csv(gcorrs,file ="C:/Users/PARTSON/gcorrs.csv",na = "",row.names = FALSE)
########################
# Prepare Table 4.7   #
# Gcorr and pvalues   #
########################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/gcorrs.rda")
load(file ="C:/Users/PARTSON/covs.rda")
covs$phase <- gsub("phase","",covs$phase)

pvalues_a = subset(covs[,c(1,3,7,8)],covparm=="UN(2,1)")%>%
na.omit()%>%arrange(covariate,phase)
gcorrs_a = gcorrs %>% arrange(covariate,phase)
#=============================================================================#
# Table 4.7: Correlation between intercept and slope                          #
#=============================================================================#
gcorrs pvalues a = data.frame(gcorrs a[,-1],pvalues_a[,-1])
gcorrs_pvalues = gcorrs_pvalues_a[,-c(4,6)]
long = gcorrs_pvalues
if(!require(dplyr)){install.packages("dplyr")}
long$is_cor = sprintf("%.4f",round(long$is_cor,4)) # round off to 4 with trailing zeros
long$pvalue = sprintf("%.4f",round(long$pvalue,4)) # round off to 4 with trailing zeros

long$pvalue = paste0("(",long$pvalue,")") # bracket the values
```

```
long$is_cor = as.character(long$is_cor)
long$is_cor_pvalue =paste(long$is_cor,long$pvalue, sep="") # combine columns
long = long[,-c(3:4)]

if(!require(reshape2)){install.packages("reshape2")}
L2W_a = dcast(long,covariate~phase,value.var = "is_cor_pvalue") # long to wide
L2W_b = L2W_a[c(19,1:18),]
L2W_c = L2W_b
L2W_c$covariate <- gsub("lymphocytes","Lymphocytes",L2W_c$covariate)
L2W_c$covariate <- gsub("basophils","Basophils",L2W_c$covariate)
L2W_c$covariate <- gsub("albumin","Albumin",L2W_c$covariate)
L2W_c$covariate <- gsub("haematocrit","Haematocrit",L2W_c$covariate)
L2W_c$covariate <- gsub("alkaline_phos","ALP",L2W_c$covariate)
L2W_c$covariate <- gsub("mcv","MCV",L2W_c$covariate)
L2W_c$covariate <- gsub("platelet","Platelet",L2W_c$covariate)
L2W_c$covariate <- gsub("potassium","Potassium",L2W_c$covariate)
L2W_c$covariate <- gsub("monocytes","Monocytes",L2W_c$covariate)
L2W_c$covariate <- gsub("protein","Total protein",L2W_c$covariate)
L2W_c$covariate <- gsub("lactate_dehyd","LDH",L2W_c$covariate)
L2W_c$covariate <- gsub("folate","Folate",L2W_c$covariate)
L2W_c$covariate <- gsub("magnesium","Magnesium",L2W_c$covariate)
L2W_c$covariate <- gsub("glucose","Glucose",L2W_c$covariate)
L2W_c$covariate <- gsub("calcium","Calcium",L2W_c$covariate)
L2W_c$covariate <- gsub("mchc","MCHC",L2W_c$covariate)
L2W_c$covariate <- gsub("red_cells","Red blood cells",L2W_c$covariate)
L2W_c$covariate <- gsub("sodium","Sodium",L2W_c$covariate)
L2W_c$covariate <- gsub("time","Time",L2W_c$covariate)
intr_slope_cor = L2W_c
save(intr_slope_cor,file ="C:/Users/PARTSON/intr_slope_cor.rda")
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(intr_slope_cor,file ="C:/Users/PARTSON/intr_slope_cor.xlsx"  ,row.names = FALSE  )

########################
# Prepare Figure 4.3   #
########################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(scales)){install.packages("scales")}
if(!require(Rcmdr)){install.packages("Rcmdr")}


save(format_plot,file ="C:/Users/PARTSON/format_plot.rda")

load(file ="C:/Users/PARTSON/covs.rda")

# UN(2,1)
format plot=(
      theme bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element text(size=9,angle=90,color="black", hjust=1, vjust=0.25),
    axis.title.y = element text(color="black", size=10, face="bold"),
    axis.text.y = element text(size=9,angle=0,color="black", hjust=1,vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

#=============================================================================#
# Plot Figure 4.3 : Proportion of variation in intercept and slope covs.   #
#=============================================================================#

ggplot(subset(covs,covparm=="UN(2,1)"& !is.na(phase)),
  aes(x=covariate, y=prop_est))+
  geom_col(aes(fill = factor(sig)),width=.6)+
          scale_fill_manual(values=c("grey70","blue"),name="Covariance parameter test:")+
  facet wrap(~phase)+
         labs( x="Covariate with random effect",
               y="Intercept & slope covariation(%)"
              )+ format_plot
```

C5: Chapter 5 codes

```
#############################################################################
#                                     #
# GENERALISED ADDITIVE MIXED MODELS   #
#                                     #
#######################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")

load(file ="C:/Users/PARTSON/data2model.rda")
datagam = data2model

if(!require(mgcv)){install.packages("mgcv")}

memory.limit()  # check memory size
memory.limit(size=10^12) # increase memory size
if(!require(itsadug)){install.packages("itsadug")} # call start_event

datagam0 <- start_event(datagam, column="time", event=c("patientid","phase"),
                        label.event="Event")
save(datagam0,file ="C:/Users/PARTSON/datagam0.rda")

#------------
# Model 1
#------------
bam_ar0 <- bam(cd4 ~
           s(red_cells     ,phase,bs="fs",m=1,k=5)
          +s(haematocrit   ,phase,bs="fs",m=1,k=5)
          +s(mcv           ,phase,bs="fs",m=1,k=5)
          +s(mchc          ,phase,bs="fs",m=1,k=5)
          +s(platelet      ,phase,bs="fs",m=1,k=5)
          +s(lymphocytes   ,phase,bs="fs",m=1,k=5)
          +s(monocytes     ,phase,bs="fs",m=1,k=5)
          +s(basophils     ,phase,bs="fs",m=1,k=5)
          #+s(glucose,phase, bs="fs", m=1, k=4)
          +s(alkaline_phos,phase,bs="fs",m=1,k=5)
          +s(calcium       ,phase,bs="fs",m=1,k=5)
          +s(magnesium     ,phase,bs="fs",m=1,k=5)
          +s(potassium     ,phase,bs="fs",m=1,k=5)
          +s(sodium        ,phase,bs="fs",m=1,k=5)
          +s(protein       ,phase,bs="fs",m=1,k=5)
          +s(albumin       ,phase,bs="fs",m=1,k=5)
          +s(lactate_dehyd,phase,bs="fs",m=1,k=5)
          +s(folate        ,phase,bs="fs",m=1,k=5)
          ,data = datagam0) # Ignore warning!
save(bam_ar0,file ="C:/Users/PARTSON/bam_ar0.rda")

# Term significance
load(file ="C:/Users/PARTSON/bam_ar0.rda")


#Inspection of the residuals of an AR0 model
if(!require(Rcmdr)){install.packages("Rcmdr")}
par(mfrow=c(1,2), cex=1.1)
acf(resid_gam(bam_ar0),main="")

# Account for autocorrelation
# we also ask to plot the ACF by specifying plot (FALSE by default):
bam_rho1 <- start_value_rho(bam_ar0, plot=TRUE);bam_rho1
save(bam_rho1,file ="C:/Users/PARTSON/bam_rho1.rda")

#------------
# Model 2
#------------
bam_ar1 <- bam(cd4 ~
           s(red_cells     ,phase,bs="fs",m=1,k=5)
          +s(haematocrit   ,phase,bs="fs",m=1,k=5)
          +s(mcv           ,phase,bs="fs",m=1,k=5)
          +s(mchc          ,phase,bs="fs",m=1,k=5)
          +s(platelet      ,phase,bs="fs",m=1,k=5)
          +s(lymphocytes   ,phase,bs="fs",m=1,k=5)
          +s(monocytes     ,phase,bs="fs",m=1,k=5)
          +s(basophils     ,phase,bs="fs",m=1,k=5)
          #+s(glucose,phase, bs="fs", m=1, k=4)
          +s(alkaline_phos,phase,bs="fs",m=1,k=5)
```

```
          +s(calcium       ,phase,bs="fs",m=1,k=5)
          +s(magnesium     ,phase,bs="fs",m=1,k=5)
          +s(potassium     ,phase,bs="fs",m=1,k=5)
          +s(sodium        ,phase,bs="fs",m=1,k=5)
          +s(protein       ,phase,bs="fs",m=1,k=5)
          +s(albumin       ,phase,bs="fs",m=1,k=5)
          +s(lactate_dehyd,phase,bs="fs",m=1,k=5)
          +s(folate        ,phase,bs="fs",m=1,k=5)
          ,rho = bam_rho1, AR.start = datagam0$start.event, data = datagam0)
AIC(bam ar0,bam ar1)

save(bam_ar1,file ="C:/Users/PARTSON/bam_ar1.rda")

# Add overall covariate smoothing
load(file ="C:/Users/PARTSON/datagam0.rda")
load(file ="C:/Users/PARTSON/bam_rho1.rda")

#------------
# Model 3
#------------
bam ar1s <- bam(cd4 ~
           s(red_cells     ,m=1,k=5)
          +s(haematocrit   ,m=1,k=5)
          +s(mcv           ,m=1,k=5)
          +s(mchc          ,m=1,k=5)
          +s(platelet      ,m=1,k=5)
          +s(lymphocytes   ,m=1,k=5)
          +s(monocytes     ,m=1,k=5)
          +s(basophils     ,m=1,k=5)
          #+s(glucose,phase,m=1,k=5)
          +s(alkaline phos,m=1,k=5)
          +s(calcium       ,m=1,k=5)
          +s(magnesium     ,m=1,k=5)
          +s(potassium     ,m=1,k=5)
          +s(sodium        ,m=1,k=5)
          +s(protein       ,m=1,k=5)
          +s(albumin       ,m=1,k=5)
          +s(lactate dehyd,m=1,k=5)
          +s(folate        ,m=1,k=5)   #17

          +s(red_cells     ,phase,bs="fs",m=1,k=5)
          +s(haematocrit   ,phase,bs="fs",m=1,k=5)
          +s(mcv           ,phase,bs="fs",m=1,k=5)
          +s(mchc          ,phase,bs="fs",m=1,k=5)
          +s(platelet      ,phase,bs="fs",m=1,k=5)
          +s(lymphocytes   ,phase,bs="fs",m=1,k=5)
          +s(monocytes     ,phase,bs="fs",m=1,k=5)
          +s(basophils     ,phase,bs="fs",m=1,k=5)
          #+s(glucose,phase, bs="fs", m=1, k=4)
          +s(alkaline_phos,phase,bs="fs",m=1,k=5)
          +s(calcium       ,phase,bs="fs",m=1,k=5)
          +s(magnesium     ,phase,bs="fs",m=1,k=5)
          +s(potassium     ,phase,bs="fs",m=1,k=5)
          +s(sodium        ,phase,bs="fs",m=1,k=5)
          +s(protein       ,phase,bs="fs",m=1,k=5)
          +s(albumin       ,phase,bs="fs",m=1,k=5)
          +s(lactate_dehyd,phase,bs="fs",m=1,k=5)
          +s(folate        ,phase,bs="fs",m=1,k=5)
          ,rho = bam_rho1, AR.start = datagam0$start.event, data = datagam0)
load(file ="C:/Users/PARTSON/bam_ar0.rda")
AIC(bam_ar0,bam_ar1,bam_ar1s)

save(bam_ar1s,file ="C:/Users/PARTSON/bam_ar1s.rda")

#------------
# Model 4
#------------
# Add parametric terms
bam_ar1sp <- bam(cd4 ~
           red cells
          +haematocrit
          +mcv
          +mchc
          +platelet
          +lymphocytes
          +monocytes
          +basophils
```

```
          #+glucose,
          +alkaline phos
          +calcium
          +magnesium
          +potassium
          +sodium
          +protein
          +albumin
          +lactate dehyd
          +folate          #17

          +s(red_cells     ,m=1,k=5)
          +s(haematocrit   ,m=1,k=5)
          +s(mcv           ,m=1,k=5)
          +s(mchc          ,m=1,k=5)
          +s(platelet      ,m=1,k=5)
          +s(lymphocytes   ,m=1,k=5)
          +s(monocytes     ,m=1,k=5)
          +s(basophils     ,m=1,k=5)
          #+s(glucose      ,m=1,k=5)
          +s(alkaline phos,m=1,k=5)
          +s(calcium       ,m=1,k=5)
          +s(magnesium     ,m=1,k=5)
          +s(potassium     ,m=1,k=5)
          +s(sodium        ,m=1,k=5)
          +s(protein       ,m=1,k=5)
          +s(albumin       ,m=1,k=5)
          +s(lactate_dehyd,m=1,k=5)
          +s(folate        ,m=1,k=5)

          +s(red_cells     ,phase,bs="fs",m=1,k=5)
          +s(haematocrit   ,phase,bs="fs",m=1,k=5)
          +s(mcv           ,phase,bs="fs",m=1,k=5)
          +s(mchc          ,phase,bs="fs",m=1,k=5)
          +s(platelet      ,phase,bs="fs",m=1,k=5)
          +s(lymphocytes   ,phase,bs="fs",m=1,k=5)
          +s(monocytes     ,phase,bs="fs",m=1,k=5)
          +s(basophils     ,phase,bs="fs",m=1,k=5)
          #+s(glucose      ,phase,bs="fs",m=1, k=4)
          +s(alkaline_phos,phase,bs="fs",m=1,k=5)
          +s(calcium       ,phase,bs="fs",m=1,k=5)
          +s(magnesium     ,phase,bs="fs",m=1,k=5)
          +s(potassium     ,phase,bs="fs",m=1,k=5)
          +s(sodium        ,phase,bs="fs",m=1,k=5)
          +s(protein       ,phase,bs="fs",m=1,k=5)
          +s(albumin       ,phase,bs="fs",m=1,k=5)
          +s(lactate_dehyd,phase,bs="fs",m=1,k=5)
          +s(folate        ,phase,bs="fs",m=1,k=5)
          ,rho = bam_rho1, AR.start = datagam0$start.event, data = datagam0)
load(file ="C:/Users/PARTSON/bam_ar0.rda")
load(file ="C:/Users/PARTSON/bam_ar1.rda")
load(file ="C:/Users/PARTSON/bam_ar1s.rda")
AIC(bam ar0,bam ar1,bam ar1s,bam ar1sp)
save(bam_ar1sp,file ="C:/Users/PARTSON/bam_ar1sp.rda")


#------------
# Model 5
#------------
# Tune basis dimensions
load(file ="C:/Users/PARTSON/datagam0.rda")
load(file ="C:/Users/PARTSON/bam_rho1.rda")


bam ar1sk <- bam(cd4 ~
           s(red_cells     ,m=1,k=5)
          +s(haematocrit   ,m=1,k=5)
          +s(mcv           ,m=1,k=10)
          +s(mchc          ,m=1,k=5)
          +s(platelet      ,m=1,k=10)
          +s(lymphocytes   ,m=1,k=10)
          +s(monocytes     ,m=1,k=5)
          +s(basophils     ,m=1,k=10)
          #+s(glucose      ,m=1,k=5)
          +s(alkaline phos,m=1,k=5)
          +s(calcium       ,m=1,k=5)
          +s(magnesium     ,m=1,k=5)
```

```
          +s(potassium     ,m=1,k=5)
          +s(sodium        ,m=1,k=5)
          +s(protein       ,m=1,k=10)
          +s(albumin       ,m=1,k=10)
          +s(lactate_dehyd,m=1,k=5)
          +s(folate        ,m=1,k=10)

          +s(red_cells     ,phase,bs="fs",m=1,k=5)
          +s(haematocrit   ,phase,bs="fs",m=1,k=5)
          +s(mcv           ,phase,bs="fs",m=1,k=5)
          +s(mchc          ,phase,bs="fs",m=1,k=5)
          +s(platelet      ,phase,bs="fs",m=1,k=5)
          +s(lymphocytes   ,phase,bs="fs",m=1,k=5)
          +s(monocytes     ,phase,bs="fs",m=1,k=5)
          +s(basophils     ,phase,bs="fs",m=1,k=5)
          #+s(glucose      ,phase,bs="fs",m=1, k=4)
          +s(alkaline_phos,phase,bs="fs",m=1,k=5)
          +s(calcium       ,phase,bs="fs",m=1,k=5)
          +s(magnesium     ,phase,bs="fs",m=1,k=5)
          +s(potassium     ,phase,bs="fs",m=1,k=5)
          +s(sodium        ,phase,bs="fs",m=1,k=5)
          +s(protein       ,phase,bs="fs",m=1,k=5)
          +s(albumin       ,phase,bs="fs",m=1,k=5)
          +s(lactate_dehyd,phase,bs="fs",m=1,k=5)
          +s(folate        ,phase,bs="fs",m=1,k=5)
          ,rho = bam_rho1, AR.start = datagam0$start.event, data = datagam0)
load(file ="C:/Users/PARTSON/bam_ar0.rda")
load(file ="C:/Users/PARTSON/bam_ar1.rda")
load(file ="C:/Users/PARTSON/bam_ar1s.rda")
load(file ="C:/Users/PARTSON/bam_ar1sp.rda")
AIC(bam ar0,bam ar1,bam ar1s,bam ar1sp,bam ar1sk)

save(bam_ar1sk,file ="C:/Users/PARTSON/bam_ar1sk.rda")

##############################
# Prepare Table 5.1          #
##############################
# Test to select the best models
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/bam_ar0.rda")
load(file ="C:/Users/PARTSON/bam_ar1.rda")
load(file ="C:/Users/PARTSON/bam_ar1s.rda")
load(file ="C:/Users/PARTSON/bam_ar1sp.rda")
load(file ="C:/Users/PARTSON/bam_ar1sk.rda")
load(file ="C:/Users/PARTSON/bam_rho1.rda")
bam_rho1=round(bam_rho1,4)
if(!require(dplyr)){install.packages("dplyr")}
if(!require(itsadug)){install.packages("itsadug")}
#============================================================================#
# Table 5.1                                                                  #
#============================================================================#
model selection = data.frame(
      Model       = c("1","2","3","4","5"),
      Terms       = c("s0+s(x,phase)","s0+s(x,phase)+AR","s0+s(x)+s(x,phase)+AR",
                      "s0+x+s(x)+s(x,phase)+AR", "s0+s(x)+s(x,phase)+AR"),
      DF          = (AIC(bam_ar0,bam_ar1,bam_ar1s,bam_ar1sp,bam_ar1sk))[1],
      AIC         = c(AIC(bam_ar0)  ,AIC(bam_ar1),AIC(bam_ar1s)
                     ,AIC(bam ar1sp),AIC(bam ar1sk)),
      R2          = c(summary(bam_ar0)$r.sq  ,summary(bam_ar1)$r.sq,
                      summary(bam_ar1s)$r.sq ,summary(bam_ar1sp)$r.sq,summary(bam_ar1sk)$r.sq)
                      ,Dev.Expl    = c(summary(bam_ar0)$dev.expl  ,summary(bam_ar1)$dev.expl,
                       summary(bam_ar1s)$dev.expl,summary(bam_ar1sp)$dev.expl,
                       summary(bam_ar1sk)$dev.expl),
      Rho         = c("",bam_rho1,bam_rho1,bam_rho1,bam_rho1)
       )%>% mutate_if(is.numeric, function(x) {
                     round(x,4)
                           })
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(model_selection,file ="C:/Users/PARTSON/model_selection.xlsx",row.names = FALSE )
#============================================================================#
# Figure 5.1: Inspection of autocorrelation                                  #
#============================================================================#

rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
load(file ="C:/Users/PARTSON/bam_ar0.rda")
load(file ="C:/Users/PARTSON/bam_ar1s.rda")
if(!require(itsadug)){install.packages("itsadug")} # call start_event
```

```
if(!require(Rcmdr)){install.packages("Rcmdr")}

#Inspection of the residuals of an AR1 model
#Uncorrected versus corrected residuals:
par(mfrow=c(2,2), cex=1.1)
acf_resid(bam_ar0, main="Uncorrected ACF") #Uncorrected
start_value_rho(bam_ar0, plot=TRUE, main="ACF coeff (rho)")
acf_resid(bam_ar1s, main="Corrected ACF") #corrected
acf resid(bam ar1s, split pred="AR.start")

# Prepare Table 5.2

gam.check(bam_ar1s)
                # copy and paste to Notepad and read as csv
# Term significance
summary(bam_ar1s)$p.table
summary(bam_ar1s)$s.table
                # copy and paste to Notepad and read as csv
#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
s table txt <- read.csv("C:/Users/PARTSON/s table.txt", sep="")
gam_check_txt <- read.csv("C:/Users/PARTSON/gam_check.txt", sep="")
#=============================================================================#
# Table 5.2                                                                   #
#=============================================================================#
s_table_a = merge(s_table_txt, gam_check_txt[,c(1,2,4)], by=c("Covariate"),sort = FALSE,
all.x=TRUE) # bring all from "left" FIRST and update with "right"

if(!require(dplyr)){install.packages("dplyr")}
s_table_a$p.value = sprintf("%.4f",round(s_table_a$p.value ,4)) # round off to 4 with trailing
zeros
s_table_a$F       = sprintf("%.4f",round(s_table_a$F ,4)) # round off to 4 with trailing zeros
s_table_a$edf     = sprintf("%.4f",round(s_table_a$edf ,4)) # round off to 4 with trailing
zeros
s_table_a$k.index = sprintf("%.2f",round(s_table_a$k.index ,2)) # round off to 4 with trailing
zeros

s_table_a$Covariate <- gsub("lymphocytes","Lymphocytes",s_table_a$Covariate)
s_table_a$Covariate <- gsub("basophils","Basophils",s_table_a$Covariate)
s_table_a$Covariate <- gsub("albumin","Albumin",s_table_a$Covariate)
s_table_a$Covariate <- gsub("haematocrit","Haematocrit",s_table_a$Covariate)
s_table_a$Covariate <- gsub("alkaline_phos","ALP",s_table_a$Covariate)
s_table_a$Covariate <- gsub("mcv","MCV",s_table_a$Covariate)
s_table_a$Covariate <- gsub("platelet","Platelet",s_table_a$Covariate)
s_table_a$Covariate <- gsub("potassium","Potassium",s_table_a$Covariate)
s_table_a$Covariate <- gsub("monocytes","Monocytes",s_table_a$Covariate)
s_table_a$Covariate <- gsub("protein","Total protein",s_table_a$Covariate)
s_table_a$Covariate <- gsub("lactate_dehyd","LDH",s_table_a$Covariate)
s_table_a$Covariate <- gsub("folate","Folate",s_table_a$Covariate)
s_table_a$Covariate <- gsub("magnesium","Magnesium",s_table_a$Covariate)
s_table_a$Covariate <- gsub("glucose","Glucose",s_table_a$Covariate)
s_table_a$Covariate <- gsub("calcium","Calcium",s_table_a$Covariate)
s_table_a$Covariate <- gsub("mchc","MCHC",s_table_a$Covariate)
s_table_a$Covariate <- gsub("red_cells","Red blood cells",s_table_a$Covariate)
s_table_a$Covariate <- gsub("sodium","Sodium",s_table_a$Covariate)
s_table = s_table_a[,c(1,6,2,4,5,7)]

if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(s_table,file ="C:/Users/PARTSON/s_table.xlsx"  ,row.names = FALSE  )


# Prepare Figure 5.2

# Figure 5.2a: get QQ and res density
par(mfrow=c(2,2), cex=1.1)
check_resid(fit, split_pred="AR.start",select = c(2,4,1,3), ask = FALSE)
     #save as tiff

# Figure 5.2b: get response vs fitted
par(mfrow=c(2,2), cex=1.1)
gam.check(bam_ar1s,old.style=T,
          type=c(
             "deviance"
            ,"pearson"
            ,"response"
                ),
          k.sample=5000,k.rep=200,
```

```
          rep=0, level=.9, rl.col=2, rep.col="gray80")
      #save as tiff

if(!require(magick)){install.packages("magick")}
if(!require(rsvg)){install.packages("rsvg")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

Figure5.2a = image_read("C:/Users/Partson/Figure5.2a.tiff");Figure5.2a
Figure5.2a_crop =image_crop(Figure5.2a, "255x");Figure5.2a_crop # crop from right

Figure5.2b = image_read("C:/Users/Partson/Figure5.2b.tiff");Figure5.2b
Figure5.2b_crop =image_crop(Figure5.2b, "-295x");Figure5.2b_crop # crop from left
#=============================================================================#
# Figure 5.2 : Combine Figure 5.1a&b: crop and merge graphs                   #
#=============================================================================#
Figure5.2_crop = image_append(c(Figure5.2a_crop , Figure5.2b_crop), stack =
FALSE);Figure5.2_crop
image_write(Figure5.2_crop,"C:/Users/Partson/Figure5.2_crop.tiff", format = "tiff")

################################################
## Define macro for the GAM random smooth plots   #
## Preparing Figure 5.3 to 5.8                    #
################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ;options(prompt = "R>")
if(!require(gtools)){install.packages("gtools")}
#
gam.plot= defmacro( x         =  ,
                    x_lab     =  ,
                    legendbox = F, # check whether header needed in first row
# functions to be evaluated start here
    expr={

# load the required packages
if(!require(Rcmdr)){install.packages("Rcmdr")}
if(!require(itsadug)){install.packages("itsadug")}

if (legendbox==F) {
inspect_random(bam_ar1s,select=x,fun=mean,col="yellow",lwd=2,xlab="",ylab="",main="",h0=0)
           title(xlab=x_lab, ylab=bquote("CD4"^"+"*"count [s(x)+s(x,z)]"),line=2.2)
inspect_random(bam_ar1s,select=x,add=TRUE)

} else if(legendbox==T){

plot(0,type='n',axes=FALSE,ann=FALSE) # draw empty plot
legend("top",inset=c(0,0),legend = c("Overall   [s(x)]",
                                     "2-Acute [z:g=1]",
                                     "3-Early   [z:g=2]",
                                     "4-Est      [z:g=3]",
                                     "5-ART    [z:g=4]")
            , title="INFECTION PHASE" # setup legend
            ,bty = "n", col = c("yellow","black","red","green","blue")
            ,horiz=FALSE,lty = c(1,1,2,3,6), lwd = c(2,1,1,1,1),cex=.9,xpd=TRUE)
}
})

load(file ="C:/Users/PARTSON/bam_ar1s.rda")

#=============================================================================#
# Figure 5.3 --- Significant difference -overall upward trend                 #
#=============================================================================#
par(mfrow=c(2,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=    ,x_lab  =                    , legend = T)
gam.plot( x= 23 ,x_lab  =bquote("Lymphocytes(x10"^"9"*"/L)"), legend = F)
gam.plot( x= 19 ,x_lab  ="Haematocrit(Hct/100)" , legend = F)
gam.plot( x= 22 ,x_lab  =bquote("Platelet(x10"^"9"*"/L)")  , legend = F)
gam.plot( x= 32 ,x_lab  ="Albumin(g/L)"     ,legend = F)
gam.plot( x= 26 ,x_lab  ="ALP(IU/L)"        ,legend = F)
#=============================================================================#
# Figure 5.4 --- Significant difference -overall downward trend               #
#=============================================================================#
par(mfrow=c(1,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=   ,x_lab  =                   ,legend = T)
gam.plot( x= 31,x_lab  ="Total protein(g/L)" , legend = F)
gam.plot( x= 30,x_lab  ="Sodium(mEq/L)"    ,legend = F)
```

```
#=============================================================================#
# Figure 5.5 --- Significant difference -overall irregular trend             #
#=============================================================================#
par(mfrow=c(1,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=   ,x_lab  =                    ,legend = T)
gam.plot( x= 33,x_lab  ="LDH(U/L)"          ,legend = F)
gam.plot( x= 18 ,x_lab  =bquote("Red blood cells(x10"^"6"*"cells/mm"^"3"*""),legend = F)
#=============================================================================#
# Figure 5.6---Insignificant difference-overall upward trend                 #
#=============================================================================#
par(mfrow=c(1,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=    ,x_lab  =                    , legend = T)
gam.plot( x= 20 ,x_lab  ="MCV(fL)"           , legend = F)
gam.plot( x= 25 ,x_lab  =bquote("Basophils(x10"^"9"*"/L)") , legend = F)
#=============================================================================#
# Figure 5.7 ---Insignificant difference-general downward trend              #
#=============================================================================#
par(mfrow=c(1,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=    ,x_lab  =                    , legend = T)
gam.plot( x= 24 ,x_lab  =bquote("Monocytes(x10"^"9"*"/L)"),legend = F)
gam.plot( x= 34,x_lab  ="Folate(nmol/L)"    ,legend = F)
#=============================================================================#
# Figure 5.8 ---Insignificant terms                                          #
#=============================================================================#
par(mfrow=c(2,3),mar = c(3.2,3.2,1.5,1),cex=0.7)
gam.plot( x=   ,x_lab  =                    ,legend = T)
gam.plot( x= 29,x_lab  ="Potassium(mmol/L)",legend = F)
gam.plot( x= 28,x_lab  ="Magnesium(mmol/L)",legend = F)
plot(0,type='n',axes=FALSE,ann=FALSE) # draw empty plot
gam.plot( x= 27,x_lab  ="Calcium(mmol/L)"   ,legend = F)
gam.plot( x= 21 ,x_lab  ="MCHC(g/dL)"       ,legend = F)
```

C6: Chapter 6 codes

```
#############################################################################
#                                  #
#       SEGMENTED                  #
#                                  #
####################################

#----------------------------
# 2-Acute
#----------------------------
rm(list= ls()[!(ls() %in% c(""))])  # clear environment and leave the selected

cat("\014")                         # clear console
options(prompt = "R>")              # customize prompt
load(file ="C:/Users/PARTSON/data2model.rda")
if(!require(segmented)){install.packages("segmented")}


out.glm<-glm(cd4~
 red cells     + haematocrit   + mcv       + mchc      +
 platelet      + lymphocytes   + monocytes + basophils +
 glucose       + alkaline_phos + calcium   + magnesium +
 potassium     + sodium        + protein   + albumin   +
 lactate_dehyd + folate,
 data=subset(data2model,phase=="2-Acute"))
glm2 = out.glm
save(glm2,file ="C:/Users/PARTSON/glm2.rda")
#plot(out.glm)

seg.control = seg.control(
  toll = 1e-04  ,it.max  =10    ,display  = FALSE, stop.if.error = TRUE
  ,K    = 2     ,quant   =FALSE ,last     = TRUE , maxit.glm     = 25
  ,h    = 500   ,n.boot  =20    ,size.boot= NULL , gap           = FALSE
  ,jt   = FALSE ,nonParam=TRUE  ,random   =TRUE  , powers        = c(1,1)
  ,seed= 100    ,fn.obj  =NULL  ,digits   =NULL)

o<-segmented(out.glm,seg.Z=~
 red cells     + haematocrit   + mcv       + mchc      +
 platelet      + lymphocytes   + monocytes + basophils +
 glucose       + alkaline_phos + calcium   + magnesium +
```

```
 potassium     + sodium        + protein    + albumin    +
 lactate dehyd + folate
,psi=list(
          red_cells     =c(4.551,4.772)
         ,haematocrit   =c(0.327,0.420)
         ,mcv           =(90.348 )
         ,mchc          =c(32.529,32.714) #Already a straight line though
         ,platelet      =c(287.424,487.099)
         ,lymphocytes   =c(0.725,3.459 )
         ,monocytes     =c(0.4375) #misbehaving-work better fixed with all covariates
         ,basophils     =c(0.035,0.0795) # rely on auto
         #,glucose       =NA#c( )
         ,alkaline_phos=c(97.249,110.982)
         ,calcium       =c(2.518) #2.2, work better fixed with all covariates
         ,magnesium     =c(0.75,0.875,0.945) #work better fixed with all covariates
         ,potassium     =c(3.084,3.586) #work better fixed with all covariates
         ,sodium        =c(133.,140) #work better fixed with all covariates
         ,protein       =c(70) #,82.5work better fixed with all covariates
         ,albumin       =c(36.5)#29.920,35.685
         ,lactate_dehyd=c(327.809,607.043)
         ,folate        =c(17.5,25)#work better fixed with all covariates
         ),visual=TRUE
         ,control = seg.control
          )
#***
o2 = o
save(o2,file ="C:/Users/PARTSON/o2.rda")
#*** -----> change end too

#-common---------------------------
psi_aa =confint.segmented(o,level=0.95, rev.sgn=FALSE, var.diff=FALSE,
      digits=max(4, getOption("digits") - 1))
psi_a =do.call(rbind.data.frame, psi_aa) # convert list to data frame
psi_b = tibble::rownames_to_column(psi_a,"covariate")
psi c = psi b
psi c$covariate[psi c$covariate=="red cells"]     <-"red cells.psi1.red cells"
psi c$covariate[psi c$covariate=="haematocrit"]   <-"haematocrit.psi1.haematocrit"
psi c$covariate[psi c$covariate=="mcv"]           <-"mcv.psi1.mcv"
psi_c$covariate[psi_c$covariate=="mchc"]          <-"mchc.psi1.mchc"
psi_c$covariate[psi_c$covariate=="platelet"]      <-"platelet.psi1.platelet"
psi c$covariate[psi c$covariate=="lymphocytes"]   <-"lymphocytes.psi1.lymphocytes"
psi c$covariate[psi c$covariate=="monocytes"]     <-"monocytes.psi1.monocytes"
psi c$covariate[psi c$covariate=="basophils"]     <-"basophils.psi1.basophils"
psi_c$covariate[psi_c$covariate=="alkaline_phos"]<-"alkaline_phos.psi1.alkaline_phos"
psi_c$covariate[psi_c$covariate=="calcium"]       <-"calcium.psi1.calcium"
psi_c$covariate[psi_c$covariate=="magnesium"]     <-"magnesium.psi1.magnesium"
psi c$covariate[psi c$covariate=="potassium"]     <-"potassium.psi1.potassium"
psi c$covariate[psi c$covariate=="sodium"]        <-"sodium.psi1.sodium"
psi_c$covariate[psi_c$covariate=="protein"]       <-"protein.psi1.protein"
psi_c$covariate[psi_c$covariate=="albumin"]       <-"albumin.psi1.albumin"
psi_c$covariate[psi_c$covariate=="lactate_dehyd"]<-"lactate_dehyd.psi1.lactate_dehyd"
psi c$covariate[psi c$covariate=="folate"]        <-"folate.psi1.folate"
psi d      =psi c
psi_d$Est. = sprintf("%.2f",round(psi_d$Est.,2)) # round off to 4 with trailing zeros
psi_d$Est. =ifelse((psi_d$"CI(95%).l")*(psi_d$"CI(95%).u")<0
                   ,paste0("",psi_d$Est.,"*"),psi_d$Est.)

if(!require(data.table)){install.packages("data.table")}

                        #***
setnames(psi_d,"Est.","2-Acute")
#***
psi2 = psi_d[,1:2]
#***
save(psi2,file ="C:/Users/PARTSON/psi2.rda")
#*** -----> change end too

if(!require(dplyr)){install.packages("dplyr")}
psi_e =psi_c %>% mutate_if(is.numeric, function(x) {
    sprintf("%.2f",round(x,2))})
psi_e$"CI(95%).l" = paste0("(",psi_e$"CI(95%).l"," ; ") # bracket and ;
psi_e$"CI(95%).u" = paste0("",psi_e$"CI(95%).u",")")
psi_e$CI =paste(psi_e$"CI(95%).l",psi_e$"CI(95%).u", sep="") # combine columns

       #***                    #***
psi_e$"2-Acute" =paste(psi_e$"2-Acute",psi_e$CI, sep="") # combine columns
psi_e$"2-Acute" =ifelse((psi_d$"CI(95%).l")*(psi_d$"CI(95%).u")<0
```

```
                    ,paste0("",psi_e$"2-Acute","*"),psi_e$"2-Acute")

#***
psi2apx = psi e[,1:2]
save(psi2apx,file ="C:/Users/PARTSON/psi2apx.rda")
#*** -----> change end too

slope        =slope(o) # show slope sig using C.I.
slope a      =do.call(rbind.data.frame, slope)
slope aa      = tibble::rownames to column(slope a,"covariate")
slope_b      = slope_aa
slope_b$Est. =ifelse((slope_b$"CI(95%).l")*(slope_b$"CI(95%).u")>0
              ,paste0("",slope_b$Est.,"*"),slope_b$Est.)

                         #***
setnames(slope_b,"Est.","2-Acute")

#***
slope2 = slope_b[,1:2]
#***
save(slope2,file ="C:/Users/PARTSON/slope2.rda")
#*** -----> change end too

# Appendix
if(!require(dplyr)){install.packages("dplyr")}
slope_c =slope_aa %>% mutate_if(is.numeric, function(x) {
    sprintf("%.2f",round(x,2))})
slope_c$"CI(95%).l" = paste0("(",slope_c$"CI(95%).l"," ; ") # bracket and ;
slope c$"CI(95%).u" = paste0("",slope c$"CI(95%).u",")")
slope_c$CI =paste(slope_c$"CI(95%).l",slope_c$"CI(95%).u", sep="") # combine columns

         #***                    #***
slope_c$"2-Acute" =paste(slope_c$"2-Acute",slope_c$CI, sep="") # combine columns

         #***
slope_c$"2-Acute" =ifelse((slope_b$"CI(95%).l")*(slope_b$"CI(95%).u")>0
             ,paste0("",slope c$"2-Acute","*"),slope c$"2-Acute")
                                  #***                    #***
#***
slope2apx = slope_c[,1:2]
save(slope2apx,file ="C:/Users/PARTSON/slope2apx.rda")
#*** -----> change end too

summary(o) # 'U.' is put before the name of the segmented variable to
           # mean the difference-in-slopes coefficient.
#---------------------------


#---------------------------------------------
# REPEAT ALL THE STEPS ABOVE FOR PHASES 3 to 5
#---------------------------------------------


#----------------------
# MERGE
#----------------------
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt

#psi
load(file ="C:/Users/PARTSON/psi2.rda")
load(file ="C:/Users/PARTSON/psi3.rda")
load(file ="C:/Users/PARTSON/psi4.rda")
load(file ="C:/Users/PARTSON/psi5.rda")

psi_a =Reduce(function(x, y) merge(x, y,by=c("covariate"),sort = TRUE, all=TRUE)
              , list(psi2, psi3, psi4,psi5))

if(!require(splitstackshape)){install.packages("splitstackshape")}
psi b = cSplit(psi a, "covariate", ".")

if(!require(data.table)){install.packages("data.table")}
setnames(psi_b,c("covariate_1","covariate_2"),c("covariate","changepoint"))
psi_c = psi_b[,c(5:6,1:4)]
psi_c$changepoint <- gsub('psi',"",psi_c$changepoint)
psi d = psi c
```

```
psi_d$covariate <- gsub("lymphocytes","Lymphocytes",psi_d$covariate)
psi_d$covariate <- gsub("basophils","Basophils",psi_d$covariate)
psi_d$covariate <- gsub("albumin","Albumin",psi_d$covariate)
psi_d$covariate <- gsub("haematocrit","Haematocrit",psi_d$covariate)
psi_d$covariate <- gsub("alkaline_phos","ALP",psi_d$covariate)
psi_d$covariate <- gsub("mcv","MCV",psi_d$covariate)
psi_d$covariate <- gsub("platelet","Platelet",psi_d$covariate)
psi_d$covariate <- gsub("potassium","Potassium",psi_d$covariate)
psi_d$covariate <- gsub("monocytes","Monocytes",psi_d$covariate)
psi_d$covariate <- gsub("protein","Total protein",psi_d$covariate)
psi_d$covariate <- gsub("lactate_dehyd","LDH",psi_d$covariate)
psi_d$covariate <- gsub("folate","Folate",psi_d$covariate)
psi_d$covariate <- gsub("magnesium","Magnesium",psi_d$covariate)
psi_d$covariate <- gsub("glucose","Glucose",psi_d$covariate)
psi_d$covariate <- gsub("calcium","Calcium",psi_d$covariate)
psi_d$covariate <- gsub("mchc","MCHC",psi_d$covariate)
psi_d$covariate <- gsub("red_cells","Red blood cells",psi_d$covariate)
psi_d$covariate <- gsub("sodium","Sodium",psi_d$covariate)
psi = psi_d

if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(psi,file ="C:/Users/PARTSON/psi.xlsx",row.names = FALSE, showNA=FALSE )

################################################
## Model diagnostics                           #
################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/data2model.rda")
# Load  the glm models
load(file ="C:/Users/PARTSON/glm2.rda")
load(file ="C:/Users/PARTSON/glm3.rda")
load(file ="C:/Users/PARTSON/glm4.rda")
load(file ="C:/Users/PARTSON/glm5.rda")
# Load  the segmented models
load(file ="C:/Users/PARTSON/o2.rda")
load(file ="C:/Users/PARTSON/o3.rda")
load(file ="C:/Users/PARTSON/o4.rda")
load(file ="C:/Users/PARTSON/o5.rda")
if(!require(Rcmdr)){install.packages("Rcmdr")}

#============================================================================#
# Table 6.1                                                                  #
#============================================================================#
# Diagnostics plots for comparing GLM and segmented models
if(!require(dplyr)){install.packages("dplyr")}

pcw diagnostics = data.frame(
  Phase = c("2-Acute","","3-Early","","4-Est","","5-ART",""),
  Model = c("GLM","Segmented","GLM","Segmented","GLM","Segmented","GLM","Segmented"),
  "NULL" =c(summary(glm2)$null,summary(o2)$null,summary(glm3)$null,summary(o3)$null,
            summary(glm4)$null,summary(o4)$null,summary(glm5)$null,summary(o5)$null),
  Deviance =c(summary(glm2)$deviance,summary(o2)$deviance,summary(glm3)$deviance,
              summary(o3)$deviance, summary(glm4)$deviance,summary(o4)$deviance,
              summary(glm5)$deviance, summary(o5)$deviance),
  AIC =c(summary(glm2)$aic,summary(o2)$aic,summary(glm3)$aic,summary(o3)$aic,
            summary(glm4)$aic,summary(o4)$aic,summary(glm5)$aic,summary(o5)$aic)
           )%>% mutate_if(is.numeric, function(x) {
                     round(x,2)
                                 })
if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(pcw_diagnostics,file ="C:/Users/PARTSON/pcw_diagnostics.xlsx"
          ,row.names = FALSE, showNA=FALSE
          )
#============================================================================#
# Figure 6.1                                                                 #
#============================================================================#
par(mfrow=c(2,2))#,mar = c(3.2,3.2,1.5,1),cex=0.7)
plot(glm2,which = c(2), main="Acute phase")
plot(glm3,which = c(2), main="Early phase")
plot(glm4,which = c(2), main="Est phase")
plot(glm5,which = c(2), main="ART phase")
```

```
#=============================================================================#
# Table 6.2                                                                   #
#=============================================================================#
#PSI-APPENDIX
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt

load(file ="C:/Users/PARTSON/psi2apx.rda")
load(file ="C:/Users/PARTSON/psi3apx.rda")
load(file ="C:/Users/PARTSON/psi4apx.rda")
load(file ="C:/Users/PARTSON/psi5apx.rda")

psi_apx_aa =Reduce(function(x, y) merge(x, y,by=c("covariate"),sort = TRUE, all=TRUE)
                , list(psi2apx, psi3apx, psi4apx,psi5apx))

if(!require(splitstackshape)){install.packages("splitstackshape")}
psi_apx_a = cSplit(psi_apx_aa, "covariate", ".")

if(!require(data.table)){install.packages("data.table")}
setnames(psi_apx_a,c("covariate_1","covariate_2"),c("covariate","Point"))
psi_apx_b = psi_apx_a[,c(5:6,1:4)]
psi_apx_b$changepoint <- gsub('psi',"",psi_apx_b$changepoint)
psi_apx_c = psi_apx_b

psi_apx_c$covariate <- gsub("lymphocytes","Lymphocytes",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("basophils","Basophils",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("albumin","Albumin",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("haematocrit","Haematocrit",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("alkaline_phos","ALP",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("mcv","MCV",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("platelet","Platelet",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("potassium","Potassium",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("monocytes","Monocytes",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("protein","Total protein",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("lactate_dehyd","LDH",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("folate","Folate",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("magnesium","Magnesium",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("glucose","Glucose",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("calcium","Calcium",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("mchc","MCHC",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("red_cells","Red blood cells",psi_apx_c$covariate)
psi_apx_c$covariate <- gsub("sodium","Sodium",psi_apx_c$covariate)
psi_apx_c[is.na(psi_apx_c)] ="-" # replace all NA with dash
psi_apx = psi_apx_c

if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(psi_apx,file ="C:/Users/PARTSON/psi_apx.xlsx"
          ,row.names = FALSE, showNA=FALSE
                    )

#slope
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt

load(file ="C:/Users/PARTSON/slope2.rda")
load(file ="C:/Users/PARTSON/slope3.rda")
load(file ="C:/Users/PARTSON/slope4.rda")
load(file ="C:/Users/PARTSON/slope5.rda")
slope =Reduce(function(x, y) merge(x, y,by=c("covariate"),sort = TRUE, all=TRUE)
                , list(slope2, slope3, slope4,slope5))


if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(slope,file ="C:/Users/PARTSON/slope.xlsx",row.names = FALSE, showNA=FALSE)

#=============================================================================#
# Table 6.3 - manually done from Table 6.2 and 6.4                            #
#=============================================================================#
# Table 6.4                                                                   #
#=============================================================================#

# slope appendix
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/slope2apx.rda")
load(file ="C:/Users/PARTSON/slope3apx.rda")
load(file ="C:/Users/PARTSON/slope4apx.rda")
```

```
load(file ="C:/Users/PARTSON/slope5apx.rda")
slope_apx =Reduce(function(x, y) merge(x, y,by=c("covariate"),sort = TRUE, all=TRUE)
                , list(slope2apx, slope3apx, slope4apx,slope5apx))

slope_apx$covariate <- gsub("lymphocytes","Lymphocytes",slope_apx$covariate)
slope_apx$covariate <- gsub("basophils","Basophils",slope_apx$covariate)
slope_apx$covariate <- gsub("albumin","Albumin",slope_apx$covariate)
slope_apx$covariate <- gsub("haematocrit","Haematocrit",slope_apx$covariate)
slope_apx$covariate <- gsub("alkaline_phos","ALP",slope_apx$covariate)
slope_apx$covariate <- gsub("mcv","MCV",slope_apx$covariate)
slope_apx$covariate <- gsub("platelet","Platelet",slope_apx$covariate)
slope_apx$covariate <- gsub("potassium","Potassium",slope_apx$covariate)
slope_apx$covariate <- gsub("monocytes","Monocytes",slope_apx$covariate)
slope_apx$covariate <- gsub("protein","Total protein",slope_apx$covariate)
slope_apx$covariate <- gsub("lactate_dehyd","LDH",slope_apx$covariate)
slope_apx$covariate <- gsub("folate","Folate",slope_apx$covariate)
slope_apx$covariate <- gsub("magnesium","Magnesium",slope_apx$covariate)
slope_apx$covariate <- gsub("glucose","Glucose",slope_apx$covariate)
slope_apx$covariate <- gsub("calcium","Calcium",slope_apx$covariate)
slope_apx$covariate <- gsub("mchc","MCHC",slope_apx$covariate)
slope_apx$covariate <- gsub("red_cells","Red blood cells",slope_apx$covariate)
slope_apx$covariate <- gsub("sodium","Sodium",slope_apx$covariate)

if(!require(splitstackshape)){install.packages("splitstackshape")}
slope apndx a = cSplit(slope apx, "covariate", ".")

if(!require(data.table)){install.packages("data.table")}
setnames(slope_apndx_a ,c("covariate_1","covariate_2"),c("Covariate","Slope"))
slope_apndx_b = slope_apndx_a[,c(5:6,1:4)]
slope_apndx_b$Slope <- gsub('slope',"",slope_apndx_b$Slope)
slope apndx b[is.na(slope apndx b)] ="-" # replace all NA with dash
slope apndx c = slope apndx b
slope_apndx_c$Slope = as.numeric(slope_apndx_c$Slope) -1
slope_apndx = slope_apndx_c


if(!require(xlsx)){install.packages("xlsx")}
write.xlsx(slope_apndx,file ="C:/Users/PARTSON/slope_apndx.xlsx"
          ,row.names = FALSE, showNA=FALSE
          )


################################################
## Define macro for the Segmented plots        #
## Prepare Figure 6.2 to 6.6                   #
################################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>") # customize prompt
if(!require(gtools)){install.packages("gtools")}
#
pcw.plot= defmacro( x      =  ,
                    x_lab  =  ,
                    header = F, # check whether header needed in first row
# functions to be evaluated start here
    expr={

# load the required packages
if(!require(segmented)){install.packages("segmented")}
if(!require(Rcmdr)){install.packages("Rcmdr")}

if (header==F) {
plot.segmented(o2,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2) # line controls distance
from axis
plot.segmented(o3,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab , ylab=bquote("Effect on CD4"^"+"),line=2.2)
plot.segmented(o4,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2)
plot.segmented(o5,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2)

} else if(header==T){

plot.segmented(o2,term=x
```

```
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2)
    title(main = "2-Acute",line=0.15,cex=0.4)
plot.segmented(o3,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab , ylab=bquote("Effect on CD4"^"+"),line=2.2)
    title(main = "3-Early",line=0.15,cex=0.4)
plot.segmented(o4,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2)
    title(main = "4-Established",line=0.15,cex=0.4)
plot.segmented(o5,term=x
    ,dens.rug =TRUE,col="red",interc=FALSE, xlab="",ylab="")
    title(xlab = x_lab, ylab=bquote("Effect on CD4"^"+"),line=2.2)
    title(main = "5-ART",line=0.15,cex=0.4)
}
})

# Load  the segmented models
load(file ="C:/Users/PARTSON/o2.rda")
load(file ="C:/Users/PARTSON/o3.rda")
load(file ="C:/Users/PARTSON/o4.rda")
load(file ="C:/Users/PARTSON/o5.rda")

#=============================================================================#
# Figure 6.2: Desirable effects (in at least two phases)                      #
#=============================================================================#
par(mfrow=c(4,4),mar = c(3.2,3.2,1.5,1),cex=0.7)
                    #(bottom, left, top,right)
pcw.plot(x="lymphocytes"  ,x_lab = bquote("Lymphocytes(x10"^"9"*"/L)"), header=T)
pcw.plot(x="albumin"      ,x_lab = "Albumin(g/L)", header=F)
pcw.plot(x="platelet"     ,x_lab = bquote("Platelet(x10"^"9"*"/L)"), header=F)
pcw.plot(x="basophils"    ,x_lab = bquote("Basophils(x10"^"9"*"/L)"), header=F)
# Save as Figure 6.2

#=============================================================================#
# Figure 6.3: Desirable effects (in at least one phase)                       #
#=============================================================================#
par(mfrow=c(4,4),mar = c(3.2,3.2,1.5,1),cex=0.7)
pcw.plot(x="calcium"      , x_lab = "Calcium(mmol/L)", header=T)
pcw.plot(x="red_cells"    , x_lab = bquote("RBC(x10"^"6"*"cells/mm"^"3"*")"), header=F)
pcw.plot(x="alkaline phos", x_lab = "ALP(IU/L)", header=F)
# Save as Figure 6.3

#=============================================================================#
# Figure 6.4: Alternating desirable and undesirable effects                   #
#=============================================================================#

par(mfrow=c(4,4),mar = c(3.2,3.2,1.5,1),cex=0.7)
pcw.plot(x="lactate_dehyd", x_lab = "LDH(U/L)", header=T)
pcw.plot(x="mchc"         , x_lab = "MCHC(g/dL)", header=F)
# Save as Figure 6.4

#=============================================================================#
# Figure 6.5: Undesirable effects                                             #
#=============================================================================#
par(mfrow=c(4,4),mar = c(3.2,3.2,1.5,1),cex=0.7)
pcw.plot(x="folate"       , x_lab = "Folate(nmol/L)", header=T)
pcw.plot(x="protein"      , x_lab = "Total protein(g/L)", header=F)
pcw.plot(x="sodium"       , x_lab = "Sodium(mEq/L)", header=F)
pcw.plot(x="monocytes"    , x_lab = bquote("Monocytes(x10"^"9"*"/L)"), header=F)
# Save as Figure 6.5


#=============================================================================#
# Figure 6.6: Insignificant relationships                                     #
#=============================================================================#
par(mfrow=c(4,4),mar = c(3.2,3.2,1.5,1),cex=0.7)
pcw.plot(x="mcv"          , x_lab = "MCV(fL)", header=T)
pcw.plot(x="haematocrit"  , x_lab = "Haematocrit(Hct/100)", header=F)
pcw.plot(x="magnesium"    , x_lab = "Magnesium(mmol/L)", header=F)
pcw.plot(x="potassium"    , x_lab = "Potassium(mmol/L)", header=F)
# Save as Figure 6.6
```

C7: Chapter 7 codes

```
#############################################################################
#                                     #
# PREPARING LAGGED DATA               #         SEM
#                                     #
#######################################
rm(list= ls()[!(ls() %in% c(""))])
                                              # clear environment and leave the selected
cat("\014")                                   # clear console

load(file ="C:/Users/PARTSON/data2model.rda")

# shorten variable names
sem a = data2model
if(!require(data.table)){install.packages("data.table")}
colnames(sem_a)[colnames(sem_a)=="cd4"]            <- "cd4"
colnames(sem_a)[colnames(sem_a)=="red_cells"]      <- "red"
colnames(sem_a)[colnames(sem_a)=="haematocrit"]    <- "hae"
colnames(sem_a)[colnames(sem_a)=="mcv"]            <- "mcv"
colnames(sem_a)[colnames(sem_a)=="mchc"]           <- "mch"
colnames(sem_a)[colnames(sem_a)=="platelet"]       <- "pla"
colnames(sem_a)[colnames(sem_a)=="lymphocytes"]    <- "lym"
colnames(sem_a)[colnames(sem_a)=="monocytes"]      <- "mon"
colnames(sem_a)[colnames(sem_a)=="basophils"]      <- "bas"
colnames(sem_a)[colnames(sem_a)=="glucose"]        <- "glu"
colnames(sem_a)[colnames(sem_a)=="alkaline_phos"] <- "alk"
colnames(sem_a)[colnames(sem_a)=="calcium"]        <- "cal"
colnames(sem_a)[colnames(sem_a)=="magnesium"]      <- "mag"
colnames(sem_a)[colnames(sem_a)=="potassium"]      <- "pot"
colnames(sem_a)[colnames(sem_a)=="sodium"]         <- "sod"
colnames(sem_a)[colnames(sem_a)=="protein"]        <- "pro"
colnames(sem_a)[colnames(sem_a)=="albumin"]        <- "alb"
colnames(sem_a)[colnames(sem_a)=="lactate_dehyd"] <- "lac"
colnames(sem_a)[colnames(sem_a)=="folate"]         <- "fol"


if(!require(dplyr)){install.packages("dplyr")}
data_sem_a = sem_a  %>%
                    tbl_df %>%
        group_by(patientid,phase) %>%
        mutate(timeg =1:n()) %>%
            as.data.frame()

if(!require(data.table)){install.packages("data.table")}
data_sem_b = setDT(data_sem_a[,c(1,3,6,7:25)])
data_sem_c = dcast(data_sem_b, patientid+phase ~ timeg,sep = "_",
                   value.var = names(data_sem_b[,c(4:22)]))

semdata = data sem c
save(semdata,file ="C:/Users/PARTSON/semdata.rda")
write.csv(semdata,file ="C:/Users/PARTSON/semdata.csv",na = "",row.names = FALSE )

#################################
# Prepare data sets             #
#################################
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/semdata.rda")

semdata2 = subset(semdata,phase =="2-Acute")
semdata3 = subset(semdata,phase =="3-Early")
semdata4 = subset(semdata,phase =="4-Est")
semdata5 = subset(semdata,phase =="5-ART")

save(semdata2,file ="C:/Users/PARTSON/semdata2.rda")
write.csv(semdata2,file ="C:/Users/PARTSON/semdata2.csv",na = "",row.names = FALSE     )

save(semdata3,file ="C:/Users/PARTSON/semdata3.rda")
write.csv(semdata3,file ="C:/Users/PARTSON/semdata3.csv",na = "",row.names = FALSE     )

save(semdata4,file ="C:/Users/PARTSON/semdata4.rda")
write.csv(semdata4,file ="C:/Users/PARTSON/semdata4.csv",na = "",row.names = FALSE     )
```

```
save(semdata5,file ="C:/Users/PARTSON/semdata5.rda")
write.csv(semdata5,file ="C:/Users/PARTSON/semdata5.csv",na = "",row.names = FALSE     )

#############################################################
#                                                           #
# SEM MODELLING IN AMOS:                                    #
#     *save the the datasets in SPSS                        #
#     *use them in AMOS to obtain Figures 7.3 and 7.4       #
#     *save the regression weights, covs and corr as csv    #
#############################################################

               #
               #
               #
             #  #  #
              # # #
               #

#############################################################
#                                                           #
# Load csv datasets from AMOS and merge phases              #
#     * Diagnostics                                         #
#     * Correlations                                        #
#     * Regression weights                                  #
#############################################################

#===========================================================================#
# Table 7.1 : Manually obtained from Excel spreadsheets                     #
#===========================================================================#

# load standardised regression weights
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>") # customize prompt
if(!require(readr)){install.packages("readr")}
if(!require(dplyr)){install.packages("dplyr")}
if(!require(data.table)){install.packages("data.table")}

srw2_a <- read_csv("C:/Users/Partson/srw2.csv")%>%
        setnames(.,c("X1","X3"),c("var2","var1"))%>%
        mutate(phase =rep("2-Acute"))%>%
        subset(.,select=c(5,3,1,4))
srw3_a <- read_csv("C:/Users/Partson/srw3.csv")%>%
        setnames(.,c("X1","X3"),c("var2","var1"))%>%
        mutate(phase =rep("3-Early"))%>%
        subset(.,select=c(5,3,1,4))
srw4_a <- read_csv("C:/Users/Partson/srw4.csv")%>%
        setnames(.,c("X1","X3"),c("var2","var1"))%>%
        mutate(phase =rep("4-Est"))%>%
        subset(.,select=c(5,3,1,4))
srw5_a <- read_csv("C:/Users/Partson/srw5.csv")%>%
        setnames(.,c("X1","X3"),c("var2","var1"))%>%
        mutate(phase =rep("5-ART"))%>%
        subset(.,select=c(5,3,1,4))

# Merge all regression weights
srw_a = rbind(srw2_a,srw3_a,srw4_a,srw5_a)
srw_a$Estimate <- gsub(",." , ".",srw_a$Estimate)
srw_a$Estimate = as.numeric(srw_a$Estimate)
srw_a$path =paste(srw_a$var1,srw_a$var2, sep=" -> ") # combine columns

srw_b = srw_a
srw_b$path <- gsub("cd4","CD4",srw_b$path)
srw_b$path <- gsub("red","Red blood cells",srw_b$path)
srw_b$path <- gsub("hae","Haematocrit",srw_b$path)
srw_b$path <- gsub("mcv","MCV",srw_b$path)
srw_b$path <- gsub("mch","MCHC",srw_b$path)
srw_b$path <- gsub("pla","Platelet",srw_b$path)
srw_b$path <- gsub("lym","Lymphocytes",srw_b$path)
srw_b$path <- gsub("mon","Monocytes",srw_b$path)
srw_b$path <- gsub("bas","Basophils",srw_b$path)
srw_b$path <- gsub("glu","Glucose",srw_b$path)
srw_b$path <- gsub("alk","ALP",srw_b$path)
srw_b$path <- gsub("cal","Calcium",srw_b$path)
srw_b$path <- gsub("mag","Magnesium",srw_b$path)
srw_b$path <- gsub("pot","Potassium",srw_b$path)
srw_b$path <- gsub("sod","Sodium",srw_b$path)
srw_b$path <- gsub("pro","Total protein",srw_b$path)
srw_b$path <- gsub("alb","Albumin",srw_b$path)
```

```
srw_b$path <- gsub("lac","LDH",srw_b$path)
srw_b$path <- gsub("fol","Folate",srw_b$path)

srw = srw b
save(srw,file ="C:/Users/PARTSON/srw.rda")



# load correlations
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
if(!require(readr)){install.packages("readr")}
if(!require(dplyr)){install.packages("dplyr")}

cor2_a <- read_csv("C:/Users/Partson/cor2.csv")%>%
        setnames(.,c("X1","X3","P"),c("var1","var2","pvalue"))%>%
        mutate(phase =rep("2-Acute"))%>%
        subset(.,select=c(6,1:5))

cor3_a <- read_csv("C:/Users/Partson/cor3.csv")%>%
        setnames(.,c("X1","X3","P"),c("var1","var2","pvalue"))%>%
        mutate(phase =rep("3-Early"))%>%
        subset(.,select=c(6,1:5))

cor4_a <- read_csv("C:/Users/Partson/cor4.csv")%>%
        setnames(.,c("X1","X3","P"),c("var1","var2","pvalue"))%>%
        mutate(phase =rep("4-Est"))%>%
        subset(.,select=c(6,1:5))

cor5_a <- read_csv("C:/Users/Partson/cor5.csv")%>%
        setnames(.,c("X1","X3","P"),c("var1","var2","pvalue"))%>%
        mutate(phase =rep("5-ART"))%>%
        subset(.,select=c(6,1:5))

# Merge all regression weights
cor_a = rbind(cor2_a,cor3_a,cor4_a,cor5_a)
cor_a$Estimate <- gsub("," , ".",cor_a$Estimate)
cor a$Estimate = as.numeric(cor a$Estimate)
cor_a$corr =paste(cor_a$var1,cor_a$var2, sep=" <-> ") # combine columns
cor_a$pvalue <- gsub("," , ".",cor_a$pvalue)
cor_a$pvalue <- gsub("\\*\\*\\*",0.001,cor_a$pvalue) # \\ forces * one at a time
cor a$pvalue = as.numeric(cor a$pvalue)
cor_b = cor_a[,c(1,2,4,7,5,6)]
cor_b$sig =
          ifelse(cor_b$pvalue <0.05 ,"Sig.","Insig.")
cor = cor_b
save(cor,file ="C:/Users/PARTSON/cor.rda")

#
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/cor.rda")

cor_c = cor

cor_c$var_a = cor_c$var1
cor_c$var_b = cor_c$var2

if(!require(splitstackshape)){install.packages("splitstackshape")}
cor_d = cSplit(cor_c, c("var1","var2"), "_")

if(!require(data.table)){install.packages("data.table")}
setnames(cor_d,c("var1_2"   ,"var2_2"   , "var1_1", "var2_1"),
                 c("var1_time","var2_time", "var1"  , "var2"))
cor_e = cor_d[,-c(6,7)]

if(!require(data.table)){install.packages("data.table")}
cor_e$corr <- gsub("_","",cor_e$corr)

cor_e$corr <- gsub("cd4","CD4_",cor_e$corr)
cor_e$corr <- gsub("red","Red blood cells_",cor_e$corr)
cor_e$corr <- gsub("hae","Haematocrit_",cor_e$corr)
cor_e$corr <- gsub("mcv","MCV_",cor_e$corr)
cor_e$corr <- gsub("mch","MCHC_",cor_e$corr)
cor_e$corr <- gsub("pla","Platelet_",cor_e$corr)
cor_e$corr <- gsub("lym","Lymphocytes_",cor_e$corr)
cor_e$corr <- gsub("mon","Monocytes_",cor_e$corr)
```

```
cor_e$corr <- gsub("bas","Basophils_",cor_e$corr)
cor_e$corr <- gsub("glu","Glucose_",cor_e$corr)
cor_e$corr <- gsub("alk","ALP_",cor_e$corr)
cor_e$corr <- gsub("cal","Calcium_",cor_e$corr)
cor_e$corr <- gsub("mag","Magnesium_",cor_e$corr)
cor_e$corr <- gsub("pot","Potassium_",cor_e$corr)
cor_e$corr <- gsub("sod","Sodium_",cor_e$corr)
cor_e$corr <- gsub("pro","Total protein_",cor_e$corr)
cor_e$corr <- gsub("alb","Albumin_",cor_e$corr)
cor_e$corr <- gsub("lac","LDH_",cor_e$corr)
cor_e$corr <- gsub("fol","Folate_",cor_e$corr)

cordata = cor_e
save(cordata,file="C:/Users/Partson/cordata.rda")


##############################################
# Plotting of the SEM AMOS results           #
##############################################

#-----------------------------------------------
# PARALLEL COORDINATE PLOTS OF REGRESSION WEIGHTS
#-----------------------------------------------

# Standardised regression weights
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/srw.rda")

data srw a = srw
if(!require(data.table)){install.packages("data.table")}
data srw b = setDT(data srw a[,])
data_srw_c = dcast(data_srw_b, phase ~ path,
                   value.var = names(data_srw_b[,c(4)]))

data srw = data_srw_c
dim(data_srw)
if(!require(lattice)){install.packages("lattice")}


#=============================================================================#
# Figure 7.5                                                                  #
#=============================================================================#
parallelplot(~data_srw[,c(2:23,28:51)] | factor(phase), #2:23,28:51 #zeros 24:27,52:55
        data_srw,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="Standardised regression weights",
        ylab="Regression paths",
        main=""
        )
#=============================================================================#
# Figure 7.6                                                                  #
#=============================================================================#
parallelplot(~data_srw[,c(56:93)] | factor(phase), # 56:93  #zeros 24:27,52:55
        data_srw,
        groups = phase,
        scales=list(cex=.7),
        layout = c(4, 1),
        xlab="Standardised regression weights",
        ylab="Regression paths",
        main=""
        )

#-----------------------------------------------------------------=
# Multiple bar charts of Standardised regression weights > 0.40
#-----------------------------------------------------------------=
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file ="C:/Users/PARTSON/srw.rda")
```

```
format_plot_srw=(
      theme_bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element_text(size=9,angle=90,color="black", hjust=1, vjust=0.25),
    axis.title.y = element_text(color="black", size=10, face="bold"),
    axis.text.y = element_text(size=8,angle=0,color="black", hjust=0.5,vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

srw_prep = subset(srw, abs(Estimate)>.40) %>% arrange(Estimate)
#============================================================================#
# Figure 7.7                                                                 #
#============================================================================#
ggplot(srw_prep[,] , aes(fill=phase, y=Estimate, x=path)) +
    geom_bar(aes(fill = factor(phase)),width=.75,position="dodge", stat="identity")+
    scale_fill_manual(values=c("blue","green1","yellow","red"),name="Phase:")+
    coord_flip()+
    labs( x="Regression paths (  Time)",
              y='Standardised regression weights > 0.40' ,
              title=""
             )+ format_plot_srw


#-------------------------------------------------------------=
# Multiple bar charts of correlations > 0.40
#-------------------------------------------------------------=
rm(list= ls()[!(ls() %in% c(""))]) ; cat("\014") ; options(prompt = "R>")              #
customize prompt
load(file="C:/Users/Partson/cordata.rda")

if(!require(ggplot2)){install.packages("ggplot2")}
if(!require(scales)){install.packages("scales")}
if(!require(Rcmdr)){install.packages("Rcmdr")}


format_plot_cor=(
      theme_bw()   +        # Background fill
      theme(                # Format all the items on the graph
    plot.title = element_text(color="black", size=9, face="bold",hjust = 0.5),
    axis.title.x = element_text(color="black", size=10, face="bold",angle=0),
    axis.text.x = element_text(size=9,angle=90,color="black", hjust=1, vjust=0.25),
    axis.title.y = element_text(color="black", size=10, face="bold"),
    axis.text.y = element_text(size=8,angle=0,color="black", hjust=0.5,vjust=0.25),
    legend.position="top",
    legend.background = element_rect(fill="bisque",size=0.2,linetype="solid",colour
="lightblue"),
    legend.title = element_text(colour="black",size=9,face="bold"),
    legend.text = element_text(colour="black",size=8,face="plain")
     ))

if(!require(dplyr)){install.packages("dplyr")}
cordata$var1 = as.character(cordata$var1)
cordata$var2 = as.character(cordata$var2)
#============================================================================#
# Figure 7.8: correlations of the same variable                              #
#============================================================================#
corsame_prep = subset(cordata, abs(Estimate)>.40 &
                        var1==var2   &
                        sig=="Sig."
                       ) %>% arrange(Estimate)
ggplot(corsame_prep[,] , aes(fill=phase, y=Estimate, x=corr)) +
    geom_bar(aes(fill = factor(phase)),width=.75,position="dodge", stat="identity")+
    scale_fill_manual(values=c("blue","green1","yellow","red"),name="Phase:")+
    coord_flip()+
    labs( x="Same covariate at different time points",
              y='Significant correlations > 0.40' ,
              title=""
             )+ format_plot_cor
```

```
#=============================================================================#
# Figure 7.9: correlations of different variables                             #
#=============================================================================#
cordiff_prep = subset(cordata,
                      var1!=var2   &
                      sig=="Sig."
                     ) %>% arrange(Estimate)
ggplot(cordiff_prep[,] , aes(fill=phase, y=Estimate, x=corr)) +
    geom_bar(aes(fill = factor(phase)),width=.75,position="dodge", stat="identity")+
    scale_fill_manual(values=c("blue","green1","yellow","red"),name="Phase:")+
    coord_flip()+
    labs( x="Different covariates at different time points",
              y='Significant correlations' ,
              title=""
             )+ format_plot_cor
```