University of KwaZulu-Natal

Statistical methods for longitudinal binary data structure with applications to antiretroviral medication adherence

By

DIKOKOLE MAQUTU

2010

Statistical methods for longitudinal binary data structure with applications to antiretroviral medication adherence

By

DIKOKOLE MAQUTU

Submitted in fulfilment of the academic requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Applied Statistics

in the

School of Statistics and Actuarial Science

University of KwaZulu-Natal

Pietermaritzburg

2010

Dedication

To my husband, Mpaiphele Maqutu
and our loving children, Thembiso and Lungile

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Sciences, University of Kwazulu-Natal, Pietermaritzburg, under the supervision of Prof. Temesgen Zewotir.

I, Dikokole Maqutu, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

December, 2010.

_____ _____

Mrs. Dikokole Maqutu Date

_____ _____

Prof. Temesgen Zewotir Date

Acknowledgements

This work could not have been accomplished without assistance of a number of people, and I would like to take this opportunity to thank them for their help. First and foremost, I would like to express my deep and sincere gratitude to my supervisor, Professor Temesgen Zewotir, whose guidance and stewardship anchored me through to the very end. He has taught me both consciously and unconsciously how sound application of statistics can contribute to scientific knowledge in other disciplines. I appreciate all his contributions of time and ideas, which made my PhD experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational to me, even during tough times in the PhD pursuit.

I wish to express my warm and sincere thanks to Professor Delia North, whose contribution to this work has been valuable. Her encouragement and kind support has been of great value. I would also like to thank Professor Henry Mwambi for reading the first draft of this thesis.

Many thanks to CAPRISA for the pre-doctoral fellowship and more importantly the data, without which this work may not have seen the light of day. Being introduced to the handling of large data sets by the CAPRISA Statistics Department proved invaluable. Special thanks to my mentor, Anneke Grobler who has been very supportive throughout my fellowship at CAPRISA.

To friends and colleagues at UKZN and beyond, a very big thank you to you all.

This vote of thanks would not be complete without paying a special tribute to my entire family for their unrelenting support. Key to my sanity was my husband and two children, who kept the family unit intact, thus leaving me to focus squarely on my research work.

Abstract

Longitudinal data tend to be correlated and hence posing a challenge in the analysis since the correlation has to be accounted for to obtain valid inference. We study various statistical methods for such correlated longitudinal binary responses. These models can be grouped into five model families, namely, marginal, subject-specific, transition, joint and semi-parametric models. Each one of the models has its own strengths and weaknesses. Application of these models is carried out by analyzing data on patient's adherence status to highly active antiretroviral therapy (HAART). One other complicating issue with the HAART adherence data is missingness. Although some of the models are flexible in handling missing data, they make certain assumptions about missing data mechanisms, the most restrictive being missing completely at random (MCAR). The test for MCAR revealed that dropout did not depend on the previous outcome.

A logistic regression model was used to identify predictors for the patients' first month's adherence status. A marginal model was then fitted using generalized estimating equations (GEE) to identify predictors of long-term HAART adherence. This provided marginal population-based estimates, which are important for public health perspective. We further explored the subject's specific effects that are unique to a particular individual by fitting a generalized linear mixed model (GLMM). The GLMM was also used to assess the association structure of the data. To assess whether the current optimal adherence status of a patient depended on the previous adherence measurements (history) in addition to the explanatory variables, a transition model was fitted. Moreover, a joint modeling approach was used to investigate the joint effect of the predictor variables on both HAART adherence status of patients and duration between successive visits. Assessing the association between the two outcomes was also of interest. Furthermore, longitudinal trajectories of observed data may be very complex especially when dealing with practical applications and as such, parametric statistical models may not be flexible enough to capture the main features of the longitudinal profiles, and so a semi-parametric approach was adopted. Specifically, generalized additive mixed models were used to model the effect of time as well as interactions associated with time non-parametrically.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Highly active antiretroviral therapy (HAART) has dramatically reduced morbidity and mortality among HIV-infected individuals (Kalichman, Ramachandran & Catz, 1999; Berg et al., 2004; Gill et al., 2005), and requires strict adherence to attain optimal clinical and survival benefits (Paterson et al., 2000). Patients who take 95% or more of their prescribed medication benefit more from treatment than those who take less than 95% (Paterson et al., 2000). Identifying and overcoming factors that reduce adherence to HAART is therefore critical if optimal clinical and survival benefits are to be attained. Optimal adherence to HAART is often influenced by a variety of factors, comprising of social, demographic, economic and behavioural issues (Chesney, 2000; Ferguson et al., 2002). There have however been inconsistent findings regarding the association between adherence, demographic and economic factors. For instance, some studies have shown an association between adherence and age (Penedo et al., 2003; Murphy et al., 2004), while others did not find such a relationship (Roca, Lapuebla & Vidal-Tegedor, 2005). Income has also been associated with HAART adherence in some studies (Kleeberger et al., 2001; Laniece et al., 2003) while others have concluded that there was no link between the adherence and income (Mohammed et al., 2004). Social factors that include family support have been reported to have an association with adherence (Ammassari et al., 2002). The fact that an association between adherence and demographic and economic factors are observed inconsistently, highlights the need for an evaluation and understanding of how these factors interact among themselves, and how they interact with other social, clinical and behavioural factors.

Moreover, adherence to medication generally is problematic when therapeutic regimens are employed for prolonged periods (Godin et al., 2005; Mehta et al., 1997). Adherence to antiretroviral medication is no exception; patients are required to take several pills each day for an undefined period of time (Godin et al., 2005). Therefore, not only high levels of adherence at a point in time are required, but also sustained adherence to antiretroviral medication is critical. This necessitates evaluating determinants of optimal HAART adherence over time, which gives rise to

a longitudinal study, in this case with a binary response defined by whether a patient is or is not optimally adherent to medication.

In a longitudinal study, individuals are observed over a period of time, and for each individual, data are collected at multiple time points. That is, the defining feature of a longitudinal study is that multiple or repeated measurements of the same variables are made for each individual in the study over a period of time. For example, with the HAART adherence study, information related to each patient's adherence is recorded at every (monthly) visit to the clinic until the patient ceases to visit the clinic. A key characteristic of longitudinal data is that observations within the same individual may be correlated, and this motivates most of the statistical methods for the analysis of longitudinal data (Diggle, et al., 2002; Fitzmaurice, Laird and Ware, 2004). Note that longitudinal data sets differ from time series data sets in that longitudinal data usually consists of a large number of short series of time points whereas time series data sets consist of a single, long series of time points (Diggle, et al., 2002). Longitudinal studies allow the direct study of change over time and the factors that influence this change, as well as assessing within-subject changes (Lindsey, 1999; Twisk, 2003; Fitzmaurice et al., 2004). Moreover, longitudinal data can also provide information about individual change. Statistical estimates of individual trends can be used to better understand heterogeneity in the population, the determinants of growth and change at the individual level. Furthermore, in some longitudinal studies, although one time-varying outcome may be of primary interest, several related processes may also be measured, and the association between a primary outcome and another related outcome can reveal a great deal of insight about the mechanism of behavioural change.

Despite the strengths of a longitudinal study, there are challenges in the analysis that need to be addressed accordingly. The set of measurements on one subject tends to be correlated, measurements on the same subject close in time tend to be more highly correlated than measurements far apart in time, and the variances of longitudinal data often change with time (Diggle et al., 2002; Fitzmaurice et al., 2004). These potential patterns of correlation and variation may combine to produce a complicated covariance structure. Accordingly, this covariance structure must be taken into account in order to draw reliable conclusions from the data Thus, unlike

in the classical setting where there exists a single source of variation between subjects, the heterogeneity between clusters introduces an additional source of variation and complicates the analysis. That is, the analysis is complicated by the presence of the within subject correlation among the repeated observations on the same subject. Therefore, standard regression models may produce invalid results because two of the parametric assumptions (independent observation and equal variances) may not be valid. More complex statistical models have to be used to account for the dependence in the data. This leads to parameter estimation that can be computationally intensive. Sometimes there is a lack of available computer software for the application of these more complex statistical models, or the level of statistical sophistication required of the user is beyond the typical level of the practitioner (Hedeker and Gibbons, 2006).

Note that if observations are positively correlated, which often occurs with longitudinal data (Crowder and Hand, 1990; Davis, 2002), then variances of time independent variables (variables that estimate group effect or between-subject effect) are underestimated if the data are analyzed as though the observations are independent. That is the Type I error rate (rejecting the null hypothesis which is true, i.e. a false positive) is inflated for these variables (Dunlop, 1994). For time-dependent predictor variables (variables that measure the time effect or within-subject effect), ignoring positive correlation leads to a variance estimate that is too large. That is, the Type II error rate (failing to reject the null hypothesis when it is false, i.e. a false negative) is inflated for these variables (Dunlop, 1994). Because the variances of the group effects will be underestimated and the variance of the time effects will be overestimated if positive correlation is ignored, it is again evident that correlated outcomes must be accounted for to obtain valid analyses.

Missing data are a common problem in longitudinal studies. Study participants do not always appear for a scheduled observation or simply leave the study before its completion. When some observations are missing, data are necessarily unbalanced over time since not all individuals have the same number of repeated measurements obtained at a common set of occasions. One of the consequences of lack of balance and/or missing data is that it requires some care to recover within-individual change (Fitzmaurice et al., 2004). For instance, when data are missing, especially

when there is attrition of subjects whose responses are different from those who remain in the study, then the mean response over time can be misleading; changes over time may reflect the pattern of missingness or attrition, and not within-individual change. Thus with missing data, one will need to examine assumptions about the reasons for missingness and the appropriateness of the analysis to determine the validity inferences. Also the data may be unbalanced due to mistimed measurements, as a result, models used for analysis must be able to handle data which are unbalanced.

One other thing to consider is that the longitudinal trajectories of observed data may be very complex especially when dealing with practical applications. Although the parametric regression models provide a powerful tool for modeling the relationship between a response variable and the covariates in longitudinal studies, they suffer from inflexibility in modeling complicated relationships between the response and covariates in various applications of practical longitudinal data. Consequently, semi-parametric or nonparametric statistical models become an attractive alternative for many applications. However, the presence of the within-subject correlation among repeated measures over time presents major challenges in developing nonparametric regression methods for longitudinal data.

Although many approaches to the analysis of longitudinal data have been studied, most are restricted to the setting in which the response variable is normally distributed. Methods for continuous normal data are the best developed and the linear mixed model (Laird and Ware, 1982) has played a prominent role in extending the general linear model to handle correlated continuous data. Many of the earlier linear models for analysis of longitudinal data were based on the analysis of variance (ANOVA) techniques (Fitzmaurice et al., 2004; Longford, 1993). These included the univariate 'Mixed Model' ANOVA (univariate repeated-measures ANOVA) and the Multivariate Analysis of Variance (MANOVA) for repeated measures. These methods had drawbacks that restricted their effectiveness in many applications. The univariate repeated measures ANOVA assumes a compound symmetry form for the covariance structure, i.e. that variances are constant across time and the correlation between any pair of measurements is the same regardless of the time interval between measurements. Repeated measures ANOVA also assume sphericity,

which is a more general form of compound symmetry. It relates to the equality of the variances of the differences between the levels of the repeated measures factor (Anderson, 1958) and Mauchly's test is commonly used to test this assumption. Moreover, measurements should be made at a common set of occasions for all individuals, all covariates must be discrete factors and data must be complete. Even though the repeated MANOVA does not make restrictive assumptions on the covariance among the longitudinal responses on the same individual, it cannot be used when the design is unbalanced over time (Hedeker and Gibbons, 2006). Both these procedures focus on the estimation of group trends over time but provide little help in understanding how specific individuals change over time (Hedeker and Gibbons, 2006).

Due to the above mentioned limitations and other reasons, linear mixed models have recently provided an alternative for analysis of longitudinal data. With these models, the inclusion of random subject effects in the model account for the influence on individuals on their repeated observations. In addition, linear mixed models are even more appealing in the analysis of longitudinal data because individuals are not assumed to be measured on the same number of time-points; as a result, individuals with incomplete data across time are included in the analysis. Again, since time is treated as a continuous variable, individuals do not have to be measured at the same time-points. Both time-invariant and time-varying covariates can be included in the model (Hedeker and Gibbons, 2006). Owing to the elegant properties of the multivariate normal distribution, linear mixed models' theory and implementation are greatly simplified. Software programs, such as SAS use the procedure MIXED (Littell et al., 2006) to fit this kind of model.

One other attractive feature of mixed models is that they have a close connection with smoothing splines. Because linear mixed models are so well developed, researchers have exploited the connection with splines by incorporating a nonparametric time function in linear mixed models (Zeger and Diggle, 1994; Zhang et al., 1998; Verbyla et al., 1999) to capture more complex longitudinal trajectories.

In contrast, when the longitudinal response is discrete (e.g. binary), the issue that arises is the lack of discrete analogue to the multivariate normal distribution.

Complete specification of the joint distribution of the response vector becomes more problematic and fully likelihood-based methods are generally awkward. Also modeling time non-parametrically for non-Gaussian outcomes has been a challenge. Consequently, the main of objective of this study is to critically assess the binary longitudinal methodologies and apply these methods to identify the causal factors for optimal HAART adherence among HIV positive adults who are on treatment. Specifically, the study seeks to survey statistical methods for longitudinal binary responses with a view to highlighting issues that arise at implementation and employ these methods to determine predictors of long-term optimal HAART adherence, and assess whether factors affecting baseline adherence also influence long-term HAART adherence.

Since many longitudinal measurements in health sciences and other fields that include social sciences are discrete (binary, count etc), and unsuitable for linear modeling, the results from this study can assist data analysts in choosing alternative appropriate models that are within reach. Moreover, the use of most statistical models is usually illustrated using simple and 'convenient' datasets, e.g. balanced data with a small number of subjects and a few covariates. In practice, researchers face 'messy' data that have a large number of subjects together with a large number of covariates (to control for potential confounders) in addition to unbalancedness, irregularly spaced observations and missing values. In such instances, the choice of the statistical procedure as well as its implementation may be far from obvious. Demonstrating analyses and implementation of the longitudinal statistical methods for binary data using practical data with all features mentioned will highlight important issues that a data analyst should be aware off. To this end, the results of our study can help clarify the substantive questions which data analysts can address with each approach while at the same time being fully conscious of the merits and limitations of each method of analysis.

With application of the surveyed statistical methods for analysis of longitudinal binary data, predictors of HAART adherence will be identified. The knowledge and understanding of such factors is particularly important if increased enrolments to treatment are to be maintained in therapy, as well as aiding in the avoidance of starting patients with the 2nd line regimen, which is not affordable to governments in

developing countries. In addition, the findings will be useful in developing tools to assist clinicians in the identification of factors related to poor adherence prior to initiating treatment as well as during therapy. That is, HAART programmes would need to take the identified factors into account in the design and implementation of short- and long-term adherence strategies. The findings can also help identify specific groups or sub-populations of patients at risk of less than optimal adherence and they can be targeted with long-term adherence boosting sessions.

The rest of the thesis is organized as follows. In Chapter 2, we give a full description of the HAART adherence data. We further undertake the exploratory work where adherence trend and data dependence are assessed. Moreover, the issue of missing data is dealt with in detail. Chapter 3 presents the exploration of factors affecting initial optimal HAART adherence using the generalized linear models. In Chapter 4, we present marginal models for longitudinal data where the emphasis is on the generalized estimating equations (GEEs) method and application of this method to optimal HAART adherence. Chapter 5 presents a comprehensive review of the generalized linear mixed models (GLMMs) as a special case of random effects models. Moreover, GLMMs are fitted to the HAART adherence data to explore subject specific effects of long-term optimal adherence. In Chapter 6, a review and fitting of transition models to HAART adherence data is presented. Review and fitting of joint modeling of adherence and visit interval is presented in Chapter 7. Chapter 8 presents the semi-parametric approaches to longitudinal data, specifically generalized additive mixed models (GAMMs). Finally, in Chapter 9, the discussions and conclusions as well as implications and avenues for future research are presented.

Chapter 2

The data

The data used in this study are secondary data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA). CAPRISA started a HAART rollout programme in 2004. The CAPRISA AIDS Treatment (CAT) Programme offers HIV care services at two sites in KwaZulu-Natal, South Africa, namely the eThekwini Clinical Research site located adjacent to the Prince Cyril Zulu Communicable Disease Clinic in the center of Durban, and the Vulindlela Clinical Research site, located in a rural area outside the town of Howick, approximately 95 km from Durban. The programme started providing free HAART through a President's Emergency Plan for AIDS Relief (PEPfAR) grant at a time when access to HAART in the public sector was limited. Adult patients with a CD4+ count below 200 cells/μL, or patients with World Health Organisation (WHO) stage 4 of the HIV disease, were eligible for HAART initiation.

During the first month, patients visited the clinic once a week for the first two weeks and again two weeks later for intensive clinical monitoring. Thereafter, patients visited the treatment sites monthly to collect their treatment and to undergo a clinical examination. Prior to HAART initiation, all patients received three sessions of adherence education, motivation and preparedness training. All patients were on regimens containing two nucleoside reverse transcriptase inhibitors and one non-nucleoside reverse transcriptase inhibitor. Patients in the urban clinic received Efavirenz (EFV), Lamivudine (3TC) and Didanosine (ddI or ddI-EC). This regimen was chosen as it can be co-administered with anti-tuberculosis (TB) medication. The regimen in the rural clinic consisted of EFV, 3TC and Stavudine (d_4T), which is recommended according to the South African HIV treatment guidelines (South African National Department of Health, 2004). A few pregnant patients (3.8%) received Nevirapine (NVP) rather than EFV.

Patient information was recorded on data collection sheets at the clinics; it underwent two levels of quality control, and was faxed to a central data management

centre. Approval for the data collection and analysis was obtained from the University of KwaZulu-Natal Biomedical Research Ethics Committee.

The variables used for analyses in this study include optimal HAART adherence status of the patients, which is the outcome variable. Optimal adherence has been measured using the pill counts data. Adherence at every visit for all the drugs is calculated as the total number of drugs dispensed, minus the total number of drugs returned, divided by the total number of days between clinic visits, times the daily dose. Patients are then classified as optimally adherent if they took at least 95% of the prescribed drugs in a given regimen (Paterson et al., 2000), otherwise they are considered to be non-adherent. At each visit, the response variable is binary, indicating whether a patient is optimally adherent or not.

Independent covariates comprise of baseline demographic and socio-economic variables that include age (in years); gender (1 = female, 0 = male); educational status (2 = no schooling, 1 = primary and 0 = secondary and higher); treatment site (1 = urban, 0 = rural); whether or not a patient lived with a partner (1 = living with a partner, 0 = not living with a partner); whether or not the patient was the source of household income (1 = source of household income, 0 = not a source of household income); access to tap water (1 = yes, 0 = no) and electricity (1 = yes, 0 = no), and whether a patient owned a cell phone (1 = yes, 0 = no). Other variables recorded at baseline and included in the analysis were World Health Organisation (WHO) HIV stages (3 = stage 1, 2 = stage 2, 1 = stage 3 and 0 = stage 4), CD4+ cell count (cells/μL), and patient's weight (in kilograms). Patients were asked why they did an HIV test and their responses included being unwell, testing for no specific reason, testing because a partner died of HIV, being ill and unfaithfulness. Reason for testing was therefore classified as follows: (2 = possible exposure to HIV, 1 = no specific reason and 0 = unwell).

Since adherence has been monitored very closely after initiation of HAART, compared to subsequent visits, the baseline optimal adherence (first month optimal adherence) was treated as a covariate in the longitudinal analyses. One of the advantages of including a baseline response as a covariate in a longitudinal study is that it permits the use of each subject as their own control to assess the effect of

treatment over time (Milliken and Johnson, 2002; Littell et al., 2006). Consequently, baseline adherence was classified as follows: (1 = optimally adherent at baseline; 0 = not optimally adherent at baseline).

Time-varying covariates included time, which was measured as a continuous variable representing monthly follow-up visits to the treatment site. The variable time starts with the value 1 for the first follow-up visit, 2 for the second visit, up to 17 for the seventeenth follow-up visit. Weight (kg) was measured at every follow-up visit and was thus modeled as a time-varying covariate.

2.1. Baseline characteristics of study population

The data analyzed consisted of a retrospective review of patients' records in the CAT programme between June 2004 and September 2006. Only patients with pill count data for the initial visit, and at least one other clinic visit for the defined study period, were included in the analysis. During the said period, 1,184 patients were enrolled in the CAT programme, 411 (35%) at the urban site, and 773 (65%) at the rural site. A total of 688 patients, 369 (54%) from the urban site and 319 (46%) from the rural site were included in the analysis. There were no differences between those included in the study and those excluded with regard to age (mean: included=34.1 years, excluded=34.0 years; t-value=0.13, p=0.90), gender (males: included=30.0%, excluded=31.8%; chi-square-value = 0.43, p=0.51) and baseline CD4+ cell count (mean: included=107.6 cells/μL, excluded=111.5 cells/μL; t-value=0.72, p=0.47). Furthermore, power calculations were performed for the available sample size (688 patients) to detect a difference in proportion of adherent patients between the first follow-up visit and the final follow-up visit of 0.27. With a sample size of 600, this gave more than 90% power when a test of proportions was done. The baseline socio-demographic and clinical characteristics of patients included in the analysis are presented in Table 2.1.

Table 2. 1: Baseline socio-demographic and clinical characteristics of the HAART patients (*n* = 688)

Characteristic	Median (Q1-Q3)	*n* (%)
Age (years)	32.5 (28–38)	
Gender:		
Men		206 (30%)
Women		482 (70%)
Education:		
No schooling		74 (12%)
Primary school		116 (19%)
Secondary school or higher		429 (69%)
Treatment site:		
Urban		369 (54%)
Rural		319 (46%)
Living with or without a partner:		
Living with a partner		168 (25%)
Living without a partner		510 (75%)
Contribution to household income:		
Source of income		186 (28%)
Not source of income		489 (72%)
WHO stage of HIV disease:		
Stage 1		71 (10%)
Stage 2		121 (16%)
Stage 3		438 (64%)
Stage 4		58 (8%)
Baseline CD4+ count (cells/μL)	108 (52–159)	
Baseline weight (kg)	60 (53–69)	
Reason for taking HIV test:		
Unwell		374 (56%)
No specific reason		170 (26%)
Possible exposure to HIV		121 (18%)
Household access to tap water:		
Yes		611 (91%)
No		59 (9%)
Household access to electricity:		
Yes		607 (91%)
No		63 (9%)
Cell phone ownership:		
Yes		281 (42%)
No		389 (58%)
First-month optimal HAART adherence:		
Optimally adherent		546 (79%)
Not optimally adherent		142 (21%)

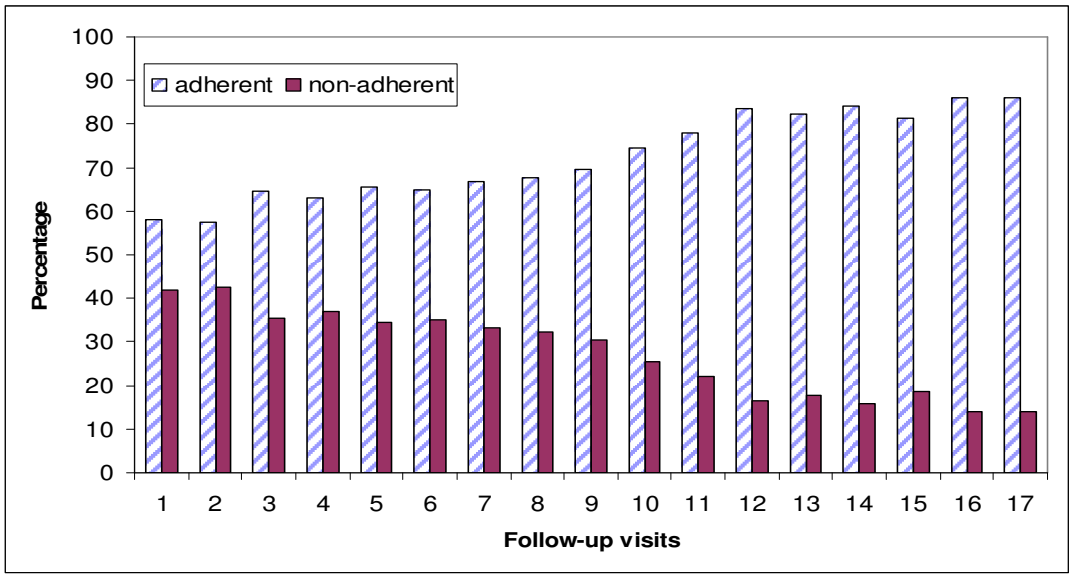
Table 2.1 shows that the median age of patients was 32.5 years (lower (Q1) and upper (Q3) quartiles were 28 and 38 years respectively), 30% were male and 75% were not living with a partner. Over two thirds of the patients had attained secondary or higher level of education (69%), and 28% of patients were classified as sources of their household income. Over 90% of the patients stayed in households that had access to tap water and electricity, while 42% of the households had cell phones. At

enrolment, the median weight was 60kg (Q1 and Q3 were 53 and 69kgs respectively), median CD4+ cell count was 108 cells/μL (Q1 and Q3 were 52 and 159 cells/ μL respectively) and 64% of patients were classified as WHO stage 3. Over half of the patients (56%) reported to have taken an HIV test as they were not well, while 26% reported no specific reason for testing and 18% took an HIV test as they were concerned that they had been exposed to HIV. In the initial month of treatment, 79% of the patients were at least 95% adherent to HAART.

2.2 Exploring HAART adherence over time

The optimal HAART adherence and non-adherence rates over the follow-up visits are presented in Figure 2.1.

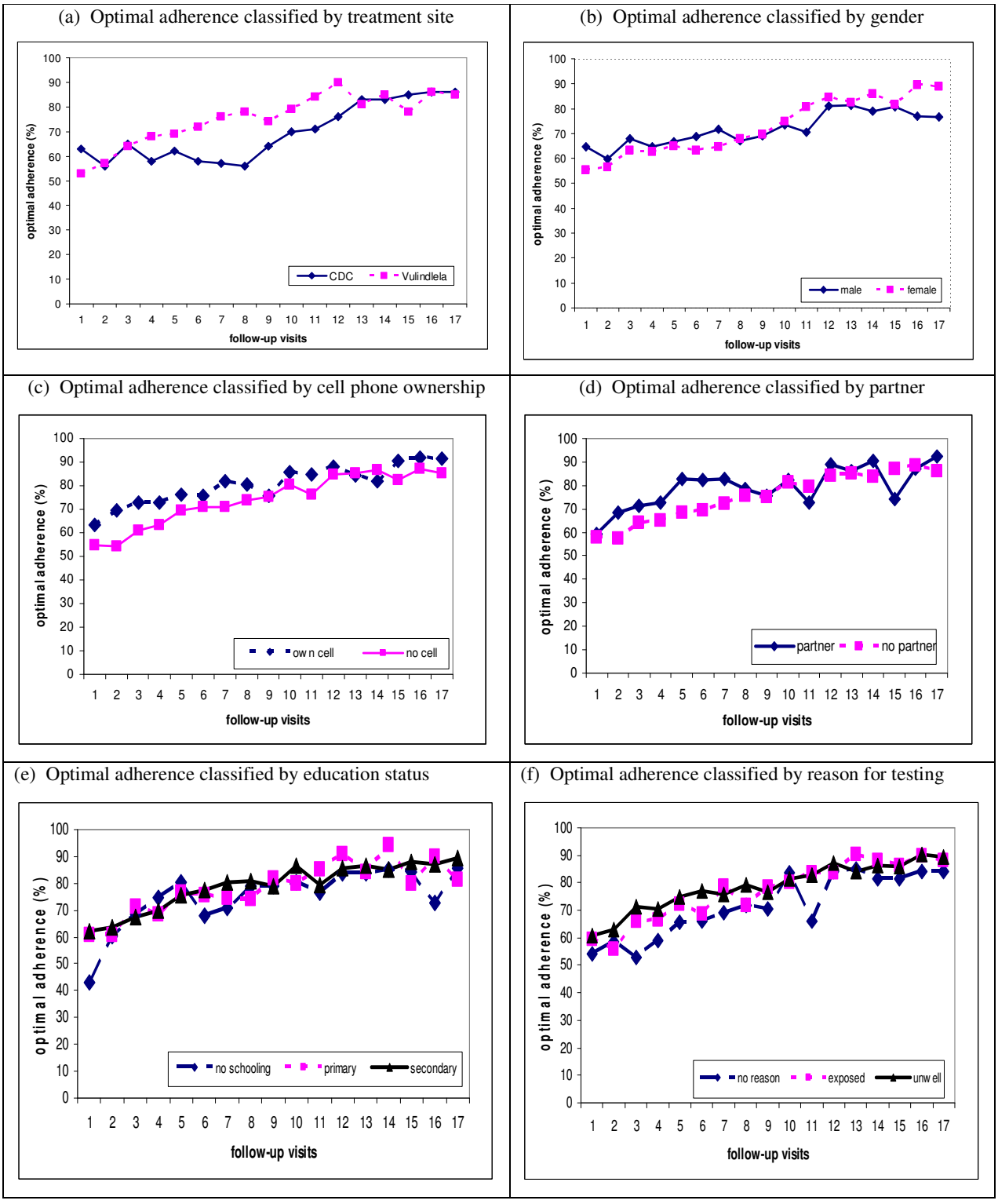
Figure 2. 1: Optimal HAART adherence and non-adherence rates over the follow-up visits



It is indicated that in general, optimal adherence has been increasing over the follow-up visits. To this end, Figure 2.1 shows that the proportion of patients who were at least 95% adherent (optimally adherent) to HAART increased from 58% at the first follow-up visit to 86% at the last follow-up visit. This observed trend of optimal adherence over the follow-up visits was tested using the Cochran-Armitage test for trend (Agresti, 2002) and the results provide a strong evidence of an increasing optimal HAART adherence rate over time (Z-statistic=17.52 with p-value<0.0001).

Moreover, Figure 2.2 presents optimal adherence over the follow-up visits classified by some of the explanatory variables that include treatment site, gender, cell phone ownership, living with/without a partner as well as educational attainment and reasons for taking an HIV test. We again observe an increasing trend in optimal adherence over time although the rate of increase seems to differ for the different groups.

Figure 2. 2: Optimal adherence over time classified by some of the explanatory variables



2.3. Exploring the dependence of HAART adherence data

As was mentioned earlier, in a longitudinal study, repeated responses are obtained on the same individuals over time; as a result, responses from the same individual tend to be correlated. Although in most longitudinal studies, the main interest is in the changes in the mean response over time and how these changes depend upon covariates, correlation among the repeated responses cannot be ignored. Failure to adequately account for the correlation among repeated measures leads to misleading inferences. For instance, suppose we are to estimate the change in the mean response over time considering only two responses from the same individual. The estimate of the change in the mean response over time is given by

$$\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1,$$

where

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij}\,.$$

In order to obtain the standard errors, we need to estimate the variability of this estimate of change in the mean response and is given by

$$var\left(\hat{\delta}\right) = var\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_{i2} - y_{i1}\right)\right\} = \frac{1}{N}\left(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}\right).$$

The inclusion of the last term accounts for the correlation among the two repeated measures. If we assume that the two repeated measures are uncorrelated, when in fact, there is a strong positive correlation among them, we would obtain incorrect estimates of the variance. This would lead to overestimation of the variability of the estimate of change in the mean response. This shows that in general, failure to account for the correlation among repeated measures, specifically for contrasts, leads to incorrect standard errors, that is, standard errors that are too large. With incorrect standard errors, test statistics and p-values will also be incorrect, which leads to incorrect inferences about the regression parameters (Fitzmaurice et al., 2004; Weiss, 2005). Thus, accounting for the correlation among the repeated measures increases the efficiency or precision with which the regression parameters can be estimated. It is therefore important to model the covariance structure as accurately as possible. Accordingly, exploring dependence among outcomes in the same subject might be a useful guide for selecting an appropriate covariance structure to be used for further analysis. In order to explore the degree of

association in the longitudinal data, the effects of explanatory variables are usually removed by first regressing the response, y_{ij}, on the explanatory variables, $\mathbf{x}_{ij}$, to obtain residuals, $r_{ij} = y_{ij} - \mathbf{x}'_{ij}\hat{\beta}$ (Diggle et al., 2002).

With continuous outcomes, association can be explored in terms of correlations using a scatterplot matrix in which r_{ij} is plotted against r_{ik} for all $j < k = 1,\ldots,n$ when observations are made at equally spaced times (Diggle et al., 2002). Alternatively, a sample variogram can be used to describe the association among repeated values for unequally spaced observations (Diggle et al., 2002). In this case, the residuals are extracted from the model for all points i; the observed half-squared differences between pairs of residuals from the relevant regression model at time j and k (where $j < k$) are calculated

$$v_{ijk} = \frac{1}{2}\left(r_{ij} - r_{ik}\right)^2$$

and the corresponding time intervals are obtained as

$$u_{ijk} = t_{ij} - t_{ik}.$$

For a binary outcome, a correlation-based approach is not feasible as the range of correlation is constrained by the means. However, association between binary variables can be modelled using the pairwise odds ratio Ψ, which are not constrained (Diggle et al., 2002; Fitzmaurice et al., 2004). For two binary outcomes, say y_1 and y_2, the odds ratio is given by

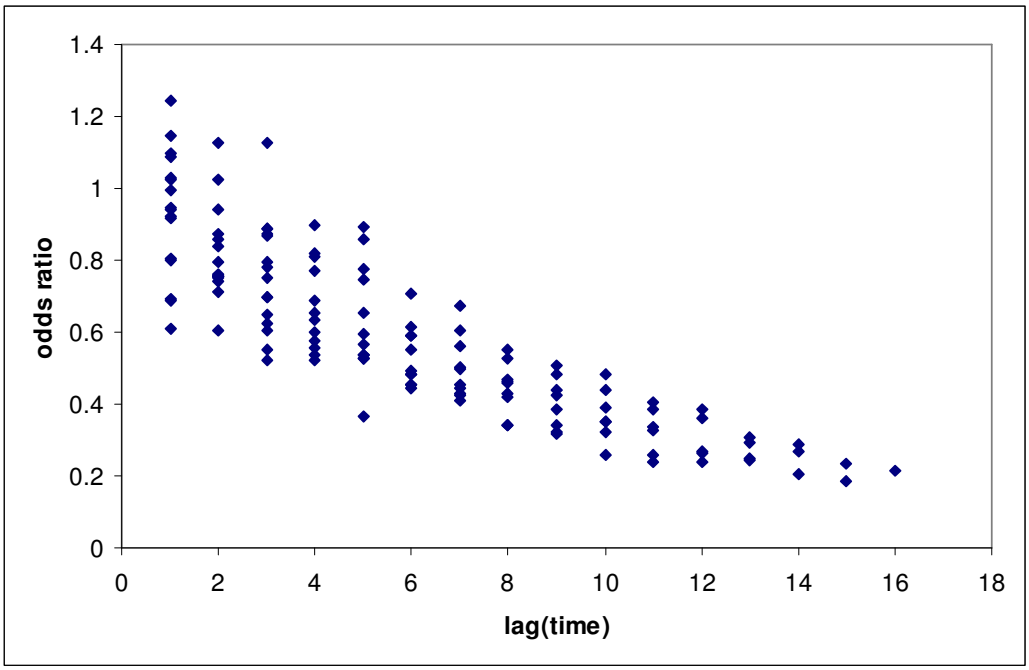
$$\Psi(y_1, y_2) = \left[\frac{\Pr(y_1 = 1, y_2 = 1)\Pr(y_1 = 0, y_2 = 0)}{\Pr(y_1 = 1, y_2 = 0)\Pr(y_1 = 0, y_2 = 1)}\right].$$

The advantage of the odds ratio is that they are strictly positive and unbounded. The logarithm of the odds ratio is taken to yield the entire real line as the range of possible outcomes. For a longitudinal sequence $y_{i1},\ldots,y_{in_i}$ with measurement times $t_{i1},\ldots,t_{in_i}$, Heagerty and Zeger (1998) proposed using the marginal pairwise log-odds ratio to describe the serial dependence for binary responses where they define a 'Lorelogram' as

$$LOR(t_{ij}, t_{ik}) = \log \Psi(y_{ij}, y_{ik}).$$

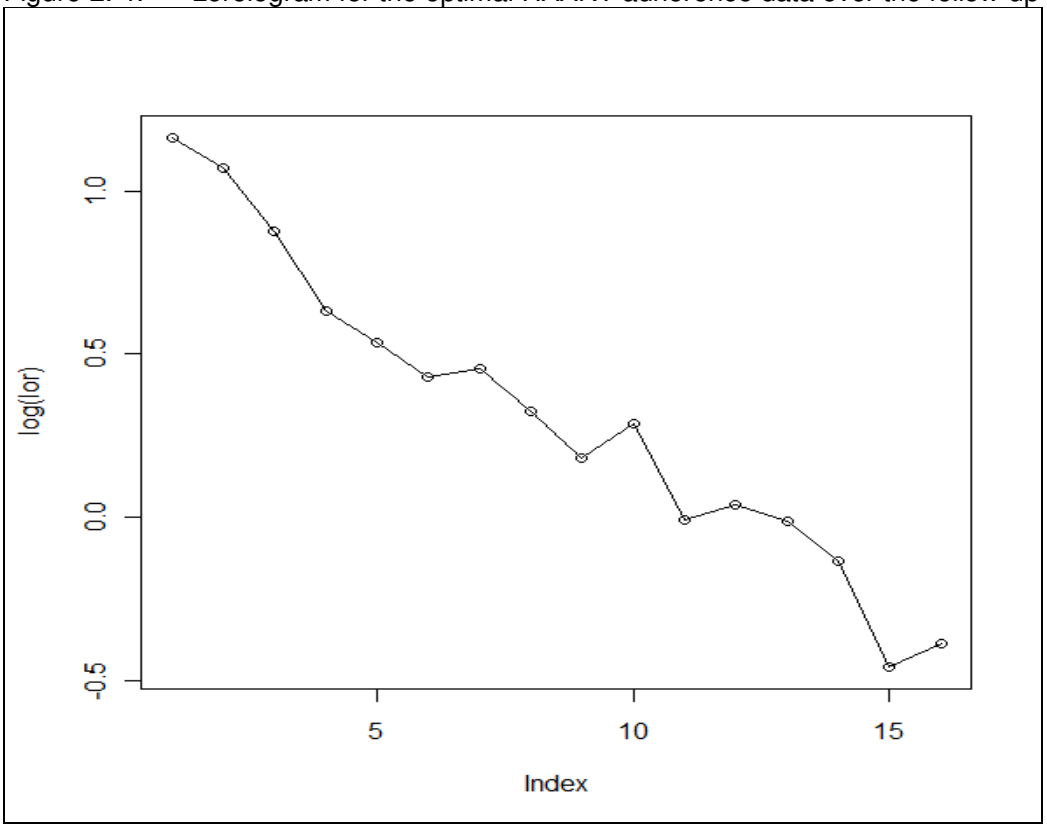
To explore dependence in the adherence data, the crude odds ratios from 2x2 tables with different time lags were calculated and the results are presented in Figure 2.3. The pattern of the crude odds ratio (Figure 2.3) suggests a decreasing association between measurements as the time separation increases.

Figure 2. 3: Crude odds ratios from 2x2 tables for all patients



Moreover, a Lorelogram was drawn using the adherence data and is presented in Figure 2.4. The x-axis (index) is the time lag between two measurements. From Figure 2.4, the log odds ratio also appears to decrease with increasing lag between repeated responses. Thus the First-Order Autoregressive (AR-1) correlation structure may be appropriate for describing the relationship between adherence scores at different visits.

Figure 2. 4: Lorelogram for the optimal HAART adherence data over the follow-up visits

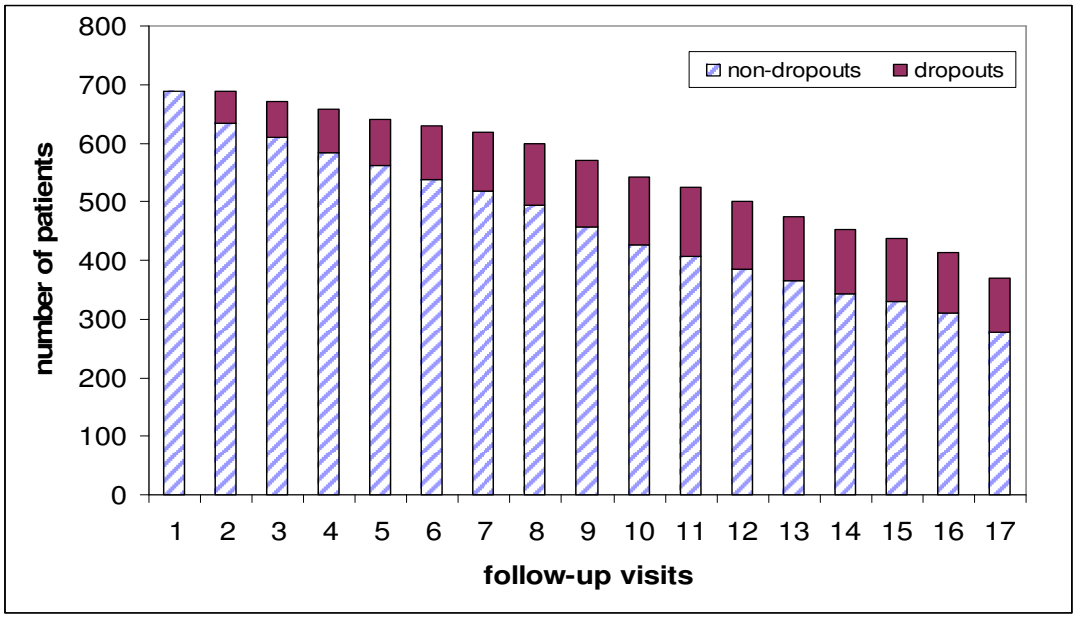


2.4 Missingness

The problem of dealing with missing values is a challenge in any analysis of data, and is almost always present in the analysis of longitudinal data. Although investigators may devote substantial effort to minimize the number of missing values, some amount of missing data is inevitable when studies are designed to collect data on every individual in the sample at each time of follow-up period. Missing data are characterized by missing data patterns and mechanisms. There are different missingness patterns and they can be classified into three categories, namely, monotone (dropout), intermittent and mixed. Monotone (dropout) missingness occurs when the data are available at every assessment until a time the patient drops out and provides no further assessment. This is the most frequently encountered pattern of missingness in longitudinal data (Fitzmaurice et al., 2004). Intermittent missingness occurs if there is a missing observation in between assessments. A mixed pattern occurs when a period of intermittent missingness is followed by monotone missingness.

With regard to adherence data, the number of follow-up visits differed per patient, because some patients started treatment earlier and therefore had more visits. Further, the adherence data like most repeated measurements data had missing data. Figure 2.5 indicates that the missingness pattern of adherence was characterized by dropout, i.e where a patient never came back to the clinic for the monthly review (at least in the period of study).

Figure 2. 5: Total number of patients expected at every visit classified by the number of patients who actually attended the clinic (non-dropouts) and those who dropped out



It is shown in Figure 2.5 that the proportion of patients who dropped out increased gradually over the follow-up period. That is, the proportion of dropout increased from 8% at follow-up visit 2 to 24% in the 17th follow-up visit.

Fitzmaurice et al., (2004) argue that missing data have three important implications for longitudinal analysis. First, when longitudinal data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. As a result, methods of analysis that require balanced data cannot be used when data are missing. Second, when data are missing, there is also some loss of information. That is, missing data cause a reduction in efficiency or drop in the precision with which changes in the

mean response over time can be estimated. Third, under certain circumstances, missing data can introduce bias and thereby led to misleading inferences about changes in the mean response. Consequently, the reasons for any missing data, often referred to as the 'missingness mechanism', must be carefully considered. In fact, reasons for missing data may or may not be related to the outcome of interest. When it is unrelated to the outcome of interest, the impact of missing data is relatively mild and does not complicate the analysis. When it is related to the outcome, greater care is required because there is potential for bias when individuals with missing data differ in important ways from those with complete data.

The missing data mechanism characterizes the reasons for missing data. That is, the mechanism addresses the basic question of why the data are missing. In order to obtain valid inferences from incomplete longitudinal data, the nature of missing data mechanism should be considered. Since missing data mechanism is not under the control of the investigators and is often not well understood, assumptions are made about the missing data mechanism and the validity of the analysis depends on whether these assumptions hold (Fitzmaurice et al., 2004; Hedeker and Gibbons, 2006). Rubin (1976) and Little and Rubin (2002) give important distinctions between different missing data mechanisms. These are missing completely at random (MCAR), which refers to missingness if the missing values are independent of both unobserved and observed data. There is also missing at random (MAR) and occurs if conditional on observed data, the missingness is independent of the unobserved measurements. The other mechanism of missingness is referred to as missing not at random (MNAR). In this case the missing data process is neither MCAR nor MAR but is non-random.

Looking at these three missing data mechanisms further, assume that for subject i in the study, a sequence of measurements y_{ij} is designed to be measured at occasions $j = 1,\ldots,n_i$. Then the outcomes can be grouped into a vector $\mathbf{y}_i = \left(y_{i1},\ldots,y_{in_i}\right)'$. Furthermore, for each occasion j, the missing data indicators can be defined as follows:

$$r_{ij} = \begin{cases} 1 \text{ if } y_{ij} \text{ is observed} \\ 0 \quad \text{otherwise} \end{cases} .$$

(2.1).

The indicators can also be grouped into a vector $\mathbf{r}_i$ which is of the same length as $\mathbf{y}_i$. The vector $\mathbf{y}_i$ can be partitioned into two sub-vectors such that $\mathbf{y}_i^o$ is the vector containing those y_{ij} for which $r_{ij} = 1$ and $\mathbf{y}_i^m$ contains the remaining components.

When the data are incomplete, statistical modeling begins by considering the full data density

$$f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}),$$

where $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{W}_i$ are design matrices for fixed effects, random effects and missing data process respectively; and $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are vectors that parameterize the joint distribution. Let $\boldsymbol{\theta} = (\beta', \alpha')'$ (fixed effects and covariance parameters) and $\boldsymbol{\psi}$ describe the measurement and missingness process respectively.

The classification of missingness mechanism by Rubin (1976) and Little and Rubin (2002) is based on the *selection modeling* framework where factorization equals

$$f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{W}_i, \boldsymbol{\psi}) \qquad (2.2).$$

The first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. The second factor describes one's self-selection mechanism to either continue or leave the study. The differences in the missing data mechanism is specified through the second factor of (2.2), that is

$$f(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{W}_i, \boldsymbol{\psi}) = f(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{W}_i, \boldsymbol{\psi}).$$

Under MCAR mechanism, the probability of an observation being missing is independent of responses:

$$f(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{W}_i, \boldsymbol{\psi}) = f(\mathbf{r}_i \mid \mathbf{W}_i, \boldsymbol{\psi}),$$

and therefore (2.2) simplifies to

$$f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i \mid \mathbf{W}_i, \boldsymbol{\psi}),$$

implying that both components are independent. The implication is that the joint distribution of $\mathbf{y}_i^o$ and $\mathbf{r}_i$ becomes

$$f\left(\mathbf{y}_i^o, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = f\left(\mathbf{y}_i^o \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}\right) f\left(\mathbf{r}_i \mid \mathbf{W}_i, \boldsymbol{\psi}\right).$$

In this case, the processes of generating missing values can be ignored regardless of whether the data are analyzed using a frequentist, likelihood or Bayesian procedure. It should be noted that the above definition is conditional on covariates.

Under MAR mechanism, the probability of an observation being missing is conditionally independent of the unobserved outcomes, given the values of the observed outcomes:

$$f\left(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{W}_i, \boldsymbol{\psi}\right) = f\left(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}\right).$$

Again, the joint distribution of the observed data can be partitioned:

$$f\left(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = f\left(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}\right) f\left(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}\right),$$

and hence at the observed data level:

$$f\left(\mathbf{y}_i^o, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = f\left(\mathbf{y}_i^o \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}\right) f\left(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}\right).$$

In the MNAR case, neither MCAR nor MAR holds. Under the MNAR, the probability of a measurement being missing depends on unobserved outcomes. The joint distribution of measurements and the missingness process is written as

$$f\left(\mathbf{y}_i^o, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = \int f\left(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}\right) f\left(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{W}_i, \boldsymbol{\psi}\right) d\mathbf{y}_i^m,$$

and no simplification of this joint distribution is possible.

Implications for analysis of longitudinal data when the missing data mechanism is MCAR, is that individuals with missing data are a random subset of the sample. In this case the observed values of the responses are a random subsample of all values of the responses and no bias will arise with almost any method of analysis of the data (Fitzmaurice et al., 2004). When the missing data mechanism is MAR, individuals with missing data are no longer a random subset of the sample. Likewise, the observed values are not necessarily a random subsample of all responses. This implies that analysis restricted to data from completers will yield biased estimates. In fact, when data are MAR, not MCAR, complete case methods yield biased estimates. In contrast, likelihood-based methods that correctly specify the entire distribution of the responses yield valid estimates when missing data are MAR. Molenberghs and Verbeke (2005) show through consideration of the

likelihood that under MAR, much of the missing value problem tends to disappear. That is, the full data likelihood contribution for unit i takes the form:

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathbf{y}_i, \mathbf{r}_i) \propto f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}).$$

Since inference has to be based on what is observed, the full data likelihood L^* needs to be replaced by the observed data likelihood L:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathbf{y}_i, \mathbf{r}_i) \propto f(\mathbf{y}_i^o, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}),$$

with

$$\begin{aligned} f(\mathbf{y}_i^o, \mathbf{r}_i \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{y}_i^m \\ &= \int f(\mathbf{y}_i^o, \mathbf{y}_i^m \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}) d\mathbf{y}_i^m. \end{aligned}$$

Under an MAR process, we obtain

$$\begin{aligned} f(\mathbf{y}_i^o, \mathbf{r}_i \mid \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{y}_i^o, \mathbf{y}_i^m \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}) d\mathbf{y}_i^m \\ &= f(\mathbf{y}_i^o \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i \mid \mathbf{y}_i^o, \mathbf{W}_i, \boldsymbol{\psi}). \end{aligned}$$

The likelihood factors into two components of the same functional form as the general factorization (2.2) of the complete data. If further, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint in the sense that the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is the product of the parameter spaces of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, then inferences can be based solely on the marginal observed data density. This requirement is referred to as separability condition. In essence, under the assumption of MAR and the mild separability condition, likelihood based analysis is valid, provided all available data are analyzed (Molenberghs and Kenward, 2007). On the other hand, since the standard GEE is not a likelihood-based method, the MAR mechanism will not generally hold for the observed data, compromising the validity of the analysis. Nonetheless, Robins, Rotnitzky and Zhao (1995) have proposed a class of weighted estimating equations to allow for MAR, extending GEE. The basic idea of weighted generalized estimating equations (WGEE) is to weight each subject's contribution in the GEEs by the inverse probability that a subject drops out at the time he dropped out. For more details refer to Molenberghs and Verbeke, 2005; Jansen et al., 2006). Furthermore, when missing data are NMAR, multiple imputation techniques with pattern mixture models can be used, among others.

Testing MCAR assumption

Since any method of analysis will yield valid inferences when missing data are MCAR, it would be quite useful to test the validity of MCAR assumption in HAART adherence data. The distinction between MCAR and MAR is that with MCAR, missingness cannot depend on the observed values of the dependent variable, $\mathbf{y}_i^o$ but it can with MAR. As a result, tests of whether MCAR is reasonable or not can therefore be based on analyses involving $\mathbf{y}_i^o$.

There have been a number of tests proposed to test the MCAR assumption when a missingness pattern is characterized by dropout and these include tests proposed by Little (1988) and Diggle (1989). The basic idea behind these tests is that if dropout happens completely at random at each time point, the group of subjects who dropout represent the random sample selected from the set of all subjects still in the study at that time. To illustrate these tests, consider an example with two data points for each subject and all subjects have data at time 1, but some are missing at time 2. Let us define $\mathbf{r}_i = 1$ for subjects with data at both time-points and $\mathbf{r}_i = 0$ for those that only have data for the first time-point. Then a simple t-test can be used to compare the y_{i1} means between the two groups (that is, $\mathbf{r}_i = 1$ versus $\mathbf{r}_i = 0$). If the missing data are MCAR, the means for the two groups will not differ. More generally, because MCAR allows missingness to depend on covariates, the following regression can be performed (Hedeker and Gibbons, 2006)

$$y_{i1} = \beta_0 + \beta_1 \mathbf{r}_i + \beta_2 \mathbf{x}_i + \varepsilon_i,$$

where β_2 is the vector of regression coefficients for the set of covariates included in $\mathbf{x}_i$. We can also form interactions of the dropout variable with the covariates to yield

$$y_{i1} = \beta_0 + \beta_1 \mathbf{r}_i + \beta_2 \mathbf{x}_i + \beta_3 (\mathbf{r}_i \times \mathbf{x}_i) + \varepsilon_i,$$

where β_3 is the vector of regression coefficients for the interactions of dropout with covariates. In this model MCAR would specify that $\beta_1 = \beta_3 = 0$, which means that dropout does not depend on the observed response.

Ridout (1991) noted that it is helpful to turn this question around and to specifically model dropout in terms of logistic regression. The model could be given as

$$\log\left(\frac{P(r_{i1}=0)}{1-P(r_{i1}=0)}\right)=\alpha_0+\alpha_1 y_{i1}+\alpha_2\mathbf{x}_i+\alpha_3(y_{i1}\times\mathbf{x}_i),$$

where α_2 and α_3 represent vectors of regression coefficients for the set of covariates $\mathbf{x}_i$ and their interaction with y_{i1}. MCAR would specify that $\alpha_1=\alpha_3=0$. This logistic regression model can be generalized to more than two time points. Assume that dropout probability at occasion j depends on the previous outcome y_{ij-1} in place of y_{i1}. This leads to the following model:

$$\log\left(\frac{P(r_{ij}=0\mid\mathbf{y}_i)}{1-P(r_{ij}=0\mid\mathbf{y}_i)}\right)=\phi_0+\phi_1 y_{ij-1}+\phi_2\mathbf{x}_i+\phi_3(y_{ij-1}\times\mathbf{x}_i),$$

where the missing data indicators for each j occasion has been defined in (2.1) as

$$r_{ij}=\begin{cases}1 \text{ if } \mathrm{y}_{ij} \text{ is observed}\\ 0 \quad \text{otherwise}\end{cases}$$

The probability $P(r_{ij}=0\mid\mathbf{y}_i)$ is the conditional probability of subject i dropping out at time j. Again, here ϕ_2 and ϕ_3 represent vectors of regression coefficients for the set of covariates $\mathbf{x}_i$ and their interaction with y_{ij-1}. Here $\phi_1=\phi_3=0$ indicates MCAR, i.e. that dropout does not depend on the previous outcome.

In order to assess whether dropout depends on the previous outcome for adherence data, a logistic regression model for the dropout indicator proposed by Ridout (1991) and illustrated by Hedeker and Gibbons (2006) is adapted. The logistic regression model is built in terms of the previous response adjusting for baseline age and gender as well as treatment site and time. That is, we consider a dropout model at occasion j as follows:

$$\log\left[\frac{P\left(r_{ij}=0\mid\mathbf{y}_i\right)}{1-P\left(r_{ij}=0\mid\mathbf{y}_i\right)}\right]=\beta_0+\beta_1 y_{ij-1}+\beta_2 A_i+\beta_3 G_i+\beta_4 S_i+\beta_5 t_{ij},$$

$$\text{with } r_{ij}=\begin{cases}1 \text{ if } y_{ij} \text{ is observed}\\ 0 \quad \text{otherwise}\end{cases}.$$

where y_{ij-1} is the previous outcome at which the jth measurement is taken for the ith subject, A_i is the baseline age for subject i, G_i is the indicator for gender (male, female) for subject i, S_i is the indicator of the treatment site (urban, rural) for subject

i and t_{ij} is the time point at which the jth measurement is taken for the ith subject. MCAR would dictate that $\beta_1 = 0$.

In addition to all the main effects in the model, 2-way interactions of the previous outcome with the other covariates were added to the model one at a time and their significance assessed. None of the interactions were significant at 5% level and therefore we present the results for the main effects only in Table 2.2.

Table 2. 2: Parameter estimates (standard errors) for a logistic regression model to describe dropout.

Parameter	Estimate	Standard error	p-values
Intercept	-3.480	0.446	<.0001
Previous outcome (ref=not adherent)			
Adherent	0.0664	0.208	0.7482
Gender (ref=female)			
Male	-0.0366	0.199	0.8543
Treatment site (ref=rural)			
Urban	0.0733	0.182	0.6867
Age	-0.0059	0.011	0.6028
Time	-0.0398	0.021	0.0577

The results in Table 2.2 show that after controlling for gender, site, age and time, previous outcome is not significant. This implies that there is no evidence against MCAR in favor of MAR. It thus, can be concluded that MCAR assumption holds for missingness in adherence data, and as a result, any method of analysis for longitudinal data will yield valid estimates.

2.5 Summary

The exploratory data analysis suggests that overall, optimal HAART adherence is increasing over time, even though the rate at which it is increasing might be different for the different sub-populations. Moreover, the exploration of the dependence structure suggests that AR-1 might be the appropriate covariance structure for these data. Further, it has also been established that the adherence data is characterized by dropout and the evaluation of the impact of dropout indicated that the data are

MCAR. This result paves the way for longitudinal models that will be used in subsequent chapters for analyses of adherence data.

Since the commonly used longitudinal methods for discrete (e.g binary) data are a direct extension of generalized linear models for independent observations to the context of correlated data, a review of these models is provided in the next chapter to lay a foundation for longitudinal models to be considered in later chapters. At the same time, these models will be fitted to the adherence data to identify factors that affect initial (one month after initiation of HAART) optimal HAART adherence, as the need for strict HAART adherence is required from the onset of treatment in order to maintain prolonged clinical benefits.

Chapter 3

Initial adherence assessment using generalized linear models

3.1 Introduction

The generalized linear model (GLM) is a generalization of the linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). In its simplest form, a linear model specifies the linear relationship between a response variable and a set of predictor (independent) variables. Consider a linear model of the form

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad (3.1)$$

where $\mathbf{y}$ is an $(n\times 1)$ vector of values of the response variable, $\mathbf{X}$ is an $(n\times p)$ matrix of the known explanatory variables generally including a column of ones, β is a $(p\times 1)$ vector of unknown parameters to be estimated and ε is an $n\times 1$ vector of independent variables assumed to be normally distributed with mean zero and constant variance $\sigma^2\mathbf{I}$. Despite the fact that linear models (3.1) have a number of advantages, including computational simplicity, they make a set of restrictive assumptions. These models assume that (a) observations are independent, (b) the mean of the observations is a linear function of explanatory variables and parameters, and (c) the observations are normally distributed with a constant variance (Jiang, 2007). GLMs relax the assumptions (b) and (c). These models allow response variables that have distributions other than the normal distribution; they may even be categorical rather than continuous (Lindsey, 1997; Fahrmeir and Tutz, 1994). In fact, one of the advances in statistical theory has been the recognition that many of the 'nice' properties of the normal distribution are shared by a wider class of distributions known as the exponential family of distributions (Dobson, 2001). Thus GLMs involve a variety of distributions selected from the exponential family of distributions. In addition, the variance of observations is no longer constant but is a function of the mean. Moreover, the relationship between the mean of the observations and explanatory variables need not be of the simple linear form in GLMs. They involve the transformation of the mean, through what is called a 'link function', linking the regression part to the mean of one of the

distributions from the exponential family (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

Since one of the objectives is to explore determinants of HAART adherence after the first month (baseline) of initiation into therapy, data analysis using GLMs seems to be appropriate. This is because we are not only faced with a binary outcome but we also have one datum point at baseline (response) for each patient (cross-sectional data) where we assume that their responses are independent.

3.2 Generalized linear models

A complete specification of a generalized linear model $y_1,\ldots,y_n$ involves three components, namely:

a. the distribution (belonging to the exponential family of distributions) of the responses,
b. the systematic component, as well as
c. link function

and they are discussed below.

a) <u>The exponential family of distributions</u>

A probability distribution is said to be a member of the exponential family if its probability density function can be written in the form

$$f(y) = f(y \mid \theta,\phi) = \exp\left\{\frac{y\theta - \psi(\theta)}{\phi} + c(y,\phi)\right\} \qquad (3.2)$$

where $\psi(\cdot)$ and $c(\cdot,\cdot)$ are specific functions; θ and ϕ are unknown parameters commonly known as a natural location parameter and a scale or dispersion parameter respectively. The mean and variance of this distribution is derived by making use of the property $\int f(y \mid \theta,\phi)dy = 1$ and taking the first- and second-order derivatives with respect to θ from both sides of the equation (3.2) so that we obtain

$$\int [y - \psi'(\theta)] f(y \mid \theta,\phi)dy = 0,$$

$$\int \left\{\phi^{-1}[y - \psi'(\theta)]^2 - \psi''(\theta)\right\} f(y \mid \theta,\phi)dy = 0.$$

It therefore follows that the mean, $\mu = E(y)$ is equal to $\psi'(\theta)$ and the variance $\sigma^2 = var(y)$ is given by $\phi\psi''(\theta)$. This implies that in general, the mean and variance are related through $\sigma^2 = \phi\psi''\left[\psi'^{-1}(\mu)\right] = \phi\upsilon(\mu)$ for an appropriate function $\upsilon(\mu)$ known as the variance function (Molenberghs and Verbeke, 2005).

In some cases, a quasi-likelihood approach is used. Even though we have seen that the relation between the mean and variance immediately follows from the density (3.2), in the quasi-likelihood perspective, one starts from specifying a mean and a variance function,

$$E(y) = \mu,$$

$$var(y) = \phi\upsilon(\mu).$$

The variance function $\upsilon(\mu)$ is chosen in accordance with a particular member of the exponential family. Since the distributional assumptions are not specified, parameters cannot be estimated using maximum likelihood principles. Instead, a set of estimating equations needs to be specified, the solution of which is referred to as the quasi-likelihood estimates (Molenberghs and Verbeke, 2005). One such approach is the method of generalized estimating equations (GEE) which is discussed in detail in Chapter 4.

b) Systematic component

The systematic component of the generalized linear model specifies the effects of the explanatory variables, on the mean through a linear predictor. Associated with each response y_i is a vector $\mathbf{x}_i = (x_{i1}, x_{12}, \cdots, x_{ip})'$ of values of p explanatory variables, then the distribution of the response variable y_i depends on $\mathbf{x}_i$ through the linear predictor η_i where

$$\begin{aligned}\eta_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\ &= \sum_j x_{ij}\beta_j \\ &= \mathbf{x}_i'\beta\end{aligned}$$

and $\beta = (\beta_1, \ldots \beta_p)'$ is a vector of fixed but unknown parameters describing the dependence of y_i on $\mathbf{x}_i$. This shows that a linear predictor is simply a linear combination of the unknown regression parameters and the explanatory variables. It describes how the location of the response distribution changes with the explanatory variables (McCullagh and Nelder, 1989; Dobson, 2001; Fahrmeir and Tutz 1994).

c) The link function

For specifying the pattern of dependence of the response variable on the explanatory variables, GLMs take a suitable transformation of the mean response and relate the transformed mean response to the explanatory variables. The link between the distribution of y_i and the linear predictor η_i is provided by the link function $g(.)$ such that, $\eta_i = g(\mu_i)$, $i = 1, \cdots, n$ where $\mu_i = E(y_i)$, $i = 1, \cdots, n$. As a result, the dependence of the distribution of the response on the explanatory variables is given by

$$g(\mu_i) = \eta_i = \mathbf{x}_i'\beta \qquad i = 1, \ldots, n.$$

This shows that the link function describes the relation between μ_i, the mean of y_i and the linear predictor η_i. That is, the use of the link function in GLMs allows the model parameters to be included in the model linearly just as in the ordinary linear models (Brown and Prescott, 2006; Collett, 2003). Moreover, Fitzmaurice et al., (2004) show that the primary motivation for considering link functions other than the identity is to ensure that the linear predictor produces predictions of the mean response that are within the allowable range. Generally, any function $g(.)$ can be chosen to link the mean of y_i to the linear predictor. However, every distribution that belongs to the exponential family has a special link function called the canonical link function. The canonical link function is defined as $g(\mu_i) = \theta_i$, where θ_i is the natural location parameter in (3.2). Thus, the canonical link function is that function which transforms the mean to a canonical location parameter of the exponential family of distributions. Although other links are possible, the canonical links are the most commonly used in practice.

3.2.1. Logistic regression for binary data

The logistic regression model is a member of the generalized linear models that are used to model binary data. Binary data can be specified either as a series of zeros and ones (Bernoulli form) or aggregated as frequencies of successes out of a certain number of trials (Binomial form). We will illustrate generalized linear model formulation and interpretation using the logistic regression model because of its relevance to HAART adherence data used in this study.

Let y be Bernoulli distributed with success probability $P(y=1)=\mu$, then the density takes the form:

$$f(y)=\exp\left(y\ln\left(\frac{\mu}{1-\mu}\right)+\ln(1-\mu)\right).$$

This is in the form of (3.2), which follows that the Bernoulli distribution belongs to the exponential family with the natural parameter $\theta=\ln\frac{\mu}{1-\mu}$, scale parameter $\phi=1$ and with mean, $E(y)=\mu$ and variance function $v(\mu)=\mu(1-\mu)$. The canonical link function is the logistic or logit link function and this gives the logistic regression model as follows

$$logit(\mu)=\log\left(\frac{\mu}{1-\mu}\right)=\mathbf{x}_i'\beta, \qquad (3.3)$$

where $y_i \sim Bernoulli(\mu_i)$, $\mathbf{x}_i'$ $(i=1,\ldots,n)$ is the ith row of the covariates matrix $\mathbf{X}$ and β is a vector of unknown parameters associated with $\mathbf{X}$. The logit transformation ensures that the probabilities lie within the interval (0, 1) for any values of $\mathbf{x}_i'\beta$ from $-\infty$ to ∞. If we solve for μ in (3.3), the result is

$$\mu=\frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)},$$

which indicates the marginal mean as a function of the covariates. The function $\exp(.)/1+\exp(.)$ is the inverse link function, that is, the inverse logit. The other link functions (non-canonical) for binary data not considered here are probit and complementary log-log functions.

Next we consider the interpretation of the logistic model coefficients. Let us consider a model given by

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x . \qquad (3.4)$$

The logistic regression intercept, β_0 has interpretation as the log odds of success $(y=1)$ when $x=0$ and β_1 represents the logistic regression slope. Suppose a predictor variable, x is a dichotomous variable taking values of 0 and 1, the logistic regression coefficient β_1 associated with x compares the log odds of success when $x=1$ to the log odds of success when $x=0$. That is

$$logit(\mu \mid x=1) - logit(\mu \mid x=0) = (\beta_0 + \beta_1(1)) - (\beta_0 + \beta_1(0)) = \beta_1 .$$

Thus $\exp(\beta_1)$ has interpretation as the odds ratio of the response for two possible values of the covariates. This is obtained by converting the log odds model to the odds model by taking the anti-log of $\log\frac{\mu}{1-\mu}$ as follows. The odds that $y=1$ for $x=1$ is $\exp(\beta_0+\beta_1)$ and the odds that $y=1$ for $x=0$ is $\exp(\beta_0)$. So the model is, $odds(y=1 \mid x) = \exp(\beta_0 + \beta_1 x)$. If we construct a ratio of these odds, we obtain

$$OR = \frac{odds(x=1)}{odds(x=0)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1) .$$

That is, the odds ratio (OR) depends only on model parameter β_1. Note that when there is no group effect, $\beta_1 = 0$ and $OR = 1$. When subjects in group 1 $(x=1)$ have increased odds of the event of interest $(y=1)$, then $OR > 1$ and $\beta_1 > 0$. Likewise, when group 1 subjects have less odds of the event of interest $(y=1)$ then $OR < 1$ and $\beta_1 < 0$.

For a continuous predictor variable x, the logistic regression coefficient β_1, associated with x has interpretation as the change in the log of odds of success for a unit change in x. The relationship between the response and predictor can again be expressed as a ratio of odds

$$OR = \frac{odds(X = x+1)}{odds(X = x)} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1) .$$

Clearly when $\beta_1 > 0$ there is increasing odds as x increases, when $\beta_1 < 0$ there is decreasing odds as x increases, and when $\beta_1 = 0$ there is no relationship between the predictor x and the response.

If we extend model (3.4) by including another predictor variable such that the logistic regression model is written as

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ,$$

where x_1 and x_2 are dichotomous predictor variables both taking only values 0 and 1. β_1 can be interpreted as the change in the log odds ratio as x_1 changes from 0 to 1 while holding x_2 constant. Likewise β_2 can be interpreted as the change in the log odds ratio as x_2 changes from 0 to 1 while holding x_1 constant. Using the odds, the logistic regression model is of the form: $odds(y=1) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$. The odds ratio for x_1 when x_2 is fixed is given as

$$OR = \frac{odds(x_1 = 1, x_2)}{odds(x_1 = 0, x_2)} = \frac{\exp(\beta_0 + \beta_1(1) + \beta_2 x_2)}{\exp(\beta_0 + \beta_1(0) + \beta_2 x_2)} = \exp(\beta_1).$$

Note that the odds ratio for x_1 does not depend on the value of x_2. If we include an $x_1 x_2$ interaction term in the model, we obtain a logistic regression model of the form $odds(y=1) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$. In this case the odds ratio for x_1 is given as

$$OR = \frac{odds(x_1 = 1, x_2)}{odds(x_1 = 0, x_2)} = \frac{\exp(\beta_0 + \beta_1(1) + \beta_2 x_2 + \beta_3 x_2)}{\exp(\beta_0 + \beta_1(0) + \beta_2 x_2)} = \exp(\beta_1 + \beta_3 x_2).$$

Contrary to when there is no interaction term between the two predictor variables, the odds ratio for x_1 is a function of x_2 and thus the relationship between x_1 and the response depends on the level of x_2. The results can be generalized to cases where predictor variables have more than two categories and where there are continuous predictor variables.

3.3 Parameter estimation and inference

The method of maximum likelihood estimation is usually used for estimating the parameters of the generalized linear models. The following derivation follows that of Molenberghs and Verbeke (2005). Assuming independence of the observations, the log-likelihood can be expressed as follows

$$\ell(\beta,\phi)=\frac{1}{\phi}\sum_{i=1}^{N}[y_i\theta_i-\psi(\theta_i)]+\sum_i c(y_i,\phi).$$

The score equations obtained from equating the first order derivatives of the log-likelihood to zero take the form

$$S(\beta)=\sum_i \frac{\partial\theta_i}{\partial\beta}[y_i-\psi'(\theta_i)]=0.$$

Since $\mu_i=\psi'(\theta_i)$ and $v_i=v(\mu_i)=\psi''(\theta_i)$, we have that

$$\frac{\partial\mu_i}{\partial\beta}=\psi''(\theta_i)\frac{\partial\theta_i}{\partial\beta}=v_i\frac{\partial\theta_i}{\partial\beta},$$

which implies the following score equations

$$S(\beta)=\sum_i \frac{\partial\mu_i}{\partial\beta}v_i^{-1}(y_i-\mu_i)=0. \qquad (3.5)$$

These score equations are solved iteratively. That is, an initial solution of the equations $\hat{\beta}^{(0)}$ is guessed and then updated until iterative algorithm converges to the solution $\hat{\beta}$, called the maximum likelihood estimate of β. The two most popular and widely used iterative algorithms for the maximum likelihood estimation are the Fisher's scoring and Newton-Raphson. The Fisher's scoring method is equivalent to the iterative reweighted least squares (Schabenberger and Pierce, 2002; Kutner et al., 2005). The Newton-Raphson method solves maximum likelihood estimates iteratively using the standard least-squares methods (McCullagh and Nelder, 1989). These iterative algorithms are available in most statistical packages, such as SAS, STATA, GenStat etc.

Once the maximum likelihood estimates have been obtained, classical inference based on asymptotic likelihood theory becomes available, including Wald-type tests, likelihood ratio tests and the score tests, all asymptotically equivalent (Molenberghs and Verbeke, 2005). Moreover, with some models such as the logistic regression model, ϕ is a known constant. In other models, such as the

linear normal model, estimation of ϕ may be required to estimate the standard errors of the elements in β. There are several ways of estimating ϕ, one of which is given by

$$\hat{\phi} = \frac{1}{N-p}\sum_{i}(\mathbf{y}_i - \hat{\mu}_i)^2 / v_i(\hat{\mu}_i)$$

where N is the total number of observations and p is the number of parameters in the model.

3.4 Model selection and diagnostics

Model Selection

Since there can be a number of models that describe a given data set, it is important to select the simplest reasonable model that adequately describes such data (Lindsey, 1997). There are three procedures that are commonly used for the selection of variables that enter the model and these are the backward, forward and stepwise methods. Backward selection starts with a saturated model (a model with all explanatory variables) and drops one explanatory variable at a time while forward selection starts with the null model (no explanatory variables) and enters one explanatory variable at a time. The stepwise selection procedure uses the same approach as the forward selection, but has the advantage over the forward selection in that the variables already in the model are considered for exclusion each time another variable enters the model. Consequently, when there are many variables under consideration, the stepwise is mostly preferred because it has an advantage of minimizing the chances of keeping redundant variables and leaving out important variables in the model.

With all the procedures, a variable that leads to a significant change in the deviance (measure of goodness-of-fit described subsequently) when added to or dropped from the model is retained, otherwise it is dropped. The contribution of each variable to the deviance reduction is given by the type 1 and type 3 analyses of effects. The type 1 analysis of effects depends on the sequence in which variables enter the model, whilst type 3 considers the overall model and assesses the contribution of each variable to the deviance reduction irrespective of the sequence in which variables enter the model. This method of model

selection is referred to as deviance analysis and is used to test the model for goodness-of-fit.

Model Checking

The main tools used for assessing the goodness-of-fit of a fitted generalized linear model are the log-likelihood ratio (deviance) and the Pearson's chi-square statistics (Jiang, 2001; Kutner et al., 2005). They measure the discrepancy of fit between the maximum log-likelihood achievable and the achieved log-likelihood by the fitted model. The deviance is presented below to illustrate the use of these measures. It is given by

$$D(\mathbf{y},\hat{\mu}) = 2\{\ell(\mathbf{y};\mathbf{y}) - \ell(\hat{\mu};\mathbf{y})\},$$

where $\ell(\mathbf{y};\mathbf{y})$ is the log-likelihood under the maximum achievable (saturated) model and $\ell(\hat{\mu};\mathbf{y})$ is the log-likelihood under the current model. The aim is to minimize D (i.e. $D(\mathbf{y},\hat{\mu})$) by maximizing $\ell(\hat{\mu};\mathbf{y})$. The hypothesis about the goodness-of-fit of the model is given by

H_0 : model is adequate $\quad vs \quad$ H_1 : model is not adequate.

H_0 will be rejected if $D > \chi^2_{n-p,\alpha}$, where n is the number of observations, p is the number of parameters and α is the given level of significance. Note that the deviance D cannot be used as a measure of goodness-of-fit for ungrouped binary data (Collett, 2003). However, it can still be used to identify important predictors as discussed earlier. The appropriate test in this case is the Hosmer-Lemeshow goodness-of-fit test. For this test, firstly, the predicted probabilities $(\hat{\mu}_i's, i = 1,\ldots,n)$ obtained using the current model being checked are used to form g groups with approximately n/g subjects. One grouping strategy is the percentile strategy and it is given as (Hosmer and Lemeshow, 2000):

i) Group 1 subjects are approximately n/g subjects whose $\hat{\mu}_i's$ are less than or equal to the $\frac{100}{g}th$ percentile of all the $\hat{\mu}_i's$.

ii) Group 10 subjects are approximately n/g subjects whose $\hat{\mu}_i's$ are more than $\left(1-\frac{1}{g}\right)\times 100th$ percentile of all the $\hat{\mu}_i's$.

iii) For $j = 2,3,\ldots,g-1$, group j subjects are approximately n/g whose $\hat{\mu}_i's$ are greater than the $\frac{j-1}{g}\times 100th$ percentile and less than or equal to the $\frac{j}{g}\times 100th$ percentile of all the $\hat{\mu}_i's$.

For large n (number of subjects), the frequently recommended g is 10 (Hosmer and Lemeshow, 2000; Dobson, 2001; and Vittinghoff et al., 2005) in order for the different analysts to get consistent conclusions.

Secondly, for each group the observed and expected frequencies of the responses $y = 0$ and $y = 1$ are determined as follows (Hosmer and Lemeshow, 2000). For the $j = 1,2,\ldots,g$,

i) The respective observed frequencies of the responses $y = 1$ and $y = 0$ are $O_{1j} =$ number of subjects with responses $y = 1$ and $O_{0j} = n/g - O_{1j}$.

ii) The respective expected frequencies of the responses $y = 1$ and $y = 0$ are $E_{1j} = n/g \times$ average of $\hat{\mu}_i's$ in group j and $E_{0j} = n/g - E_{1j}$

Consequently, the Hosmer-Lemoshow statistic X^2_{HL} for testing the goodness-of-fit of the model is given by

$$X^2_{HL} = \sum_{j=1}^{g}\sum_{i=0}^{1}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The X^2_{HL} has a chi-square distribution with $g-2$ degrees of freedom. As a result, the statistic X^2_{HL} is compared with the critical value of the chi-square distribution with $g-2$ degrees of freedom $(\chi^2_{(g-2,\alpha)})$ for checking goodness-of-fit of the model. Thus, if the X^2_{HL} is statistically significant, then it indicates lack-of-fit of the model, whereas a non-significant one indicates goodness-of-fit of the model.

The appropriateness of the link function can be assessed by refitting the model with the linear predictor obtained from the original model and the square of the linear predictor as explanatory variables (Vittinghoff et al., 2005). If the linear

predictor is statistically significant and the squared linear predictor term is insignificant, then the link function is appropriate. This means that, prediction given by the linear predictor is not improved by adding the squared linear predictor term which is basically used to evaluate the null hypothesis that the model is adequate. Alternatively, the original model can be estimated with an extra constructed variable, where for an adequate model the extra variable will be statistically insignificant (Williams, 1987). Moreover, the appropriateness of the link function can also be checked graphically by plotting the residuals against the fitted values and for an appropriate link, the plot should not exhibit any systematic pattern (Collett, 2003).

Outliers, influential and high-leverage points also have to be assessed. An outlier is a datum point that differs from the general trend of the data and is not necessarily influential (Lindsey, 1997). With an influential point, a slight change or omission of such an observation leads to a substantial effect on parameter estimates of the model. The magnitude of influence is measured by the leverage (denoted by h_{ii}), which is the ith diagonal element of the hat-matrix ($\mathbf{H}$) (Kutner et al., 2005; Lindsey, 1997). For generalized linear models, the hat-matrix is given by

$$\mathbf{H} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{X}'\mathbf{W}^{-\frac{1}{2}},$$

where $\mathbf{X}$ is the design matrix of the known covariates and $\mathbf{W}$ is a diagonal weight matrix with ith diagonal element given by

$$w_i = \frac{1}{var(y_i)[g'(\mu_i)]^2}.$$

The commonly used measure for detection of influential data points is the Cook's distance, which is approximated by

$$C_i = \frac{r_{p_i}^2 h_{ii}}{(1-h_{ii})^2} = \frac{r_{p_{is}}^2 h_{ii}}{(1-h_{ii})},$$

where $r_{p_i} = (1-h_{ii})y_i$ is the Pearson's residual and $r_{p_{is}} = \frac{r_{p_i}}{\sqrt{1-h_{ii}}}$ is the standardized Pearson's residual (Moeti, 2007). A large C_i implies that the ith observation has undue influence on the set of parameter estimates. The most widely used cut-off value for C_i is 1, however, some authors (Rousseeuw and Leroy, 2003; Skovgaard

and Ritz, 2007) suggest considering the data points that depart from the rest as influential when examining the index plot of C_i.

3.5 Fitting the logistic regression model to adherence data

Let the response $y_i = 1$ if the ith patient $(i = 1,\ldots,688)$ has been optimally adherent to HAART and $y_i = 0$ otherwise. Then the fitted logistic regression model (3.3) is given as

$$logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i'\beta\,,$$

where $\mu_i = E(y_i) = P(y_i = 1)$, $\mathbf{x}_i'$ is a vector of appropriately coded values of the explanatory variables and β is a vector of unknown parameters. The 13 baseline variables presented in Chapter 2 were used together with their interaction terms as explanatory variables in the model. The deviance analysis was used for model selection. To control for potential confounding, all the main effects were retained in the model. It was then assessed as to whether any interaction terms needed to be incorporated into the model. This was examined by fitting each of the product terms formed from all the predictor variables, one at a time, to the model that had all the 13 variables.

Three interactions reduced the deviance by a relatively large amount: age*cell phone ownership led to a deviance reduction of 6.79 on 1 degree of freedom and a corresponding p-value of 0.009, gender*reported reason for taking an HIV test and treatment site*source of household income reduced the deviance by 11.08 (p-value = 0.004) and 8.38 (p-value = 0.004) on 2 and 1 degrees of freedom respectively. To see if all three could be retained in the model, they were all fitted at the same time and they reduced the deviance by 25.5 on 4 degrees of freedom (p-value = <0.001). Since the reduction in the deviance is significant, all the three interactions were then retained in the model. Consequently, the final model included all the main effects and the three interaction terms. The results from type 3 analysis of the effects for this model are presented in Table 3.1.

Table 3. 1: Type 3 analysis of effects for the logistic model

Effect	DF	Wald Chi-square	p-values
Age	1	2.8226	0.0929
Gender	1	4.6331	0.0314
Educational attainment	2	4.0392	0.1327
Treatment site	1	19.339	<.0001
Living with/without partner	1	0.1218	0.7271
Contribution to household income	1	6.0316	0.0141
WHO staging of disease	3	2.1298	0.5459
Baseline CD4+ count (cells/μL)	1	5.4553	0.0195
Baseline weight (kg)	1	0.0440	0.8338
Reason for taking HIV test	2	4.2281	0.1207
Household access to tap-water	1	0.0975	0.7548
Household access to electricity	1	0.0021	0.9632
Cell-phone ownership	1	6.1334	0.0133
Age*cell-phone ownership	1	5.4662	0.0194
Gender*reason for taking the test	2	8.5540	0.0139
Treatment site*household income	1	7.5991	0.0058

Before the final model was accepted, diagnostics were performed to see whether it fits the data well. The goodness-of-fit was assessed using the Hosmer-Lemeshow test. The observed and expected frequencies are given in Table 3.2. The goodness-of-fit statistic was 6.45 with 8 degrees of freedom and the corresponding p-value of 0.597. The very large p-value for this test shows that the model fits the data well (i.e. the predicted probabilities correspond with the observed values).

Table 3. 2: Partition for the Hosmer-Lemeshow Goodness-of-Fit test

Group	Total	Event = adherent		Non-event = non adherent	
		Observed	Expected	Observed	Expected
1	54	23	25.27	31	28.73
2	54	34	34.56	20	19.44
3	54	40	39.39	14	14.61
4	54	46	42.82	8	11.18
5	54	44	44.65	10	9.35
6	54	49	46.01	5	7.99
7	54	46	47.73	8	6.27
8	54	47	49.04	7	4.96
9	54	52	50.02	2	3.98
10	54	53	54.52	4	2.48

The appropriateness of the link function was tested by refitting the model with a linear predictor and its square (i.e. the squared linear predictor) as independent variables and the results are given in Table 3.3.

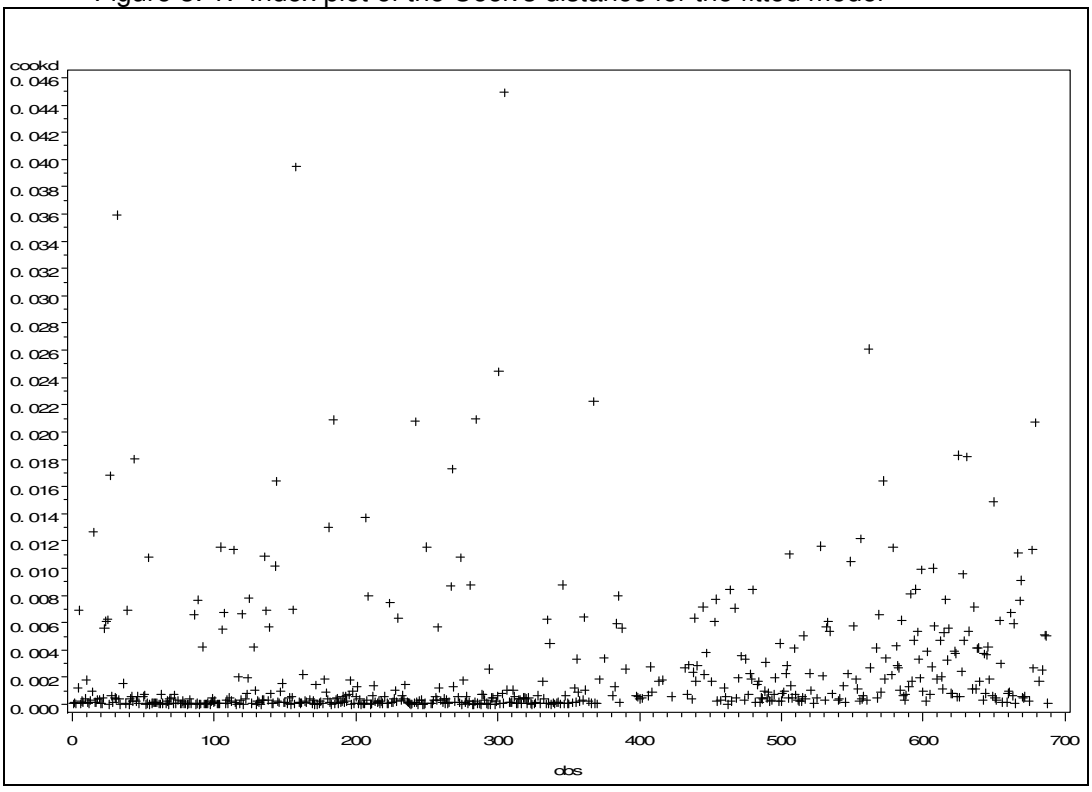
Table 3. 3: The Logit link function test

Variable	DF	Chi-square	p-value
Constant	1	0.0104	0.919
Linear predictor	1	67.609	<.0001
Squared linear predictor	1	0.0643	0.800

The very small p-value for the linear predictor and a very large *p*-value for the squared linear predictor variables in Table 3.3 suggest that the link is appropriate, and this confirms the goodness-of-fit test that the model fits the data well.

Influential observations were assessed by plotting the Cook's distance statistic against the observations and this index plot is presented in Figure 3.1.

Figure 3. 1: Index plot of the Cook's distance for the fitted model



Note: 'Cookd' = Cook's distance and 'obs' = observation number

The plot indicates that none of the Cook's distance values for the fitted model are greater than 1, which suggests that there are no observations with undue influence on parameter estimates. To confirm this result, the three observations with the largest Cook's distance values were further investigated by re-fitting the model with each one of them deleted one at the time (referred to as a single-case deletion). These observations were numbers 32, 158 and 305 with Cook's distance values of 0.0359257227, 0.0394773227 and 0.0449326732 respectively. When these numbers were deleted one at a time, the results were the same as those obtained from fitting the model to the full data set. This reaffirms that these three observations with the largest Cook's distance values do not have undue influence on the parameter estimates of the model.

The final chosen model is as follows:

$$\begin{aligned} logit\,[P(y_i = 1 \mid \mathbf{x}_{ij}) = {} & \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 educ\_no_i + \beta_4 educ\_p_i + \beta_5 site_i \\ & + \beta_6 partner_i + \beta_7 income_i + \beta_8 WHOstage\ 1_i + \beta_9 WHOstage\ 2_i + \beta_{10} WHOstage\ 3_i \\ & + \beta_{11} base\_cd4_i + \beta_{12} baseweight_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i \\ & + \beta_{15} tapwater_i + \beta_{16} electricity_i + \beta_{17} cellphone_i + \beta_{18} age_i cellphone_i \\ & + \beta_{19} gender_i reason\_no_i + \beta_{20} gender_i reason\_risk_i + \beta_{21} site_i income_i . \end{aligned}$$

The parameter estimates and the corresponding confidence limits for the above model are presented in Table 3.4. Gender, treatment site, source of income, cell phone ownership and baseline CD4+ cell count are all found to be significant main effects. There are three significant interaction terms: between age and cell phone ownership; between gender and reported reason for taking an HIV test; and between treatment site and source of household income. All significant main effects, except baseline CD4+ cell count, were involved in significant interaction terms (Table 3.4). For a unit (cells/μL) increase in CD4+ cell count, the odds of HAART adherence decreased by 5% (adjusted Odds ratio [aOR = 0.995 ($e^{-0.005}$), p=0.019]).

Table 3. 4: Parameter estimates for the chosen logistic regression model

Effect	Estimates	Std errors	p-value	95% C.I.	
				Lower	Upper
Intercept	-0.862	1.150	0.454	-3.116	1.392
Age	0.046	0.028	0.093	-0.008	0.101
Gender (ref=male)					
female	0.712	0.331	0.031	0.064	1.361
Education (ref=sec & higher)					
No schooling	-0.520	0.390	0.182	-1.284	0.243
Primary	0.394	0.365	0.280	-0.321	1.110
Treatment site (ref=rural)					
Urban	1.470	0.334	<.0001	0.826	2.125
Living with partner (ref=no)					
Yes	-0.096	0.275	0.727	-0.635	0.443
Income (ref=not source)					
Source of income	1.342	0.547	0.014	0.271	2.414
WHO staging (ref=stage4)					
Stage1	0.454	0.655	0.488	-0.829	1.737
Stage2	-0.187	0.562	0.739	-1.289	0.915
Stage3	-0.151	0.506	0.766	-1.143	0.842
Baseline CD4+ count	-0.005	0.002	0.019	-0.008	-0.001
Baseline weight	0.002	0.010	0.834	-0.018	0.022
Reason for testing (ref=unwell)					
No specific reason	-0.488	0.417	0.242	-1.304	0.329
Exposed to the risk	1.200	0.830	0.148	-0.427	2.826
Access to tapwater (ref=yes)					
No	0.141	0.455	0.755	-0.755	1.037
Househld with electricity (ref=yes)					
No	-0.019	0.408	0.963	-0.819	0.781
Cell phone ownership (ref=yes)					
No	2.744	1.108	0.0133	0.572	4.915
Age*cell-phone (ref=yes)					
No cell phone	-0.076	0.033	0.019	-0.140	-0.012
Gender*reason (ref=male & unwell)					
Female*no specific reason	0.879	0.584	0.132	-2.266	2.024
Female*exposed to the risk	-1.921	0.895	0.0318	-3.674	-0.160
Site*income (ref=rural & not source)					
Urban*source of income	-1.756	0.637	0.0058	-3.004	-0.507

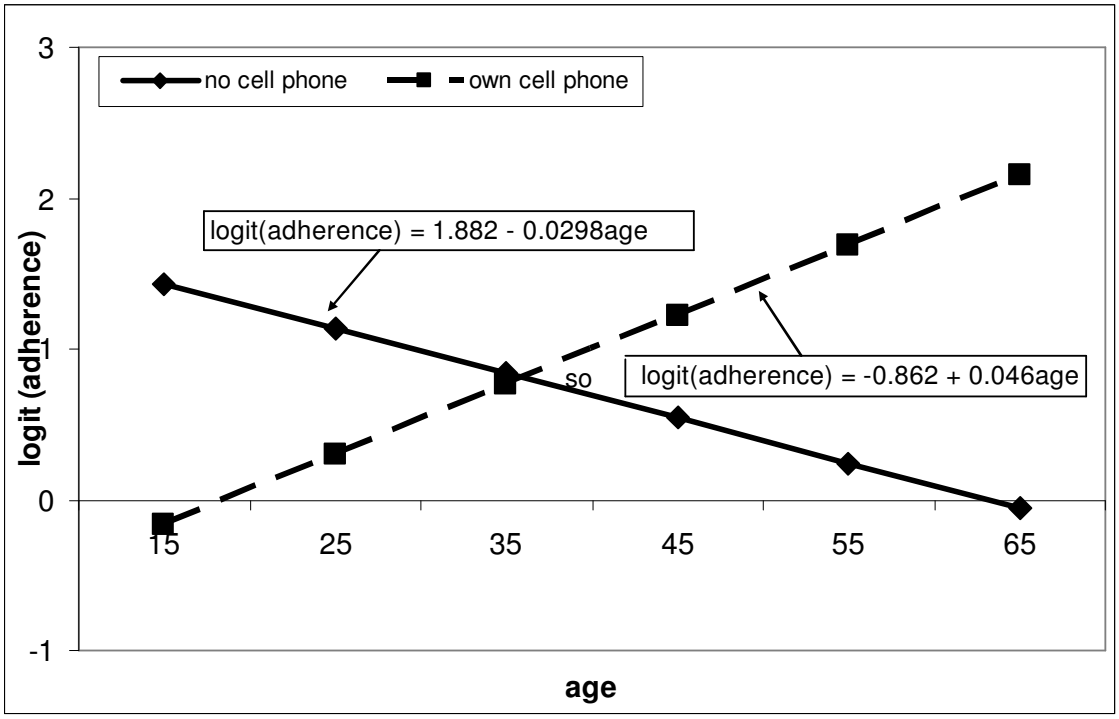
The interaction effects are presented below. It should be noted that for the interaction terms that involve two categorical variables, the meaningful odds ratios for comparison need to be further calculated from the parameter estimates in Table 3.4. Thus post-hoc effects of interactions between gender and reported reason for taking an HIV test as well as between treatment site and source of household income are reported in Tables 3.5 and 3.6 respectively.

a) Interaction between patient age and cell phone ownership

As age of patients increased, optimal HAART adherence was less likely for patients without cell phones than those with cell phones [aOR = 0.927 ($e^{-0.076}$), p-

value=0.019] (Table 3.4). More specifically, optimal HAART adherence rate increased with age for patients with cell phones whereas it decreased as age increased for patients without cell phones (Figure 3.2). Figure 3.2 further shows that the gap in optimal adherence between patients with and without cell phones widened with increasing age.

Figure 3. 2: Log odds ratio associated with optimal HAART adherence and age for patients with and without cell phones



b) Interaction between gender and the reported reason for taking an HIV test

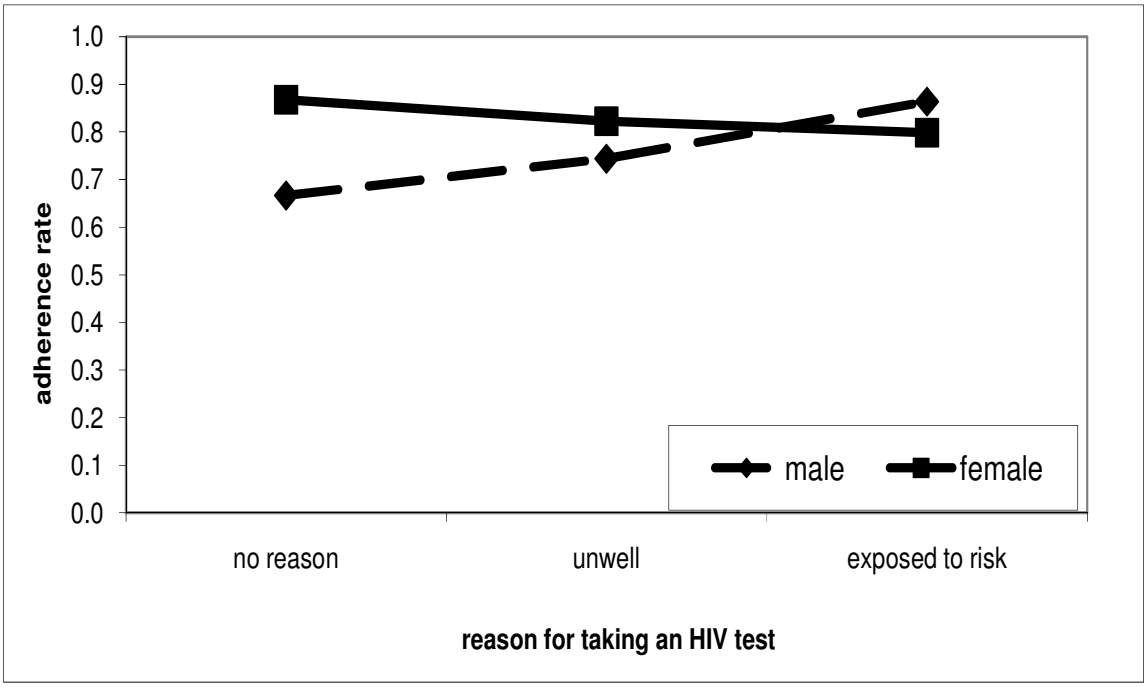
Optimal HAART adherence was significantly higher for females than males with patients who reported no specific reason for taking an HIV test [aOR = 4.911, p-value=0.001] as well as those who reported to have tested because they were not well [aOR = 2.039, p-value=0.031] (Table 3.5). There was however, no significant difference in optimal HAART adherence between females and males who reported to have tested for HIV because they felt exposed to the risk of contracting the disease [aOR = 0.299, p-value=0.145] (Table 3.5). It is also shown that for males, HAART adherence was significantly lower for patients who reported no specific reason for taking an HIV test than those who tested because they felt exposed to risk of contracting the HIV disease [aOR = 0.185, p-value=0.049] (Table 3.5). These

results confirm the observed proportions of optimal HAART adherence for gender classified by reported reason for taking an HIV test depicted in Figure 3.3.

Table 3. 5: Post-hoc effects of the interaction between gender and reported reason for taking an HIV test (adjusted odds ratio [AOR] with 95% confidence interval [CI])

Interaction effect	estimate	aOR	p-value	95% CI (OR)	
				lower	Upper
Women versus men					
No specific reason	1.5915	4.911	<.0001	1.892	12.75
Risk of exposure	-1.2073	0.299	0.145	0.059	1.519
Unwell	0.7125	2.039	0.031	1.066	3.900
Men					
No specific reason versus risk of exposure to HIV	1.6874	0.185	0.0429	0.035	0.993
No specific reason versus unwell	-0.4878	0.614	0.242	0.271	1.389
Risk of exposure to HIV versus unwell	1.1997	3.319	0.148	0.653	16.88
Women					
No specific reason versus risk of exposure to HIV	-0.8871	0.446	0.383	0.073	2.739
No specific reason versus unwell	0.3914	1.479	0.345	0.657	3.330
Risk of exposure to HIV versus unwell	1.1997	3.319	0.148	3.319	16.88

Figure 3.3: Percentage adherence associated with gender and reported reason for taking an HIV test (based on observed proportions of optimal HAART adherence)



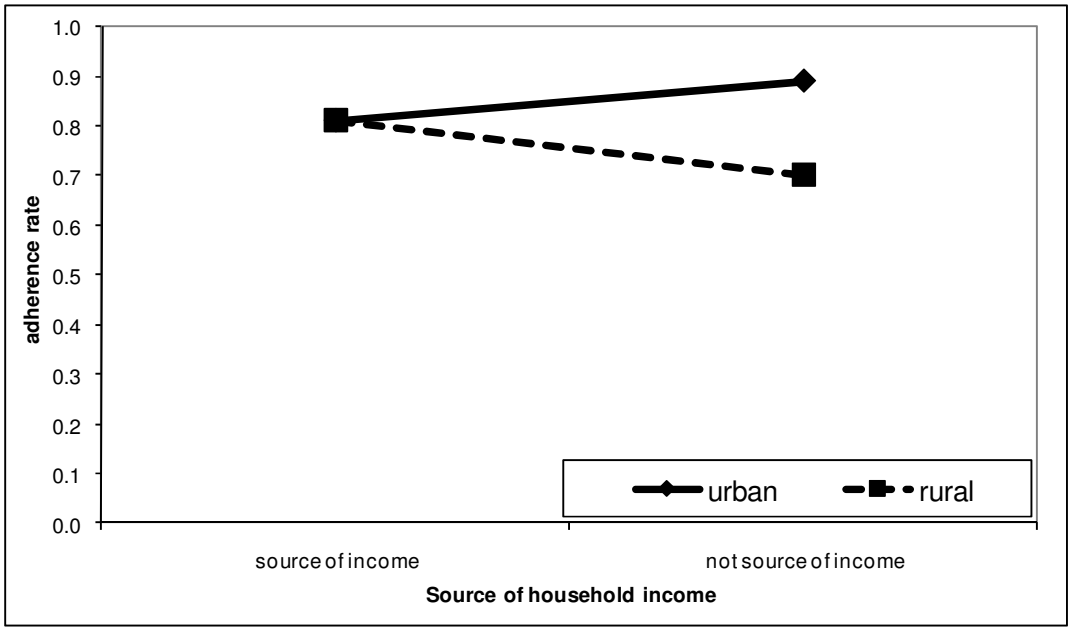
c) Interaction between treatment site and whether patient is a source of household income

Optimal HAART adherence was significantly higher in the urban treatment site than in the rural treatment site for patients who were not sources of household income [aOR = 4.347, p-value=<0.001] (Table 3.6), whereas, with patients who were sources of household income, there was no difference between treatment sites [aOR = 0.751, p-value=0.628] (Table 3.6). For the rural treatment site, optimal HAART adherence was significantly higher for patients who were a source of household income than those who were not a source of household income [aOR = 3.828, p-value=0.014] (Table 3.6). These results confirm the observed proportions of optimal HAART adherence for treatment site classified by contribution to household income depicted in Figure 3.4.

Table 3. 6: Post-hoc effects of the interaction between HAART treatment site and patient's contribution to household income (adjusted odds ratio [aOR] with 95% confidence interval [CI])

Interaction effect	estimate	aOR	p-value	95% CI (OR)	
				lower	Upper
Urban vs rural treatment site					
Source of household income	-0.2863	0.751	0.628	0.237	2.385
Not source of household income	1.4694	4.347	<.001	2.258	8.369
Source vs not source of household income					
Urban treatment site	-0.4125	0.662	0.219	0.342	1.278
Rural treatment site	1.3423	3.828	0.014	3.828	11.17

Figure 3. 4: Percentage adherence associated with treatment site and whether a patient is a source of household income (based on observed proportions of optimal HAART adherence)



3.6 Summary

The generalized linear models using logistic regression provided a tool for assessing factors that affect initial HAART adherence. The results, revealed three significant two-way interaction terms between a) age and cell phone ownership, b) gender and reported reason for taking an HIV test, as well as c) treatment site and household source of income. The baseline CD4+ count was also shown to be negatively associated with optimal HAART adherence. Maqutu et al., (2010a) provides a full discussion of these results (attached in the Appendix A). The next step leads us to survey statistical longitudinal approaches for binary responses to be used for the evaluation of adherence data in order to gain greater insight into the long-term predictors of optimal HAART adherence. Specifically, we begin by reviewing marginal models for longitudinal settings and use them to evaluate the explanatory variables of long-term optimal HAART adherence in Chapter 4.

Chapter 4
Marginal models application to treatment adherence

4.1 Introduction

One of the aims of this study is to identify predictors of long-term optimal HAART adherence, and assess whether factors affecting initial adherence also influence long-term HAART adherence. The result is that a longitudinal study and a straight forward application of GLMs is no longer appropriate because of a lack of independence among observations obtained from the same subject. It is therefore, necessary to extend these models in order to account for correlated observations. For the generalized outcomes in a longitudinal study, one must distinguish between three broad model families, namely, marginal models, subject-specific models and transition models. A marginal model is one in which marginal probabilities of the response are directly modeled. That is, responses are modeled marginalized over all other responses, and the association structure is typically captured using a set of association parameters, such as correlations, odds ratios, etc. These models are also called population-averaged models because the parameters characterize the marginal expectation. The marginal models include an extensive range of possibilities, especially, those that use methods based on likelihood such as the *Bahadur model*, or log-linear models (Molenberghs and Verbeke, 2005). In this thesis, we will focus on marginal models in which the dependence between observations within each individual is modeled using correlation structure matrices, and generalized estimating equations (GEE) to obtain estimates. Their theoretical aspects and characteristics as well as their application to HAART adherence are the focus of this chapter.

4.2 Marginal models for longitudinal data

When inferences based on the mean parameters or population-average are of interest, a marginal model is adequate to analyze the data originating from a longitudinal study. A marginal model specifies the mean of the response variable, or marginal expectation and the association structure separately. The specification of a marginal model can be summarized as follows (Diggle et al., 2002; Fitzmaurice et

al., 2004). Consider the outcome, y_{ij} for the i^{th} subject in the j^{th} measurement. The general form of the model relating to the response and its mean can be defined as

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}.$$

Then, a marginal model can be formulated as follows:

a) The marginal expectation of the response variable $\mu_{ij} = E(y_{ij})$ and a linear combination of the covariates are related by means of

$$g(\mu_{ij}) = g[E(y_{ij})] = \mathbf{x}'_{ij}\beta,$$

where

y_{ij} is the response for subject i at time j,

$\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates associated with subject i at time j,

β is a $p \times 1$ vector of unknown regression coefficients, and

$g(.)$ is a known link function.

b) The marginal variance of y_{ij} is described as a function of the mean

$$var(y_{ij}) = \phi v(\mu_{ij}),$$

where

$v(\cdot)$ is the variance function,

ϕ is the dispersion parameter

c) The correlation between observations on the same subject, $corr(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$ depends on the marginal means, and on a parameter vector α. This correlation is modeled using a correlation structure matrix called a 'working correlation matrix' $\mathbf{R}_i(\alpha)$ for each $\mathbf{y}_i$. It is assumed that the correlation matrix $\mathbf{R}_i(\alpha)$ depends on a vector of association parameters, α and is also assumed to be the same for all subjects.

The first two components of the marginal model specification correspond to the GLM for independent data discussed in Chapter 3 but without distributional assumptions about the responses. The extension of longitudinal data is represented by the third component which incorporates the correlation among the repeated responses. Note that full distributional assumptions about the vector of the responses can still be

made with marginal models such that the likelihood can be specified and the method of maximum likelihood can be applied for estimation and inference. The association parameters from such models are however constrained by the marginal probability, and are dependent on high order associations that are usually not of interest (Pendergast et al., 1996). That is, unlike a continuous response that has a multivariate normal distribution where the joint distribution of responses is fully specified by the mean, variance and the correlations; this is not the case with discrete data. Instead, the joint distribution requires the specification of the mean vector, two-way associations, as well as three- and higher- way associations among repeated responses (Fitzmaurice et al., 2004; Molenberghs and Verbeke, 2005). As a result, the number of association parameters increases as the number of responses increase which will often far exceed the number of subjects enrolled in the study. For instance, suppose $\mathbf{y}_i$ is a vector of binary responses with $n_i = 10$, the joint distribution of $\mathbf{y}_i$ has 1, 013 $(2^{10} - 10 - 1)$ two-way, three-way, four-way and higher-way association parameters. In general, specification of joint distribution for discrete longitudinal data is difficult, even in cases where it might be possible to specify the joint distribution, the likelihood can be too complicated to evaluate.

Alternatively, the distributional assumptions about the vector of responses can be avoided and a marginal model based only on the mean response, the variances and pairwise associations can be specified. This leads to the method of generalized estimating equations.

4.3 Generalized Estimating Equations (GEE)

When interest is in the first-order marginal parameters, McCullagh and Nelder (1989) have shown that a full likelihood procedure can be replaced by quasi-likelihood based methods. Wedderburn (1974) shows that the likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood estimating equations provide consistent estimates of regression parameters β in any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families. Consequently, Liang and Zeger (1986) proposed the method of generalized estimating equations (GEE) as an extension of GLM to accommodate correlated data using quasi-likelihood approach. Rather than

assuming a particular distribution for the response, GEE method requires a correct specification of the mean as well as how the variance depends on the mean. One of the desirable properties of the GEE method is that it yields consistent and asymptotically normal solutions even with the misspecification of the covariance structure (Liang and Zeger, 1986; Zeger and Liang, 1986; Davis, 2002). The covariance structure is treated as a nuisance.

For the ith subject, let $\mathbf{A}_i$ be the $n_i \times n_i$ diagonal matrix with $v(\mu_{ij})$ as the jth diagonal element. Also let $\mathbf{R}_i(\alpha)$ be the $n_i \times n_i$ 'working' correlation matrix for the ith subject. Then the working variance-covariance matrix of $\mathbf{y}_i$ is given by

$$\mathbf{V}_i = \phi \mathbf{A}_i(\beta)^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i(\beta)^{1/2}. \qquad (4.1)$$

Then GEE estimate of β is the solution of

$$\sum_i \mathbf{D}_i' (\mathbf{V}_i)^{-1} (\mathbf{y}_i - \mu_i) = \mathbf{0}, \qquad (4.2)$$

where $\mathbf{D}_i = \partial \mu_i / \partial \beta$. Note that (4.2) is an extension of the estimating equations for β in GLM, which is given by (3.5) where the variance-covariance matrix of $\mathbf{y}_i$ can be re-written in the form

$$\mathbf{V}_i = \phi \mathbf{A}_i(\beta)^{1/2} \mathbf{I}_{n_i} \mathbf{A}_i(\beta)^{1/2}. \qquad (4.3).$$

That is, an extension of (3.5) that would account for the correlation is obtained by replacing the identity matrix $\mathbf{I}_{n_i}$ in (4.3) by a correlation matrix $\mathbf{R}_i(\alpha)$ to obtain (4.1). Note that although $\mathbf{A}_i(\beta)$ follows directly from the marginal mean model, β commonly contains absolutely no information about $\mathbf{R}_i$ hence the reason why $\mathbf{R}_i$ had to be parameterized by an additional parameter vector: $\mathbf{R}_i = \mathbf{R}_i(\alpha)$. Thus while the first moment completely specifies the second and higher order moments, this is only partially so in the correlated data setting, variances are still specified by the marginal means but the correlations are not. This demonstrates the key difference between the correlated data and their univariate counterparts in a generalized linear model setting.

Recall that our aim is to restrict model specification to the first moments only but we are faced with the second moments. If we would model the second moments, we

would have to address the third and fourth moment as well. This then leads to the full specification of the joint distribution which we have been trying to avoid. In order to overcome this problem, Liang and Zeger (1986) suggest that while still acknowledging the need for $\mathbf{R}_i(\alpha)$ in $\mathbf{V}_i$ (4.1), one is allowed to specify an incorrect structure or the so called 'working correlation' matrix.

The estimating equations given by (4.2) depends on unknown parameters β, α and ϕ. Liang and Zeger (1986) propose moment-based estimates for α and ϕ. These are estimated using functions of standardized Pearson residuals

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}. \qquad (4.4)$$

Some of the popular choices for the working correlations and their moment based estimators are presented in Table 4.1. Similarly, the dispersion parameter can be estimated by

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}^2 . \qquad (4.5)$$

Table 4. 1: Popular choices of working correlation assumptions in standard GEE and moment based estimators

Structure	$Corr(y_{ij}, y_{ik})$	Estimator
Independence	0	-
Exchangeable	α	$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq k} e_{ij}e_{ik}$
AR(1)	$\alpha^{\lvert j-k\rvert}$	$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(n_i-1)}\sum_{j\leq n_i-1} e_{ij}e_{i,j+1}$
Unstructured	α_{jk}	$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} e_{ij}e_{ik}$

Some of the commonly used matrices of $\mathbf{R}_i(\alpha)$ under each covariance structure are illustrated by considering a study with 4 measurements $(n_i = 4)$ taken through time. The $\mathbf{R}_i(\alpha)$ for each subject i are given by the following:

Independence structure:

$$\mathbf{R}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The independence model adopts the working assumption that repeated observations for a subject are independent. In this case the GEE simplifies to the GLM estimating equations in (3.5).

Exchangeable or Compound Symmetry structure:

$$\mathbf{R}_i = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}.$$

The exchangeable working correlation specification assumes a constant correlation between any two measurements within a subject, regardless of the time interval between the measurements. This structure may not be appropriate in a longitudinal study where the correlations are expected to decay with increasing separation in time (Fitzmaurice et al., 2004). The exchangeable structure may be more appropriate with data sets that have clustered observations such that there may be no logical ordering of observations within a cluster (Horton and Lipsitz, 1999).

First-order auto-regressive (AR-1) structure:

$$\mathbf{R}_i = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}.$$

The AR-1 correlation structure depends on the distance between the measures. The correlations decline over time as the separation between pairs of repeated measures increases. This structure assumes that the measurement occasions are equally spaced (Molenberghs and Verbeke, 2005; Fitzmaurice et al., 2004).

m-dependent structure:

$$\mathbf{R}_i = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_1 & 1 & \alpha_1 & 0 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}.$$

With the m-dependent structure, the correlations depend on the distances between measures; eventually they diminish to zero for $t \geq m$. With mistimed measurements, this structure may be reasonable to consider since the correlation is a function of the time between observations (Stokes, Davis and Koch, 2000; Horton and Lipsitz, 1999).

Unstructured:

$$\mathbf{R}_i = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{pmatrix}.$$

When the correlation matrix is completely unspecified, there are $n_i(n_i - 1)/2$ parameters to be estimated. It provides the most efficient estimator for β but is useful only when the number of observation times is relatively few (Stokes et al., 2000). One of the disadvantages of assuming the unstructured covariance is that the number of parameters to be estimated increases with the number of measurement occasions. Consequently, the estimates tend to be unstable when the number of covariance parameters to be estimated is large relative to the sample size (Fitzmaurice et al., 2004). That is, with too many covariance parameters to be estimated from the limited amount of data available affects negatively the precision with which the regression parameters of interest will be estimated. Moreover, when there are missing data and/or irregularly measured occasions, an estimation of a complete correlation structure may result in a non-positive definite matrix and parameter estimation may not proceed (Stokes et al., 2000).

In order to estimate β from the score equations in (4.2), estimates of α and ϕ are required whereas, the moment based estimates for α (Table 4.1) and the

expression (4.5) for ϕ depend on β (since $e_{ij} = e_{ij}(\beta)$ through $\mu_{ij} = \mu_{ij}(\beta)$ and therefore also through $v(\mu_{ij})$). In order to overcome this circularity stumbling block, an iterative procedure is used for estimation. Therefore, Liang and Zeger (1986) suggest computing the GEE estimates using a standard iterative procedure as follows:

i. Compute initial estimates for β using a univariate GLM, that is, assuming independence among the n_i responses for subject i.
ii. Compute standardized residuals e_{ij} using (4.4).
iii. Compute estimates for α (some examples are given in Table 4.1).
iv. Compute an estimate of ϕ using (4.5).
v. Compute $\mathbf{R}_i(\alpha)$ under a given assumption of the correlation structure.
vi. Compute $\mathbf{V}_i(\beta, \alpha, \phi) = \phi \mathbf{A}_i(\beta)^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i(\beta)^{1/2}$.
vii. Update the estimate for β:

$$\beta^{(t+1)} = \beta^{(t)} - \left[\sum_{i=1}^{N} \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}' \right]^{-1} \times \sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mu_i).$$

viii. Repeat 2 – 7 until convergence is reached.

Currently, there are many statistical software packages for fitting GEE models for longitudinal data, such as SAS PROC GENMOD procedure, gee and geepack R libraries, STATA, SPSS and others.

Let $\hat{\beta}$ be the solution to the generalized estimating equations (4.2), then $\hat{\beta}$ has the following properties:

a). $\hat{\beta}$ is a consistent estimator of β. That is, with very high probability, $\hat{\beta}$ is close to the population regression parameters β in large samples (Fitzmaurice et al., 2004). Note that $\hat{\beta}$ has a very appealing robustness property of producing a consistent estimate of β even if the working correlation matrix is specified as long as the model for the mean response has been correctly specified.

b). The estimator $\hat{\beta}$ is asymptotically normally distributed with mean β and asymptotic variance-covariance matrix:

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} \mathbf{D}'\mathbf{V}_i^{-1}\mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{D}_i'\mathbf{V}_i^{-1} var(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i \right) \left(\sum_{i=1}^{N} \mathbf{D}'\mathbf{V}_i^{-1}\mathbf{D}_i \right)^{-1}, \qquad (4.6)$$

which can be presented notationally as follows

$$var(\hat{\beta}) = \mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1}.$$

The $var(\mathbf{y}_i)$ in (4.6) is typically replaced by $(\mathbf{y}_i - \mu_i)(\mathbf{y}_i - \mu_i)'$. The variance estimator, $var(\hat{\beta})$ expressed by (4.6) is commonly referred to as the sandwich or empirical or robust estimator. Regardless of whether or not the working correlation structure is correct, the point estimates and standard errors based on (4.6) are asymptotically correct (Molenberghs and Verbeke, 2005). If the variance matrix $\mathbf{R}_i$ is correctly specified, then the variance-covariance matrix $var(\hat{\beta})$ reduces to

$$var(\hat{\beta}) = \left(\sum_{i=1}^{N} \mathbf{D}'\mathbf{V}_i^{-1}\mathbf{D}_i \right)^{-1} = \mathbf{M}_0^{-1}, \qquad (4.7)$$

and is usually referred to as a 'model-based' estimator. With the misspecification of the covariance structure, the standard errors based on the model based estimator (4.7) are not valid and therefore cannot be used. In fact, Molenberghs and Verbeke (2005), suggest that model based standard errors can be used as an indication of the 'distance' between the working assumptions for the correlation and the true structure. When both standard errors are far apart, this can be seen as an indication of a poor choice of working correlation assumptions.

4.4 Model selection

Model selection is an essential part of any practical data analysis. A common challenge encountered by many researchers dealing with regression is the selection of variables to be included in the model. That is, faced with a large number of covariates that include higher order terms, the challenge is selecting a subset to be included in the regression model. For instance, in observational studies, excluding some important risk factors (i.e. confounders) may result in misleading estimates of the effects of other risk factors while on the other hand, including all covariates as

well as higher order terms may lead to a too complex model which would be difficult to interpret and with less precise parameter estimates (Pan, 2001).

With repeated measures data, model selection becomes even more complicated because one does not only deal with variable selection but also a selection of a covariance structure among many available structures to account for the fact that data might be correlated. For instance, it has been noted earlier that the GEE method offers asymptotically consistent parameter estimates even if the covariance structure of repeated measurements is not correctly specified (Liang and Zeger, 1986; Fitzmaurice et al, 2004). However, if the working correlation matrix is correctly specified, the resulting parameter estimates are efficient (Hardin and Hilbe; 2003). In other words, a poor working correlation assumption is not wrong but may hinder efficiency of the obtained estimators, thus there is a need to choose a working correlation that approximates as much as possible the true structure. This clearly poses a model-selection challenge in selecting the working correlation structure for GEE models.

There are a number of techniques available to compare models based on the likelihood and asymptotic properties of maximum likelihood estimator (MLE). These include the likelihood ratio test (LRT) and Akaike's Information Criterion (AIC) (Akaike, 1974) for nested and non-nested models respectively. Since the GEE method is based on quasi-likelihood procedures, there is no associated likelihood underlying the model and thus LRT and AIC are not directly applicable. To overcome this problem, a selection criterion termed 'quasi-likelihood under the independence model criterion' (QIC) was proposed by Pan (2001) to select the best fitting model in the GEE analysis. The QIC is an extension of AIC measure to GEE. Since in general, the GEE estimator has different asymptotic properties from those of the MLE, a modification to the penalty term in the usual AIC is necessary.

The following summarizes QIC (Pan, 2001) starting with a brief review of the quasi-likelihood approach. Given a generalized linear model, $g(\mu) = \mathbf{x}_i'\beta$, the quasi-likelihood takes the following form (McCullagh and Nelder, 1989)

$$Q(\mu) = \int_y^{\mu} \frac{y-t}{\phi v(t)} dt ,$$

where $\mu = E(y)$ and $var(y) = \phi v(\mu)$ with ϕ being the dispersion parameter. Examples of the variance function $v(\mu)$ for the commonly used distributions and the quasi-likelihood $Q(\mu)$ that follows are presented in Table 4.2.

Table 4.2: The variance function and quasi-likelihood of commonly used distributions in the exponential family

Distribution	$v(\mu)$	$\phi Q(\mu)$
Normal	1	$-\frac{1}{2}(y-\mu)^2$
Bernoulli	$\mu(1-\mu)$	$y\ln[\mu/(1-\mu)] + \ln(1-\mu)$
Poisson	μ	$y\ln(\mu) - \mu$

The Akaike information criterion may be calculated based on (ML or REML) log-likelihood, LL of a fitted model as follows

$$AIC = -2LL + 2p ,$$

where p is the total number of parameters being estimated in the model. The AIC penalizes the fit of a model for the number of parameters being estimated by adding $2p$ to the log-likelihood. The QIC is a modification to AIC for the GEE method in that the likelihood function value of the AIC is replaced by the quasi-likelihood function value obtained under independence correlation structure and the penalty term is adjusted. It is defined as

$$QIC = -2Q(\hat{\mu}; \mathbf{I}) + 2\, trace\left(\Omega_{\mathbf{I}}^{-1}\hat{\mathbf{V}}_{\mathbf{R}}\right), \qquad (4.8)$$

where $\mathbf{I}$ represents the independent correlation structure and $\mathbf{R}$ is the specified working correlation structure (Pan, 2001). The p-dimensional matrices $\hat{\Omega}_{\mathbf{I}}^{-1}$ and $\hat{\mathbf{V}}_{\mathbf{R}}$ are variance estimators of the regression coefficients under the correlation structures $\mathbf{I}$ and $\mathbf{R}$ respectively. The QIC value is computed based on the quasi-likelihood estimate $\hat{\mu}$. Here $Q(.)$ is the quasi-likelihood function and examples from different distributions are presented in Table 4.2. To calculate the QIC value, one needs to run the GEE model twice, one with the independent correlation structure $\mathbf{I}$ and the other with the correlation structure $\mathbf{R}$ in order to obtain the generalized least

squares estimator $\hat{\Omega}_{\mathbf{I}}^{-1}$ and the empirical variance estimator $\hat{\mathbf{V}}_{\mathbf{R}}$ (Cui and Feng, 2009; Cui and Qian; 2007).

The QIC value in (4.8) can be used to select the best fitting correlation structure as well as the best model in terms of the subset of variables to be included in the model. The best correlation is usually selected first and is done based on the full model with all the explanatory variables (Hardin and Hilbe, 2003). The correlation structure with the smallest QIC value is selected as the most optimal correlation structure. Then, based on the chosen correlation structure, the mean response model with the smallest value of QIC is regarded as the most appropriate GEE model.

Note that in general, the choice of the correlation structure should be guided by the subject matter, for instance, there are time dependent correlation structures (e.g autoregressive) and those that are not (e.g exchangeable). Then, if there are competing correlation structures, the QIC measure can be used to determine the appropriate one (Hardin and Hilbe, 2003).

Also of note is the fact that although the QIC method was published in 2001 and included in the book by Hardin and Hilbe (2003), few applications of this method have been published. One possible reason is that the issue of model selection in GEE has been largely neglected (Pan, 2001; Cantoni, Flemming and Ronchetti, 2005). The other possible reason is that perhaps no commercial software such as SAS, Stata, SPSS or SPLUS had a program to calculate the QIC value at the time when Pan's (2001) article was published (Cui and Qian, 2007). Nonetheless, in recent years there seem to be an increasing use of QIC for selection of the best model in GEE. Examples include Kuchibhatla and Fillenbaum (2003), Ballinger (2004), Martus et al (2004), Hwang and Takane (2005), Cui and Qian (2007), Lin and Chen (2009) and Cui and Feng (2009). Again in most commercial software that includes SAS and Stata, calculation of the QIC has been implemented.

The generalized Wald tests can be used to compare models with different subsets of the regression parameters. That is, one can use the generalized Wald tests to test

the joint null hypothesis that a set of regression parameters βs are equal to zero (Hedeker and Gibbons 2006). In general, for any matrix $\mathbf{L}$ a test for hypothesis can be written as follows

$$H_0 : \mathbf{L}\beta = \mathbf{0} \quad \text{versus} \quad H_A : \mathbf{L}\beta \neq \mathbf{0},$$

where $\mathbf{L}$ is a $q \times p$ indicator matrix of ones and zeros. Here, p is equal to the number of regressors in the full model (including the intercept) and q equals the number of parameters in the generalized Wald test (that is, the difference in regressors between the full and reduced models). The Wald statistic is a quadratic form defined as follows

$$\mathbf{W}^2 = \hat{\beta}'\mathbf{L}'\left(\mathbf{L}var(\beta)\mathbf{L}'\right)^{-1}\mathbf{L}\hat{\beta}, \qquad (4.9)$$

and is distributed as χ^2 with q degrees of freedom under the null hypothesis. With a single regression coefficient to be tested, the dimension of matrix $\mathbf{L}$ is $1 \times p$ thus the equation (4.9) reduces to

$$\mathbf{W}^2 = \left[\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}\right]^2,$$

with one degree of freedom.

4.5 Evaluation of HAART adherence predictors using the GEE method

The aim is to determine the predictors of long-term optimal adherence, and evaluate whether factors affecting first month (initial) adherence also influence long-term HAART adherence. Letting $y_{ij} = 1$ if the ith patient is classified adherent to medication at the jth follow-up visit, and $y_{ij} = 0$ otherwise, we assume that the marginal probability of adherence at each visit follows the logistic model

$$logit(\mu_{ij}) = \mathbf{x}'_{ij}\beta,$$

where $\mu_{ij} = E(y_{ij}) = P(y_{ij} = 1)$, $\mathbf{x}'_{ij}$ is a vector of explanatory variables and β is a vector of unknown parameters. In addition to the 13 baseline variables used for the evaluation of initial HAART adherence (Chapter 3), optimal HAART adherence status at baseline was now used as a covariate. Moreover, the other additional

variables included weight at every visit and time (follow-up visits). They have all been described in Chapter 2. This specifies the first component of the marginal model, the model for the mean response. Next we assume that

$$var(y_{ij}) = \mu_{ij}(1-\mu_{ij}),$$

which specifies the second component, the variance function and known scale parameter ($\phi_j = 1, j = 1,\ldots,17$). Moreover, an assumed correlation structure between observations on the same subject has to be chosen.

Since model selection criteria for the mean response depend upon the correct specification of the model for the covariance, the first step is to choose a suitable model for the covariance. Because the observations are measured repeatedly on the same subject over time, the choice of the working correlation is based on the correlation structures that have time dependence. For this reason, the *m*-dependent structure and the Autoregressive (AR-1) structure were used. The two correlation structures were also contrasted with the independence structure where the assumption is made that the repeated observations for a subject are independent. The different correlation structures were fitted using the same mean structure composed of all the main effects. Then QIC (Pan 2001) was used to select the model with the best fitting correlation structure and the results are as follows:

Correlations	QIC
Independent	22328.3090
AR-1	22312.3316
1-dependent	22318.9656
2-dependent	22317.1215
3-dependent	22318.6152

The QIC measure leads to the selection of AR-1 correlation structure since it has the smallest QIC value.

The next step is to evaluate the mean structure. In order to check whether factors that affect baseline adherence are still important when adherence is studied over time, a model with all the main effects was compared with a model that contained all

the main effects plus the three two-way interactions that were significant in the analysis of baseline adherence in Section 3.4. The two-way interaction terms were between age and cell phone ownership, gender and patient's reported reason for taking an HIV test as well as treatment site and whether or not a patient is the source of household income. The generalized Wald test was used to test the joint null hypothesis that the three interaction terms are equal to zero. The test produced a Wald statistic of 4.78 with 4 degrees of freedom (p-value=0.3105) and is not significant. Additionally, none of the two-way interactions of the three variables were significant (age*cell phone: p-value=0.6613; gender*reason: p-value=0.4434; treatment site*income: p-value=0.0710).

A new model was then built by fitting all the main effects and then each of the two-way interaction terms formed from the predictor variables were added to the model one at a time. The significance of each interaction term was assessed using the Wald test. Only five of the two-way interactions were significant, of which three of the two-way interactions involved time. Three variables which interact with time were treatment site, gender and reason for taking an HIV test. The other two interaction terms involved age with gender and with education. Then the generalized Wald test was used to test the joint null hypothesis that the five interaction terms are equal to zero. The test produced a Wald statistic of 45.38 with 7 degrees of freedom (p-value = <.0001), which is significant. We further checked if higher order interactions were significant. However, the Wald test has never favoured three-way and higher interaction terms. Consequently, the final model contained five two-way interaction terms and all the main effects. The results from the type 3 analysis of effects for this model are presented in Table 4.3.

Table 4. 3 : Type 3 analysis of effects for the GEE model

Effect	DF	Wald Chi-square	p-values
Age	1	3.93	0.0474
Gender	1	12.28	0.0005
Educational attainment	2	10.03	0.0066
Treatment site	1	52.7	<.0001
Living with/without partner	1	8.33	0.0039
Contribution to household income	1	0.01	0.9338
WHO staging of disease	3	4.86	0.1825
Baseline CD4+ count (cells/μL)	1	0.39	0.5320
Baseline weight (kg)	1	0.34	0.5570
Reason for taking HIV test	2	7.89	0.0194
Household access to tap-water	1	0.52	0.4689
Household access to electricity	1	0.01	0.9751
Cell-phone ownership	1	6.64	0.0100
Baseline adherence	1	1.72	0.1903
Weight	1	0.02	0.8902
Time	1	145.59	<.0001
Time*gender	1	13.56	0.0002
Time*treatment site	1	8.18	0.0042
Time*reason for test	2	7.74	0.0209
Age*gender	1	6.68	0.0098
Age*educational attainment	2	6.58	0.0372

The final model was re-checked for the correlation structure choice and the model re-affirmed the AR-1 structure. The final accepted model is given by

$$\begin{aligned}\log it[P(\mathbf{y}_i = 1 \mid x_{ij})] &= \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 educ\_no_i + \beta_4 educ\_no_i + \beta_5 site_i + \beta_6 partner_i \\ &+ \beta_7 income_i + \beta_8 WHOstage1_i + \beta_9 WHOstage2_i + \beta_{10} WHOstage3_i + \beta_{11} base\_cd4_i \\ &+ \beta_{12} baseweight_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} tapwater_i + \beta_{16} electricity_i \\ &+ \beta_{17} cellphone_i + \beta_{18} adh\_month1_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} time_i gender_i \\ &+ \beta_{22} time_i site_i + \beta_{23} time_i reason\_no_i + \beta_{24} time_i reason\_risk_i + \beta_{25} age_i gender_i \\ &+ \beta_{26} age_i educ\_no_i + \beta_{27} age_i educ\_p_i \,. \end{aligned} \tag{4.10}$$

All the results are presented in Table 4.4. The significant main effects (that were not involved in significant two-way interaction terms) were cell phone ownership and living with a partner. The interpretation of results on all the significant main effects

that are involved in significant interaction terms will be restricted to post-hoc probing of the interaction effects.

Table 4. 4: Adjusted odds ratio (aOR) from the GEE model with associated 95% confidence intervals (CI)

Parameter	Estimates	aOR	p-value	95% C.I. (aOR)	
				Lower	Upper
Intercept	0.589	1.802	0.146	0.832	3.903
Age	-0.012	0.988	0.147	0.972	1.004
Gender (ref=male)					
female	-3.507	0.030	0.001	0.155	0.591
Education (ref=sec & higher)					
No schooling	1.617	5.038	0.002	1.837	15.82
Primary	0.099	1.104	0.856	0.378	3.223
Treatment site (ref=rural)					
Urban	1.173	3.233	<.001	2.355	4.438
Staying with partner (ref=no)					
Yes	0.289	1.335	0.004	1.097	1.625
Income (ref=not source)					
Source of income	0.008	1.008	0.934	0.828	1.229
WHO staging (ref=stage4)					
Stage1	-0.334	0.716	0.091	0.486	1.054
Stage2	-0.251	0.778	0.147	0.554	1.092
Stage3	-0.108	0.898	0.492	0.660	1.221
Baseline CD4+ count	0.001	1.000	0.532	0.998	1.001
Baseline weight	-0.004	0.996	0.557	0.981	1.010
Reason for testing (ref=unwell)					
No specific reason	-0.027	0.973	0.877	0.685	1.381
Exposed to the risk	-0.486	0.615	0.007	0.431	0.877
Access to tapwater (ref=no)					
Yes	0.108	1.114	0.469	0.831	1.494
Household with electricity (ref=no)					
Yes	0.004	1.004	0.975	0.786	1.283
Cell phone ownership (ref=no)					
Yes	0.231	1.260	0.010	1.056	1.503
Initial adherence (ref=not adherent)					
Adherent	-0.140	0.869	0.190	0.704	1.072
Weight at follow-up	0.001	1.001	0.890	0.987	1.015
Time	0.100	1.105	<.001	1.059	1.153
Time*gender (ref=male)					
Female	0.071	1.074	<.001	1.034	1.116
Time*treatment site					
Urban site	-0.057	0.945	0.004	0.908	0.982
Time*reason (ref= unwell)					
No specific reason	-0.011	0.989	0.624	0.947	1.033
Exposed to the risk	0.056	1.058	0.018	1.010	1.108
age*gender (ref=male)					
Female	0.024	1.024	0.010	1.006	1.043
Age*education (ref=secondary)					
No schooling	-0.033	0.968	0.012	0.943	0.993
Primary	-0.002	0.998	0.875	0.972	1.025

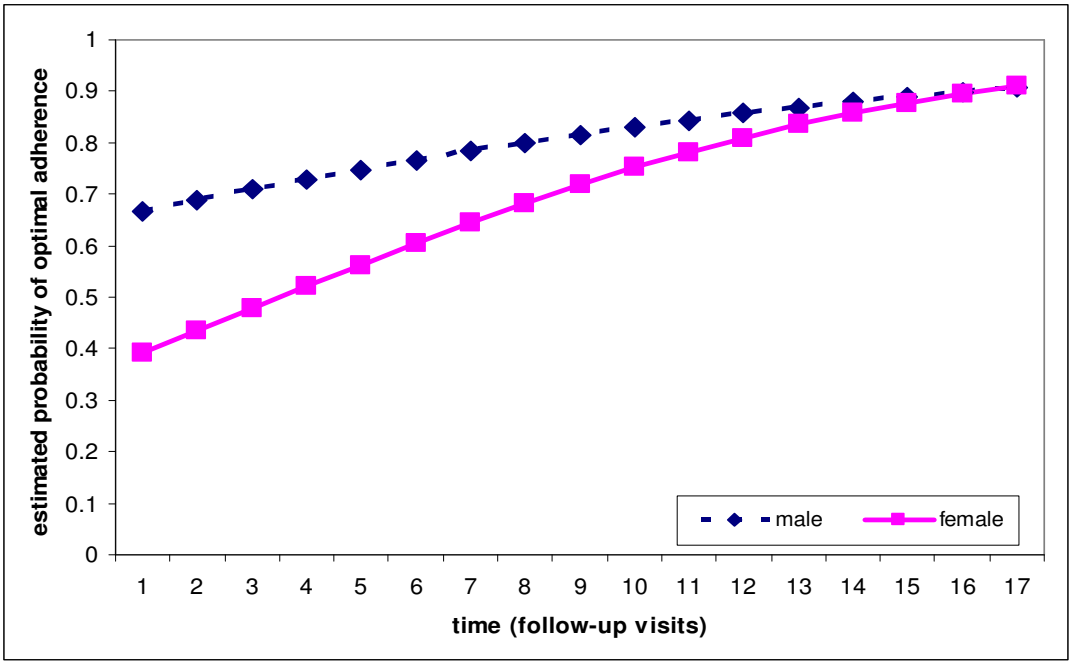
The results show that after controlling for other variables in the model, optimal adherence was significantly higher when patients had cell phones than when they did not have cell phones [aOR = 1.260, p-value= 0.010] and when they lived with a partner compared to when they did not live with a partner [aOR = 1.335, p-value = 0.004] (Table 4.4).

The results further reveal that optimal HAART adherence increased on average over time, however, the rate at which optimal adherence increased differed by treatment site, gender and the patient's reported reason for taking an HIV test. Age interacted significantly with gender and education. The interaction effects are presented below.

a) Interaction between gender and time

Optimal HAART adherence increased over time for both males and females. However, the rate of increase was not the same for males and females after controlling for other covariates in the model. The rate of increase was 7.4% higher for females than for males [aOR = 1.074, p-value<0.001] (Table 4.4). The estimated probabilities for this interaction are presented in Figure 4.1.

Figure 4. 1: Estimated probability of optimal adherence interaction between gender and time

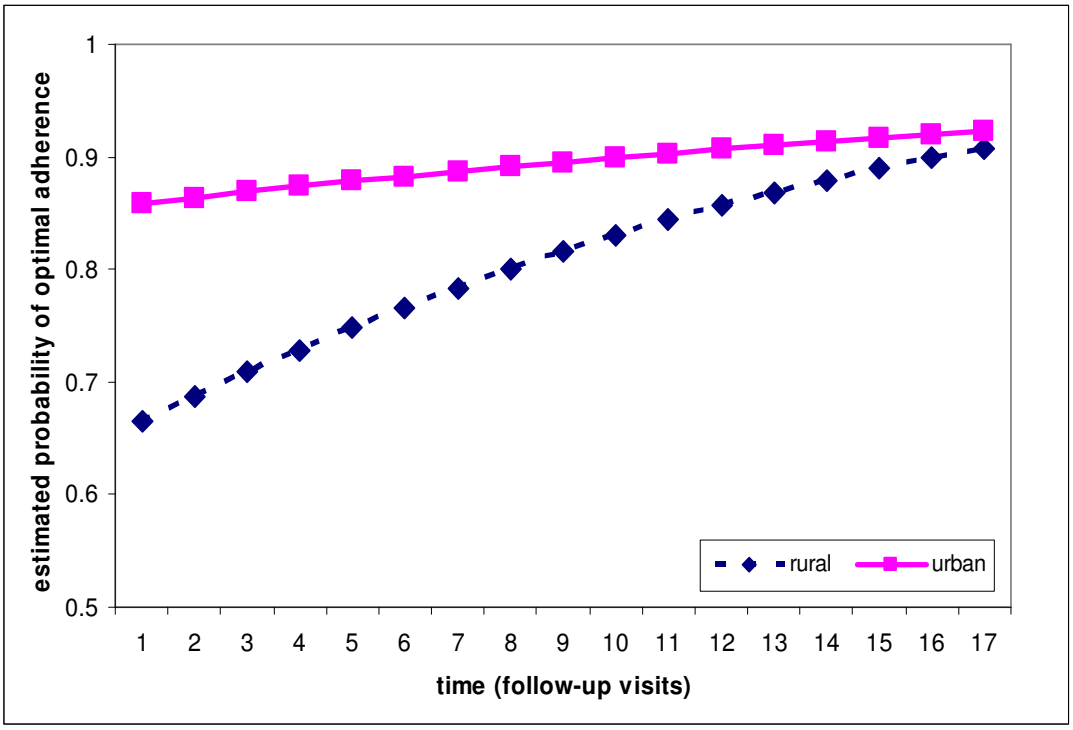


It is shown in Figure 4.1 that the estimated probabilities of optimal adherence were higher for males at the beginning of the follow-up visits, but by the end of the study period (17th follow-up visit), they were similar for both groups.

b) Interaction between treatment site and time

The rate at which optimal adherence increased over time differed in the urban and rural treatment sites. After controlling for other variables in the model, the rate of increase in optimal adherence was 6% higher in the rural treatment site than in the urban treatment site [aOR = 1.06, p-value=0.004] (Table 4.4). The estimated probability of optimal adherence for this interaction is presented in Figure 4.2. It is shown (Figure 4.2) that the estimated probability of optimal adherence at the first follow-up visit was 66% at the rural site and 86% at the urban site. Since the rate of increase was higher in the rural site relative to the rate of increase in the urban site, the gap in adherence between the treatment sites gradually decreased over time until, by the end of the study, the estimated optimal adherence probabilities were similar (at 91% and 92%, respectively).

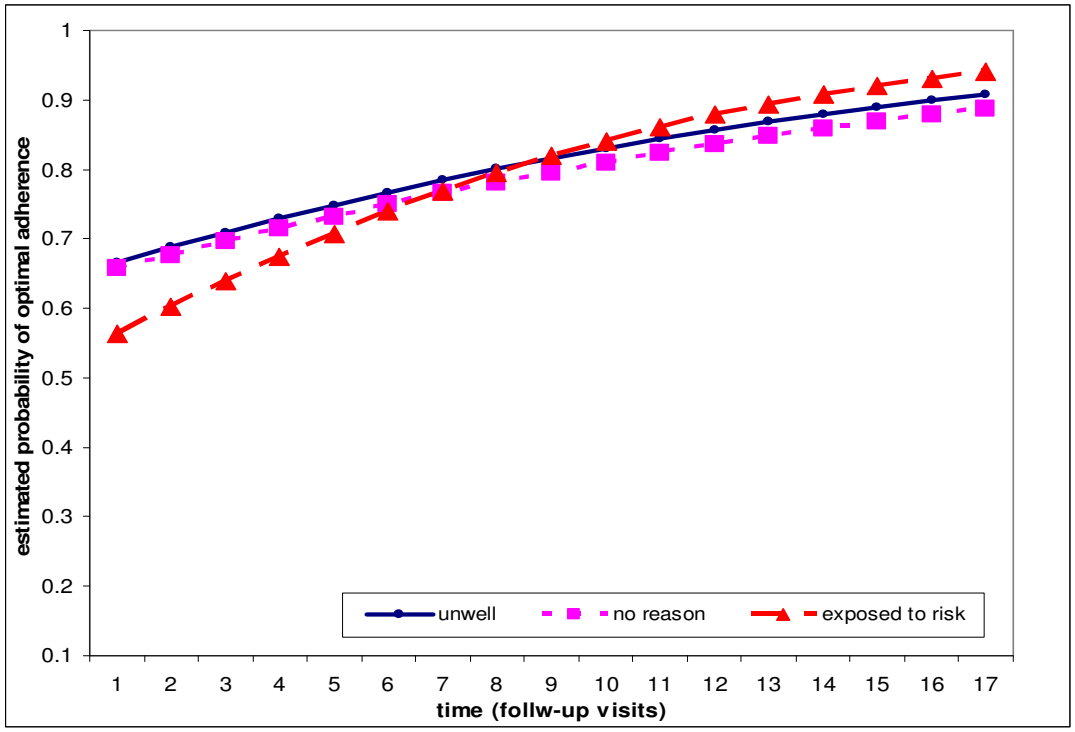
Figure 4. 2: Estimated probability of optimal adherence interaction between treatment site and time



c) Interaction between reason for taking an HIV test and time

Optimal adherence increased over time, but the rate at which it increased differed with the patient's reported reason for taking an HIV test. The rate of increase in optimal adherence was 5.8% higher over the study period for patients who tested due to possible exposure to HIV, than for patients who tested because they were unwell [aOR=1.058, p-value=0.016] (Table 4.4). There was however no significant difference in the rate of change of optimal adherence between patients who tested because they were unwell and those who reported no specific reason for taking an HIV test [p-value=0.666] (Table 4.4). Further analysis revealed that the rate of increase in optimal adherence over the study period was 7% higher for patients who tested due to possible exposure to HIV, than for those who reported no specific reason for taking an HIV test [aOR=1.069, 95% CI: (1.016, 1.126), p-value=0.0107). Figure 4.3 presents the estimated probabilities for this interaction.

Figure 4. 3: Estimated probability of optimal adherence interaction between reason for taking an HIV test and time



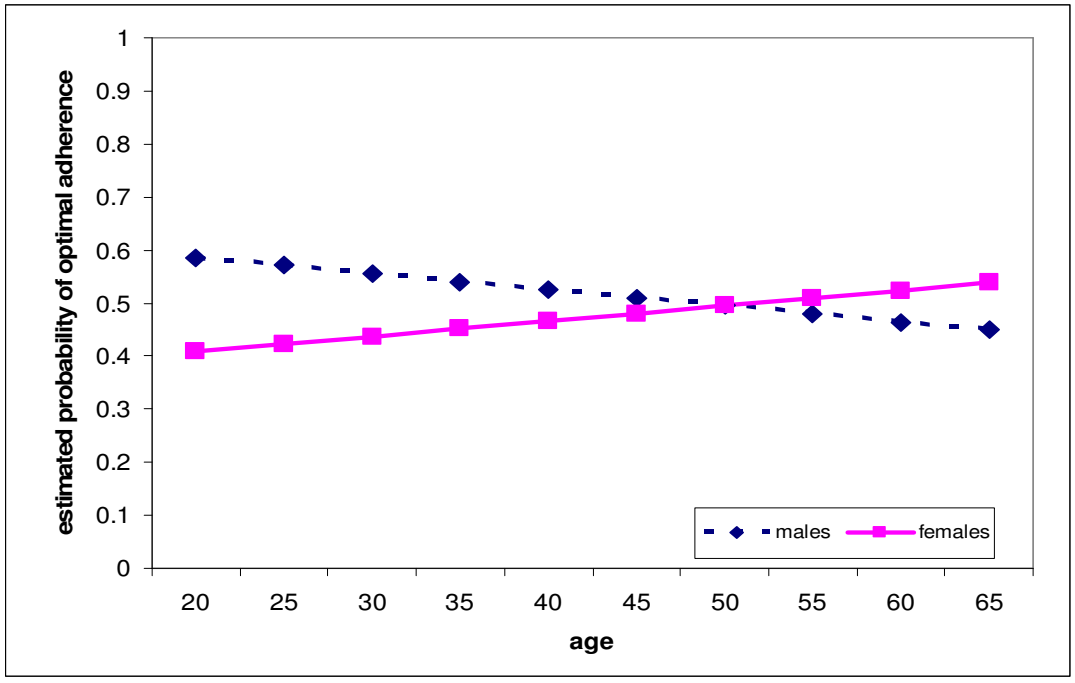
It is shown in Figure 4.3 that estimated probabilities of optimal adherence for patients who tested because they were unwell, and those who reported no specific reason for taking an HIV test, were similar throughout the study. It is again shown in Figure 4.3

that at the beginning of the follow-up period, estimated probabilities of optimal adherence for patients who tested due to possible exposure to HIV were less than the estimated probabilities of patients who reported no specific reason for taking an HIV test, as well as those who tested as they were unwell. However, during the middle of the follow-up visits, estimated probabilities were similar for all the reported reasons. Towards the end of the study, the probabilities of optimal adherence were higher for patients who tested due to possible exposure to HIV than those who reported no specific reason for testing for HIV, or those who tested as they were unwell.

d) Interaction between age and gender

Optimal HAART adherence differed by age for males and females. As the age of patients increased, females tend to adhere better to HAART than males [aOR: 1.024; p-value=0.010] and estimated probabilities are presented in Figure 4.4.

Figure 4. 4: Estimated probability of optimal adherence between age and gender



It is shown in Figure 4.4 that the estimated probabilities of optimal adherence were higher with younger males than with younger females, whereas with older patients, estimated probabilities were higher with females than males.

e) Interaction between age and education

Optimal HAART adherence differed by age at different education levels. Among older patients, those with no schooling were less likely to achieve optimal HAART adherence than those with secondary and higher education [OR=0.97, p-value=0.012] (Table 4.4). There was, however, no significant difference in optimal HAART adherence between patients with secondary education and patients with primary education, regardless of age [p-value=0.875)] (Table 4.4). Further analysis revealed that as patients got older, those with primary education were more likely to achieve optimal adherence than those with no schooling [aOR=1.03, 95% CI: (1.002, 1.070); p-value=0.048]. The estimated probabilities are presented in Figure 4.5.

Figure 4. 5: Estimated probability of optimal adherence interaction between age and educational attainment

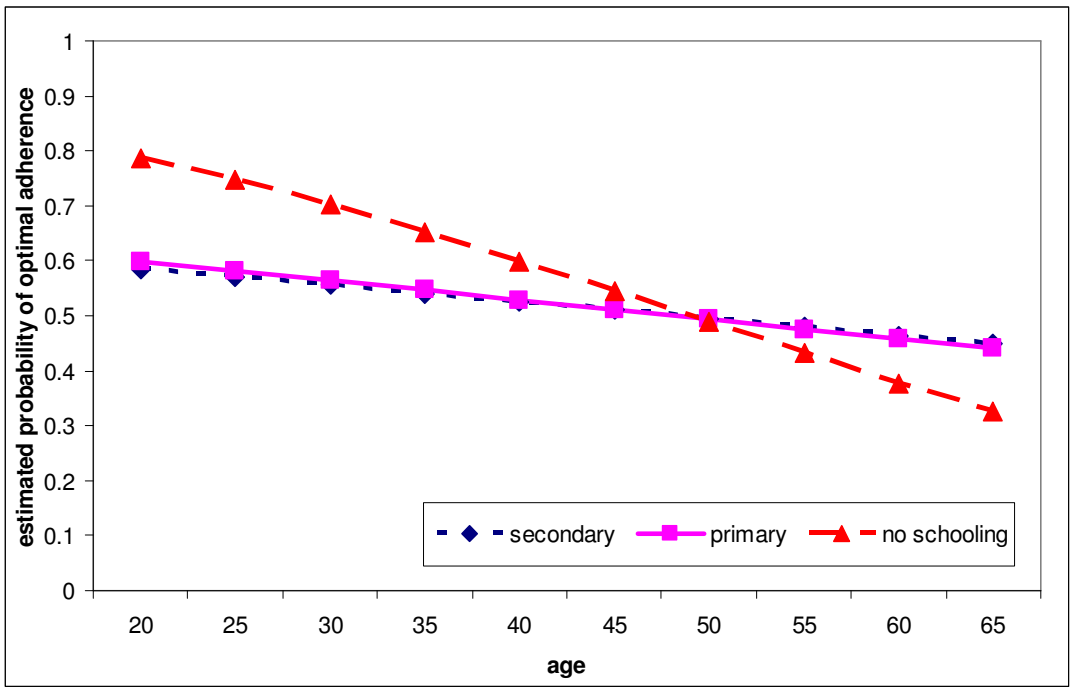


Figure 4.5 indicates that the estimated probabilities of optimal adherence by patients decreased with age. More specifically, the probability of optimal adherence for patients with secondary and primary education was similar for all ages, whereas probabilities of optimal adherence for those with no schooling were higher for younger ages and lower for older ages, relative to those with primary and secondary schooling.

4.6 A note on GEE extensions

A number of extensions to the standard GEE or (GEE1) introduced by Liang and Zeger (1986) have been proposed. We have seen that the standard GEE methodology combines an estimating equation for regression (first-moment) parameters with moment-based estimators for the association (second-moment) parameters α. Prentice (1988) and Prentice and Zhao (1991) extends Liang and Zeger's work by replacing the moment-based approach to estimating second-moment parameters with an 'ad-hoc' estimating equations for these quantities. That is, this method allows for estimation of both parameter vectors, β and α but proposes estimating equations for both sets of parameters. Second order GEE (GEE2) have been proposed by Zhao and Prentice (1990), using correlations, and by Liang, Zeger and Qaqish (1992), using odds ratios. They consider an alternative equation for simultaneous estimation of the regression parameters β and covariance parameters α. This requires modeling the third and fourth moments of y_{ij} instead of just the mean and variance.

Lipsitz, Laird and Harrington (1991) proposed a different approach to modeling the GEE1 association parameters. For binary outcomes, they used the odds ratio as the measure of association instead of Pearson's correlation coefficient. Liang et al., (1992) did the same for GEE2. Carey, Zeger and Diggle (1993) developed an approach for binary repeated measurements similar to that of Lipsitz et al. (1991). The alternating logistic regression (ALR) methodology Carey et al. (1993) simultaneously regresses the response on explanatory variables as well as modeling the association among responses in terms of pairwise odds ratios. The ALR algorithm iterates between a logistic regression using first-order generalized estimating equations to estimate regression coefficients and a logistic regression of each response on others from the same subject using an appropriate offset to update the odds-ratio parameters. The ALR algorithm is now implemented in the SAS GENMOD procedure.

4.7 Summary

The GEE approach is appealing for analysis of binary (discrete) data because of its computational simplicity compared to the maximum likelihood-based approaches. The marginal GEE model for adherence data was fitted with ease using the SAS GENMOD procedure. However, because there is no likelihood function, likelihood-based methods are not available for testing fit, comparing models and conducting inference about parameters. Instead inference can only use Wald statistics constructed with asymptotic normality of the estimators together with their estimated covariance matrix. Moreover, even though GEE estimates are consistent with misspecification of the covariance structure, it is important to choose the covariance structure that closely approximates the true underlying one for greater efficiency. However, because of the lack of the likelihood function, tests such as AIC and likelihood ratio test cannot be used to guide the selection of the appropriate form of non-nested and nested covariance structures respectively. Nonetheless, the quasi-likelihood information criterion (QIC) (Pan, 2001), has been advocated with the GEE for choosing a reasonable working correlation structure. QIC is a modified AIC for GEE, where the likelihood is replaced by the quasi-likelihood and the penalty also takes a modified form. The QIC macro implemented within the GENMOD procedure was used to select the best fitting covariance structure among the competing ones in the adherence data.

Application of the GEE method to adherence data to evaluate long-term predictors of optimal adherence revealed that five of the two-way interactions were significant. These were between: time and treatment site, time and gender, time and reported reason for taking an HIV test, age and gender as well as age and educational level. Also, cell phone ownership and living with a partner were positively associated with optimal adherence. A detailed discussion of these results can be found in Maqutu et al., (2010b) attached in Appendix B. Next, in Chapter 5 we explore subject-specific effects that are associated with optimal adherence by reviewing and fitting generalized linear mixed models.

Note that GEEs are flexible in handling missing data in the sense that there are no restrictions on the number of observations per individual; subjects who are missing

at a given assessment are not excluded from the analysis. However, these models make a very restrictive assumption about the missing data mechanism, i.e., that data are missing completely at random (MCAR). Evaluation of the impact of dropout in the adherence data suggests that there is no evidence against MCAR, and we therefore conclude that the GEE estimates are valid.

Chapter 5

Generalized linear mixed models

5.1 Introduction

In the previous chapter, we adopted a GEE modeling approach for analysis of adherence data. This entailed a population-averaged modeling formulation that provided population average estimates of overall trends in adherence. In this chapter, the goal of the analyses is to determine subject specific changes in medication adherence over time and explore factors that influence such adherence. In addition, the association structure of data is of interest, that is, we seek to examine the correlation structure among the repeated measures. Much as the GEE method provided marginal population averaged evolutions, which are important from the public health perspective, it cannot provide evolution of each subject separately. Moreover, in the GEE approach, the correlation assumptions are allowed to be incorrect, which does not affect the validity of the estimates but hinders formal inferences about the correlation structure. Moreover, the GEEs by themselves do not help separate out different sources of variation. It is often an advantage to be able to attribute variation as being associated with different factors (McCulloch, 2003). In this chapter, we present an alternative approach using subject-level terms in the model. The inclusion of subject-specific term leads us to random effects models which have conditional interpretations (referred to as subject-specific), that are a direct contrast with the GEE models which have population-averaged interpretations.

Linear mixed models as a special case of random effects models are well established. By contrast, only recently have random effects been used much in models for discrete data (Molenberghs and Verbeke, 2005). In this chapter, we extend generalized linear models (presented in Chapter 3) to include random effects. This leads to the generalized linear mixed model (GLMM), which is a special case of the random effects models and is the most frequently used model for analysis of discrete data. The GLMMs are basically suitable for analysis of longitudinal data where the objective is to study how subject specific effects change over time and

what characteristics influence such changes. The advantages of GLMMs over GEEs are that they are more robust in cases where there are missing data and where there are unbalanced clusters. GLMMs can also estimate variances at different levels.

5.2 Model formulation

Suppose y_{ij} is the jth response for subject i, $i = 1,\ldots,N$, $j = 1,\ldots,n_i$ and $\mathbf{y}_i$ is the n_i-dimensional vector of all measurements available for subject i. Conditionally on random effects $\mathbf{b}_i$, it is assumed that the elements of y_{ij} for $\mathbf{y}_i$ are independent, following a generalized linear model, but with the linear predictor appended with subject-specific regression parameters $\mathbf{b}_i$. Specifically, it is assumed that all y_{ij} have densities of the form

$$f_i\left(y_{ij} \mid \mathbf{b}_i, \beta, \phi\right) = \exp\left\{\frac{y_{ij}(\theta_{ij}) - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \qquad (5.1)$$

where μ_{ij}, the conditional mean of y_{ij} is modeled through a linear predictor containing fixed regression parameters β as well as subject specific parameters $\mathbf{b}_i$, that is

$$g(\mu_{ij}) = g[E(y_{ij} \mid \mathbf{b}_i)] = \mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i, \qquad (5.2)$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are p-dimensional and q-dimensional vectors of known covariate values corresponding to the fixed and random effects β and $\mathbf{b}_i$ respectively, through a known link function $g(.)$. Furthermore, conditionally on random effects $\mathbf{b}_i$, the responses y_{ij} are independent and it is also assumed that the random effects $\mathbf{b}_i$, are $N(\mathbf{0}, \mathbf{G})$. The variance of observations, $\mathbf{y}_i$, conditional on the random effects, $\mathbf{b}_i$ is given by

$$var\left(y_{ij} \mid \mathbf{b}_i\right) = a_{ij}(\phi)\upsilon\left(\mu_{ij}\right),$$

where $\upsilon(.)$ is the variance function that relates the conditional means and variances, and ϕ is a dispersion parameter.

It is important to note that the model specification in (5.1) and (5.2) is made conditional on the value $\mathbf{b}_i$. The consequences of including the random effects in the model can better be appreciated by studying the first two moments of the

marginal distribution of y_{ij}. With linear mixed models, the marginal mean of $\mathbf{y}_i$ coincide with the conditional mean given that $\mathbf{b}_i = 0$. However, this property is not necessarily true in GLMMs. The marginal mean, variance and co-variances are given (McCulloch and Searle, 2001) as follows.

The marginal mean of y_{ij} is given by

$$\begin{aligned} E(y_{ij}) &= E\left[E\left(y_{ij} \mid \mathbf{b}_i\right)\right] \\ &= E\left[\mu_{\text{ij}}\right] \\ &= E\left[g^{-1}\left(\mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i\right)\right] \end{aligned}$$

which in general cannot be simplified due to the nonlinear function g^{-1}. In a linear mixed model, the induced marginal mean is reduced to $E(y_{ij}) = \mathbf{x}_{ij}'\beta$. The marginal variance of y_{ij} induced by the random effects has the following expression

$$\begin{aligned} var(y_{ij}) &= var\left(E\left[y_{ij} \mid \mathbf{b}_i\right]\right) + E\left(var\left[y_{ij} \mid \mathbf{b}_i\right]\right) \\ &= var\left(\mu_{ij}\right) + E\left[\phi a_i v\left(\mu_{ij}\right)\right] \\ &= var\left(g^{-1}\left[\mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i\right]\right) + E\left\{\phi a_i v\left(\mu_{ij}\right)\left[g^{-1}\left(\mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i\right)\right]\right\} \end{aligned}$$

this again cannot be simplified in most cases without making specific assumptions about the form of $g(.)$ and/or the conditional distribution of y_{ij}. The induced marginal variance in a linear mixed model is reduced to $var(y_{ij}) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. Finally, assuming conditional independence of $\mathbf{y}_i$, the marginal covariance is given by

$$\begin{aligned} cov\left(y_{ij}, y_{ik}\right) &= cov\left(E\left[y_{ij} \mid \mathbf{b}_i\right], E\left[y_{ik} \mid \mathbf{b}_i\right]\right) + E\left[cov\left(y_{ij}, y_{ik} \mid \mathbf{b}_i\right)\right] \\ &= cov\left[g^{-1}\left(\mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\mathbf{b}_i\right), g^{-1}\left(\mathbf{x}_{ik}'\beta + \mathbf{z}_{ik}'\mathbf{b}_i\right)\right] \end{aligned}$$

Similarly, in a linear mixed model, the above equation is reduced to $cov\left(y_{ij}, y_{ik}\right) = \mathbf{G}_i\mathbf{z}_{ij}\mathbf{z}_{ik}'$. If the linear mixed model has only a random intercept, the marginal covariance is $cov\left(y_{ij}, y_{ik}\right) = \sigma_b^2$ where σ_b^2 is the variance of the random intercept.

For illustration, we consider a Poisson GLMM with a random intercept distributed normally with zero mean and variance σ_b^2. Suppose we have a log link so that

$g(\mu) = \log(\mu)$ and $g^{-1}(x) = exp\{x\}$. Then the induced marginal mean of y_{ij} is equal to

$$E(y_{ij}) = E[exp\{\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}\}]$$
$$= exp\{\mathbf{x}'_{ij}\beta\}E[exp\{\mathbf{z}'_{ij}\mathbf{b}\}]$$
$$= exp\{\mathbf{x}'_{ij}\beta\}exp\{\sigma_b^2/2\}$$

and the marginal variance and covariance leads to

$$var(y_{ij}) = var\left(E\left[y_{ij} \mid \mathbf{b}_i\right]\right) + E\left(var\left[y_{ij} \mid \mathbf{b}_i\right]\right)$$
$$= E[y_{ij}](exp\{\mathbf{x}'_{ij}\beta\}[exp\{3\sigma_b^2/2\} - exp\{\sigma_b^2/2\}] + 1)$$

and

$$cov\left(y_{ij}, y_{ik}\right) = cov\left(E\left[y_{ij} \mid \mathbf{b}_i\right], E\left[y_{ik} \mid \mathbf{b}_i\right]\right) + E\left[cov\left(y_{ij}, y_{ik} \mid \mathbf{b}_i\right)\right]$$
$$= exp(\mathbf{x}'_{ij}\beta)exp(\mathbf{x}'_{ik}\beta)(exp(2\sigma_b^2) - exp(\sigma_b^2))$$

respectively. This shows that although the conditional distribution of y_{ij} given $\mathbf{b}_i$ is Poisson, the marginal distribution cannot be. For instance, the term in parentheses for the marginal variance is greater than 1, therefore, the variance is larger than the mean. Thus, although $y_{ij} \mid \mathbf{b}_i$ follows a regular Poisson distribution, the marginal distribution of y_{ij} is over-dispersed. Again, the marginal covariance expression shows that due to random intercept, observations on the same subject are no longer independent.

In general, it should be appreciated that the consequence of adding random effects in a generalized linear model complicates the evaluation of the first two moments of the marginal distribution.

<u>The random intercept model</u>

The random intercept model is the simplest case of a mixed effects model. In the case of non-normal data, a random-intercept model is simply a generalized linear model with randomly varying subject effect. In this model, each subject is assumed to have an underlying level of response that persists over time. Assume that there are $i = 1, \cdots, N$ subjects and $j = 1, \cdots, n_i$ repeated observations for each subject. A

random-intercept model augments the linear predictor with a single random effect for subject i,

$$\eta_{ij} = \mathbf{x}'_{ij}\beta + b_i$$

where b_i is a random effect for each subject. These random effects represent the influence of subject i on his/her repeated observations that is not captured by the observed covariates. The random effects are commonly assumed to be normally distributed with mean zero and constant variance σ_b^2. The parameter σ_b^2 indicates the variance in the population distribution and therefore the degree of heterogeneity of subjects.

The generalized linear mixed model for a binary response is used here as an illustration of a random intercepts model because of its relevance to the adherence data in this study. Suppose that y_{ij} is a binary response, taking values of 0 or 1. A logistic mixed effects model for y_{ij} is given as follows (Fitzmaurice et al., 2004):

a) Conditional on single random effects b_i, the y_{ij} are independent and have a Bernoulli distribution, with $var(y_{ij} \mid b_i) = E(y_{ij} \mid b_i)\{1 - E(y_{ij} \mid b_i)\}$ with $\phi = 1$.

b) The conditional mean of y_{ij} depends upon fixed and random effects via the following linear predictor

$$\eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}b_i = \mathbf{x}'_{ij}\beta + b_i,$$

where $\mathbf{z}_{ij} = 1$ for all $i = 1,\ldots,N$, and $j = 1,\ldots,n_i$ with

$$\log\left\{\frac{\Pr(y_{ij} = 1 \mid b_i)}{\Pr(y_{ij} = 0 \mid b_i)}\right\} = \eta_{ij} = \mathbf{x}'_{ij}\beta + b_i.$$

That is, the conditional mean of y_{ij} is related to the linear predictor by a logit link function.

c) The single random effect b_i is assumed to have a univariate normal distribution, with zero mean and the variance, say g_{11} because in this case $\mathbf{G}$ is of dimension 1×1.

This example illustrates a simple logistic regression model with randomly varying intercepts. The model shows that there is a natural heterogeneity in individuals'

propensity to respond positively that persists throughout all the binary responses obtained on any individual.

The randomly varying intercepts and slopes model

The random intercepts model can be extended to include multiple random effects. Denote $\mathbf{z}_{ij}$ as the $r \times 1$ vector of variables having random effects (a column of ones is usually included for the random intercept). The vector of random effects $\mathbf{b}_i$ is assumed to follow a normal distribution with the mean vector $\mathbf{0}$ and the variance-covariance matrix $\mathbf{G}$. The linear predictor is now written as

$$\eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}_i,$$

where now the conditional mean is specified in terms of the vector of random effects, i.e $E\left[y_{ij} \mid \mathbf{b}_i, \mathbf{x}_{ij}\right]$. For instance, it is common to have a random subject intercept and a random slope in longitudinal problems. Our earlier illustration of the random intercepts model for binary data can be extended to incorporate random slopes as follows:

a) Conditional on a vector of random effects $\mathbf{b}_i$, the y_{ij} are independent and are assumed to have a Bernoulli distribution, with

$$var(y_{ij} \mid \mathbf{b}_i) = E\left(y_{ij} \mid \mathbf{b}_i\right)\left\{1 - E(y_{ij} \mid \mathbf{b}_i)\right\} \text{ with } \phi = 1.$$

b) The conditional mean of y_{ij} depends upon fixed and random effects via the following linear predictor

$$\eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}_i,$$

where $\mathbf{x}'_{ij} = \mathbf{z}'_{ij} = \left(1, t_{ij}\right)$ for all $i = 1, \ldots, N$, and $j = 1, \ldots, n_i$ with

$$\log\left\{\frac{\Pr(y_{ij} = 1 \mid \mathbf{b}_i)}{\Pr(y_{ij} = 0 \mid \mathbf{b}_i)}\right\} = \eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}_i$$

$$= \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij}$$

$$= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t_{ij}.$$

c) The random effects are assumed to have a bivariate normal distribution, with zero mean and a 2×2 covariance matrix $\mathbf{G}$.

This example illustrates a logistic regression model with randomly varying intercepts and slopes. The model shows that there is a natural heterogeneity among individuals in both their baseline level and changes in the expected outcome over time.

5.3 Interpretation of model parameters

Fixed part: conditional and marginal relationships

Although the introduction of random effects can simply be thought of as a means of accounting for correlation among longitudinal responses, it has important implications for the interpretation of the regression coefficients in generalized linear mixed models (Fitzmaurice et al., 2004). The regression parameters, β, have somewhat different interpretations than the regression parameters in the marginal models considered in Chapter 4. Molenberghs and Verbeke (2005) also show that the severe differences in results obtained from marginal and random effects follow from the fact that parameters in both models have completely different interpretations. To appreciate the nature of the difference between the two models where we have a non-linear link function, let us consider an example using a binary outcome variable and assume a random-intercepts logistic model with linear predictor $logit\left[P\left(y_{ij}=1\mid b_i\right)\right]=\beta_0+\beta_1 t+b_i$ where t is the time covariate. The conditional means $E\left(y_{ij}\mid b_i\right)$, as a functions of t, are given by

$$E\left(y_{ij}\mid b_i\right)=\frac{\exp\left(\beta_0+\beta_1 t+b_i\right)}{1+\exp\left(\beta_0+\beta_1 t+b_i\right)}.$$

The model assumes that the conditional means all satisfy a logistic model, with the same slope β_1 but with different intercepts β_0+b_i for all subjects. The marginal average evolution is as follows

$$\begin{aligned}E\left(y_{ij}\right)&=E\left[E\left(y_{ij}\mid b_i\right)\right]\\&=E\left[\frac{\exp\left(\beta_0+\beta_1 t+b_i\right)}{1+\exp\left(\beta_0+\beta_1 t+b_i\right)}\right]\neq\frac{\exp\left(\beta_0+\beta_1 t+b_i\right)}{1+\exp\left(\beta_0+\beta_1 t+b_i\right)}.\end{aligned}$$

This shows that there is no straightforward relationship between estimated parameters in both random effects models and marginal models.

Consequently, parameter interpretation in generalized linear mixed models has subject specific interpretations. That is, they represent the influence of covariates on a specific subject's mean response. In particular, the regression coefficients are interpreted in terms of the effects of covariates on changes in an individual's transformed mean response, while holding the remaining covariates constant (Fitzmaurice et al., 2004).

Random part:

(a) Measures of heterogeneity

It has been suggested that one way of interpreting estimated standard deviations of the random effects is to produce percentiles of the effects based on the normality assumption (Hedeker and Gibbons, 2006; Rabe-Hesketh and Skrondal, 2008). For instance, for a covariate with an estimated fixed coefficient $\hat{\beta}_k$ and a random coefficient b_{ik}, form the approximate 2.5th and 95.5th percentiles using $\hat{\beta}_k \pm 1.96\sqrt{\hat{g}_{kk}}$ to express the range of effects that are likely to occur. In a random intercept model, for instance, $\hat{\beta}_0 \pm 1.96\sqrt{\hat{g}_{00}}$ gives a range of intercepts which can be converted to a range of conditional expectations μ_{ij} (e.g probabilities for dichotomous responses) by plugging particular covariate values into the linear predictor and applying the inverse link function. The same approach can be used in random coefficient models by using the standard deviation of $\mathbf{z}_{ij}'\mathbf{b}_i$ at particular values of $\mathbf{z}_{ij}'$.

(b) Measures of dependence

The introduction of random subject effect, can be seen to induce correlation (within-subject dependence) among the repeated measures. For a random intercept model, the between subject variance is often expressed in terms of an intraclass correlation. Given that the covariance between any pair of repeated measurements is σ_b, the correlation is

$$Corr(y_{ij}, y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2},$$

also known as the intra-class correlation. This can also be interpreted as the proportion of the total residual variance that is due to unobserved between-subject heterogeneity. In linear random coefficient models, the covariance matrix of the total

residuals is a function of time and there is no simple measure. For instance, in the model with randomly varying intercepts and slopes, the covariance between any pair of repeated measurements depend on the measurements of time (Fitzmaurice et al., 2004; Longford, 1993), that is

$$Corr(y_{ij}, y_{ik}) = g_{11} + (t_{ij} + t_{ik})g_{12} + t_{ij}t_{ik}g_{22}.$$

However, for GLMMs, the correlations $Corr(y_{ij}, y_{ik})$ generally depends on covariates even if only a random intercept is included and are therefore not useful measures of dependence. Fortunately, the correlations of latent responses in logit and probit random intercept models are constant and obtained by simply substituting either 1 (probit) or $\pi^2/3$ (logit) for σ_b^2 (Rabe-Hesketh and Skrondal, 2008).

5.4 Estimation of model parameters

To fit a GLMM, it is possible to use two alternatives: Bayesian approach and maximum likelihood estimation approach. In the Bayesian approach, it is necessary to specify the prior densities for β, $\mathbf{G}$ and ϕ denoted by $f(\beta)$, $f(\mathbf{G})$ and $f(\phi)$ respectively. Once priors have been specified, the posterior distribution can be found (Molenberghs and Verbeke, 2005). It is pointed out that an attractive feature of Bayesian approach is its flexibility for full assessment of the uncertainty in the estimated random effects and functions of model parameters, however, the potential drawbacks include the intensive computations (which require fairly sophisticated computer programs) and questions about when the sampling process has achieved equilibrium (Breslow and Clayton, 1993; Agresti et al., 2000).

On the other hand, it has been argued that maximum likelihood is a well-established and well-respected method of estimation that has a variety of optimality properties and as such it is usually the default technique for estimating parameters (Searle, Casella and McCulloch, 2006). Our focus will be on this method of estimation. We have seen that the estimation method of fixed effects in GLMs is based on the well defined log-likelihood and it is simple to construct an objective function based on the independence of the data (Section 3.3). In linear mixed models, estimation of parameters is based on the marginal likelihood of the data and can be evaluated analytically (Verbeke and Molenberghs, 2000). With GLMMs, to obtain maximum

likelihood estimates, one would maximize the marginal likelihood, obtained by integrating over the random effects.

The complete likelihood function for GLMM can be obtained from the product of the known distributions of $\mathbf{y} \mid \mathbf{b}$ and $\mathbf{b}$. The true likelihood function can be written as follows (Demidenko, 2004)

$$L(\mathbf{y}, \mathbf{b}) = L(\mathbf{y} \mid \mathbf{b}) L(\mathbf{b}) \tag{5.3}$$

If we assumed that $\mathbf{b}$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{G})$, then

$$L(\mathbf{y}, \mathbf{b}) \propto L(\mathbf{y} \mid \mathbf{b}) \mid \mathbf{G} \mid^{-½} \exp\left(-\frac{1}{2} \mathbf{b}' \mathbf{G}^{-1} \mathbf{b} \right). \tag{5.4}$$

Based on (5.4), the marginal likelihood of $\mathbf{y}$ could be obtained by integrating over the random effects $\mathbf{b}$. However, this is not always possible because the resultant integral of the right side of (5.4) does not always have an analytical solution, that is, it must be solved using numerical methods.

5.4.1 Approximations to the likelihood function in GLMM

The GLMM can be fitted by maximizing the marginal likelihood obtained by integrating over the random effects. The contribution of the ith subject to the likelihood is given by (Molenberghs and Verbeke, 2005)

$$L_i = f_i(\mathbf{y}_i \mid \beta, \mathbf{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid \mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i \mid \mathbf{G}) d\mathbf{b}_i \tag{5.5}$$

Thus, the likelihood function L can be written as:

$$L = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid \mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i \mid \mathbf{G}) d\mathbf{b}_i \tag{5.6}$$

where $\mathbf{y}_i$ is the $n_i -$dimensional vector containing all the measurements available for the ith subject. Notice that (5.5) has the form of (5.3). This integration is done over the $q -$dimensional distribution of $\mathbf{b}$. In some cases, (5.6) can be worked out analytically, for example probit-normal model (Molenberghs and Verbeke, 2005). In other cases, such as the logistic mixed model, the integral in (5.6) cannot be evaluated in closed form. Therefore, numerical approximation is necessary in order

to evaluate the likelihood. The numerical approximations for the integral of the marginal likelihood in (5.6) can be divided into those based on the *approximation of the integrand,* those based on the *approximation of the integral itself* and those based on *an approximation of the data*.

Approximation of the integrand: Laplace Approximation

Laplace approximation is a well known method to approximate integrals where the exact likelihood is difficult to evaluate. To illustrate how Laplace approximation works, let us suppose that we want to approximate integrals of the form (Molenberghs and Verbeke, 2005)

$$I = \int e^{-Q(\mathbf{b})} d\mathbf{b},$$

and that $\hat{\mathbf{b}}$ is the value of $\mathbf{b}$ for which Q is minimized. Then, the second-order Taylor expansion of $Q(\mathbf{b})$ around $\hat{\mathbf{b}}$ is of the form

$$Q(\mathbf{b}) \approx Q(\hat{\mathbf{b}}) + \frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})' Q''(\mathbf{b})(\mathbf{b}-\hat{\mathbf{b}}) \qquad (5.7)$$

where $Q''(\hat{\mathbf{b}})$ is equal to the Hessian of Q, i.e the matrix of the second order derivatives of Q, evaluated at $\hat{\mathbf{b}}$. The integral I can be approximated using (5.5), thus

$$I \approx (2\pi)^{q/2} \left| Q''(\hat{\mathbf{b}}) \right|^{-1/2} e^{-Q(\hat{\mathbf{b}})}. \qquad (5.8)$$

In this case, it is considered that $Q(.)$ is unimodal. When Q is bimodal, it is necessary to use an improved Laplace approximation (Demidenko, 2004). In this method, the approximation to the integral uses many different estimates of $\mathbf{b}$ as necessary according to the different modes of the Q function. The integral in (5.6) is proportional to an integral I in (5.8), for a $Q(\mathbf{b})$ function given by

$$Q(\mathbf{b}) = (a_i \phi)^{-1} \sum_{j=1}^{n_i} \left[y_{ij} \left(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b} \right) - \psi\left(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b} \right) \right] - \frac{1}{2}\mathbf{b}'\mathbf{G}^{-1}\mathbf{b},$$

such that Laplace's method can be applied here.

Approximation to the integral: Gauss-Hermite quadrature

The Gauss-Hermite quadrature is often used for numerical integration in statistics, because of its relation to Gaussian densities (Liu and Pierce, 1994). In a particular

context of random effects models, the Gauss-Hermite quadrature can be used for performing maximum likelihood estimation. We will consider the classical Gaussian and adaptive Gaussian quadratures, designed to approximate integrals of the form

$$\int h(s)c(s)\,ds\,, \tag{5.9}$$

for a known function $h(s)$ and for $c(s)$ the density of the univariate or multivariate standard normal distribution. Thus, random effects have to be standardized such that they get the identity covariance matrix. Let $\boldsymbol{\delta}_i$ be equal to $\boldsymbol{\delta}_i = \mathbf{G}^{-1/2}\mathbf{b}_i$. Then $\boldsymbol{\delta}_i$ is normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{I}$, and the linear predictor becomes $\theta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{G}^{1/2}\boldsymbol{\delta}_i$. Consequently, the variance components in $\mathbf{G}$ are now contained in the linear predictor. Then the likelihood contribution for subject i is given by

$$f_i(\mathbf{y}_i \mid \beta, \mathbf{G}, \phi) = \int \prod_{i=1}^{n_i} f_{ij}(y_{ij} \mid \mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i \mid \mathbf{G}) d\mathbf{b}_i$$

$$f_i(\mathbf{y}_i \mid \beta, \mathbf{G}, \phi) = \int \prod_{i=1}^{n_i} f_{ij}(y_{ij} \mid \boldsymbol{\delta}_i, \beta, \phi) f(\boldsymbol{\delta}_i \mid \mathbf{G}) d\boldsymbol{\delta}_i\,, \tag{5.10}$$

where the random effects $\mathbf{b}_i$ are assumed to be normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{G}$. Expression (5.10) is the form of (5.9) as required to apply the Gaussian quadratures (Molenberghs and Verbeke, 2005; Antonio and Beirlant, 2007).

Classical Gaussian quadrature approximates an integral of the form (5.9) by a weighted sum, namely

$$\int h(s)c(s)\,dz \approx \sum_{q=1}^{Q} w_q h(s_q)\,. \tag{5.11}$$

where Q is the order of the approximation, the s_q are solutions of the Qth order Hermite polynomial and w_q are corresponding weights. The nodes or quadrature points s_q and the weights w_q are reported in Tables (e.g Abramowitz and Stegun, 1972). Alternatively, an algorithm to compute s_q and w_q for any value Q is given in Press et al., (1992). The quadrature points used in (5.11) do not depend on $h(s)$ such that it is possible that only very few nodes lie in the region where most of the mass of $h(s)$ is, which would lead to poor approximations (Antonio and Beirlant,

2007). However, if adaptive Gaussian quadrature rule is used, the nodes would be rescaled and shifted such that the integrand is sampled in a suitable range. That is, the quadrature points are rescaled as if the $\int h(s)c(s)\,ds$ is a normal distribution with the mean of this distribution being the mode $\hat{z}$ of $\ln[h(s)c(s)]$ and the variance would equal to

$$\left[-\frac{\partial^2}{\partial s^2}\ln[h(s)c(s)]\;\Big|_{s=\hat{s}}\right]^{-1}.$$

Then the new quadrature points are given by

$$s_q^+ = \hat{s} + \left[-\frac{\partial^2}{\partial s^2}\ln[h(s)c(s)]\;\Big|_{s=\hat{s}}\right]^{-1/2} s_q$$

with corresponding weights

$$w_q^+ = \left[-\frac{\partial^2}{\partial s^2}\ln[h(s)c(s)]\;\Big|_{s=\hat{s}}\right]^{-1/2} \frac{c\left(s_q^+\right)}{c\left(s_q\right)} w_q .$$

In this case, then the integral is approximated by

$$\int h(s)c(s)ds \approx \sum_{q=1}^{Q} w_q^+ h\left(s_q^+\right). \qquad (5.12)$$

(Molenberghs and Verbeke, 2005).

It has been shown that when (5.12) is applied with only one node, the result is equivalent to approximating the integrand using the Laplace approximation (Liu and Pierce, 1994). Some simulation results suggest that in the classical Gaussian quadrature, a larger number of quadrature points (100 or more) are necessary to obtain high accuracy while the adaptive quadrature provides good accuracy with 20 or fewer quadrature points (Diggle et al., 2002). Nonetheless, the adaptive Gaussian quadrature is much more time consuming than the classical Gaussian quadrature. This is due to the fact that the adaptive Gaussian quadrature requires calculation of $\hat{s}$ for each unit in the dataset, hence the numerical maximization of N functions of the form (5.9) (Molenberghs and Verbeke, 2005). Moreover, since these functions (5.9) depend on the unknown parameters β, $\mathbf{G}$ and ϕ, the quadrature points as well as the weights used in the adaptive Gaussian quadrature depend on those parameters, and hence the need to be updated in every step of the iterative procedure (Molenberghs and Verbeke, 2005).

Once the problem of the intractable integral is solved, the actual maximization of the likelihood is carried out using algorithms such as Newton-Raphson and Fisher scoring (Tuerlinckx et al., 2006). The numerical integration methods work relatively well with GLMMs that have low-dimensional random effects distributions such as single random effect or two or three nested random effects (Diggle et al., 2002). However, none of the numerical methods have been made computationally practical for models with random effects distributions with $q > 5$. This limitation makes numerical integration using quadrature prohibitive for GLMMs that have serial random effects for instance, which greatly limits its application to categorical longitudinal data analysis (Diggle et al., 2002).

Approximation to the data: penalized and marginal quasi-likelihood

For GLMs, quasi-likelihood is attractive because of its ability to generate highly efficient estimators without making precise distributional assumptions (McCulloch and Searle, 2001). The quasi-likelihood does not specify a distribution, only the mean and the variance (Agresti, 2002). For GLMMs, optimization of the quasi-likelihood function (which involves the first- and second-order conditional moments) is augmented with a penalty term on the random effects and therefore, it is called Penalized Quasi-likelihood (PQL) (Molenberghs and Verbeke, 2005). PQL can be considered as an approximated inference method in GLMM. There have been a number of proposed versions of this approximate method. We will confine our discussion to the widely used methods, namely, algorithms proposed by Breslow and Clayton (1993), Schall (1991) and Wolfinger and O'Connell (1993). The similarities and differences between these methods will be highlighted.

The method proposed by Breslow and Clayton (1993) can be summarized as follows (Jiang, 2007). In the GLM context, and based on (5.3), for any type of GLMM, the quasi-likelihood form is given by

$$QL(\mathbf{y}, \beta, \mathbf{b}) = QL(\mathbf{y} \mid \mathbf{b}) L(\mathbf{b})$$

and the integrated quasi-likelihood function is given as

$$L_Q \propto (2\pi)^{-q/2} |\mathbf{G}|^{-1/2} \int \exp\left[-\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i; \mu_i) - \frac{1}{2} \mathbf{b}' \mathbf{G}^{-1} \mathbf{b} \right] d\mathbf{b} \qquad (5.13)$$

where the subscript Q indicates quasi-likelihood, and

$$d_i(y_i;\mu_i) = -2\int_{y_i}^{\mu_i} \frac{\mathbf{y}_i - u}{a_i(\phi)v(u)}du \qquad (5.14)$$

is known as the (quasi-) deviance. If y is Gaussian and g^{-1} is the identity, the integral in (5.14) is normal and may be evaluated in closed form. If not, this expression contains integrals that must be solved using numerical methods.

Expression (5.13) has the form $c\,|\mathbf{G}|^{-\frac{1}{2}}\int e^{-Q(\mathbf{b})}d\mathbf{b}$, with a constant term c and Q is a function of $\mathbf{b}$. Therefore the Laplace method can be used to approximate the integral in (5.13). Using the result in (5.8), the logarithm of L_Q, denoted by l_Q is given by

$$l_Q \approx c - \frac{1}{2}\log|\mathbf{G}| - \frac{1}{2}\log\left|q''(\hat{\mathbf{b}})\right| - q(\hat{\mathbf{b}}),$$

where $q(\mathbf{b})$ is defined as

$$q(\mathbf{b}) = \frac{1}{2}\left(\sum_{i=1}^{n} d_i(y_i;\mu_i) + \mathbf{b}'\mathbf{G}^{-1}\mathbf{b}\right).$$

Typically, $\hat{\mathbf{b}} = \hat{\mathbf{b}}(\beta,\theta)$ is the solution to

$$q'(\mathbf{b}) = \mathbf{G}^{-1}\mathbf{b} - \sum_{i=1}^{n}\frac{\mathbf{y}_i - \mu_i}{a_i(\phi)v(\mu_i)g'(\mu_i)}\mathbf{z}_i = \mathbf{0}$$

that minimizes $q(\mathbf{b})$. In addition, the second derivative of q is given by

$$q''(\mathbf{b}) = \mathbf{G}^{-1} + \sum_{i=1}^{n}\frac{\mathbf{z}_i\mathbf{z}_i'}{a_i(\phi)v(\mu_i)\{g'(\mu_i)\}^2} + r \approx \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1},$$

where the term r has expectation 0, $\mathbf{Z}$ is the matrix whose ith row is $\mathbf{z}_i'$ and $\mathbf{W}$ is the $N \times N$ diagonal matrix with diagonal terms $w_i = \left\{\phi a_i v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\right\}^{-1}$. Therefore, the log quasi-likelihood is given by

$$l_Q \approx -\frac{1}{2}|\mathbf{G}| - \frac{1}{2}\log|\mathbf{I} + \mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{G}| - \frac{1}{2\phi}\sum_i d_i(y_i,\hat{\mu}_i) - \frac{1}{2}\hat{\mathbf{b}}'\mathbf{G}^{-1}\hat{\mathbf{b}},$$

where $\hat{\mathbf{b}}$ is chosen to maximize the last two terms, $\mathbf{I}$ is the identity matrix. This expression leads to the PQL algorithm to estimate β and θ. Differentiating with respect to β and $\mathbf{b}$, the score equations for the mean parameters are

$$\sum_{i=1}^{n} \frac{(\mathbf{y}_i - \mu_i)\mathbf{x}_i}{a_i(\phi)v(\mu_i)g'(\mu_i)} = \mathbf{0}, \qquad (5.15)$$

and

$$\sum_{i=1}^{n} \frac{(\mathbf{y}_i - \mu_i)\mathbf{z}_i}{a_i(\phi)v(\mu_i)g'(\mu_i)} - \mathbf{G}^{-1}\mathbf{b} = \mathbf{0}. \qquad (5.16)$$

Breslow and Clayton (1993) propose an iterative procedure for solving (5.15) and (5.16). The attractive feature of the proposed procedure is that it exploits a close correspondence with well known mixed model equations (Henderson et al., 1959), which leads to the best linear unbiased prediction (BLUP) in linear mixed models. If we define the working vector $\mathbf{y}_i^*$ with components $\mathbf{y}_i^* = \eta_i + g'(\mu_i)(\mathbf{y}_i - \mu_i)$, and where η_i and μ_i are evaluated at the current estimators of β and $\mathbf{b}$. Then the solution to (5.15) and (5.16) using Fisher's scoring may be expressed as the iterative solution to the system

$$\begin{pmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{W}\mathbf{y}^* \\ \mathbf{Z}'\mathbf{W}\mathbf{y}^* \end{pmatrix}. \qquad (5.17)$$

Expression (5.17) is the same as in the linear mixed models equations (see Verbeke and Molenberghs, 2000). It should be noted that because $\mathbf{W}$ depends on β and $\mathbf{b}$, it has to updated at each iteration. Equivalently, the solution to (5.17) may be expressed as follows:

$$\beta = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}^*,$$
$$\mathbf{b} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y}^* - \mathbf{X}\beta\right),$$

where $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$.

The other version of PQL algorithm was developed by Schall (1991) in the context of longitudinal data. This method proposes to use a linearization of the conditional mean as a function of fixed and random effects. This method is based on a decomposition of the data into the conditional mean and an appropriate error term with Taylor series expansion of the mean that is a non-linear function of the linear predictor (Molenberghs and Verbeke, 2005). Consider the decomposition

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = g^{-1}\left(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}_i\right) + \varepsilon_{ij} \qquad (5.18)$$

where y_{ij} is the j^{th} outcome for the i^{th} subject. The $var(y_{ij}) = \phi v(\mu_{ij})$ for $v(\cdot)$ is the variance function.

Consider a linear Taylor expansion of (5.18) (summarized by Molenberghs and Verbeke, 2005) around current estimates $\hat{\beta}$ and $\hat{\mathbf{b}}_i$ of the fixed effects and random effects respectively. Thus,

$$\begin{aligned} y_{ij} &\approx g^{-1}\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right) \\ &\quad + (g^{-1})'\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right)\mathbf{x}'_{ij}\left(\beta - \hat{\beta}\right) \\ &\quad + (g^{-1})'\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right)\mathbf{z}'_{ij}\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) + \varepsilon_{ij} \\ &= \hat{\mu}_{ij} + v\left(\hat{\mu}_{ij}\right)\mathbf{x}'_{ij}\left(\beta - \hat{\beta}\right) + v\left(\hat{\mu}_{ij}\right)\mathbf{z}'_{ij}\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) + \varepsilon_{ij} \end{aligned}$$

where $\hat{\mu}_{ij}$ equals the current predictor $g^{-1}\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right)$ for the conditional mean $E(y_{ij} \mid \mathbf{b}_i)$. Rewriting the above expression in vector notation, it becomes

$$\mathbf{y}_i \equiv \hat{\mu}_i + \hat{\mathbf{V}}_i\mathbf{X}_i\left(\beta - \hat{\beta}\right) + \hat{\mathbf{V}}_i\mathbf{Z}_i\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) + \varepsilon_i .$$

For appropriate design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$, and with $\hat{\mathbf{V}}_i$ equal to the diagonal matrix with diagonal entries equal to $v\left(\hat{\mu}_{ij}\right)$. Reordering the terms yields

$$\mathbf{y}_i^* \equiv g'(\hat{\mu}_i)(\mathbf{y}_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}_i \approx \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i^* , \qquad (5.19)$$

where $\hat{\mathbf{V}}_i^{-1} = g'(\hat{\mu}_i)$ and $\varepsilon_i^* = g'(\mu_i)\varepsilon_i$, with zero mean. Expression (5.19) can be viewed as a linear mixed model for the pseudo data $\mathbf{y}_i^*$ with fixed effects β, random effects $\mathbf{b}_i$ and the error terms ε_i^*. This result yields an algorithm for fitting a GLMM. Given the starting values for the parameters β, $\mathbf{G}$ and ϕ in the marginal likelihood, empirical Bayes estimates are calculated for $\mathbf{b}_i$ and pseudo data $\mathbf{y}_i^*$ are computed. Then the approximate linear mixed model presented in (5.19) is fitted (for fitting of linear mixed models, see Verbeke and Molenberghs, 2000), and then the estimates of the parameters are updated. This process is iterated until convergence is reached. The resulting estimates are called *penalized quasi-likelihood* estimates because they are obtained from optimizing the quasi-likelihood function using approximations of first and second order.

An alternative approximation is very similar to PQL method, but is based on a linear Taylor expansion of the mean μ_{ij} around the current estimates of $\hat{\beta}$ for the fixed

effects and around $\mathbf{b}_i = \mathbf{0}$ for the random effects. This produces similar expressions to PQL but $\hat{\mu}_{ij}$ takes the form $g^{-1}\left(\mathbf{x}'_{ij}\hat{\beta}\right)$ rather than $g^{-1}\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right)$. The pseudo data are now of the form $\mathbf{y}_i^* \equiv g'(\hat{\mu}_i)(\mathbf{y}_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\beta}$, and satisfy the approximate linear mixed model

$$\mathbf{y}_i^* \approx \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i^*.$$

The resulting estimates are called *marginal quasi-likelihood* estimates (MQL). Similar to PQL estimates, MQL can be obtained by optimizing a quasi-likelihood function which only involves first- and second-order moments, but now evaluated at the linear predictor $\left(\mathbf{x}'_{ij}\hat{\beta}\right)$ rather than the conditional linear predictor $\left(\mathbf{x}'_{ij}\hat{\beta} + \mathbf{z}'_{ij}\hat{\mathbf{b}}_i\right)$.

A similar approximation method was proposed by Wolfinger and O'Connell (1993). They introduced pseudo-likelihood (PL) and restricted pseudo-likelihood (REPL) procedures for estimation. Their approach to PL/REPL is also based on a Gaussian approximation and Taylor's linearization theorem, already discussed in Schall's (1991) approach. That is, the non-linearity of a GLMM is removed by applying a first-order Taylor expansion to $g^{-1}\left(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{b}_i\right)$ about the current values of β and $\mathbf{b}_i$. This results in a weighted linear mixed model (5.19) that can be estimated with the standard linear mixed model methods. The estimates of β and $\mathbf{b}_i$ are used to update the pseudo-data and the process is repeated until convergence.

Wolfinger and O'Connell (1993) show that implementation of PQL and MQL procedures by Breslow and Clayton (1993) may be achieved with their algorithm. In fact Littell et al., (2006) argue that Breslow-Clayton and Wolfinger-O'Connell's procedures are similar in that they both use the generalized mixed model equations (5.19) for solutions of β and $\mathbf{b}_i$. Nonetheless, one of the differences between the two procedures is primarily based on the fact that Breslow and Clayton motivate their estimation procedures from a quasi-likelihood viewpoint using approximations based on Laplace's method whereas Wolfinger and O'Connell's approach to PL/REPL is based on a Gaussian approximation and Taylor's theorem (Wolfinger and O'Connell, 1993). Moreover, Littell et al., (2006) points out that the other difference between what Breslow and Clayton (1993) term PQL and what Wolfinger and O'Connell

(1993) term PL/REPL lies in the estimation of the parameter ϕ. In the Breslow-Clayton procedure, the scale parameter is fixed at $\phi = 1$, whereas in the Wolfinger-O'Connell procedure, it is always estimated.

5.4.2 A note on modeling serial correlation in GLMM

Using the linearization approach to estimation (Schall, 1991; Wolfinger and O'Connell, 1993) the linear mixed model for the pseudo data given in (5.19) is

$$\mathbf{y}_i^* \equiv g'(\hat{\mu}_i)(\mathbf{y}_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}_i \approx \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i^*,$$

where $\varepsilon_i^* = g'(\hat{\mu}_i)\varepsilon_i$, which has mean zero. The $E(\mathbf{y}_i^* \mid \mathbf{b}_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i$ and $var(\mathbf{y}_i^* \mid \mathbf{b}_i) = g'(\hat{\mu}_i)\boldsymbol{\Sigma}_i g'(\hat{\mu}_i)$ where $\boldsymbol{\Sigma}_i = var(\mathbf{y}_i - \hat{\mu}_i)$. The variance function can be expressed such that

$$var(\varepsilon_i) = \Phi^{1/2}\mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}\Phi^{1/2},$$

where Φ is the diagonal matrix with over dispersion parameters along the diagonal, $\mathbf{A}_i$ is a diagonal matrix containing the variances from the model specification of y_{ij} given the random effects $\mathbf{b}_i$ and $\mathbf{R}_i$ is the correlation matrix. The distribution of ε_i^* with a Gaussian distribution has approximately the same first two moments as the distribution of ε_i (Wolfinger and O'Connell, 1993). In particular, it is assumed that ε_i^* is Gaussian with mean $\mathbf{0}$ and variance $\Phi^{1/2}\mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2}\Phi^{1/2}$. Then the marginal variance of the linear pseudo-model is defined as

$$var(\mathbf{y}^*) = \mathbf{V} = \mathbf{ZGZ}' + \Phi^{1/2}\mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i\mathbf{A}_i^{\frac{1}{2}}\Phi^{1/2}.$$

That is, the pseudo response variable $\mathbf{y}^*$ takes the form of a weighted linear mixed model with the diagonal weight matrix $\hat{\mathbf{W}} = \mathbf{A}^{-1}[g'(\hat{\mu})]^{-2}$. Then an iterative algorithm in which a linear mixed model is fitted to get estimates of β and $\mathbf{b}$ is used. The use of this estimation algorithm allows the introduction of autocorrelation at the level of the linear predictor in modeling the pseudo response variable $\mathbf{y}_i^*$ at each step of the iterative process. This takes the advantage of well-established correlation structures used with linear mixed models, which include an AR-1 structure if measurements are equally spaced (see Chapter 4) and spatial correlation models which are appropriate when time points do not occur at pre-determined intervals (Brown and Prescott, 2006). With spatial correlation models, the correlation is specified as a continuous

function of the absolute difference in times between two observations (Weiss, 2005). There are a number of ways of defining covariances from the time interval. For instance, the spatial power and spatial Gaussian correlations for four time points are defined as

$$cov(\mathbf{y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \rho^{|t_1-t_4|} \\ & 1 & \rho^{|t_2-t_3|} & \rho^{|t_2-t_4|} \\ & & 1 & \rho^{|t_3-t_4|} \\ & & & 1 \end{pmatrix}$$

and

$$cov(\mathbf{y}_i) = \sigma^2 \begin{pmatrix} 1 & e^{\left(\frac{|t_1-t_2|^2}{\rho^2}\right)} & e^{\left(\frac{|t_1-t_3|^2}{\rho^2}\right)} & e^{\left(\frac{|t_1-t_4|^2}{\rho^2}\right)} \\ & 1 & e^{\left(\frac{|t_2-t_3|^2}{\rho^2}\right)} & e^{\left(\frac{|t_2-t_4|^2}{\rho^2}\right)} \\ & & 1 & e^{\left(\frac{|t_3-t_4|^2}{\rho^2}\right)} \\ & & & 1 \end{pmatrix}$$

respectively.

Note that as much as the random effects and residual correlation enter the pseudo data in (5.19) side by side, the residual error of the pseudo data is now a transformed version of the original error ε_i. Therefore the autocorrelation parameter obtained using this approach has to be considered only as an approximate of the underlying correlation structure of its working variate (Molenberghs and Verbeke, 2005).

5.5 Model selection

The primary objective of model selection is to choose the simplest model that provides the best fit to the data. West, Welch and Galecki; (2007) pointed out that the process of building a model with both fixed and random effects given a set of longitudinal or clustered data is an iterative one, a series of model fitting steps and investigations are required, which include the selection of appropriate mean and covariance structures for the observed data. Model building typically involves a balance of statistical and subject matter considerations; there is no single strategy that applies to every application (West et al., 2007). A generalized linear mixed

model can be fitted with a variety of possible covariance structures for $\mathbf{G}$ and $\mathbf{R}_i$. Unless robust inference is used, an appropriate covariance model is essential to obtain valid inferences for the parameters in the mean structure, hence the need for careful model building.

The recommended general guidelines for model building (Verbeke and Molenberghs, 2000) are as follows. First, a preliminary mean structure has to be selected by fitting an over-elaborated model. This is to remove the systematic trends in the data that cannot be explained by the covariance structure. Second, a preliminary random effects structure is then selected. The hierarchical nature of random effects has to be taken into account when a model is fitted. For instance, the design matrix of the random effects should not contain a polynomial effect if not all hierarchically inferior terms are included. Third, a residual covariance structure is then selected conditional on the preliminary structure of the selected random effects. Accordingly, based on the selected residual covariance structure, the need for random effects should be re-assessed. With the selected covariance structure, tests for the fixed effects can then be carried out.

In the process of model selection, one has to be wary of over-parameterization of the model structure that would lead to inefficient estimation and potentially poor assessment of standard errors for the estimates (Verbeke and Molenberghs, 2000). As a result, a model reduction exercise of both the fixed and random parameters is essential in order to achieve a parsimonious model. Model reduction is commonly done by comparing tests of significance in different models. With model selection of fixed effects, the likelihood ratio (LR) test is commonly used for comparison of two models when one model is a special case of the other. That is, it compares two models with different mean structure but with the same covariance structure. The LR test for two nested models is constructed by comparing the maximized log-likelihoods for the full and reduced models respectively and the test statistic is defined as

$$-2\ln\lambda_N = -2\ln\left[\frac{L_{ML}(\hat{\theta}_{ML,0})}{L_{ML}(\hat{\theta}_{ML})}\right] \qquad (5.20)$$

where $\hat{\theta}_{ML,0}$ and $\hat{\theta}_{ML}$ are respective maximum likelihood estimates which maximize the ML likelihood functions of the reduced and full models. The asymptotic null distribution of the LR test statistic is a chi-square distribution with the degrees of freedom equal to the difference between the number of parameters in the two models. Small values of $-2\ln\lambda_{\mathrm{N}}$ are obtained when the reduced model is similar to the full model, indicating that the reduced model is a good one.

Similarly, with hypothesis testing of variance components, the LR test can be derived for comparing nested models with different covariance structures but with the same mean. Also, the asymptotic null distribution of the LR test statistic is a chi-square distribution with the degrees of freedom equal to the difference between the number of parameters in the two models. However, the normal approximation fails if the variance parameter to be tested takes values on the boundary of the parameter space. Self and Liang (1987) and Stram and Lee (1994, 1995) show that the asymptotic null distribution for the LR test statistic for testing the hypothesis of the need for random effects is often a mixture of chi-squared distributions rather than the classical single chi-squared distribution. This is derived under the assumption of conditional independence assumption. The specific LR tests are presented as follows (Verbeke and Molenberghs, 2000):

Case 1 - No random effects versus one random effect: For testing $H_0 : \mathbf{G} = 0$ versus $H_A : \mathbf{G} = g_{11}$, where g_{11} is non-negative scalar that represents the variance component of the random effect. Then the asymptotic null distribution of $-2\ln\lambda_{\mathrm{N}}$ is a mixture of χ_1^2 and χ_0^2 with equal weights of 0.5. The χ_0^2 distribution is the distribution that gives probability mass 1 to value 0.

Case 2 – One versus two random effects: In the case where one wishes to test

$$H_0 : \mathbf{G} = \begin{pmatrix} g_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

for strictly positive g_{11}, versus H_A that $\mathbf{G}$ is a (2×2) positive semi-definite matrix. The asymptotic null distribution of $-2\ln\lambda_{\mathrm{N}}$ is a mixture with equal weights of 0.5 for χ_2^2 and χ_1^2.

Case 3 - q versus $q+1$ random effects: In the case where one wishes to test

$$H_0 : \mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix},$$

for which $\mathbf{G}_{11}$ is a $(q \times q)$ positive definite matrix, versus H_A that $\mathbf{G}$ is a general $((q+1)\times(q+1))$ positive semi-definite matrix. The large sample behaviour of the null distribution of $-2\ln\lambda_{\mathrm{N}}$ is a mixture of χ^2_{q+1} and χ^2_q again with equal weights of 0.5.

Case 4 - q versus $q+k$ random effects: In this case, one wishes to test the H_0 in case 3 versus

$$H_A : \mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}'_{12} & \mathbf{G}_{22} \end{pmatrix},$$

in which $\mathbf{G}$ is a general $((q+k)\times(q+k))$ positive semi-definite matrix. The null distribution of $-2\ln\lambda_{\mathrm{N}}$ is a mixture of χ^2 random variables formed by the lengths of projections of multivariate normal random variables upon curved as well as flat surfaces.

It should be noted that if the classical null distribution is used for the cases presented above, then all the p-values would be overestimated. Consequently the null hypothesis would be accepted too often, resulting in incorrectly simplifying the covariance structure of the model, thus invalidating inferences (Altham, 1984). The correction for the boundary parameter values under the null hypotheses therefore reduces the p-values in order to protect against the use of an oversimplified or a too parsimonious covariance structure.

A set of useful tools in model selection of non-nested models are usually referred to as Information Criteria (IC). Recall that the idea behind the LR test for comparing model A to a more extended model B is to select model A if the increase in likelihood under model B is small compared to an increase in complexity. Likewise, with the comparison of non-nested models, the model with the largest likelihood is selected provided it is not too complicated. Under the IC method, the model with the highest penalized log-likelihood $l - \Gamma(\#\theta)$ for some penalty function $\Gamma(.)$ dependent on the

number of parameters, $\#\theta$ is selected. That is, the IC provides a way to assess the fit of the model based on its optimum log-likelihood value after applying a penalty for the number of parameters that are estimated in the model (West et al.; 2007). Different forms of $\Gamma(.)$ lead to different criteria and some commonly used functions are presented in Table 5.1 (Verbeke and Molenberghs, 2000).

Table 5. 1 Commonly used information criteria

Criteria	Definition of $\Gamma(.)$
Akaike (AIC)	$\Gamma(\#\theta) = \#\theta$
Schwarz (BIC)	$\Gamma(\#\theta) = (\#\theta \ln n^*)/2$
Hannan and Quinn (HQIC)	$\Gamma(\#\theta) = \#\theta \ln(\ln n^*)$
Bozdogan (CAIC)	$\Gamma(\#\theta) = \#\theta(\ln n^* + 1)/2$

n* is equal to the total number $n = \sum_{i=1}^{N} n_i$ of observations or equal to $n - p$, depending on whether ML or REML estimation was used in the calculations respectively.

Note from Table 5.1 that, except for the Akaike information criterion (AIC), the other IC involve the sample size. This shows that differences in likelihood need to be viewed, not only relative to the differences in the numbers of parameters but also relative to the number of observations included in the analysis. As the sample size increases, more severe increases in the likelihood are required before a complex model will be preferred over a simple model. It should be emphasized that IC are not formal testing procedures, they only provide rules of thumb to discriminate between several statistical models (Verbeke and Molenberghs, 2000).

Note however, that with the linearization estimation techniques where parameters are estimated by fitting linear mixed models to pseudo-data (e.g PQL, MQL), nested as well as non-nested models cannot be compared using values based on the likelihood computation (Molenberghs and Verbeke, 2005). This is because when one changes the error structure, then the linearized response also changes. Since the values for the likelihood ratio test and information criteria are based on the likelihood, they cannot be used to compare models with the linearization estimation techniques. One can only compare likelihood values across models when the response variable for each model is the same. Nonetheless, tests that include

approximate Wald tests can still be used to test hypotheses about both fixed effects and random effects as discussed in the next section.

5.6 Inference for fixed and random effects

The ultimate goal in statistical analysis is to draw inferences about the parameters in a model in order to generalize results obtained from a specific sample to the general population from which the sample was drawn. Since fitting of GLMMs is based on maximum likelihood principles, inferences for parameters are obtained from classical maximum likelihood theory. Assuming that the fitted GLMM is appropriate, the obtained estimators are asymptotically normally distributed with the correct values as means, and with the inverse Fisher information matrix as covariance matrix (Molenberghs and Verbeke, 2005). As a result, Wald-type tests can be performed for comparing standardized estimates to the standard normal distribution. Also composite hypotheses can be tested using the more general formulation of the Wald statistic, which is a standardized quadratic form that is compared to the chi-squared distribution. Alternatively, likelihood ratio and score test can be used.

As discussed in Section 5.4, the parameters in GLMMs are often estimated by fitting linear mixed models to pseudo-data. Therefore, precision estimates for the fixed effects and for the random effects are often calculated using linear mixed model methodology (Molenberghs and Verbeke, 2005). Accordingly, tests for the fixed effects include the approximate Wald test (also referred to as the Z-test) and the approximate t-tests and F-tests and they are summarized by Verbeke and Molenberghs (2000) as follows. For fixed effects, approximate Wald test (also known as the Z-test) can be obtained from approximating the distribution of $(\hat{\beta}_j - \beta_j)/\text{s.e.}(\hat{\beta}_j)$ by a standard univariate normal distribution for each parameter $\beta_j, j = 1, \cdots, p$. More generally, it may be of interest to construct confidence intervals and tests of hypotheses about certain linear combinations of the components of β. For instance, given any known matrix $\mathbf{L}$, a test for the hypothesis

$$H_0 : \mathbf{L}\beta = \mathbf{0} \quad \text{versus} \quad H_0 : \mathbf{L}\beta \neq \mathbf{0} \tag{5.21}$$

follows from the fact that the distribution of

$$(\hat{\beta} - \beta)' \mathbf{L}' \left[\mathbf{L} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_i^{-1}(\alpha) \mathbf{X} \right)' \mathbf{L}_i' \right]^{-1} \mathbf{L}(\hat{\beta} - \beta)$$

follows a chi-squared distribution with rank $(\mathbf{L})$ degrees of freedom. However, it should be noted that the Wald test statistic is based on estimated standard errors which underestimate the true variability of $\hat{\beta}$. This is because the variability introduced by estimating the variance parameters is not taken into account. This downward bias can be resolved by using approximate t- and F-tests for testing the hypothesis about β. For each parameter β_j in vector β, $j = 1, \cdots, p$, an approximate *t*-test and associated confidence interval can be obtained by approximating the distribution of $(\hat{\beta}_j - \beta_j)/\text{s.e.}(\hat{\beta}_j)$ by an appropriate t-distribution. Testing general linear hypotheses of the form (5.21) is thus based on an F-approximation to the distribution of

$$F = \frac{(\hat{\beta} - \beta)' \mathbf{L}' \left[\mathbf{L} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_i^{-1}(\hat{\alpha}) \mathbf{X}_i \right)^{-1} \mathbf{L}' \right]^{-1} \mathbf{L}(\hat{\beta} - \beta)}{rank(\mathbf{L})},$$

with the numerator degrees of freedom equal to rank($\mathbf{L}$), while the denominator degrees have to be established from the data. There are several methods that are available for estimating the denominator degrees of freedom; one of which is the Satterthwaite approximation. Note that even though different methods for the degrees of freedom may lead to severe differences in the resulting p-values, in the context of longitudinal data, whatever estimation method used, lead to very similar p-values. This is due to the fact that in a longitudinal setting, different subjects contribute independent information, which results in numbers of degrees of freedom which are typically large enough to lead to very similar p-values.

With regard to inference for variance components $\mathbf{G}$, classical Wald, likelihood ratio and score tests can be used, as long as the hypotheses to be tested are not on the boundary of the parameter space (that is, testing the hypothesis that the variance of the population distribution is zero). Because variances cannot be negative, zero is at the boundary of the parameter space and as such, only a one-sided hypothesis test can be carried out (Verbeke and Molenberghs, 2003). For instance, the classical

Wald test (based on standard normal approximation) would no longer be valid if one were to test whether the variance g_{11} of a single random effect is equal to zero, i.e $H_0 : g_{11} = 0$ versus $H_A : g_{11} \geq 0$. Thus, under H_0 the Wald statistic is asymptotically equivalent and χ_1^2 in 50% of the cases, and equal to zero in the other 50% of the cases. Hence the null distribution is a mixture of the χ_0^2 (with all probability mass at zero) and χ_1^2, with equal probability 0.5 (Molenberghs and Verbeke, 2005). Consequently, the p-value is obtained by halving the classical p-value that is obtained when testing a hypothesis for a single parameter. Similar properties can be obtained for the one-sided likelihood ratio test (Self and Liang, 1987; Stram and Lee, 1994) and the one-sided score test (Silvapulle and Silvapulle, 1995; Verbeke and Molenberghs, 2003). Moreover, even with the information criteria, there are still concerns about the boundary effects and estimation of degrees of freedom for random effects (Vaida and Blanchard, 2005).

5.7 Evaluating adherence using GLMMs

The conditional independence model was fitted to the data first. The conditional independence model refers to the context where it is assumed that the correlation between measurements on the same subjects is modeled only by the random effects while conditionally upon them, the repeated measurements are assumed to be independent. This model has been presented in Section 5.2, expression (5.2). With both fixed and random effects in the model, model selection requires a series of model fitting steps and investigations on the appropriateness of mean and covariance structure of the observed data. In fitting such a model, a preliminary covariance structure of the random effects was evaluated using the preliminary mean structure which includes all the main effects (fixed) in the data. All the main effects were used in the preliminary structure in order to remove, as much as possible, the systematic trends in the data that cannot be explained by the covariance structure. Once the preliminary covariance structure was selected, the mean structure was evaluated by carrying out tests for fixed effects. Then with the selected mean structure, the need for random effects was re-assessed.

The results of fitting models to the adherence data using different maximum likelihood estimation techniques as discussed in Section 5.4, are presented and

compared. In particular, the models are fitted with penalized quasi likelihood (PQL), with Laplace approximation as well as classical Gaussian and adaptive Gaussian quadrature. The data is analyzed using SAS (version 9.1.3) procedures; GLIMMIX and NLMIXED. PQL is an approximation method to the data and has been implemented in the SAS procedure GLIMMIX. The classical Gaussian and adaptive Gaussian quadrature, as approximations to the integral in the marginal likelihood have been implemented in the SAS procedure NLMIXED. Model fitting based on the Laplace approximation for the integrals in the marginal likelihood can be specified by choosing adaptive Gaussian quadrature with one quadrature point. One advantage of estimation methods implemented in NLMIXED is that they offer a true log-likelihood whereas the methods implemented in GLIMMIX offer pseudo-likelihood. The advantage of GLIMMIX on the other hand, is that it allows greater flexibility in the types of models that can be estimated and the number of random effects that can be specified. For instance, GLIMMIX is flexible enough to fit complex models that accommodate serial correlation in addition to random effects, whereas NLMIXED cannot accommodate such models.

Evaluation of random effects structure

For adherence data, the hypothesis of interest is that random intercepts and random slopes for the linear time effect are needed in the model. We tested our hypothesis by adding one random effect at a time to the mean model. The following models (fixed effects only, fixed effects plus random intercepts and fixed effects, random intercepts and random slopes) are considered:

$$y_{ij} \mid \mathbf{b}_i \sim \text{Bernoulli}(\mu_{ij})$$

$$\begin{aligned}\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\ &+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\ &+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\ &+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij}\end{aligned} \tag{5.22}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i$$
$$+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i$$
$$+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i$$
$$+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + b_{0i} \quad (5.23)$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i$$
$$+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i$$
$$+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i$$
$$+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + b_{0i} + b_{1i} time_{ij} \quad (5.24)$$

where $y_{ij} = 1$ if the ith $(i = 1,\ldots,688)$ patient adheres to medication in the jth time (follow-visits) $(j = 1,\ldots,17)$ and $y_{ij} = 0$ otherwise. Fixed effects are represented by the $\beta_1, \beta_2, \ldots, \beta_{20}$, b_{0i} represents random intercepts and b_{1i} represents random slopes. In model (5.23), it is assumed that random intercepts b_{0i} are normally distributed with mean zero and variance g_{11} whereas in model (5.24), it is assumed that the random intercepts and slopes $\mathbf{b}_i = (b_{0i}, b_{1i})'$ have a bivariate normal distribution with zero mean and 2×2 covariance matrix $\mathbf{G}$.

Results from integral approximation methods

We first present the results from fitting the models (5.22), (5.23) and (5.24) using the classical Gaussian and adaptive Gaussian quadrature as well as the Laplace approximation. Recall that with these methods, the likelihood obtained is based on the numerical integration of the actual observed data; therefore the likelihood ratio test can be used to compare the nested models. However, it should be noted that with classical Gaussian and adaptive Gaussian quadrature, the obtained log-likelihood value equals the maximum of the approximation to the model likelihood, which implies that log-likelihoods corresponding to different quadrature points are not

necessarily comparable. As a result, for model building purposes, we need to assess the effect of the different numbers of quadrature points used on the estimates with a view to selecting a specific number of quadrature points that provide good accuracy in order to compare different nested models.

The impact of the number of quadrature points used in the estimates of a fitted model when using classical Gaussian and adaptive Gaussian quadrature was assessed by fitting a single model (5.23) for varying numbers of quadrature points, Q = 3, 5, 10 and 20. The different values for quadrature points, Q did not lead to considerable differences in the parameter estimates nor the standard errors. As the quadrature points, Q, were increased from 3 to 5, a minimal change in estimates and their corresponding standard errors was observed with adaptive Gaussian quadrature. However, the estimates and their standard errors were similar with 5, 10 and 20 quadrature points. On the other hand, with classical Gaussian quadrature, there were slight differences in estimates when quadrature points were increased from 3 to 5. For instance, the increase of Q from 3 to 5 increased the estimate for the variance of the random intercepts, g_{11} from 0 to 0.2852. Nonetheless, the estimates and their corresponding standard errors were similar for 10 and 20 quadrature points. Consequently, for model selection, 10 and 20 quadrature points were then used for adaptive Gaussian and classical Gaussian quadrature respectively.

As was discussed earlier, it is worth noting that our results confirm to a certain degree the fact that with classical Gaussian quadrature, a larger number of quadrature points are necessary to obtain high accuracy while the adaptive Gaussian quadrature provides good accuracy with fewer quadrature points. On the other hand, in fitting the model (5.23), there was a huge difference in terms of the time taken by the two estimation methods to converge. The adaptive Gaussian quadrature was much more time consuming than the classical Gaussian quadrature. As the number of quadrature points were increased, the time consumed by adaptive Gaussian quadrature escalated compared to classical Gaussian quadrature. This was not surprising because as discussed in Section 5.4, the adaptive Gaussian

quadrature requires calculation of the mode ($\hat{s}$) for each unit in the dataset, hence the numerical maximization of N functions of the form (5.9).

The different random effects models were fitted and their associated minus twice log-likelihood values from Laplace approximation, adaptive and classical Gaussian quadrature are presented in Table 5.2 whereas the likelihood ratio tests for the different models are presented in Table 5.3. Since the log-likelihood values for adaptive Guassian and classical Guassian quadrature are the same (Table 5.2), the likelihood ratio test was computed using only one of them, which is the adaptive Guassian quadrature.

Table 5. 2: Random effects models with associated value for the log-likelihood value for Laplace, adaptive Gaussian and classical Gaussian quadrature estimation techniques

Random effects	-2 Log-likelihood values		
	Laplace approximations	Gaussian quadrature	
		Adaptive (Q=10)	Classical (Q=20)
Model 5.22: _____	4542.3	4542.3	4542.3
Model 5.23: Intercepts	4504.2	4503.1	4503.1
Model 5.24: Intercepts, time	4650.5	4668.2	4668.2

Table 5. 3: Likelihood ratio statistics for comparing random effects models and the associated null distribution (a mixture of chi-squared distribution) for Laplace approximation and adaptive Gaussian quadrature

Hypothesis	Likelihood ratio statistics		
	Null distribution	Laplace	Adaptive Gaussian
Model 5.22 vs Model 5.23	$\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$	38.1 (p<0.001)	39.2 (p<0.001)
Model 5.23 vs Model 5.24	$\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$	33.5 (p<0.001)	34.9 (p<0.001)

The need for the random intercepts in the model was assessed by carrying out tests that compare models 5.22 and 5.23. The likelihood ratio tests from Laplace and adaptive Gaussian quadrature (computed from the minus twice log-likelihood values in Table 5.2) yield values of 38.1 and 39.2 respectively (Table 5.3). These are highly significant (p<0.0001) indicating that the random intercepts are needed in the model.

Next, we evaluated whether in addition to the random intercepts, the inclusion of the random slopes in the model is necessary, i.e model 5.23 versus model 5.24. The likelihood ratio tests from Laplace and adaptive Gaussian quadrature yielded statistics of 33.5 and 34.9 respectively (see Table 5.3). These are highly significant ($p<0.0001$). The other half of the null distribution (Table 5.3) with this test has two degrees of freedom because we are testing the null hypothesis that the slope variance g_{22} and slope-intercept covariance g_{12} are jointly equal to zero. In this case the null hypothesis is rejected, indicating that both random intercepts and slopes are needed in the model.

Penalized quasi-likelihood results

With the PQL estimates, the approximate Wald test is used to test the significance or otherwise of the variance components. In this instance, the likelihood ratio tests cannot be used because the likelihood computation is based on the linear mixed models for pseudo-data. To evaluate whether the random intercepts are necessary, the hypothesis to be tested is $H_0 : g_{11} = 0$ versus $H_A : g_{11} \geq 0$. The Wald statistic for random intercepts is 29.6 ($p<0.0001$) which is significant, indicating that the random effects are necessary. It should be noted that since the null hypothesis of the variance component to be tested was on the boundary of the parameter space, a mixture of chi-square distribution as illustrated in Table 5.3 (for Laplace and classical Gaussian quadrature) was used to carry out this test.

Then, we evaluated whether in addition to the random intercepts, the inclusion of the random slopes in the model is necessary, i.e model 5.23 versus model 5.24. More specifically, the hypothesis to be tested is $H_0 : g_{12} = g_{22} = 0$, which is still on the boundary of the parameter space. The observed value for the Wald statistic equals

$$\begin{pmatrix} \hat{g}_{12} & \hat{g}_{22} \end{pmatrix} \begin{pmatrix} var(\hat{g}_{12}) & cov(\hat{g}_{12}, \hat{g}_{22}) \\ cov(\hat{g}_{12}, \hat{g}_{22}) & var(\hat{g}_{22}) \end{pmatrix}^{-1} \begin{pmatrix} \hat{g}_{12} \\ \hat{g}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} -0.06628 & 0.01131 \end{pmatrix} \begin{pmatrix} 0.000377 & -0.00004 \\ -0.00004 & 0.000006 \end{pmatrix}^{-1} \begin{pmatrix} -0.06628 \\ 0.01131 \end{pmatrix}$$

$$= 21.97$$

which is significant when compared to a mixture of two chi-squared distributions with 1 and 2 degrees of freedom, with equal weights of 0.5 ($p<0.0001$).

All the estimation techniques suggest a similar preliminary conclusion, i.e that a model with randomly varying intercepts and slopes fitted the data well.

Evaluation of the mean structure

The mean structure is examined by first, evaluating whether factors that affect initial adherence are still important when adherence is studied over time at a subject specific level. Recall that the final model for initial adherence (Section 3.4) contained all the main effects as well as three two-way interaction terms, which were between age and cell phone ownership, gender and patient's reported reason for taking an HIV test as well as treatment site and whether or not a patient is the source of household income. To evaluate whether the three two-way interactions in the baseline model are still significant, the following model was compared with model (5.24).

$$\begin{aligned}\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\ &+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\ &+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\ &+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} age_i cellphone_i \\ &+ \beta_{22} gender_i reason\_no_i + \beta_{23} gender_i reason\_risk_i + \beta_{24} site_i income_i + b_{0i} + b_{1i} time_{ij}\end{aligned} \quad (5.25)$$

Likelihood ratio tests were calculated using the Laplace approximation, classical Gaussian and Adaptive Gaussian quadrature. Let model (5.24) be denoted by M_0 and model (5.25), with additional interactions from the initial analysis model, be denoted by M_1. To compare the two models, the likelihood ratio test statistic is computed by calculating the difference $(-2l_0)-(-2l_1)$ where l_0 is the log-likelihood of model M_0 and l_1 is the log-likelihood of model M_1. In addition, the degrees of freedom $m_1 - m_0$ are calculated where m_1 and m_0 are the respective number of parameters in the models M_1 and M_0. The likelihood ratio statistics is then compared with the chi-squared value χ^2_{α, m_1-m_0}.

The results showed that the likelihood ratio tests from all the estimation methods were insignificant. The likelihood statistics with their corresponding *p*-values from Laplace was 5.5 (p=0.103) whereas the likelihood statistics for adaptive Gaussian and classical Gaussian quadrature was 5.6 (p=0.105) at 4 degrees of freedom. Again the results from classical ($Q = 20$) and adaptive ($Q = 10$) Gaussian quadrature were the same.

With the penalized quasi-likelihood estimation method, the multi-parameter or generalized Wald test is computed to test the joint null hypothesis that the three interaction terms are equal to zero. That is, to compare models (5.24) and (5.25), the null hypothesis being tested is

$$H_0 = \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = 0.$$

In order to carry out the generalized Wald test, let us define a $q \times p$ indicator matrix $\mathbf{C}$ of ones and zeros to select the parameters of interest. Here p equals the number of regressors in model (5.25), i.e., the full model (including the intercept) and q equals the number of parameters in the multi-parameter test (i.e., the difference in regressors between model (5.25) and model (5.24). Each row of $\mathbf{C}$ contains a 1 in only one location, and zeros elsewhere, in order to select one of the parameters that comprise this test. In fact, there are q rows in this matrix because each row is used to uniquely select one of the q parameters. For comparing models (5.25) and (5.24), q = 4, p = 24 and the indicator matrix is given as

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then the generalized Wald test equals

$$\mathbf{X}^2 = \hat{\beta}'\mathbf{C}'(\mathbf{C}Var(\hat{\beta})\mathbf{C}')^{-1}\mathbf{C}\hat{\beta},$$

which is distributed as χ^2 with q degrees of freedom under the null hypothesis. The test produced a Wald statistic of 5.20 with 4 degrees of freedom (p=0.2653) and is not significant. Additionally, none of the two-way interactions of the three variables

were significant (age*cell phone: p=0.5841; gender*reason: p=0.2560; treatment site*income: p=0.0880).

All the estimation techniques give similar results, that is, the additional two-way interaction terms that were significant when adherence was modeled at month one of initiation into therapy are no longer significant when adherence is modeled over time and at a subject specific level.

We then proceeded to assess whether factors that determined adherence at the population level using the GEE are still valuable when adherence is assessed over time at a subject specific level. The results from the GEE model (4.10) includes all the main effects and five two-way interaction terms. Three of the two-way interaction terms involved time with gender, treatment site and patient's reported reason for taking an HIV test. The other two interactions were between age and gender as well as age and educational level. A following random effects model was fitted with the same mean structure from the marginal model results (GEE results):

$$\begin{aligned}\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\ &+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\ &+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\ &+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} time_i gender_i \\ &+ \beta_{22} time_i site_i + \beta_{23} time_i reason\_no_i + \beta_{24} time_i reason\_risk_i + \beta_{25} age_i educ\_no_i \\ &+ \beta_{26} age_i educ\_p_i + \beta_{27} age_i gender_i + b_{0i} + b_{1i} time_{ij}\end{aligned} \tag{5.26}$$

For evaluating whether factors that affect adherence at the population level still affect adherence at a patient specific level, models (5.26) and (5.24) were compared using results from the different estimation methods. With the Laplace approximation, classical Gaussian and adaptive Gaussian quadrature, likelihood ratio tests (computed in the same way as in the section above) were used to compare the models (5.26) and (5.24). The likelihood ratio statistics from Laplace was 25 (p=0.0003) and adaptive Gaussian and classical Gaussian quadrature have the same value of 26.4 (p=0.0004), all at 5 degrees of freedom. These likelihood ratio

tests indicate that the additional interactions from the population average model are significant.

With the penalized quasi-likelihood estimation method, the multi-parameter or generalized Wald test (computed in the same manner as in the section above) is used to test the joint null hypothesis that the five interaction terms are equal to zero. That is, to compare models (5.24) and (5.26), the null hypothesis being tested is

$$H_0 = \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = \beta_{27} = 0,$$

where the parameters are in model (5.5). The test produced a Wald statistic of 48.26 with 5 degrees of freedom ($p<0.0001$), which is significant. Further, the results show that individual two-way interactions were significant (time*gender: $p<0.0001$; time*site: $p<0.0001$; time*reason: $p=0.0255$; age*gender: $p=0.0125$) except for the interaction between age and educational level ($p=0.2555$).

Consequently, the model with the GEE results (5.26) was the preferred model since all the estimation techniques give similar results, that is, the additional two-way interaction terms that were significant when adherence was modeled at the marginal level are still significant when adherence is modeled over time and at a subject specific level.

Re-assessment of the preliminary covariance structure

All the estimation techniques give similar results, indicating that the additional two-way interactions in the GEE model are significant. Consequently, the model with the GEE results was the preferred model. Then the preliminary covariance structure (random intercepts and slopes) was re-assessed with the now confirmed mean structure (GEE results). This was done by re-fitting model (5.26) but with the randomly varying slopes using different estimation techniques. The likelihood ratio tests from Laplace and adaptive Gaussian quadrature yielded statistics of 26.5 and 25.4 respectively. These were highly significant ($p<0.0001$), and therefore re-affirmed the presence of the random intercepts and random slopes in the model (5.26).

With the PQL estimates, we evaluated whether in addition to the random intercept, the inclusion of the random slopes in the model is necessary by testing the hypothesis that $H_0 : g_{12} = g_{22} = 0$, which is still on the boundary of the parameter space. The observed value for the Wald statistic equals

$$\begin{pmatrix} \hat{g}_{12} & \hat{g}_{22} \end{pmatrix} \begin{pmatrix} var(\hat{g}_{12}) & cov(\hat{g}_{12}, \hat{g}_{22}) \\ cov(\hat{g}_{12}, \hat{g}_{22}) & var(\hat{g}_{22}) \end{pmatrix}^{-1} \begin{pmatrix} \hat{g}_{12} \\ \hat{g}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} -0.05508 & 0.009395 \end{pmatrix} \begin{pmatrix} 0.000346 & -0.00004 \\ -0.00004 & 0.000005 \end{pmatrix}^{-1} \begin{pmatrix} -0.05508 \\ 0.009395 \end{pmatrix}$$

$$= 9.94$$

which is significant when compared to a mixture of two chi-squared distributions with 1 and 2 degrees of freedom, with equal weights of 0.5 (p=0.004).

All the estimation techniques suggest a similar finding, i.e that the GEE model with randomly varying intercepts and slopes fits the data well.

Interpretation of the results

The results of the final model (fixed effects and covariance parameters) from three estimation methods, Laplace approximations, adaptive Gaussian quadrature and PQL are presented in Table 5.4. There was virtually no difference with the parameter estimates between the numerical based methods, classical Gaussian ($Q = 10$) and adaptive Gaussian ($Q = 20$) quadrature, therefore only the results of adaptive Gaussian quadrature are presented. The estimation methods provided similar results in terms of the significance or otherwise of the estimates. It is however noted that generally PQL provided smaller estimates while adaptive Gaussian quadrature has the largest estimates, even though the margin of the difference is minimal. PQL and Laplace methods are known to perform poorly in cases with a relatively small number of repeated binary observations available for all subjects (Wolfinger, 1998). But in this case, both methods' performance was not relatively poor when compared to adaptive Gaussian quadrature. This might be due to sufficiently large number of repeated measurements per patient. Nonetheless, the

interpretation of the results will be done using estimates from an adaptive Gaussian quadrature.

There are four (two-way) significant interaction terms and two significant main effects (not involved in significant interaction). Three of the significant interactions involved time, namely time and gender, time and treatment site as well as time and reason for taking an HIV test. The other significant interaction is between age and gender. The significant main effects are cell phone ownership and whether or not a patient stays with a partner.

The results indicate that a typical patient (i.e. a patient with a random effect of zero) with a cell phone tends to adhere to medication 1.25 ($e^{0.233}$) times more than a typical patient without a cell phone. Again, it is shown that a typical patient who does not stay with a partner is 0.45 ($e^{-0.7524}$) times less likely to adhere to medication than a typical patient who stays with a partner. The results also show that conditional on random effects, adherence to medication is generally increasing, though the rate of increase differs by gender, treatment site and the patient's reported reason for taking an HIV test. The interaction terms are summarized below.

Table 5. 4: Parameter estimates (standard errors) for conditional independence model on HAART adherence

		Estimate (std error)		
Effect	**Parameter**	**Laplace Approx**	**adaptive Gaussian**	**PQL**
Intercept	β_0	0.880 (0.301)	1.6064 (0.306)	0.852 (0.510)
Gender (ref = male)				
Female	β_1	-1.563 (0.535)*	-1.555 (0.546)*	-1.269(.428)*
Education (ref= sec+)				
No schooling	β_2	1.692 (0.522)*	1.505 (0.534)*	1.538 (.643)*
Primary	β_3	-0.112 (0.699)	-0.108 (0.510)	0.064 (0.551)
Treatment Site (ref=rural)				
Urban site	β_4	1.283 (0.201)*	1.290 (0.204)*	1.258 (.165)*
Income (ref= not source of income)				
Source of income	β_5	0.053 (0.116)	0.053 (0.118)	0.038 (0.098)
Access to tap water (ref = No)				
Yes	β_6	0.089 (0.182)	0.089 (0.185)	0.044 (0.153)
Having electricity (ref = No)				
Yes	β_7	0.053 (0.156)	0.055 (0.159)	0.001(0.139)
Cell phone ownership (ref = No)				
Yes	β_8	0.224 (0.104)*	0.223 (0.106)*	0.255 (.089)*
WHO staging of disease (ref = stage 4)				
Stage 1	β_9	-0.459 (0.256)	-0.482 (0.280)	-0.351(0.228)
Stage 2	β_{10}	-0.359 (0.246)	-0.361 (0.250)	-0.250(0.199)
Stage 3	β_{11}	-0.253 (0.229)	-0.256 (0.232)	-0.091(0.180)
Partner (ref = living with partner)				
living without a partner	β_{12}	-0.023 (0.301)*	-0.5524 (0.306)*	-0.329(.102)*
Reason for taking HIV test (ref = unwell)				
VCT	β_{13}	-0.095 (0.215)	-0.093 (0.221)	-0.015(0.182)
Exposed to the virus	β_{14}	-0.534 (0.241)*	-0.535 (0.245)*	-0.485(.201)*
Initial adherence (ref = not adherent)				
Adherent	β_{15}	-0.118 (0.123)	-0.119 (0.125)	-0.138(0.105)
Time (visit)	β_{16}	0.133 (0.025)*	0.133 (0.025)*	0.123 (.021)*
Age	β_{17}	-0.021 (0.014)	-0.021 (0.014)	-0.015(0.011)
Baseline CD4 Cell Count	β_{18}	-0.001 (0.001)	-0.001 (0.001)	-.0003(0.001)
Baseline weight	β_{19}	-0.008 (0.008)	-0.008 (0.008)	-0.006(0.006)
Weight at follow-up visit	β_{20}	0.004 (0.005)	0.004 (0.005)	0.004 (0.006)

* significant at 5% level

Table 5.4: (cont) Parameter estimates for standard GLMM on HAART adherence

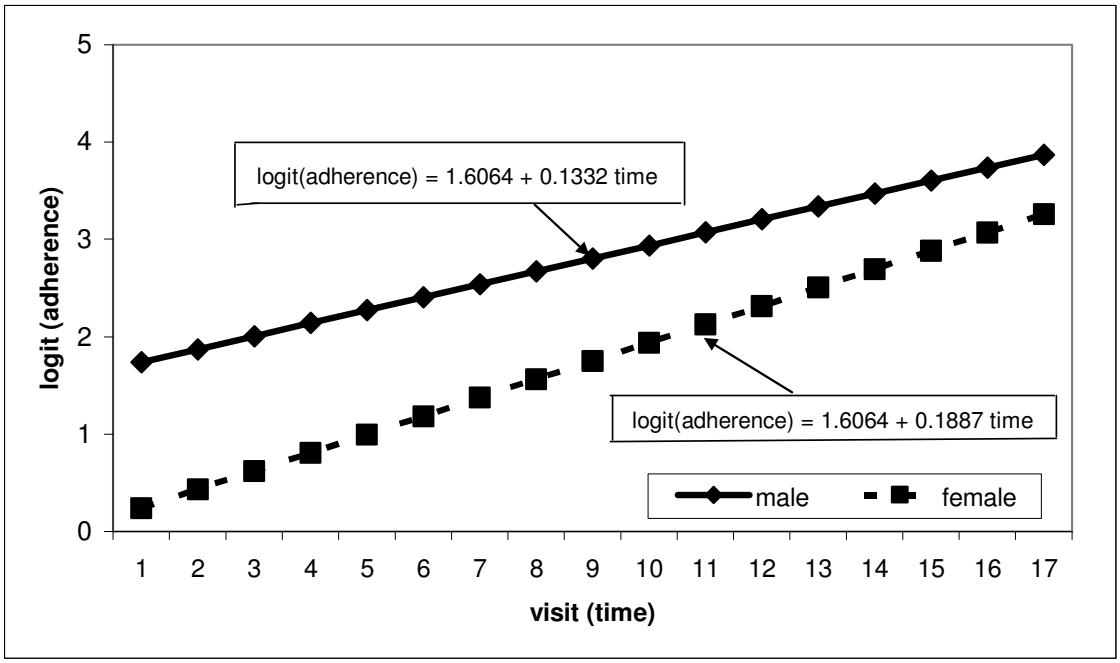
Effect	Parameter	Estimate (std error) Laplace Approx	adaptive Gaussian	PQL
Time*gender (ref = male)				
Female	β_{21}	0.053 (0.024)*	0.055 (0.024)*	0.069 (.020)*
Time*treatment site (ref = rural)				
Urban	β_{22}	-0.055 (0.023)*	-0.055 (0.023)*	-0.050(.020)*
Time*reason for taking the test (ref = unwell)				
VCT	β_{23}	-0.001 (0.026)	-0.001 (0.025)	-0.014(0.022)
Exposed to the risk of HIV	β_{24}	0.065 (0.030)*	0.065 (0.030)*	0.056 (.025)*
Age*education (ref= sec+)				
No schooling	β_{25}	-0.035 (0.189)	-0.036 (0.019)	-0.030(0.015)
Primary	β_{26}	0.002 (0.019)	0.002 (0.019)	-0.001(0.015)
Age*gender (ref = male)				
Female	β_{27}	0.035 (0.146)*	0.035 (0.015)*	0.026 (.011)*
Random Effects				
$var(b_{0i})$	g_{11}	0.680 (0.228)	0.565 (0.246)	0.562 (0.150)
$var(b_{1i})$	g_{22}	0.010 (0.003)	0.011 (0.003)	0.009 (0.002)
$cov(b_{0i}, b_{i1})$	g_{12}	-0.065 (0.024)	-0.054 (0.036)	-0.055(0.018)
$corr(b_{0i}, b_{i1})$		-0.810	-0.810	-0.553

* significant at 5% level

a) Interaction between gender and time

The results show that the rate of increase in adherence is higher with females than males conditional on the random effects (Table 5.4). This is illustrated graphically in Figure 5.1. As time increases by one additional follow-up visit, the odds of adherence for females increases by 1.06 ($e^{0.0555}$) times more than the odds of adherence for males. Therefore, the wide gap in adherence between males and females observed at the beginning of the follow-up period narrows down as the number of follow-up visits increases.

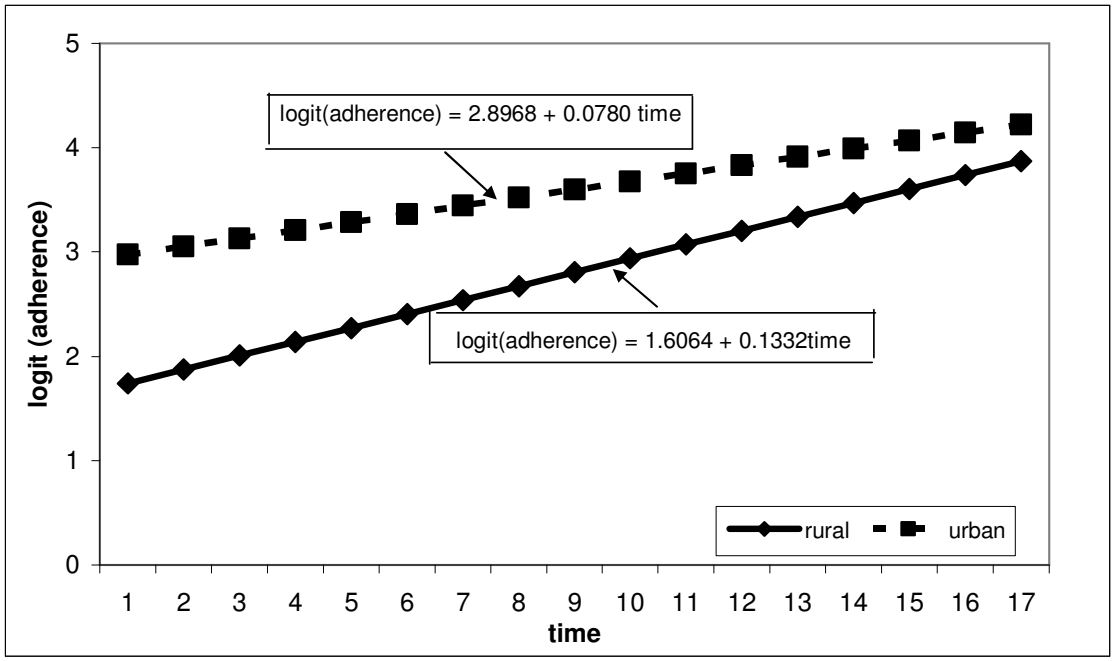
Figure 5. 1: Log of odds of adherence to HAART between the interaction of gender and time



b) Interaction between treatment site and gender

Conditional on random effects, adherence is increasing both in the urban and rural treatment site, but the rate of increase is higher in the rural treatment site than the urban treatment site (Figure 5.2). With one additional follow-up visit, the odds of adherence in the rural treatment site increases by 1.055 ($e^{0.0552}$) times more than the odds of adherence in the urban treatment site (Table 5.5). Therefore, as the number of visits increases, the gap in adherence between both treatment sites narrows down.

Figure 5. 2: Log odds of HAART adherence between the interaction of treatment site and time



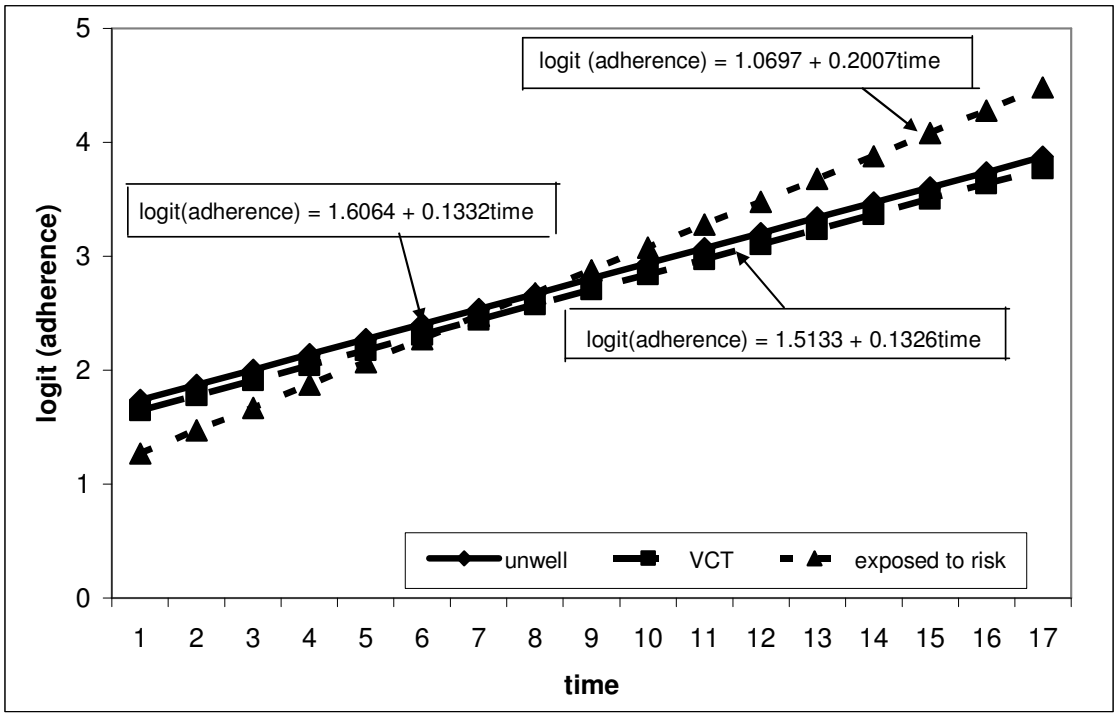
c) Interaction between reported reason and time

Adherence has been increasing over time; however, the rate of increase is highest with patients who reported to have tested for HIV because they felt exposed to the risk of contracting the disease. With the patients who reported to have been exposed to the risk of contracting HIV, the rate of increase in the odds of adherence are 1.05 times more likely to adhere to medication than those who reported to have tested because they were unwell and because they had taken an HIV test for no specific reason. There is no significant difference in the odds of adherence between patients who have reported to have tested because they were unwell and those who have tested for no specific reason (see Figure 5.3). These are conditional on random effects.

d) Interaction between age and gender

The results show that as age increases, the rate at which adherence increases differs by gender. It is shown in Table 5.4 that with one additional year in age, the rate of increase in the odds of adherence for females is 1.04 ($e^{0.037}$) times more likely than the odds of adherence for males.

Figure 5. 3: Log odds of HAART adherence between the patient's reported reason for taking an HIV test and time



Covariance parameters

The estimated covariance parameters and their standard errors are also presented in Table 5.4. Although the overall estimates of the variance components are not large, there is a noticeable individual heterogeneity in terms of both the intercepts and slopes. For instance, the estimated individual random intercepts ranged from -0.869 and 0.900, whereas the estimated individual random slopes indicated that heterogeneity in the rate of adherence over time from patient-to-patient ranged from -0.2182 and 0.1028. Moreover, the correlation between the random intercepts and slopes is very strong, with a correlation coefficient of -0.81. The negative correlation between the random intercepts and slopes suggests that there was a decline over time in adherence rates of patients who had high levels of adherence at the baseline follow-up visit and vice versa. The empirical Bayes estimates of the intercept and slope for each patient are presented in Figure 5.4

Figure 5. 4: Empirical Bayes estimates for the intercept and slope

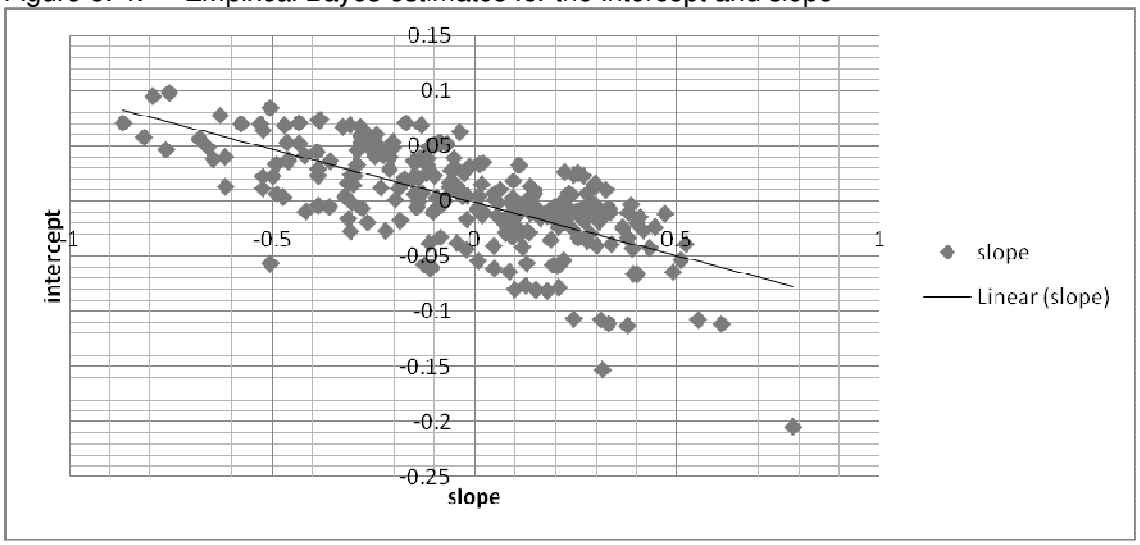


Figure 5.4 reveals the nature of the negative covariance between the random intercepts and slopes. That is, patients with more negative intercepts (less likely to adhere to medication at the baseline follow-up visit) have more positive slopes (more likely to adhere to medication over time).

Assessment of serial correlation

With the selected structure of random effects as well as the mean structure in model (5.26), different residual covariance structures were fitted with the view to selecting the best fitting one. There are many possible covariance structures that are available. Unfortunately, apart from the information criteria, there are generally no simple techniques available to compare all these models (Verbeke and Molenberghs, 2000). The method of estimation used is PQL (with linearized pseudo-data) because it accommodates models with serial correlation. As noted earlier, in the case where parameters are estimated by the linearization method, the information criteria is not applicable since the 'pseudo-information' criteria (e.g pseudo-AIC) techniques cannot be used for selection of non-nested models. Since there is no formal testing, the selection of the residual structure is based on roughly observing how the parameter estimates and their corresponding standard errors change as a consequence of the amount of variability explained by the introduction of auto-correlated errors.

Convergence difficulties were encountered when fitting the different residual structures in addition to the random effects. These include structures such as compound symmetry, unstructured, general toeplitz, spatial power as well as spatial exponential. Molenberghs and Verbeke (2005) have shown that convergence failures are relatively common when modeling of the covariance structure involves joint specification of the random effects, serial correlation and measurement error, simply because these components of variability cannot easily be disentangled. Nonetheless, models that were fitted with random effects plus AR-1, toeplitz(2) and toeplitz(3) as well as spatial Gaussian, converged successfully and the results are presented in Table 5.5. With toeplitz structures, only toeplitz(3) is presented and also the results from the conditional independence (PQL estimates in Table 5.4) are presented for ease of reference.

The results (Table 5.5) in all the fitted models showed negligible presence of residual serial correlation in addition to random effects. That is, all fitted models (random effects with AR-1, toeplitz(3) and spatial Gaussian) indicate that correlation between any two measurements one visit apart is roughly 0.03 (It should be noted that with SAS parameterization, the correlation between two measurements one unit apart for spatial Gaussian structure is obtained by $\rho = \rho(1) = exp\left(-\frac{1}{\theta}\right)$). Thus there is no impact on the parameter estimates of the fixed effects. For instance, there is no reduction in the standard errors when one compares the random effects model and the other three models with residual serial correlation. There is a slight reduction in the Toeplitz structure but it is not meaningful. One disadvantage of including residual covariance in the GLMM-based approach is that the calculation of the overall variance or the overall correlation is not straightforward because as noted earlier, the random effects and the autocorrelation structure enter at different places into the model irrespective of whether one consider the direct outcomes or the pseudo data derived from them (Molenberghs and Verbeke, 2005). In this case the conditional independence model presented is the preferred model.

Table 5. 5: Parameter estimates (standard errors) for the random intercepts and slopes (RIAS) models with different residual structures for the adherence to HAART data

Effect	RIAS	RIAS + AR(1)	RIAS + toep(3)	RIAS + spatial Gaus
Intercept	0.852 (0.510)	0.823(0.515)	0.830(0.499)	0.824(0.515)
Gender (ref = male)				
Female	-1.269(0.428)*	-1.268(0.434)*	-1.255(0.418)*	-1.269(.433)*
Education (ref= sec+)				
No schooling	1.538 (0.643)*	1.525(0.652)*	1.526(0.625)*	1.525(0.651)*
Primary	0.064 (0.551)	0.088(0.559)	0.055(0.555)	0.085(0.558)
Treatment Site (ref=rural)				
Urban site	1.258 (0.165)*	1.282(0.165)*	1.264(0.163)*	1.281(0.165)*
Income (ref= not source of income)				
Source of income	0.038 (0.098)	0.041(0.099)	0.035(0.095)	0.041(0.099)
Access to tap water (ref = No)				
Yes	0.044 (0.153)	0.035(0.156)	0.041(0.149)	0.034(0.156)
Having electricity (ref = No)				
Yes	0.001(0.139)	0.010(0.141)	0.005(0.136)	0.009(0.141)
Cell phone ownership (ref = No)				
Yes	0.255 (.089)*	0.260(0.090)*	0.255(0.086)*	0.260(0.090)*
WHO staging of disease (ref = stage 4)				
Stage 1	-0.351(0.228)	-0.360(0.230)	-0.354(0.222)	-0.360(0.230)
Stage 2	-0.250(0.199)	-0.254(0.201)	-0.252(0.194)	-0.253(0.200)
Stage 3	-0.091(0.180)	-0.090(0.181)	-0.093(0.156)	-0.090(0.181)
Partner (ref = living with partner)				
living without a partner	-0.329(.102)*	-0.335(0.103)*	-0.329(0.099)*	-0.333(.103)*
Reason for taking HIV test (ref = unwell)				
VCT	-0.015(0.182)	-0.009(0.185)	-0.010(0.181)	-0.010(0.184)
Exposed to the virus	-0.485(0.201)*	-0.485(0.203)*	-0.482(0.199)*	-0.486(\.203)*
Initial adherence (ref = not adherent)				
Adherent	-0.138(0.105)	-0.133(0.109)	-0.134(0.104)	-0.132(0.108)
Time (visit)	0.123 (0.021)*	0.130(0.022)*	0.126(0.022)*	0.130(0.022)*
Age	-0.015(0.011)	-0.016(0.011)	-0.015(0.011)	-0.016(0.011)
Baseline CD4 Cell Count	-.0003(0.001)	-.0003(0.001)	-.0003(.0005)	-.0003(0.001)
Baseline weight	-0.006(0.006)	-0.006(0.005)	-0.006(0.005)	-0.006(0.005)
Weight at follow-up visit	0.004 (0.006)	0.004(0.006)	0.004(0.006)	0.004(0.006)
Time*gender (ref = male)				
Female	0.069 (.020)*	0.069(0.020)*	0.069(0.020)*	0.069(0.020)*
Time*treatment site (ref = rural)				
Urban	-0.050(.020)*	-0.053(0.020)*	-0.051(0.019)*	-0.053(.020)*
Time*reason for taking the test (ref = unwell)				
VCT	-0.014(0.022)	-0.015(0.023)	-0.014(0.022)	-0.015(0.023)
Exposed to the risk of HIV	0.056 (0.025)*	0.055(0.026)*	0.055(0.026)*	0.055(0.026)*
Age*education (ref= sec+)				
No schooling	-0.030(0.015)	-0.030(0.015)	-0.030(0.016)	-0.030(0.015)
Primary	-0.001(0.015)	-0.002(0.015)	-0.002(0.014)	-0.002(0.015)
Age*gender (ref = male)				
Female	0.026 (0.011)*	0.025(0.012)*	0.026(0.011)*	0.025(0.011)*

* significant at 5% level

Table 5.5(cont): Parameter estimates (standard errors) for the random intercepts and slopes (RIAS) models with different residual structures for the adherence to HAART data

Effect	RIAS	RIAS + AR(1)	RIAS + toep(3)	RIAS + spatial Gaus
Random Effects				
$var(b_{0i})$	0.562 (0.150)	0.813(0.185)	0.661(0.185)	0.802(0.185)
$var(b_{1i})$	0.009 (0.002)	0.012(0.003)	0.011(0.003)	0.012(0.003)
$cov(b_{0i}, b_{i1})$	-0.055(0.018)	-0.059(0.020)	-0.068(0.020)	-0.058(0.020)
Residual				
Autocorr. par. θ		0.029(0.0158)	0.030(0.014) 0.019(0.014)	0.541(0.035)
Autocorr. σ^2		0.839(0.0150)	0.845(0.018)	0.839(0.015)

* significant at 5% level

5.8 Comparing marginal and random effects' estimates

The random effects models are conditional models (subject specific), as both within- and between-cluster effects apply conditional on the random effect value. By contrast, the effects in the marginal models are averaged over all subjects (i.e. population averaged), so those effects do not refer to a comparison at a fixed value of a random effect (Agresti, 2002). For instance, a marginal gender contrast compares the mean among men to that among women, while a conditional gender contrast compares the mean among men to that among women holding the same value of a random effect. With linear models, it is possible to formulate the two regression approaches to have coefficients with the same interpretation. That is, coefficients from random effects models can have marginal interpretations as well. To develop this idea, briefly consider the linear mixed model

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i, \qquad (5.27)$$

where $\mathbf{y}_i = (y_{i1},\ldots,y_{in_i})'$, $\mathbf{X}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i})'$, $\mathbf{Z}_i = (\mathbf{z}_{i1},\ldots,\mathbf{z}_{in_i})'$ $\varepsilon_i = (\varepsilon_{i1},\ldots,\varepsilon_{in_i})'$, and where $E(\mathbf{b}_i) = \mathbf{0}$, $cov(\mathbf{b}_i) = \mathbf{G}$, $cov(\varepsilon_i) = \sigma^2\mathbf{I}$, and $cov(\mathbf{b}_i, \varepsilon_i) = \mathbf{0}$. Note that the subject-specific coefficient for the ith individual is $\beta + \mathbf{b}_i$. Since $E(\mathbf{b}_i) = \mathbf{0}$, β has interpretation as the typical subject-specific parameter. Alternatively, (5.27) can be expressed as

$$E(\mathbf{y}_i) = \mathbf{X}_i\beta,\ cov(\mathbf{y}_i) = \sigma^2\mathbf{I} + \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i'.$$

Here, β has the interpretation as the rate of change in the population-averaged $\mathbf{y}$ with $\mathbf{X}$. The random effects in the linear mixed model do not alter the marginal expectation of $\mathbf{y}$, the marginal covariance matrix. Hence, β has both a subject-specific and population-averaged interpretation.

With non-linear link functions such as the logit, the parameters from marginal and random-effects models have completely different interpretations. To expand on this point, recall an example in Section 5.2 of a binary outcome variable that

assumes a random intercepts logistic model with a linear predictor $logit[P(\mathbf{y}_{ij}=1 \mid b_i)] = \mathbf{x}'_{ij}\beta + b_i$. The conditional means are given by

$$E(\mathbf{y}_{ij} \mid b_i) = \frac{\exp(\mathbf{x}'_{ij}\beta + b_i)}{1+\exp(\mathbf{x}'_{ij}\beta + b_i)}.$$

The marginal average evolution is

$$E(\mathbf{y}_{ij}) = E[E(\mathbf{y}_{ij} \mid b_i)] = E\left[\frac{\exp(\mathbf{x}'_{ij}\beta + b_i)}{1+\exp(\mathbf{x}'_{ij}\beta + b_i)}\right],$$

which does not have the form

$$\frac{\exp(\mathbf{x}'_{ij}\beta)}{1+\exp(\mathbf{x}'_{ij}\beta)},$$

except when b_i has a degenerate distribution, i.e. $var(b_i)=0$ (Agresti, 2002).

Nonetheless, approximate relationships exist between the estimates from the random effects and marginal models. Neuhaus, Kalbfleisch and Hauck (1991) show that if $var(\mathbf{b}_i) > 0$, then the elements of the marginal (β^M) and random effects (β^{RE}) regression vector satisfy:

a) $|\beta_k^m| \leq |\beta_k^{RE}|$, for all k.

b) Equality holds if and only if $\beta^{RE}=0$, or $\tau=0$, where τ is the standard deviation of $\mathbf{b}_i$.

c) The discrepancy between (β^M) and (β^{RE}) increases with $var(\mathbf{b}_i)$.

In particular, if $\mathbf{b}_i$ is assumed to follow a Gaussian distribution with mean zero and variance τ^2 in a logistic model with random intercept, Zeger, Liang and Albert (1988) showed that if

$$logitE(y_{ij}) \approx (c^2\tau^2+1)^{-1/2}\mathbf{x}'_{ij}\beta^{RE},$$

where $c = 16\sqrt{3}/(15\pi)$ so that

$$\beta^M \approx (c^2\tau^2+1)^{-1/2}\beta^{RE},$$

where $c^2 \approx 0.346$. Since the effect in the marginal model multiplies that of the conditional model by about d where $d = (c^2\tau^2 + 1)^{-1/2}$, β^M is typically smaller in absolute value than β^{RE} and the discrepancy increases as τ^2 (Neuhaus et al., 1991). That is, if the variance of the random effects is large, parameters from fitting marginal models and random-effects model will be different, while equal parameter values hold if the variance of the random effects is equal to zero (Molenberghs and Verbeke, 2005). Thus, the use of marginal models can be dangerous, even when marginal inferences are of interest if the data exhibits a large degree of heterogeneity between subjects (Agresti, 2002).

With HAART adherence data, the ratio between the parameter estimates of the GLMM (using the conditional independence model results from adaptive Gaussian quadrature) and the GEE (Chapter 4, Table 4.4) were calculated and are presented in Table 5.6. The overall trend of the observed ratio between the two model estimates show that the GLMM estimates are slightly higher in magnitude than the GEE estimates. The slight difference might be due to the relatively small variance components estimates in the GLMM (Table 5.4). However, there are a few cases where the differences in the size of the effects were more pronounced (e.g. source of household income and households with electricity). This suggests, for instance, that patients who are sources of household income have more heterogeneity on adherence, thus their marginal effect is smaller than that of patients who were not sources of household income.

There were also a few exceptions where the GEE estimates are slightly higher than their GLMM counterparts. It is further noted that although the GLMM estimates were generally larger, the statistical significance of the parameters (as measured by the ratio of estimate to standard error) remained the same across the analyses of the two models. Agresti (2002) argues that the choice of the model is usually not crucial to inferential conclusions because if one effect seems more important than another in a conditional model, the same is usually true with the marginal model. This was also experienced with the analyses of HAART

adherence data, the predictor variables that were important in the GEE model were also found to be important in the GLMM model.

Table 5. 6: Parameter estimates (standard errors) for GLMM (using adaptive Gaussian quadrature) and GEE as well as the ratio between the two sets of parameters on adherence data

		Estimate (std error)		
Effect	**Parameter**	**GLMM**	**GEE**	**Ratio**
Intercept	β_0	1.606 (0.306)	0.597 (0.411)	2.69
Gender (ref = male)				
Female	β_1	-1.555 (0.546)*	-1.194(.341)*	1.30
Education (ref= sec+)				
No schooling	β_2	1.707 (0.734)*	1.617 (0.515)*	1.06
Primary	β_3	-0.108 (0.710)	0.099 (0.548)	1.09
Treatment Site (ref=rural)				
Urban site	β_4	1.290 (0.204)*	1.173 (0.162)*	1.10
Income (ref= not source of income)				
Source of income	β_5	0.053 (0.118)	0.010 (0.101)	5.30
Access to tap water (ref = No)				
Yes	β_6	0.089 (0.185)	0.108 (0.150)	0.82
Having electricity (ref = No)				
Yes	β_7	0.077 (0.159)	0.010(0.125)	7.70
Cell phone ownership (ref = No)				
Yes	β_8	0.223 (0.106)*	0.231 (.090)*	0.97
WHO staging of disease (ref = stage 4)				
Stage 1	β_9	-0.482 (0.280)	-0.334(0.197)	1.44
Stage 2	β_{10}	-0.361 (0.250)	-0.251(0.173)	1.44
Stage 3	β_{11}	-0.256 (0.232)	-0.108(0.157)	2.37
Partner (ref = living with partner)				
living without a partner	β_{12}	-0.752 (0.306)*	-0.289(0.100)*	2.60
Reason for taking HIV test (ref = unwell)				
VCT	β_{13}	-0.093 (0.221)	-0.028(0.179)	3.32
Exposed to the virus	β_{14}	-0.537 (0.245)*	-0.486(0.181)*	1.11
Initial adherence (ref = not adherent)				
Adherent	β_{15}	-0.119 (0.125)	-0.141(0.108)	0.84
Time (visit)	β_{16}	0.133 (0.027)*	0.100 (.022)*	1.33
Age	β_{17}	-0.021 (0.014)	-0.012(0.008)	1.75
Baseline CD4 Cell Count	β_{18}	-0.001 (0.001)	-.0004(0.001)	2.50
Baseline weight	β_{19}	-0.008 (0.008)	-0.004(0.008)	2.00
Weight at follow-up visit	β_{20}	0.004 (0.007)	0.001 (0.007)	4.00

* significant at 5% level

Ratio= $\left|\frac{\beta^{RE}}{\beta^{M}}\right|$

Table 5.6: Parameter estimates (standard errors) for GLMM (using adaptive Gaussian quadrature) and GEE as well as the ratio between the two sets of parameters on adherence data

		Estimate (std error)		
Effect	**Parameter**	**GLMM**	**GEE**	**Ratio**
Time*gender (ref = male)				
Female	β_{21}	0.055 (0.024)*	0.072 (.019)*	0.76
Time*treatment site (ref = rural)				
Urban	β_{22}	-0.055 (0.023)*	-0.057(0.020)*	0.96
Time*reason for taking the test (ref = unwell)				
VCT	β_{23}	-0.001 (0.027)	-0.010(0.022)	0.10
Exposed to the risk of HIV	β_{24}	0.067 (0.030)*	0.056 (0.024)*	1.20
Age*education (ref= sec+)				
No schooling	β_{25}	-0.036 (0.019)	-0.033(0.013)	1.09
Primary	β_{26}	0.002 (0.019)	-0.002(0.014)	1.00
Age*gender (ref = male)				
Female	β_{27}	0.037 (0.015)*	0.024 (0.010)*	1.54
Random Effects				
$var(b_{0i})$	g_{11}	0.765 (0.246)		
$var(b_{1i})$	g_{22}	0.011 (0.003)		
$cov(b_{0i}, b_{i1})$	g_{12}	-0.074 (0.036)		
$corr(b_{0i}, b_{i1})$		-0.810		

* significant at 5% level

Ratio= $\left|\frac{\beta^{RE}}{\beta^{M}}\right|$

5.9 Summary

GLMM is considered as a straight forward extension of the generalized linear model by appending the random effects in the linear predictor. This achieves two main goals of allowing responses to be non-normal and correlated. The GLMMs have the ability to model subject-specific evolutions in addition to determining the association structure of the data. However, the estimation procedures are more complex with GLMM. For maximum likelihood, approximation methods are used for parameter estimation with each method having its own strengths and limitations. As a result, fitting data using GLMM was quite challenging. Regardless of the embedded differences between the approximation methods used to fit the adherence data, the results obtained from all methods largely agree, which gives us more confidence in all the approximation methods. Note that GLMM is more robust to missingness than GEE because it assumes that data are missing at random (MAR).

Application of GLMM to the adherence data enabled us to model the dependence among the repeated adherence observations in addition to evaluating the influence of the predictor variables on a specific patient's mean adherence response over time. The results showed that predictor variables associated with HAART adherence at the population level also have an association with HAART adherence at the subject specific level. In addition, the results revealed that patients with high adherence at the beginning of the follow-up period tend to be less adherent over time. Similarly, patients who are non-adherent at the beginning tend to improve their adherence over time. Next, is to evaluate whether the patient's past adherence outcome influences his/her current adherence outcome using transition models.

Chapter 6

Transition models

6.1 Introduction

We have so far discussed extensively two approaches (including their results) of modeling longitudinal data by extending generalized linear models (GLMs). These are marginal models using the generalized estimating equations (GEEs) and generalized linear mixed models (GLMMs) as a special case of random effects models. The third approach of handling longitudinal data and still extending the GLMs involves modeling the mean and time dependence simultaneously through conditioning an outcome on other outcomes or a subset of other outcomes (Molenberghs and Verbeke, 2005). Such models are known as transition models and they are considered as a very specific class of conditional models. Transition models are appealing for longitudinal designs due to the sequential nature of the data (Fitzmaurice and Molenberghs, 2008). In transition models, the conditional distribution of each response y_{ij} is described as an explicit function of previous outcomes or history $\mathbf{h}_{ij} = (y_{i1}, \ldots, y_{ij-1})$ and the covariates $\mathbf{x}_{ij}$ (Diggle et al., 2002). That is, the past outcomes are treated as additional predictor variables (Davis, 2002). The dependence among the repeated measures is thought of as arising due to the past values influencing the present observations. Specific classes of transition models are Markov models for which the conditional distribution of y_{ij} given $\mathbf{h}_{ij}$ depends only on the q prior observations $y_{i1}, \ldots, y_{ij-q}$. (Diggle et al., 2002; Molenberghs and Verbeke, 2005; Ramroop, 2008). The order of a transition model is the number of previous measurements (q) that is considered to influence the current one. Moreover, a model is called stationary if the functional form of the dependence is the same regardless of the actual time (Molenberghs and Verbeke, 2005; Ware, Lipsitz and Speizer, 1988).

There are several different schools of thought regarding the best approach to analysis of correlated binary data. Unlike in the normal setting, marginal, conditional (transition) and random effects approaches tend to give dissimilar results (Molenberghs and Ryan, 1999). While each one of the model families has advantages and disadvantages, model choice should primarily be based on the specific goals of the analysis. In transition models, the expectation of the response at a given occasion is modeled in terms of the responses at the other occasions, whereas in marginal models the covariates are directly related to the marginal expectations. Random-effects models differ from the other two by the inclusion of parameters that are specific to the subject (Molenberghs and Verbeke, 2005).

Due to the popularity of marginal (especially GEE) and random effects models for correlated binary data, transition models have been given relatively little attention (Molenberghs and Ryan, 1999). Diggle et al., (2002) criticized the transition approach because the interpretation of a fixed effect parameter of one response is conditional on other responses for the same subject, outcomes of other subjects and the number of repeated measures. However, transition models are likelihood based and one advantage is that efficiency can be gained over other procedures such as GEE under correct model specification. Other advantages include availability of inferential procedures such as likelihood ratio tests and robustness of likelihood-based inference to certain dropout or missing data mechanisms (Heagerty, 2002).

The rest of the chapter is organized as follows. Section 6.2 presents the transition model formulation with illustrative examples from continuous, binary and count responses. Estimation and fitting of transition models, with a particular focus on transition models for binary data is presented in Section 6.3. In Section 6.4 the results from the fitted transition models to adherence data are presented.

6.2 Model specification

A transition model specifies a GLM for the conditional distribution of y_{ij} given past responses, $\mathbf{h}_{ij}$. The form of the conditional GLM is

$$f(y_{ij} \mid \mathbf{h}_{ij}) = \exp\{[y_{ij}\theta_{ij} - \psi(\theta_{ij})]/\phi + c(y_{ij}, \phi)\} \qquad (6.1)$$

for known functions $\psi(\theta_{ij})$ and $c(y_{ij}, \phi)$. The conditional mean and variance are given by

$$\mu_{ij}^c = E(y_{ij} \mid \mathbf{h}_{ij}) = \psi'(\theta_{ij})$$

and

$$v_{ij}^c = var(y_{ij} \mid \mathbf{h}_{ij}) = \psi''(\theta_{ij})\phi$$

respectively.

Then a transition model establishes the following assumptions:

a) The conditional expectation, denoted by μ_{ij}^c, is specified by

$$g(\mu_{ij}^c) = g(E(y_{ij} \mid \mathbf{h}_{ij}) = \mathbf{x}_{ij}'\beta + \sum_{r=1}^{q} f_r(\mathbf{h}_{ij}; \alpha)$$

where $f_1, \ldots, f_q$ are functions of previous observations and, possibly, of an unknown parameter vector $\alpha' = (\alpha_1, \ldots \alpha_q)$; and $g(\cdot)$ is a known link function.

b) The conditional variance satisfies the equation:

$$v_{ij}^c = \phi v(\mu_{ij}^c)$$

where $v(\cdot)$ is a known variance function and ϕ is an overdispersion parameter.

This shows that the transition model expresses the conditional mean μ_{ij}^c as a function of both the covariates $\mathbf{x}_{ij}$ and the past responses $y_{i1}, \ldots, y_{ij-q}$. That is, the linear predictor component of the model includes the original covariates as well as additional covariates that are known functions of past responses (Diggle et al., 2002; Davis, 2002). The transition model formulation is illustrated with continuous, binary and count responses.

For a continuous response, we consider the linear regression model with autoregressive errors for Gaussian data (Diggle et al., 2002; Tsay, 1984) which is a Markov model. It has the form

$$y_{ij} = \mathbf{x}'_{ij}\beta + \sum_{r=1}^{q} \alpha_r (y_{ij-r} - \mathbf{x}'_{ij-r}\beta) + \mathbf{Z}_{ij} ,$$

where $\mathbf{Z}_{ij}$ are independent, and zero-mean Gaussian errors. This is a transition model with

$$g(\mu_{ij}^c) = \mu_{ij}^c ,\ \upsilon(\mu_{ij}^c) = 1 \text{ and } f_r(\mathbf{h}_{ij};\alpha) = \alpha_r (y_{ij-r} - \mathbf{x}'_{ij-r}\beta) .$$

It should be noted that the present observation y_{ij} is the linear function of $\mathbf{x}'_{ij}$ and of the earlier deviations $y_{ij-r} - \mathbf{x}'_{ij-r}\beta$, $r = 1,\ldots,q$. This formulation of the transition model within the linear regression model leads to $E(y_{ij}) = \mathbf{x}'_{ij}\beta$ whatever the value of the number of previous outcomes (q) is.

For binary responses, we consider a logistic regression model that consists of a first-order Markov chain (Diggle et al., 2002), which is given as

$$logit[\Pr(y_{ij} = 1 \mid \mathbf{h}_{ij})] = \mathbf{x}'_{ij}\beta + \alpha y_{ij-1} ,$$

where

$$g(\mu_{ij}^c) = logit(\mu_{ij}^c) = \log\left(\frac{\mu_{ij}^c}{1-\mu_{ij}^c}\right),\ \upsilon(\mu_{ij}^c) = \mu_{ij}^c(1-\mu_{ij}^c)$$

and

$$f_r(\mathbf{h}_{ij};\alpha) = \alpha_r y_{ij-r} , r = q = 1 .$$

A simple extension to a model of order q has the form

$$g(\mu_{ij}^c) = logit(\mu_{ij}^c) = \mathbf{x}'_{ij}\beta + \sum_{r=1}^{q} \alpha_r y_{ij-r} .$$

The interpretation of the regression coefficients depends on the order q, that is $\beta = \beta_q$.

For count responses, we consider a log-linear model discussed in Zeger and Qaqish (1988) with a first order Markov chain and it can be summarized as follows

$$log[\Pr(y_{ij} = 1 \mid \mathbf{h}_{ij}) = \mathbf{x}'_{ij}\beta + \alpha\{\log(y^*_{ij-1}) - \mathbf{x}'_{ij-1}\beta\},$$

where

$$y^*_{ij-1} = \max(y_{ij}, d);\ 0 < d < 1,$$

which leads to

$$\mu^c_{ij} = \exp(\mathbf{x}'_{ij}\beta)\left(\frac{y^*_{ij-1}}{\exp(\mathbf{x}'_{ij-1}\beta)}\right)^{\alpha}$$

and

$$f_1 = \alpha\{\log(y^*_{ij-1}) - \mathbf{x}'_{ij-1}\beta\}.$$

The constant d prevents $y_{ij-1} = 0$ from being an absorbing state, otherwise $y_{ij-1} = 0$ compels all future responses to be zero. For $\alpha < 0$, a response at time $t-1$ greater than $\exp(\mathbf{x}'_{t-1}\beta)$ decreases the expectation for the current response and when $\alpha > 0$, the opposite occurs.

It is worth noting that for outcomes of the general type (in this case logistic and log-linear models), it is difficult to formulate models in such a way that β has the same meaning for different assumptions about the time dependence. Thus, when β is of scientific interest, evaluation of the sensitivity of the results to the choice of time dependence model should be carried out.

6.3 Estimation and fitting of a transition model

Since the contribution of y_{ij} given the history $\mathbf{h}_{ij}$ leads to independent GLM contributions, estimation in transition models is quite straight forward (Molenberghs and Verbeke, 2005). The following derivation follows that of Diggle et al., (2002). In the first-order Markov model the contribution to the likelihood for the ith subject can be written as

$$L_i(y_{i1},\ldots,y_{in_i}) = f(y_{i1})\prod_{j=2}^{n_i} f(y_{ij} \mid \mathbf{h}_{ij}).$$

In a Markov model of order q, the conditional distribution of y_{ij} is

$$f(y_{ij} \mid \mathbf{h}_{ij}) = f(y_{ij} \mid y_{ij-1},\ldots,y_{ij-q}),$$

then the likelihood contribution for the ith subject becomes

$$f(y_{i1},\ldots,y_{iq})\prod_{j=q+1}^{n_i} f(y_{ij} \mid y_{ij-1},\ldots,y_{ij-q}).$$

The conditional GLM (6.1) only specifies the conditional distribution $f(y_{ij} \mid \mathbf{h}_{ij})$; the likelihood of the first q observation $f(y_{i1},\ldots,y_{iq})$ is not specified directly.

When y_{ij} given $\mathbf{h}_{ij}$ follows a Gaussian distribution and also $y_{i1},\ldots,y_{iq}$ are also multivariate Gaussian, the marginal distribution $f(y_{i1},\ldots,y_{iq})$ can be fully determined from the conditional distribution model without additional unknown parameters for a weakly stationary covariance structure of y_{ij}. Thus full maximum likelihood estimation can be used to fit Gaussian autoregressive models.

However, for models of the general type such as logistic and log-linear models, $f(y_{i1},\ldots,y_{iq})$ is not determined from the GLM assumption about the conditional model, and therefore the full likelihood is not available. In this case, β and α can be estimated by maximizing the conditional likelihood

$$\prod_{i=1}^{N} f(y_{iq+1},\ldots,y_{in_i} \mid y_{i1},\ldots,y_{iq}) = \prod_{i=1}^{N}\prod_{j=q+1}^{n_i} f(y_{ij} \mid \mathbf{h}_{ij}). \qquad (6.2)$$

There are two distinct cases to consider when maximizing the above conditional likelihood.

Case 1: If $f_r(\mathbf{h}_{ij};\alpha,\beta) = \alpha_r f_r(\mathbf{h}_{ij})$ where f_r does not depend on the parameters β, estimation proceeds as in GLMs for independent data. That is, the conditional expectation, $g(\mu_{ij}^c) = \mathbf{x}_{ij}'\beta + \alpha\sum_{r=1}^{q} f_r(\mathbf{h}_{ij})$ is a linear function of both β and $\alpha' = (\alpha_1,\ldots,\alpha_q)$. We can simply regress y_{ij} on the $(p+q)$-dimensional vector of extended explanatory variables $(\mathbf{x}_{ij}, f_1(\mathbf{h}_{ij}),\ldots, f_q(\mathbf{h}_{ij}))$.

Case 2: When f_r depends on both α and β, the conditional score function has the form

$$\mathbf{S}^c(\delta) = \sum_{i=1}^{N}\sum_{j=q+1}^{n_i} \frac{\partial \mu_{ij}^c}{\partial \delta} v_{ij}^{c^{-1}}(y_{ij} - \mu_{ij}^c) = 0, \qquad (6.3)$$

where $\delta' = (\beta', \alpha')$. This equation is the conditional analogue of the GLM score equation presented in Chapter 3. The derivative $\partial\mu_{ij}/\partial\delta$ is analogous to $\mathbf{x}_{ij}$ but it can depend on α and β. The iterative weighted least squares estimation procedure is formulated as follows. Let $\mathbf{y}_i$ be the $(n_i - q)$ vector of responses for $j = q+1,\ldots,n_i$ and μ_{ij}^c is its expectation given, $\mathbf{h}_{ij}$. Let $\mathbf{X}_i^*$ be an $(n_i - q)\times(p+q)$ matrix with the kth row $\partial\mu_{iq+k}/\partial\delta$ and $\mathbf{W}_i = diag(1/v_{ik+q}^c, k = 1,\ldots,n_i - q)$ is an $(n_i - q)\times(n_i - q)$ diagonal weighting matrix. Finally, let $\mathbf{z}_i = \mathbf{X}_i^*\hat{\delta} + (\mathbf{y}_i - \hat{\mu}_i^c)$. Then an updated $\hat{\delta}$ can be obtained by iteratively regressing $\mathbf{z}$ on $\mathbf{X}^*$ using weights $\mathbf{W}$.

It is worth noting that if q is large, relative to n_i, the use of transitional models with conditional likelihood could be inefficient.

When the correct model is assumed for the conditional mean and variance, the solution to $\hat{\delta}$ of (6.3) asymptotically follows a Gaussian distribution as N goes to infinity, with mean equal to the true value, δ, and $(p+q)\times(p+q)$ variance matrix

$$\mathbf{V}_{\hat{\delta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^{*\prime} \mathbf{W}_i \mathbf{X}_i^{*} \right)^{-1}.$$

The variance $\mathbf{V}_{\hat{\delta}}$ depends on β and α and a consistent estimate, $\hat{\mathbf{V}}_{\hat{\delta}}$, is obtained by replacing β and α by their estimates $\hat{\beta}$ and $\hat{\alpha}$. However, when the conditional mean is correctly specified and the conditional variance is not, consistent inferences about δ can still be obtained using the robust variance:

$$\mathbf{V}_{\mathbf{R}} = \left(\sum_{i=1}^{N} \mathbf{X}_i^{*\prime} \mathbf{W}_i \mathbf{X}_i^{*} \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^{*\prime} \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i \mathbf{X}_i^{*} \right) \left(\sum_{i=1}^{N} \mathbf{X}_i^{*\prime} \mathbf{W}_i \mathbf{X}_i^{*} \right)^{-1}.$$

A consistent estimate of $\hat{\mathbf{V}}_{\mathbf{R}}$ is obtained by replacing $\mathbf{V}_i = var(\mathbf{y}_{ij} \mid \mathbf{h}_{ij})$ by its estimate $(\mathbf{y}_i - \hat{\mu}_i^c)(\mathbf{y}_i - \hat{\mu}_i^c)'$.

The robust variance will often give consistent confidence intervals for $\hat{\delta}$ even when the Markov assumption is violated. In such a situation, however, the interpretation of $\hat{\delta}$ is questionable since $\mu_{ij}^c(\hat{\delta})$ is not the conditional mean of y_{ij} given $\mathbf{h}_{ij}$.

Fitting transition models in SAS is relatively straightforward. Since subsequent measurements, given their previous history, are independent of each other, standard GLM software can be used to fit these models. These include SAS procedures that comprise GENMOD and LOGISTIC. However, one must ensure that the data is organized in such way that previous measurements can be used as covariates. Molenberghs and Verbeke (2005) describe how the DROPOUT macro can be used to prepare previous responses as covariates in the longitudinally organized data set (one record per measurement rather than per subject). Besides the DROPOUT macro, the LAG function in SAS can also be used to create previous measurements as additional covariates.

For binary outcomes observed at equally spaced intervals, a first-order Markov chain is characterized by the transition matrix

$$\begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix},$$

where $\pi_{ab} = \Pr(y_{ij} = b \mid y_{ij-1} = a)$, $a, b = 0,1$. For instance, π_{01} is the probability that $y_{ij} = 1$ when the previous response is $y_{ij-1} = 0$. That is, the probability of a transition from $y_{ij} = 0$ at time $j-1$ to $y_{ij} = 1$ at time j is $\pi_{01} = \Pr(y_{ij} = 1 \mid y_{ij-1} = 0)$. Similarly, the probability of a transition from $y_{ij} = 1$ at time $j-1$ to $y_{ij} = 1$ at time j is $\pi_{11} = \Pr(y_{ij} = 1 \mid y_{ij-1} = 1)$. With adherence data, π_{01} would represent the probability that a patient adheres to medication ($y_{ij} = 1$) when they were non-adherent to medication in the previous visit ($y_{ij-1} = 0$). Likewise, π_{11} would represent the probability that a patient adheres to medication ($y_{ij} = 1$) when they were adherent to medication in the previous visit ($y_{ij-1} = 0$). It should be noted that each row of a transition matrix sums to one since

$$\Pr(y_{ij} = 0 \mid y_{ij-1} = a) + \Pr(y_{ij} = 1 \mid y_{ij-1} = a) = 1.$$

In general, a transition matrix records the probabilities of making each of the possible transitions from one visit to the next (Diggle et al., 2002). It is worth noting that the transition probabilities are conditional probabilities of going into each state, given the immediately preceding state. In a first-order Markov chain, there is dependence on the immediately preceding state but not on earlier outcomes. Moreover, higher-order sequential dependence can be incorporated, with dependence on more than an immediately preceding state (Fitzmaurice and Molenberghs, 2008).

Transition probabilities can be modeled as functions of covariates $\mathbf{x}'_{ij} = (1, x_{ij1}, x_{ij2}, \ldots, x_{ijp})$ in a regression setting. For instance, Cox (1970, 1972) showed the link between the transition probabilities for a Markov chain and a logistic regression. Korn and Whittemore (1979) applied the model to panel data on the effects of air pollution. Thus, with the adherence data, we might assume that the probability of adherence to medication for patient i at visit j has direct

dependence on whether or not a patient adhered to medication at visit $j-1$ as well as on the explanatory variables $\mathbf{x}_{ij}$ (age, gender, treatment site etc). A logistic regression model can be formulated as

$$logit[P(y_{ij}=1 \mid \mathbf{x}_{ij}, y_{ij-1}, \beta, \alpha)] = \mathbf{x}_{ij}'\beta + y_{ij-1}\alpha . \qquad (6.4)$$

This model is the stationary first order autoregressive model. The parameter $\exp(\alpha)$ is the ratio of the odds of adherence among patients who did and did not adhere to medication at the previous follow-up visit. The parameter $\exp(\beta)$ can be interpreted as the odds of adherence to medication having controlled for the outcome (adherent/not adherent) in the previous follow-up visit.

Evaluating the logistic regression equation (6.4) to $y_{ij-1}=0$ and $y_{ij-1}=1$ produces the following transition probabilities

$$\pi_{00} = \frac{1}{1+\exp(\mathbf{x}_{ij}'\beta)},\ \pi_{01} = \frac{\exp(\mathbf{x}_{ij}'\beta)}{1+\exp(\mathbf{x}_{ij}'\beta)}$$

$$\pi_{10} = \frac{1}{1+\exp(\mathbf{x}_{ij}'\beta+\alpha)} \text{ and } \pi_{11} = \frac{\exp(\mathbf{x}_{ij}'\beta)}{1+\exp(\mathbf{x}_{ij}'\beta+\alpha)} .$$

In this model (6.4), the covariates have the same effect on the response probability whether $y_{ij-1}=0$ or $y_{ij-1}=1$. In order to link covariate dependence with transition probabilities so as to analyze factors associated with such transitions, logistic regression models were employed by Muenz and Rubinstein (1985) for first-order Markov models while generalizations to higher-order were proposed by, amongst others, Islam and Crowdhury (2006). Thus, for assessing covariate dependence on the previous response, model (6.4) can be extended to include interaction of the previous response y_{ij-1} and explanatory variables, $\mathbf{x}_{ij}$. That is, it can be assumed that the effects of explanatory variables will differ depending on the previous response. In such a case, a very general model uses separate logistic regressions for $P(y_{ij}=1 \mid y_{ij-1}=0)$ and $P(y_{ij}=1 \mid y_{ij-1}=1)$. That is

$$logit[P(y_{ij}=1 \mid y_{ij-1}=0, \mathbf{x}_{ij}) = \mathbf{x}'_{ij}\beta_0,$$

$$logit[P(y_{ij}=1 \mid y_{ij-1}=1, \mathbf{x}_{ij}) = \mathbf{x}'_{ij}\beta_1,$$

where if $\beta_0 \neq \beta_1$ indicates the possibility that the effects of explanatory variables differ depending on the previous response. Diggle et al., (2002) show that the two equations above can be combined to form the model

$$logit[P(y_{ij}=1 \mid \mathbf{x}_{ij}, y_{ij-1}, \beta, \alpha)] = \mathbf{x}'_{ij}\beta + y_{ij-1}\mathbf{x}'_{ij}\alpha, \qquad (6.5).$$

where

$$\beta = \beta_0 \text{ and } \alpha = \beta_1 - \beta_0.$$

Model (6.4) is a special case of (6.5). Since these models are nested, standard statistical methods for nested models can be applied to test whether smaller models can fit the data as equally well as saturated models. For instance, likelihood ratio tests can be applied to test whether a limited number of covariates interact with the previous outcome. Moreover, we can test whether a smaller model (6.4) fit the data as equally well as the saturated model (6.5).

The covariate dependent higher order models can be fitted by extending the model for the first order Markov chain (6.5). To illustrate the extension to higher models, the second-order model is considered. The transition matrix for the second-order model is of the form

		y_{ij}	
y_{ij-2}	y_{ij-1}	0	1
0	0	π_{000}	π_{001}
0	1	π_{010}	π_{011}
1	0	π_{100}	π_{101}
1	1	π_{110}	π_{111}

where $\pi_{abc} = \Pr(y_{ij} = c \mid y_{ij-2} = a, y_{ij-1} = b)$; for example, π_{001} is the probability that $y_{ij} = 1$ given $y_{ij-2} = 0$ and $y_{ij-1} = 0$. For ease of reference regarding the previous

responses, lag 2 will be used to refer to y_{ij-2} while lag1 will refer to y_{ij-1}. With adherence data for example, π_{001} would indicate the probability that a patient adhered to medication when he/she was non-adherent in lag 2 and lag 1 follow-up visits ($y_{ij-2}=0, y_{ij-1}=0$). Similarly, π_{011} would indicate the probability that a patient adhered to medication when they were not adherent in lag 2 ($y_{ij-2}=0$) follow-up visit but adherent in lag 1 ($y_{ij-1}=1$) follow-up visit.

With the second-order transition models, four separate logistic regression models can be fitted, one for each of the four possible histories (y_{ij-2}, y_{ij-1}), which are (0,0), (0,1), (1,0) and (1,1) with regression coefficients β_{00}, β_{01}, β_{10} and β_{11} respectively. It is more concise to write a single equation as follows:

$$\begin{aligned} &logit[P(y_{ij}=1 \mid y_{ij-2}, y_{ij-1})] \\ &\quad = \mathbf{x}_{ij}'\beta + y_{ij-1}\mathbf{x}_{ij}'\alpha_1 + y_{ij-2}\mathbf{x}_{ij}'\alpha_2 + y_{ij-1}y_{ij-2}\mathbf{x}_{ij}'\alpha_3 \end{aligned} \qquad (6.7)$$

By putting in the different values for y_{ij-2} and y_{ij-1}, we obtain $\beta_{00}=\beta$, $\beta_{01}=\beta+\alpha_1$, $\beta_{10}=\beta+\alpha_2$ and $\beta_{11}=\beta+\alpha_1+\alpha_2+\alpha_3$. The following transition probabilities can be produced:

$$\pi_{001} = P(y_{ij}=1 \mid y_{ij-2}=0, y_{ij-1}=0, \mathbf{x}_{ij}) = \frac{e^{\mathbf{x}_{ij}'\beta_{00}}}{1+e^{\mathbf{x}_{ij}'\beta_{00}}} = \frac{e^{\mathbf{x}_{ij}'\beta}}{1+e^{\mathbf{x}_{ij}'\beta}},$$

$$\pi_{011} = P(y_{ij}=1 \mid y_{ij-2}=0, y_{ij-1}=1, \mathbf{x}_{ij}) = \frac{e^{\mathbf{x}_{ij}'\beta_{01}}}{1+e^{\mathbf{x}_{ij}'\beta_{01}}} = \frac{e^{\mathbf{x}_{ij}'(\beta+\alpha_1)}}{1+e^{\mathbf{x}_{ij}'(\beta+\alpha_1)}},$$

$$\pi_{101} = P(y_{ij}=1 \mid y_{ij-2}=1, y_{ij-1}=0, \mathbf{x}_{ij}) = \frac{e^{\mathbf{x}_{ij}'\beta_{10}}}{1+e^{\mathbf{x}_{ij}'\beta_{10}}} = \frac{e^{\mathbf{x}_{ij}'(\beta+\alpha_2)}}{1+e^{\mathbf{x}_{ij}'(\beta+\alpha_2)}},$$

$$\pi_{111} = P(y_{ij}=1 \mid y_{ij-2}=1, y_{ij-1}=1, \mathbf{x}_{ij}) = \frac{e^{\mathbf{x}_{ij}'\beta_{11}}}{1+e^{\mathbf{x}_{ij}'\beta_{11}}} = \frac{e^{\mathbf{x}_{ij}'(\beta+\alpha_1+\alpha_2+\alpha_3)}}{1+e^{\mathbf{x}_{ij}'(\beta+\alpha_1+\alpha_2+\alpha_3)}}.$$

It should be noted that $\pi_{000}+\pi_{001}=1$, $\pi_{010}+\pi_{011}=1$, $\pi_{100}+\pi_{101}=1$ and $\pi_{110}+\pi_{111}=1$.

Tests can again be carried out to ascertain whether a parsimonious model could fit the data well, that is, testing whether many of the components of the α_i would be equal to zero.

An important special case of (6.7) occurs when there are no interactions between the past outcomes y_{ij-2} and y_{ij-1}, and the explanatory variables, that is, when all elements of α_i are zero except the intercept term. In this case, the previous responses affect the probability of a positive outcome, but the effects of the explanatory variables are the same regardless of the history. Even in this situation, we must still choose from Markov models of a different order. For instance, we might start with a third-order model which can be written in the form

$$logit[P(y_{ij}=1 \mid y_{ij-3}, y_{ij-2}, y_{ij-1})]$$
$$= \mathbf{x}'_{ij}\beta + \alpha_1 y_{ij-1} + \alpha_2 y_{ij-2} + \alpha_3 y_{ij-3} + \alpha_4 y_{ij-1}y_{ij-2} + \alpha_5 y_{ij-1}y_{ij-3} + \alpha_6 y_{ij-2}y_{ij-3}\alpha_7 y_{ij-1}y_{ij-3} \quad (6.8)$$

A second order model can be used if the data are consistent with $\alpha_3 = \alpha_5 = \alpha_6 = \alpha_7 = 0$; a first order model is implied if $\alpha_j = 0$ for $j = 2,\ldots,7$. As with any regression coefficients, the interpretation of the value of β in (6.8) depends on the other explanatory variables in the model, in particular on which previous responses are included. When inference about β are of scientific interest, it is essential to check their sensitivity to the assumed order of the Markov regression model (Diggle et al., 2002).

6.4 Evaluating adherence data using transition models

With the data from HIV positive adults who are on HAART, the overall scientific interest has been to explore factors that affect HAART adherence over time among these adults. In this chapter, the analyses use covariates identified using the GEE model to address whether or not the patient's adherence to medication depends on the previous outcomes (history) in addition to the identified GEE covariates. Moreover, it would be interesting to also assess whether the effects

of explanatory variables are the same regardless of the history (no interactions between the previous outcomes and explanatory variables) or the effects differ depending on the history (interactions between previous responses and explanatory variables exist).

Recall the GEE model for the adherence data (Chapter 4, section 4.6)

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 educ\_no_i + \beta_4 educ\_no_i + \beta_5 site_i + \beta_6 partner_i$$
$$+ \beta_7 income_i + \beta_8 WHOstage1_i + \beta_9 WHOstage2_i + \beta_{10} WHOstage3_i + \beta_{11} base\_cd4_i$$
$$+ \beta_{12} baseweight_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} tapwater_i + \beta_{16} electricity_i$$
$$+ \beta_{17} cellphone_i + \beta_{18} adh\_month1_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} time_i gender_i$$
$$+ \beta_{22} time_i site_i + \beta_{23} time_i reason\_no_i + \beta_{24} time_i reason\_risk_i + \beta_{25} age_i educ\_no_i$$
$$+ \beta_{26} age_i educ\_p_i + \beta_{27} age_i gender_i$$

A series of transition models were fitted to this GEE model using 'Proc Logistic' and 'Proc Genmod' in SAS. First, we assessed whether there is a dependence on (a) the previous response, y_{ij-1}, (b) the response two occasions prior to the current one (y_{ij-1}, y_{ij-2}) and (c) the response three occasions prior to the current one (y_{ij-1}, y_{ij-2}, y_{ij-3}). Thus three models (first-, second- and third-order models) were fitted, where the previous responses were included as additional explanatory variables into the GEE model. The goodness-of-fit was tested using the deviance analysis which is produced by the type 3 analysis of effects. Type 3 analysis of effects considers the overall model and assesses the contribution of each variable to deviance reduction irrespective of the sequence in which the variables enter the model. The results from type 3 analysis of the effects are presented in Table 6.1.

The results show a very strong dependence on the previous outcome measurement (*p*-value = <0.001) and on the measurement two occasions prior to the current (*p*-value = <0.001). Three prior responses did not depend on the current response (*p*-value = 0.167), and for further model assessments, the third-order transition model was not considered.

Table 6. 1: Type 3 analysis of effects for first-, second- and third order models

Effects	DF	**First order**		**Second order**		**Third order**	
		Wald Chi-square	*p*-value	Wald Chi-square	*p*-value	Wald Chi-square	*p*-value
Gender	1	2.9914	0.0837	2.2137	0.1368	1.9983	0.1575
Education	2	13.6998	0.0011	11.9152	0.0026	11.5297	0.0031
Treatment site	1	17.1163	<.0001	13.1546	0.0003	12.1355	0.0005
Income	1	0.1828	0.6690	0.1761	0.6747	0.1780	0.6731
Tapwater	1	0.9473	0.3304	0.8523	0.3559	0.8094	0.3683
Electricity	1	0.3083	0.5787	0.4145	0.5197	0.4548	0.5000
Cell phone	1	2.1826	0.1396	1.5157	0.2183	1.3948	0.2376
WHO staging	3	1.3976	0.7061	1.2082	0.7510	1.1675	0.7608
Partner	1	8.2366	0.0041	6.8889	0.0087	6.5753	0.0103
Reason	2	5.4967	0.0640	4.4073	0.1104	4.2353	0.1203
Baseline_adh	1	3.4725	0.0624	2.8261	0.0927	2.7130	0.0995
Time (visit)	1	16.0774	<.0001	13.6226	0.0002	12.7552	0.0004
Age	1	2.6288	0.1049	2.3056	0.1289	2.1968	0.1383
Baseline CD4+	1	0.0706	0.7904	0.0491	0.8246	0.0474	0.8277
Baseline weight	1	0.1906	0.6624	0.1491	0.6994	0.1395	0.7088
Weight_follow-up	1	0.1527	0.6960	0.1796	0.6717	0.1853	0.6668
Time*gender	1	9.2673	0.0023	8.2419	0.0041	7.9157	0.0049
Time*site	1	0.0512	3.8019	2.9202	0.0875	2.7032	0.1001
Time*reason	2	6.8689	0.0322	5.9122	0.0520	5.7521	0.0564
Age*gender	1	1.2757	0.2587	0.8763	0.3492	0.7833	0.3761
Age*education	1	10.5237	0.0052	9.3435	0.0094	9.0683	0.0107
Lag1 (y_{ij-1})	1	59.3599	<.0001	47.9417	<.0001	46.0769	<.0001
Lag2 (y_{ij-2})	1			25.8090	<.0001	24.0398	<.0001
Lag3 (y_{ij-3})	1					1.9190	0.1660

Second, we checked whether the effects of the explanatory variables differ depending on the previous responses (history). That is, we assessed whether any interaction terms (between the previous response and explanatory variables) need to be incorporated into the second-order model. To examine this, the second-order model (in Table 6.1) was extended by fitting each of the two-way interaction terms formed from the previous responses and explanatory variables.

They were added to the model one at a time and their significance assessed. Three interactions reduced the deviance by relatively large amounts: lag1*time led to a deviance reduction of 6.48 on 1 degree of freedom and a corresponding *p*-value of 0.011, lag2*baseline adherence and lag2*cell phone reduced the deviance by 5.98 (*p*-value = 0.014) and 6.13 (*p*-value = 0.013) on 1 degree of freedom respectively. To see if all the three interactions could be retained in the model, they were fitted all at the same time and they reduced the deviance by 19.98 on 3 degrees of freedom and a corresponding *p*-value of <0.001. Finally, the model was examined to see if any interactions from the GEE model could be omitted. The interaction between age and gender increased the deviance by 2.185 on 1 degree of .freedom (*p*-value = 0.139) and was therefore dropped. As a result, the selected model was the second-order model (Table 6.1) but excluding the interaction term age*gender and including the three interaction terms: lag1*time, lag2*baseline adherence and lag2*cell phone

The goodness-of-fit was tested using Hosmer-Lemeshow test. The observed and expected frequencies are given in Table 6.2.

Table 6. 2: Partion for the Hosmer-Lemeshow Goodness-of-Fit test

Group	Total	Event = adherent		Non-event = non adherent	
		Observed	Expected	Observed	Expected
1	508	264	257.43	244	250.57
2	508	339	345.56	169	162.44
3	508	381	380.63	127	127.37
4	508	398	403.13	110	104.87
5	508	425	420.31	83	87.69
6	508	441	432.75	67	75.25
7	508	445	441.67	63	66.33
8	508	433	450.06	75	57.94
9	508	460	459.31	48	48.69
10	505	475	470.12	30	34.88

The goodness-of-fit statistics is 9.0149 with 8 degrees of freedom, and the corresponding *p*-value of 0.341. The very large *p*-value for this test shows that the model fits the data well (i.e. the predicted probabilities correspond with the observed values).

The appropriateness of the link function was tested by refitting the model using a linear predictor and a squared linear predictor and the results are given in Table 6.3.

Table 6. 3 Logit link function test

Variable	DF	Chi-square	p-value
Constant	1	0.0104	0.919
Linear predictor	1	67.609	<.0001
Squared linear predictor	1	0.0643	0.800

The very small p-value for the linear predictor and a very large p-value for the squared linear predictor variables in Table 6.3 suggest that the link is appropriate, and thus agrees with the goodness-of-fit test that the model fits the data well.

A simple check of the sensitivity of inferences about the regression coefficients to the Markov assumption, which in this case is the 2^{nd} order, was carried out by assessing whether the ordinary and robust standard errors were similar. This was carried out using the GLIMMIX procedure (to fit a generalized linear model) since it can provide empirical error estimates for the regression coefficients and the results are presented Table 6.4. The results showed both standard errors to be very similar, which suggests that the 2^{nd} order Markov assumption is valid, thus the standard errors estimated by ordinary logistic regression are valid.

Table 6. 4: Comparison of model based and empirical standard errors of the 2nd Markov model

Effect	Estimates	Model based std errors		Empirical std errors	
		Std error	Pvalue	Std error	Pvalue
Intercept	-0.2143	0.4351	0.6223	0.4378	0.6245
Gender (ref=male)					
female	-0.4148	0.1979	0.0361	0.1990	0.0372
Education (ref=sec & higher)					
No schooling	2.0950	0.5866	0.0004	0.6092	0.0006
Primary	0.2412	0.5086	0.6353	0.5180	0.6414
Treatment site (ref=rural)					
Urban	1.0166	0.1997	<.0001	0.2015	<.0001
Income (ref=not source)					
Source of income	-0.0303	0.0897	0.7353	0.0901	0.7364
Access to tapwater (ref=no)					
Yes	0.1602	0.1273	0.2082	0.1275	0.2089
Househld with electricity (ref=no)					
Yes	0.0897	0.1178	0.4469	0.1161	0.4399
Cell phone ownership (ref=no)					
Yes	-0.1686	0.1403	0.2298	0.1397	0.2278
WHO staging (ref=stage4)					
Stage1	-0.2706	0.2078	0.1929	0.2118	0.2015
Stage2	-0.1464	0.1884	0.4373	0.1937	0.4499
Stage 3	0.0581	0.1736	0.7375	0.1795	0.7459
Living with partner (ref=no)					
Yes	0.2242	0.0942	0.0173	0.0929	0.0158
Reason for testing (ref=unwell)					
No specific reason	-0.2963	0.2209	0.1798	0.2197	0.1776
Exposed to the risk	-0.5473	0.2504	0.0289	0.2512	0.0294
Baseline adherence (ref=no)					
yes	-0.5254	0.1696	0.0020	0.1669	0.0017
Time (visit)					
Age	0.0877	0.0254	0.0006	0.0261	0.0008
Baseline CD4+	-0.0002	0.0007	0.7976	0.0007	0.7997
Baseline weight	-0.0020	0.0058	0.7363	0.0063	0.7524
Weight at follow-up visits	-0.0014	0.0053	0.7944	0.0058	0.8142
Time*gender (ref=male)					
female	0.0719	0.0195	0.0002	0.0196	0.0003
Time*site (ref=rural)					
urban	-0.0508	0.0196	0.0094	0.0198	0.0102
Time*reason (ref=unwell)					
No specific reason	0.0147	0.0219	0.5022	0.0223	0.5084
Exposed to the risk	0.0606	0.0260	0.0198	0.0260	0.0197
Age*education					
No schooling	-0.0439	0.0147	0.0029	0.0153	0.0042
Primary	-0.0059	0.0131	0.6505	0.0134	0.6577
Previous1 (ref=no)					
yes	1.1597	0.1986	<.0001	0.1990	<.0001
Previous2 (ref=no)					
yes	-0.1332	0.1941	0.4927	0.1888	0.4807
Previous1*time	-0.0580	0.0213	0.0066	0.0214	0.0068
Previous2*cellphone (ref=no)					
Yes	0.4122	0.1668	0.0135	0.1670	0.0136
Previous2*baseline adh (ref=no)					
yes	0.4938	0.2013	0.0142	0.1996	0.0134

The final selected model is as follows:

$$
\begin{aligned}
logit[P(y_{ij}=1 \mid y_{ij-2}, y_{ij-1}, \mathbf{x}_{ij}) &= \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i \\
&+ \beta_6 tapwater_i + \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i \\
&+ \beta_{11} WHOstage3_i + \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i \\
&+ \beta_{15} adh\_month1_i + \beta_{16} age_i + \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} \\
&+ \beta_{20} time_{ij} + \beta_{21} time_i gender_i + \beta_{22} time_i site_i + \beta_{23} time_i reason\_no_i \\
&+ \beta_{24} time_i reason\_risk_i + \beta_{25} age_i educ\_no_i + \beta_{26} age_i educ\_p_i + \alpha_1 y_{ij-1} time_{ij} \\
&+ \alpha_2 y_{ij-2} cellphone_i + \alpha_3 y_{ij-2} adh\_month1_i
\end{aligned}
$$

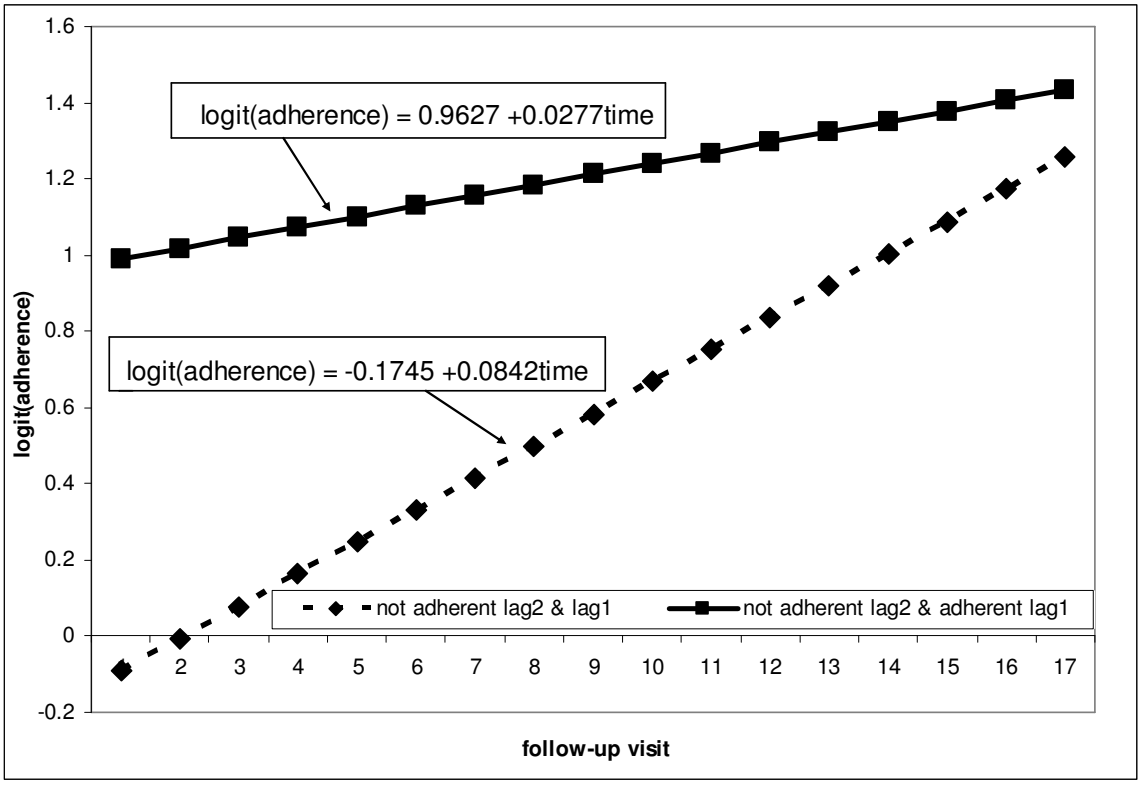
The parameter estimates and the corresponding confidence limits for the above model are presented in Table 6.5. The results show that there are three interactions between the past responses and the explanatory variables and they are lag1*time, lag2*cell phone ownership and lag2*baseline adherence.

Adherence is increasing over time but the rate of increase is different for patients who were adherent and those who were not adherent in the previous follow-up visits. With a unit increase in time (follow-up visit), the odds of adherence are 1.09 ($e^{0.0842}$) times more likely for the patients who were not adherent in both lag2 and lag1 follow-up visits ($y_{ij-2}=0$ and $y_{ij-1}=0$), whereas the odds of adherence are 1.03 ($e^{0.0277}$) times more likely for patients who were not adherent in lag2 follow-up visits ($y_{ij-2}=0$) but adherent in lag1 follow-up visits ($y_{ij-1}=1$). Thus the rate of increase in adherence is significantly higher for patients who were not adherent in two previous successive follow-up visits than for those who changed from not being adherent to being adherent in two previous successive follow-up visits (p=0.0086) (Table 6.5). Figure 6.1 depicts this interaction between the past responses and time.

Table 6. 5: Parameter estimates of the chosen 2nd order Markov model

Effect	Estimates	Std errors	p-value	95% C.I.	
Intercept	-0.1745	0.4399	0.6915	-1.0367	0.6876
Gender (ref=male)					
female	-0.3959	0.1995	0.0472	-0.7870	-0.0049
Education (ref=sec & higher)					
No schooling	2.3686	0.5948	<.0001	1.2028	3.5344
Primary	0.3411	0.5319	0.5213	-0.7014	1.3837
Treatment site (ref=rural)					
Urban	0.9974	0.2011	<.0001	0.6032	1.3915
Income (ref=not source)					
Source of income	-0.0238	0.0898	0.7906	-0.1999	0.1522
Access to tapwater (ref=no)					
Yes	0.1716	0.1324	0.1951	-0.0880	0.4311
Househld with electricity (ref=no)					
Yes	0.0952	0.1223	0.4362	-0.1445	0.3349
Cell phone ownership (ref=no)					
Yes	-0.1798	0.1407	0.2059	-0.4538	0.0978
WHO staging (ref=stage4)					
Stage1	-0.2470	0.2106	0.2408	-0.6597	0.1657
Stage2	-0.1574	0.1887	0.4040	-0.5272	0.2123
Stage 3	-0.0860	0.1742	0.6217	-0.4274	0.2555
Living with partner (ref=no)					
Yes	0.2374	0.0948	0.0123	0.0516	0.4232
Reason for testing (ref=unwell)					
No specific reason	-0.3035	0.2224	0.1724	-0.7395	0.1324
Exposed to the risk	-0.6019	0.2516	0.0168	-1.0950	-0.1087
Baseline adherence (ref=no)					
yes	-0.5425	0.1702	0.0014	-0.8761	0.2088
Time (visit)	0.0842	0.0257	0.0011	0.0338	0.1346
Age	0.0050	0.0070	0.4710	-0.0087	0.0187
Baseline CD4+	-0.0003	0.0007	0.6878	-0.0016	0.0010
Baseline weight	-0.0016	0.0059	0.7880	-0.0131	0.0099
Weight at follow-up visits	-0.0018	0.0053	0.7395	-0.0121	0.0086
Time*gender (ref=male)					
female	0.0727	0.0197	0.0002	0.0341	0.1112
Time*site (ref=rural)					
urban	-0.0484	0.0197	0.0143	-0.0484	-0.0097
Time*reason (ref=unwell)					
No specific reason	0.0145	0.0221	0.5112	-0.0288	0.0579
Exposed to the risk	0.0659	0.0262	0.0119	0.0145	0.1173
Age*education					
No schooling	-0.0521	0.0150	0.0005	-0.0815	-0.0227
Primary	-0.0080	0.0137	0.5613	-0.0347	0.0189
Previous1 (ref=no)					
yes	1.1372	0.2000	<.0001	0.7453	1.5292
Previous2 (ref=no)					
yes	-0.1267	0.1948	0.5155	-0.5084	0.2551
Previous1*time	-0.0565	0.0215	0.0086	-0.0565	-0.0144
Previous2*cellphone (ref=no)					
Yes	0.4087	0.1674	0.0146	0.0805	0.7368
Previous2*baseline adh (ref=no)					
yes	0.4920	0.2020	0.0149	0.0961	0.8879

Figure 6. 1 Log odds of adherence for the interaction between time and previous adherence responses



The odds ratios and their confidence limits for the effects of the interaction between the previous responses and cell phone ownership as well as previous responses and baseline adherence are presented in Tables 6.6 and 6.7 respectively. The contrasts were derived from the estimated model coefficients in Table 6.5.

Table 6. 6: Odds ratios and their confidence limits for the effects of the interaction between outcomes at previous visits and cell phone ownership

Contrast	Odds ratio	95% CI of odds ratio
Cell phone vs no cell phone		
$(y_{ij-2}=1 \ \& \ y_{ij-1}=0)$	1.2595	(1.0370,1.5282)
$(y_{ij-2}=0 \ \& \ y_{ij-1}=0)$	0.8369	(0.6352, 1.1027)
$(y_{ij-2}=1 \ \& \ y_{ij-1}=0)$ vs $(y_{ij-2}=0 \ \& \ y_{ij-1}=0)$		
Own a cell phone	1.3258	(0.8784, 2.0010)
Do not own a cell phone	0.8810	(0.6015, 1.2905)

The results show that patients who owned a cell phone are 1.26 times (95% CI: 1.04, 1.5) (Table 6.6) more likely to adhere to medication than patients without cell phones for patients who changed from being adherent in lag2 to being non-adherent in lag1 follow-up visits ($y_{ij-2}=1$ and $y_{ij-1}=0$). There was no significant difference in adherence between patients who owned cell phones and those without cell phones if they were not adherent in both lag2 and lag1 follow-up visits ($y_{ij-2}=0$ and $y_{ij-1}=0$) (Table 6.6). Furthermore, there was no significant difference between the patients who were adherent in lag2 but not adherent in lag1 and those who were not adherent in both lag2 and lag1 follow-up visits regardless of whether or not they owned a cell phone (Table 6.6).

Table 6. 7: Odds ratios and their confidence limits for the effects of the interaction between outcomes at previous visits and baseline adherence

Contrasts	Odds ratio	95% CI of odds ratio
Adherent vs not adherent at baseline		
($y_{ij-2}=1$ & $y_{ij-1}=0$)	0.9508	(0.7569,1.1943)
($y_{ij-2}=0$ & $y_{ij-1}=0$)	0.5813	(0.4164, 0.8115)
($y_{ij-2}=1$ & $y_{ij-1}=0$) vs ($y_{ij-2}=0$ & $y_{ij-1}=0$)	1.4410	(1.1709, 1.7735)
Adherent at baseline	0.8810	(0.6015, 1.2905)
Non adherent at baseline		

Patients who were adherent at baseline are less likely to adhere to medication than patients who were not adherent at baseline [OR=0.58, 95% CI (0.42, 0.81)] (Table 6.7) given that they were not adherent in both lag2 and lag1 follow-up visits. Moreover, with patients that were adherent to medication at baseline, the odds of adherence are 1.44 times (95% CI: 1.17, 1.77) more likely for patients who made a transition from being adherent in lag2 to not being adherent in lag1 follow-up visits ($y_{ij-2}=1$ and $y_{ij-1}=0$) compared to those who were not adherent in both lag2 and lag1 follow-up visits ($y_{ij-2}=0$ and $y_{ij-1}=0$).

From Table 6.5, there were four other significant interactions (that do not involve a previous response) and one significant main effect that was not involved in an interaction. The interaction terms were visit*gender, visit*treatment site, visit*reason for test and age*education while the main effect was living with/without a partner. The results show that after controlling for other variables as well as the outcome (adherent/not adherent) at two prior successive follow-up visits, adherence for patients who live with a partner are 1.27 (95% CI: 1.05, 1.53) times higher than those who do not live with a partner (odds ratios derived from Table 6.5).

Figure 6.2 depicts the effects of the interaction between time and gender. Having controlled for other covariates in the model including the outcome at two prior successive follow-up visits, it is shown in Figure 6.2 that adherence is increasing over time but the rate of increase is different for males and females, it was higher for females compared to males (p-value=0.0002).

Figure 6. 2: Log odds of adherence for the interaction between time and gender

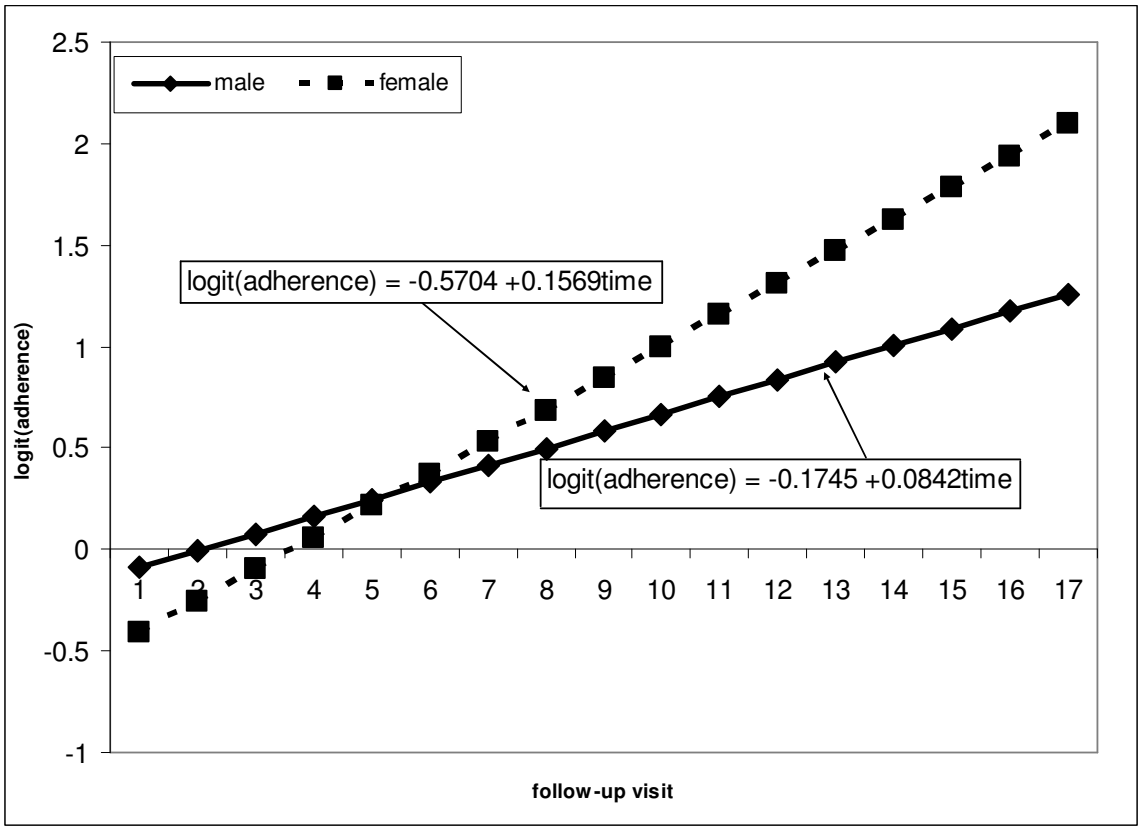
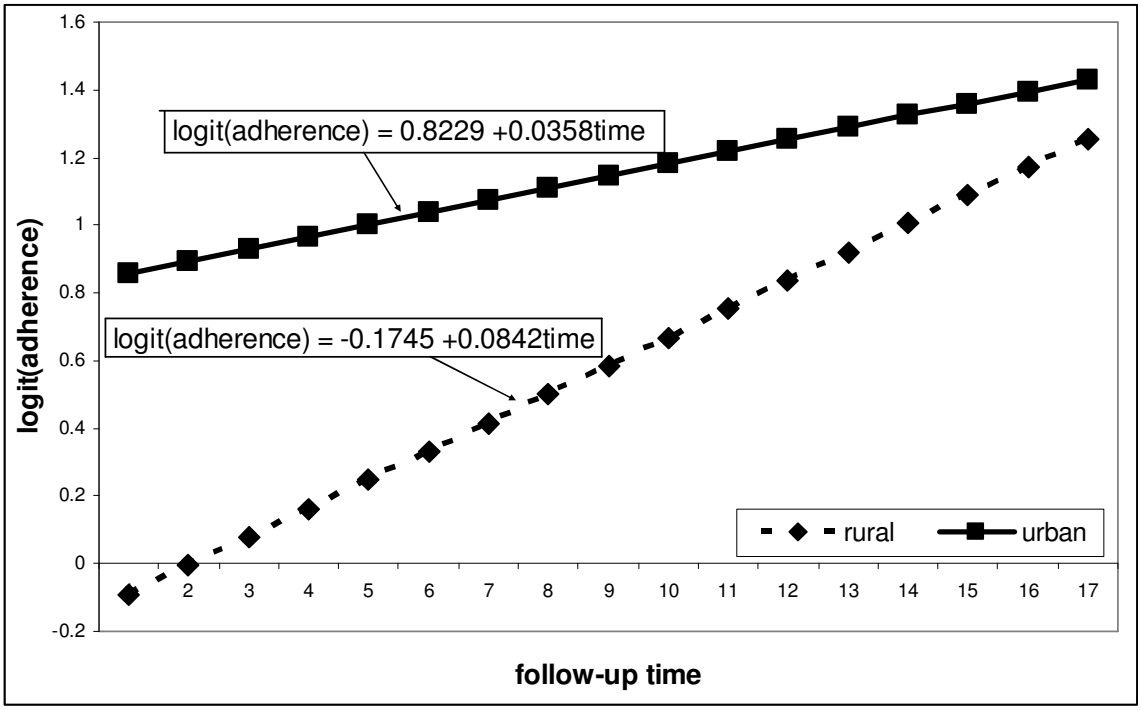


Figure 6.3 depicts the effects of the interaction between time and treatment site. It is shown in Figure 6.3 that adherence is increasing over time but the rate of increase is different for the urban and rural treatment sites after controlling for the outcome in the two prior successive follow-up visits and other covariates in the model. The rate of increase in adherence is higher in rural than the urban treatment sites (p=0.0143)(Table 6.4).

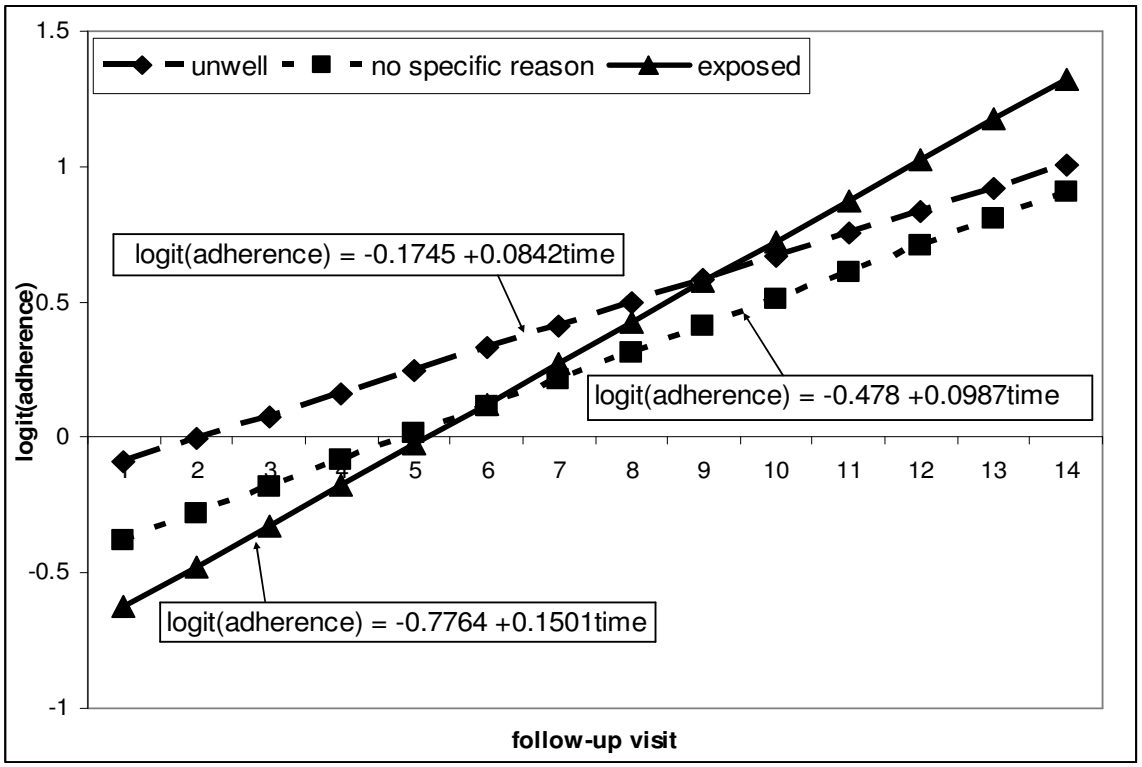
Figure 6. 3: Log odds of adherence for the interaction between time and treatment site



With regard to the interaction between time and reason for taking an HIV test, it is shown (Figure 6.4) that the rate of increase in adherence was higher for patients who have reported to have been exposed to the risk of contracting the HIV disease compared to those who reported to have taken the test because they were unwell (p=0.0119). There was however, no significant difference in the rate of adherence between patients who reported no specific reason and those who reported to be unwell (p=0.5112). Further analysis revealed that the rate of adherence between patients who reported no specific reason and those who

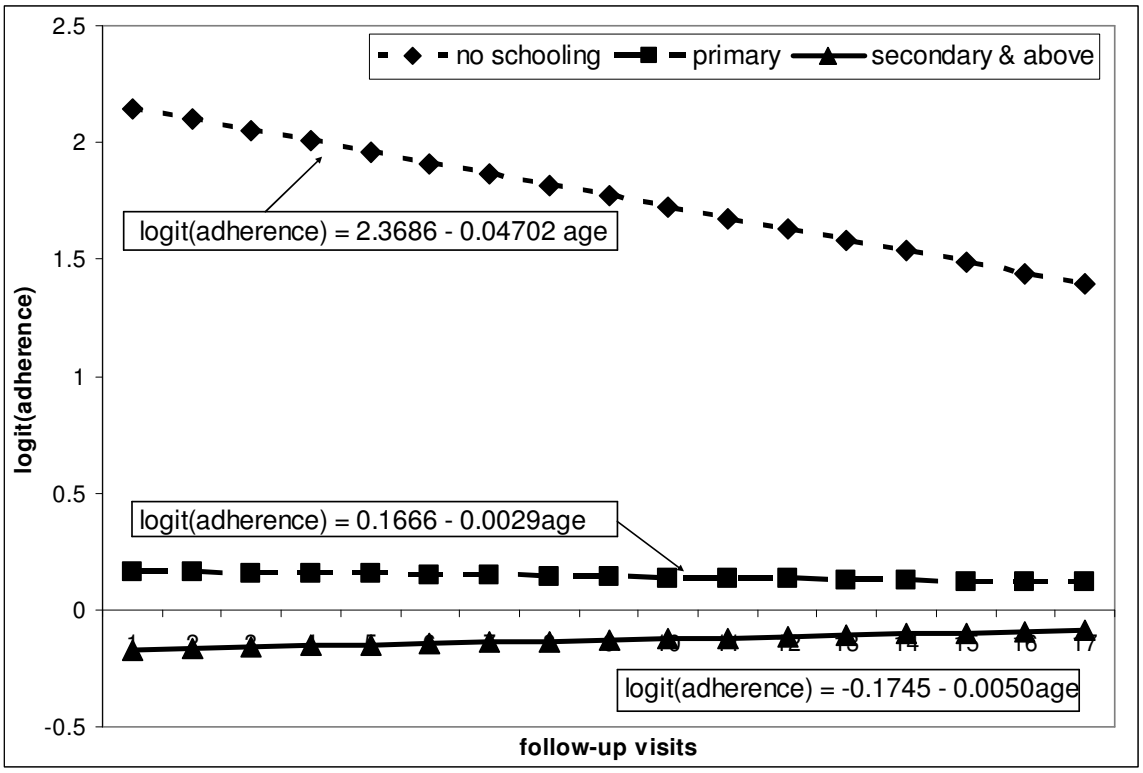
reported to have been exposed to the risk is marginally insignificant (p=0.0811). These results have been adjusted for the outcome at two prior successive follow-up visits and other explanatory variables in the model.

Figure 6. 4: Log odds of adherence for the interaction between time and reason for taking an HIV test



Having controlled for adherence outcomes in the two prior successive follow-up visits and other covariates in the model, the rate at which adherence is decreasing with age was higher with patients that have no schooling compared to patients with secondary level (p=0.0005), whilst there was no significant difference in the rate of change in adherence with age between patients with primary and secondary level of education (p=0.5613) (Table 6.5). Further analysis revealed that the rate of decrease in adherence with age is more pronounced with patients with no schooling than patients with the primary level of education (p=0.0133). The log odds of adherence between age and gender is depicted in Figure 6.5.

Figure 6. 5: Log odds of adherence for the interaction between age and education



6.5 Summary

Transition models address dependency among repeated measurements by treating previous outcomes as additional explanatory variables. Fitting transition models is relatively straightforward because subsequent measurements, given their previous history are considered to be independent; as a result, standard GLM procedures can be employed. Also transition models are likelihood-based methods, therefore inferential procedures such as a likelihood ratio test can be used. Also, because they are likelihood-based, they tend to be robust to missingness; they assume MAR. However, these models are usually criticized because of the conditional interpretation of the parameters on other outcomes and on the number repeated measures.

The findings of a transition model highlighted that in addition to the variables that have been significant in the GEE and GLMM models, optimal HAART adherence depends on the two previous outcome measurements. Furthermore, effects of

some of the explanatory variables differ depending on the previous responses. The negative relationship between baseline adherence and adherence over time observed earlier is in some ways re-affirmed by the significant interaction between previous adherence outcomes and baseline adherence where it is shown that adherent patients at baseline were less likely to adhere to medication that those who were not adherent at baseline given that they were not adherent in two previous measurements.

With all the model families fitted, namely, marginal, random effects and transition models; we had assumed that the time interval between successive follow-up visits is equal. In the next chapter, we relax the equal interval assumption by jointly modeling HAART adherence with duration between follow-up visits.

Chapter 7

Joint modeling of HAART adherence and visit interval

7.1 Introduction

We have so far explored the factors affecting adherence to antiretroviral medication over time. More specifically, we have intensively assessed the evolution of adherence and the associated factors at the population level (marginal model using GEE) and subject-specific level (random effects models using GLMM). Moreover, we evaluated whether or not the past adherence outcomes have any influence on the current response of adherence using transition models. Nonetheless, in some longitudinal studies, although one time-varying outcome may be of primary interest, several related processes are also measured (Liu, Daniels and Marcus, 2009). The association between a primary outcome and another related outcome can reveal a great deal of insight about the mechanism of behavioural change. In the adherence study, a related outcome to adherence is the time interval (duration) between successive visits, which has been measured as the number of days taken by a patient between clinic visits. In the previous chapters, the effect of predictor variables on adherence status of patients over time was evaluated assuming equal interval between successive visits. In this chapter we aim to further investigate the joint effect of these predictor variables on both adherence status of patients and duration between successive visits. More specifically, we seek to assess whether the explanatory variables that were found to be significantly related to adherence in the previous chapters would still have a significant effect on adherence even when duration between successive visits was accounted for. Assessing the association between the two outcomes (adherence and duration) is also of interest. This can be best addressed within the framework of joint modeling of the two outcomes of interest. Advantages of joint over separate fitting of models include better control over type I error rates in multiple tests,

possible gains in efficiency in the parameter estimates and the ability to answer multivariate questions (Gueorguieva, 2001).

Difficulties analyzing longitudinal data arise because of correlations usually present between observations on the same subject. In case of multiple outcomes, two types of correlations must be taken into account: correlations between measurements on different variables and correlations between measurements on the same variable within a subject (Gueorguieva, 2001). While joint models for longitudinal and time-to-event data are widely present in the literature (Henderson, Diggle and Dobson, 2000, 2002; Guo and Carlin, 2004; Tsiatis and Davidian, 2004), joint models for two longitudinal outcomes of similar or dissimilar nature are less widespread. Nonetheless, methods focusing on models that jointly analyze discrete and continuous outcomes have been explored (Fitzmaurice and Laird, 1995; Aerts et al., 2002; Molenberghs and Verbeke, 2005; Faes, Geys and Catalono, 2008). The challenge has been the lack of multivariate distributions for combining both types of outcomes; as a result, specification of a joint distribution of the responses is not straightforward. There are broadly two approaches adapted to joint modeling. A first approach avoids direct specification of a joint distribution; it is based on a conditioning argument that allows joint distribution to be factored in a marginal component and a conditional component, where the conditioning can be done either on the discrete or on the continuous outcome (Catalano and Ryan, 1992; Faes et al., 2004). A disadvantage of mixed outcome models based on conditional models is that they do not directly lead to marginal inferences (Verbeke and Davidian, 2008). Also the correlation among the two outcomes cannot be directly estimated (Faes et al., 2008).

A second approach directly formulates a joint model for both outcomes. The latent variable idea has been used to directly specify the joint distribution of the discrete and continuous outcomes based on two methods (Regan and Catalano, 2002). The first method introduces a probit approach, where an underlying

continuous variable is assumed for each binary outcome, following a normal distribution. The second method is based on a Plackett-Dale approach, assuming a Plackett latent variable to model bivariate outcomes in which one component is continuous and the other is binary (Faes et al., 2004). The two bivariate latent variable models are different in the way the association between the two outcomes is described. The probit approach uses a correlation coefficient, while the Plackett-Dale approach makes use of the odds ratio (Faes et al., 2004).

Instead of using a latent variable approach, one can directly specify the joint distribution for both outcomes through a mixed model, by specification of the marginal distribution, conditional on the correlated random effect (Faes et al., 2008). The generalized linear mixed model (GLMM) forms a very general class of subject-specific models for discrete and continuous responses in the exponential family and is used for univariate repeated measures (Fahrmeir and Tutz, 1994). In this chapter, we aim to review how the GLMM approach can be extended for multivariate longitudinal data by assuming separate random effects for each outcome variable and then combining them by imposing a joint multivariate distribution on the random effects. That is, the different response processes are associated by imposing a joint multivariate distribution on the random effects and an advantage of this approach is that additional correlation emerging from the longitudinal data structure can be modeled within the same framework (Fieuws and Verbeke, 2004; Gueorguieva, 2001).

This chapter is organized as follows. In Section 7.2, a multivariate generalized linear mixed model for two longitudinal outcomes is presented. An in-depth formulation of a joint model is demonstrated using longitudinal continuous and binary outcomes in Section 7.3. Section 7.4 presents the results of a joint model between optimal HAART adherence and duration between successive visits.

7.2 Multivariate GLMM formulation of a joint model

Joint outcomes can be measured repeatedly over time or might be observed within a hierarchical context. The generalized linear mixed model introduced in Chapter 5 can easily be adapted to situations where various outcomes of a different nature are observed (Molenberghs and Verbeke, 2005). Let us consider a conditional random effects model with a bivariate response. The following exposition is given in Gueorguieva (2001). Denote the response vector for the ith subject by $\mathbf{y}_i = (\mathbf{y}_{i1}', \mathbf{y}_{i2}')'$, where $\mathbf{y}_{i1} = (y_{i11}, y_{i12}, \ldots, y_{i1n_{i1}})'$ and $\mathbf{y}_{i2} = (y_{i21}, y_{i22}, \ldots, y_{i2n_{i2}})'$ are repeated measurements on the first and second outcomes. Assume that $\mathbf{y}_{i1j}, j = 1, \ldots, n_{i1}$, are conditionally independent given $\mathbf{b}_{i1}$ with density $f_1(.)$ in the exponential family. Likewise, we assume that $\mathbf{y}_{i2j}, j = 1, \ldots, n_{i2}$, are conditionally independent given $\mathbf{b}_{i2}$ with density $f_2(.)$ in the exponential family. Also, $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ are conditionally independent given $\mathbf{b}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2})'$ and the responses on different subjects are independent. Let $g_1(.)$ and $g_2(.)$ be appropriate link functions for f_1 and f_2. Denote the conditional means of $\mathbf{y}_{i1j}$ and $\mathbf{y}_{i2j}$ by μ_{i1j} and μ_{i2j} respectively. Let $\mu_{i1} = (\mu_{i11}, \ldots, \mu_{i1n_{i1}})'$ and $\mu_{i2} = (\mu_{i21}, \ldots, \mu_{i2n_{i1}})'$. At stage 1 of the mixed model specification, we assume

$$\mu_{i1} = g_1^{-1}(\mathbf{X}_{i1}\beta_1 + \mathbf{Z}_{i1}\mathbf{b}_{i1}) \tag{7.1}$$

$$\mu_{i2} = g_2^{-1}(\mathbf{X}_{i2}\beta_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2}) \tag{7.2}$$

where the components of the inverse link functions g_1^{-1} and g_2^{-1} are allowed to change with the outcomes $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ respectively. $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are $(n_{i1} \times p_1)$ and $(n_{i1} \times p_2)$ design matrices for fixed effects, $\mathbf{Z}_{i1}$ and $\mathbf{Z}_{i2}$ are $(n_{i1} \times q_1)$ and $(n_{i1} \times q_2)$ design matrices for the random effects and β_1 and β_2 are $p_1 -$ and $p_2 -$ vectors of unknown fixed regression coefficients respectively. At stage 2

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} \sim \text{i.i.d.}\, \mathbf{MVN}(\mathbf{0}, \mathbf{G}) = \mathbf{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{12}' & \mathbf{G}_{22} \end{bmatrix} \right),$$

where $\mathbf{G}$, $\mathbf{G}_{11}$ and $\mathbf{G}_{22}$ are in general unknown positive-definite matrices. When $\mathbf{G}_{12} = \mathbf{0}$ then the above model is equivalent to two separate GLMM's for two outcome variables, that is, both outcomes are assumed to be completely independent. If the two vectors of the random effects are perfectly correlated, that is, if $\mathbf{b}_{i1} = c\mathbf{b}_{i2}$, where c is a constant, then this model reduces to the shared-parameter model (Molenberghs and Verbeke, 2005; Faes et al., 2008).

The marginal means and the variances of $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ for the model defined in (7.1) and (7.2) are the same as those of the GLMM considering one variable at a time (refer to Chapter 5 for details)

$$E(\mathbf{y}_{i1}) = E\left[E\left(\mathbf{y}_{i1} \mid \mathbf{b}_{i1}\right)\right] = E\left[\mu_{i1}\right]$$

$$E(\mathbf{y}_{i2}) = E\left[E\left(\mathbf{y}_{i2} \mid \mathbf{b}_{i2}\right)\right] = E\left[\mu_{i2}\right]$$

and

$$var(\mathbf{y}_{i1}) = var(E(\mathbf{y}_{i1} \mid \mu_{i1}) + E(var[\mathbf{y}_{i1} \mid \mathbf{b}_{i1}) = var\left(\mu_{i1}\right) + E\left[\phi a_i \upsilon\left(\mu_{i1}\right)\right]$$

$$var(\mathbf{y}_{i2}) = var(E(\mathbf{y}_{i2} \mid \mu_{i2}) + E(var[\mathbf{y}_{i2} \mid \mathbf{b}_{i2}) = var\left(\mu_{i2}\right) + E\left[\phi a_i \upsilon\left(\mu_{i2}\right)\right],$$

where $\upsilon(\mu_{i1})$ and $\upsilon(\mu_{i2})$ denote the variance functions corresponding to the exponential family distributions for the two outcome variables. The marginal covariance matrix between $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ is found to be equal to the covariance between μ_{i1} and μ_{i2}, that is, $cov(\mathbf{y}_{i1}, \mathbf{y}_{i2}) = cov(\mu_{i1}, \mu_{i2})$. This property is a consequence of the key assumption of conditional independence between the two outcome variables, it is critical because it allows an extension of model fitting methods from the univariate to the multivariate GLMM.

These joint models can be generalized to more than two outcomes as well as random effects models with a correlated residual error structure and in models with no random effects (marginal generalized linear model). These models are demonstrated in the next section with continuous and binary endpoints.

7.3 A joint model for continuous and binary longitudinal data

In this section, we describe joint models for the specific setting of a continuous and a binary longitudinal outcome. More specifically, we formulate three possible models to account for the longitudinal structure of a joint continuous and binary outcome. These are a) a fully marginal model, b) a conditional independence random-intercepts model and c) a random-intercepts model with a correlated residual error structure. In order to demonstrate the formulation of these possible models, it is useful to start from the formulation in Chapter 5, Section 5.4.2, where both random effects and serial correlation have been allowed for. For a bivariate response vector for subject i by $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2})'$ where $\mathbf{y}_{i1} = (y_{i11}, y_{i12}, \ldots, y_{i1n_i})'$ and $\mathbf{y}_{i2} = (y_{i21}, y_{i22}, \ldots, y_{i2n_i})'$ are repeated measurements for continuous and binary outcomes respectively, we assume a general model of the form:

$$\mathbf{y}_i = \mu_i(\eta_i) + \varepsilon_i = g^{-1}(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i) + \varepsilon_i \qquad (7.3)$$

where μ_i is specified in terms of fixed and random effects and ε_i is the residual error term. The model is written in its most general form, as a decomposition of the mean and an appropriate error term, where both the mean and error term are allowed to change with the nature of the outcomes. The components of the inverse link function $g^{-1}(.)$ depend on the nature of outcomes in $\mathbf{y}_i$. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(2n_i \times p)$ and $(2n_i \times q)$-dimensional matrices of the known covariate values corresponding to subject i, and β is a p-dimensional vector of unknown fixed regression coefficients. Furthermore, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ are the q-dimensional random effects. The components of the residual error structure ε_i have the appropriate distribution with the variance depending on the mean-variance relationship of the various outcomes, and can contain in addition a correlation matrix $\mathbf{R}_i(\alpha)$ and an overdispersion parameter ϕ_i.

A general first-order approximate expression for the variance-covariance matrix of $\mathbf{y}_i$ is given as follows (Molenberghs and Verbeke, 2005; Faes et al., 2008)

$$var(\mathbf{y}_i) = \boldsymbol{\Delta}_i \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' \boldsymbol{\Delta}_i' + \mathbf{V}_i$$

with

$$\boldsymbol{\Delta}_i = \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \Bigg|_{\mathbf{b}_i = 0}$$

and

$$\mathbf{V}_i = \Phi_i^{\frac{1}{2}} \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \mathbf{A}_i^{\frac{1}{2}} \Phi_i^{\frac{1}{2}},$$

where $\mathbf{A}_i$ is a diagonal matrix containing the variances following from the generalized linear model specification of $\mathbf{y}_{ik} (k = 1,2)$ given the random effects $\mathbf{b}_i = \mathbf{0}$, that is, with diagonal elements $v(\mu_{ik} \mid \mathbf{b}_i = \mathbf{0})$. Likewise, Φ_i is a diagonal matrix with the overdispersion parameters along the diagonal. $\mathbf{R}_i(\alpha)$ is a correlation matrix.

When an exponential family specification is used for all components, with a canonical link, $\boldsymbol{\Delta}_i = \mathbf{A}_i$, the resulting GLMM has the variance-covariance matrix of $\mathbf{y}_i$ that takes the form (Molenberghs and Verbeke, 2005):

$$var(\mathbf{y}_i) = \boldsymbol{\Delta}_i \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' \boldsymbol{\Delta}_i' + \Phi_i^{\frac{1}{2}} \boldsymbol{\Delta}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \boldsymbol{\Delta}_i^{\frac{1}{2}} \Phi_i^{\frac{1}{2}}. \qquad (7.4)$$

When there are no residual correlations, that is, the matrix $\mathbf{R}_i(\alpha) = \mathbf{I}$, this results in a so-called conditional independence or a purely random effects model and the variance-covariance matrix of $\mathbf{y}_i$ in (7.4) reduces to

$$var(\mathbf{y}_i) = \boldsymbol{\Delta}_i \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' \boldsymbol{\Delta}_i' + \Phi_i^{\frac{1}{2}} \boldsymbol{\Delta}_i \Phi_i^{\frac{1}{2}}. \qquad (7.5)$$

When there are no random effects in a model (7.3), a marginal model is obtained and the variance-covariance of $\mathbf{y}_i$ in (7.4) takes the form

$$var(\mathbf{y}_i) = \mathbf{V}_i = \Phi_i^{\frac{1}{2}} \boldsymbol{\Delta}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \boldsymbol{\Delta}_i^{\frac{1}{2}} \Phi_i^{\frac{1}{2}}. \qquad (7.6)$$

For a sequence of continuous and binary outcomes, the marginal generalized linear model (MGLM), that is, a model with no random effects is of the form:

$$\begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_0 + \alpha_1 \mathbf{X}_i \\ \dfrac{\exp(\beta_0 + \beta_1 \mathbf{X}_i)}{1 + \exp(\beta_0 + \beta_1 \mathbf{X}_i)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix},$$

where the first component corresponds to the identity link function whilst the logit link is in the second component. Both the association between the two outcomes at each time point as well as the association emerging from the longitudinal structure of the data are incorporated in the residual error structure. This is done by specifying a correlation matrix $\mathbf{R}_i(\alpha)$ such as the unstructured correlation structure in order to allow each pair of outcomes to have its own correlation coefficient (Faes et al., 2008). While a marginal model with fully unstructured $2n_i \times 2n_i$ variance-covariance matrix is appealing because of its ease of interpretation, it can become computationally intensive, especially when the number of measurements per subject is large (Faes et al., 2008).

A conditional independence random-intercepts model for the continuous and binary outcomes with a general variance-covariance matrix $\mathbf{G}$ and residual correlation matrix $\mathbf{R}_i(\alpha) = \mathbf{I}$ can be written in the following form:

$$\begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_0 + \alpha_1 \mathbf{X}_i + \mathbf{b}_{i1} \\ \dfrac{\exp(\beta_0 + \beta_1 \mathbf{X}_i + \mathbf{b}_{i2})}{1 + \exp(\beta_0 + \beta_1 \mathbf{X}_i + \mathbf{b}_{i2})} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \qquad (7.7)$$

where $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$ are normally distributed as

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right\} \qquad (7.8)$$

and where ε_{i1} and ε_{i2} are independent. The random effects $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$ are used to accommodate the longitudinal structure in the data for the continuous and binary outcomes respectively. Furthermore, the correlation among the continuous and binary outcomes is induced by the incorporation of a correlation ρ among the two random effects. The variance of $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ can be derived from (7.5), in which

$$\mathbf{Z}_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{\Delta}_i = \begin{pmatrix} 1 & 0 \\ 0 & \upsilon_{i2} \end{pmatrix},$$

$$\mathbf{G} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}, \qquad \Phi_i = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

where

$\upsilon_{i2} = \mu_{i2}(\mathbf{b}_2 = \mathbf{0})[1 - \mu_{i2}(\mathbf{b}_2 = \mathbf{0})]$, with $\mu_{i2} = \exp(\beta_0 + \beta_1 \mathbf{X}_i)/[1 + \exp(\beta_0 + \beta_1 \mathbf{X}_i)]$. As a result, the approximate variance-covariance matrix of the two measurements for subject i at time point j is equal to

$$var(\mathbf{y}_i) = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2\upsilon_{i2} \\ \rho\tau_1\tau_2\upsilon_{i2} & \upsilon_{i2}^2\tau_2^2 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \upsilon_{i2} \end{pmatrix}$$

$$= \begin{pmatrix} \tau_1^2 + \sigma^2 & \rho\tau_1\tau_2\upsilon_{i2} \\ \rho\tau_1\tau_2\upsilon_{i2} & \upsilon_{i2}^2\tau_2^2 + \upsilon_{i2} \end{pmatrix}.$$

The correlation ρ_{12} among the continuous and binary outcomes is induced by the incorporation of a correlation ρ among the two random effects, and is approximately equal to

$$\rho_{12} = \frac{\rho\tau_1\tau_2\upsilon_{i2}}{\sqrt{\tau_1^2 + \sigma^2}\sqrt{\upsilon_{i2}^2\tau_2^2 + \upsilon_{i2}}}.$$

Under the conditional independence of the two random effects $(\rho \equiv 0)$, the approximate marginal correlation function ρ_{12} is equal to zero. That is with $\rho \equiv 0$, model (7.7) is equivalent to two separate GLMMs for the two outcomes. When $\rho \equiv 1$, this model reduces to the shared parameter model.

The third approach involves a random-intercepts model for each outcome with a correlated residual error structure to account for the association among the continuous and binary outcomes. With this model, a random effect is introduced at the level of the linear predictor after application of the link function, while the correlation is introduced at the level of ε_i. The model is also of the form (7.7), but where b_{i1} and b_{i2} are independent and normally distributed with mean of zero

and variance τ_1 and τ_2 respectively. The residual error is assumed to have a variance-covariance matrix

$$\mathbf{V}_i = \begin{pmatrix} \sigma^2 & \rho\sigma\sqrt{v_{i2}} \\ \rho\sigma\sqrt{v_{i2}} & v_{i2} \end{pmatrix}.$$

In this case ρ denotes the correlation between ε_1 and ε_2. The variance among the continuous and binary outcomes at time point j, for subject i, is approximately equal to

$$var(\mathbf{y}_i) = \begin{pmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 v_{i2}^2 \end{pmatrix} + \begin{pmatrix} \sigma^2 & \rho\sigma\sqrt{v_{i2}} \\ \rho\sigma\sqrt{v_{i2}} & v_{i2} \end{pmatrix}$$

$$= \begin{pmatrix} \tau_1^2 + \sigma^2 & \rho\sigma\sqrt{v_{i2}} \\ \rho\sigma\sqrt{v_{i2}} & \tau_2^2 v_{i2}^2 + v_{i2} \end{pmatrix}.$$

As a result, the approximate correlation among the two outcomes is

$$\rho_{12} = \frac{\rho\sigma\sqrt{v_{i2}}}{\sqrt{\tau_1^2 + \sigma^2}\sqrt{v_{i2}^2\tau_2^2 + v_{i2}}}.$$

Under the conditional independence $(\rho \equiv 0)$, which in this case implies that there is no correlation between ε_1 and ε_2, the approximate marginal correlation function ρ_{12} is equal to zero. Further, the marginal correlation, ρ_{12} would be reduced to ρ for the marginal generalized linear model, i.e. when there are no random effects.

Other generalizations with random effects or residual error structures are possible as well. Expression (7.4) can be used to obtained correlation structures for each setting or specific forms, such as the ones presented by expressions (7.5) and (7.6). It should be noted that although we have discussed in detail the correlation among the continuous and binary outcomes, the correlation structure among two outcomes of any nature can be derived in a similar way. For instance, when sequences of outcomes are count and binary, a generalized linear mixed model (GLMM), under conditional independence random-intercepts

model with a general variance-covariance matrix $\mathbf{G}$ and residual correlation matrix $\mathbf{R}_i(\alpha) = \mathbf{I}$ for subject i, can be written in the form:

$$\begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \end{pmatrix} = \begin{pmatrix} \exp(\alpha_0 + \alpha_1 \mathbf{X}_i + \mathbf{b}_{i1} \\ \dfrac{\exp(\beta_0 + \beta_1 \mathbf{X}_i + \mathbf{b}_{i2})}{1 + \exp(\beta_0 + \beta_1 \mathbf{X}_i + \mathbf{b}_{i2})} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix},$$

where the random effects $\mathbf{b}_{i1}$ and $\mathbf{b}_{i2}$ are normally distributed as in (7.8), and ε_{i1} and ε_{i2} are independent. It is assumed that $\upsilon_{i1} = \mu_{i1}(\mathbf{b}_{i1} = \mathbf{0})]$ and $\upsilon_{i2} = \mu_{i2}(\mathbf{b}_{i2} = \mathbf{0})[1 - \mu_{i2}(\mathbf{b}_{i2} = \mathbf{0})]$. The approximate variance-covariance matrix of the two binary measurements for subject i at time point j is equal to

$$var(y_{ij}) = \begin{pmatrix} \upsilon_{i1}^2 \tau_1^2 + \upsilon_{i1} & \rho \tau_1 \tau_2 \upsilon_{i1} \upsilon_{i2} \\ \rho \tau_1 \tau_2 \upsilon_{1j} \upsilon_{2j} & \upsilon_{i2}^2 \tau_2^2 + \upsilon_{2i} \end{pmatrix},$$

and the approximate correlation among the two outcomes is approximately

$$\rho_{12} = \frac{\rho \tau_1 \tau_2 \upsilon_{i1} \upsilon_{i2}}{\sqrt{\upsilon_{i1}^2 \tau_1^2 + \upsilon_{i1}} \sqrt{\upsilon_{i2}^2 \tau_2^2 + \upsilon_{i2}}}.$$

The parameters of these joint models can be estimated using the numerical approximation methods discussed in Chapter 5, Section 5.4. These include approximation to the integral using Gaussian quadratures or Laplace approximation. The other estimation method is based on approximation of the data using the pseudo-likelihood in which pseudo data are created based on a linearization of the mean. More specifically, the pseudo-likelihood approach can be used to estimate parameters in marginal models and random effects with or without serial correlation, whilst quadratures or Laplace approximations can only estimate parameters in the conditional independent random effects models. The SAS procedure GLIMMIX can be used for parameter estimation using a pseudo-likelihood approach whereas the NLIMIXED procedure can be used for parameter estimation using Gaussian quadratures or Laplace approximation.

7.4 Evaluation of HAART adherence data

In the adherence study, the patient's adherence status (adherent/not adherent) has been modeled as a binary variable that follows the Bernoulli distribution. Before we jointly model adherence status and duration (days between successive visits), we have to establish the theoretical framework that will inform the choice of the appropriate distribution of duration variable.

Patients in the CAT programme are normally expected to make clinic visits on a monthly (i.e 28 days) basis to collect their medication, undergo a physical examination review, etc. That is, the expected time interval between two successive clinic visits for patients is roughly 28 days. However, at times the patients are unable to keep their scheduled clinic appointments for various reasons/circumstances; they either make an early visit (less than 28 days) or a late visit (more than 28 days) to the clinic. Since such visits are arbitrary, they are considered to be random, which then changes the time interval between successive visits from planned ones to random occurrences. It is therefore useful to determine the probabilistic model which can be used to describe the occurrences of these unpredictable events (clinic visits). Random occurrences of events in time are often modeled as a Poisson process (Birnbaum, 1954; Kingman, 1993; Skogvoll and Lindqvist, 1999). A Poisson process refers to a continuous–time counting process $\{N(t), t \geq 0\}$ that possesses the following properties (Kingman, 1993):

i) $N(0) = 0$

ii) the number of events in disjoint intervals are independent ('independent increments') and the number of events in any given interval depends only on the length of that interval ('stationary increments');

iii) events occur one at a time, that is, there are no simultaneous occurrences.

Consequently, a Poisson process can be characterized in two simple alternative and equivalent ways (Birnbaum, 1954)

a) The 'waiting times' u between successive events are independently distributed with the exponential density function

$$g(u) = (1/\theta)e^{-u/\theta} \qquad \text{for } u \geq 0.$$

Here θ is the mean of the waiting times:

$$E(u) = \theta .$$

b) Let x be the number of events observed, then the increments $y = x(t_2) - x(t_1)$ of $x(t)$ on any interval of length $d = t_2 - t_1$ has the Poisson distribution

$$p(y) = e^{-d\lambda} \frac{(d\lambda)^y}{y!}, \qquad y = 0,1,\ldots$$

and the increments of $x(t)$ on non-overlapping intervals are independent. Here $d\lambda$ is the mean increment on an interval of length d. Thus λ is the mean rate of occurrences, and $\lambda = 1/\theta$. Similar formulation in a disease modeling structure has also been used by Ramroop (2008).

We can therefore model the time interval in days (duration) between successive clinic visits using two approaches; the 'waiting time' approach and the 'number of occurrences' approach. The 'waiting time' approach treats duration as the time in days between two successive clinic visits, which is a continuous variable that follows an exponential distribution. The 'number of occurrences' approach is used if we count the number of days to the next clinic visit, which is a discrete variable that follows a Poisson distribution. For purposes of this study, the 'waiting time' approach is more relevant and is therefore adopted.

For fitting of joint models with a binary and an exponential distribution, two possible models were considered in each case, these were marginal models as well as conditional independence random-effects models. Furthermore, the linear predictor used for all the fitted models consists of the same variables that have been identified in the GEE model.

Recall the linear predictor in the GEE model for the adherence data (Chapter 4, section 4.6)

$$\begin{aligned}\mathbf{x}_{ij}'\beta = \beta_0 &+ \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\ &+ \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\ &+ \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\ &+ \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} time_i gender_i \\ &+ \beta_{22} time_i site_i + \beta_{23} time_i reason\_no_i + \beta_{24} time_i reason\_risk_i + \beta_{25} age_i educ\_no_i \\ &+ \beta_{26} age_i educ\_p_i + \beta_{27} age_i gender_i \ .\end{aligned}$$

With a view that the duration outcome variable follows an exponential distribution, a preliminary exploratory analysis for this outcome variable was explored by fitting a generalized linear model with all the predictor variables using SAS PROC GENMOD. The SCALE parameter used in PROC GENMOD is the inverse of the gamma dispersion parameter, and is commonly referred to as the gamma *index parameter*. A value of 1 for the index parameter corresponds to the exponential distribution (SAS Institute Inc., 2004). With our data, the estimated value of the scale parameter was 10.2381 with a 95% profile likelihood confidence interval of (9.8843, 10.6004)**,** which did not contain 1. The hypothesis of an exponential distribution for the data was, therefore, rejected at 0.05 level and hence the duration outcome was considered to follow a gamma distribution in subsequent analyses.

The marginal model

We fitted the joint marginal model for both the binary (adherence status) and Gamma (duration) outcomes using logit and log links respectively. The linearization estimation method (discussed in Chapter 5, Section 5.4) was used as an approximation method. The SAS procedure GLIMMIX was used to fit the data; it allows us to jointly model outcomes with different distributions and/or different link functions. We first fitted a joint marginal model with a fully unstructured variance-covariance matrix in order to allow each pair of outcomes to have its own correlation coefficient. However, due to the computational

complexity, this model did not converge. Alternatively, because our main interest is in assessing the joint effect of the predictor variables, we adopted the GEE approach, where the association between the two outcomes at each time point as well as the association emerging from the longitudinal structure of the data are treated as a nuisance that has to be accounted for. The independence and AR-1 working assumptions were considered, AR-1 being the chosen structure in the GEE model, Chapter 4. Comparing model-based and empirically corrected standard errors, there was clear difference in the case of independence working assumption, but less so in the AR-1 case. Thus a joint model with the AR-1 working structure was chosen.

The results are presented in Table 7.1 and they confirm the results from the previous chapters. That is, even after accounting for the interval between successive visits (duration), cell phone ownership and living with a partner enhanced HAART adherence (Table 7.1). Moreover, the 2-way interaction terms, namely, time*gender, time*site, time*reported reason, age*gender and age*education are still significantly associated with HAART adherence (Table 7.1). In fact, the results indicate that the interaction between time and treatment site has a joint effect on HAART adherence and duration between successive visits. The rate at which optimal adherence increased over time in the rural site was higher compared to the urban site. At the same time however, the interval between successive visits was higher as the number of follow-up visits increased in the rural than in the urban treatment site.

Table 7. 1: Parameter estimates and their corresponding empirical standard errors of a joint marginal model for the adherence and duration outcomes with AR(1) working covariance structure

Effect	adherence Outcome			duration outcome		
	Estimates	**Std errors**	**p-value**	**Estimates**	**Std errors**	**p-value**
Intercept	0.923	0.393	0.019	3.171	0.055	<0.001
Gender (ref=male)						
female	-1.164	0.334	0.0005	0.020	0.044	0.649
Education (ref=sec & higher)						
No schooling	1.728	0.524	0.001	0.082	0.072	0.251
Primary	0.107	0.447	0.810	-0.013	0.064	0.845
Treatment site (ref=rural)						
Urban	1.200	0.143	<0.001	-0.004	0.017	0.817
Income (ref=not source)						
Source of income	-0.008	0.078	0.921	0.013	0.010	0.209
Access to tapwater (ref=no)						
Yes	0.120	0.118	0.309	0.016	0.107	0.948
Househld with electricity (ref=no)						
Yes	-0.007	0.107	0.948	0.020	0.017	0.238
Cell phone ownership (ref=no)						
Yes	0.238	0.070	0.001	0.005	0.009	0.586
WHO staging (ref=stage4)						
Stage1	-0.336	0.180	0.061	0.025	0.024	0.310
Stage2	-0.247	0.161	0.125	0.030	0.022	0.174
Stage 3	-0.117	0.148	0.426	0.023	0.0210	0.278
Living with partner (ref=no)						
Yes	0.298	0.080	0.002	0.0003	0.010	0.975
Reason for testing (ref=unwell)						
No specific reason	-0.040	0.156	0.796	0.004	0.023	0.852
Exposed to the risk	-0.444	0.175	0.011	-0.016	0.019	0.396
Baseline adherence (ref=no)						
yes	-0.153	0.085	0.070	0.012	0.011	0.251
Time (visit)	0.102	0.018	<0.001	0.013	0.002	<0.001
Age	-0.011	0.008	0.161	0.001	0.001	0.466
Baseline CD4+	-0.0004	0.001	0.438	0.00004	0.0001	0.595
Baseline weight	-0.001	0.006	0.795	-0.001	0.001	0.089
Weight at follow-up visits	-0.002	0.005	0.665	0.002	0.007	0.450
Time*gender (ref=male)						
female	0.071	0.016	<0.0001	0.001	0.002	0.760
Time*site (ref=rural)						
Urban	-0.061	0.016	0.0002	-0.007	0.002	0.0003
Time*reason (ref=unwell)						
No specific reason	-0.011	0.018	0.560	0.0001	0.002	0.9652
Exposed to the risk	0.045	0.021	0.032	0.0010	0.002	0.632
Age*gender						
Female	0.023	0.009	0.008	-0.0001	0.001	0.914
Age*education						
No schooling	-0.036	0.014	0.009	-0.002	0.002	0.209
Primary	-0.003	0.012	0.792	-0.0004	0.002	0.798

The conditional independence random-effects model

The model considered was as follows:

y_{i1j} - j th adherence status on the i th patient,

y_{i2j} - j th duration measurement on the i th patient,

$y_{i1j} \setminus b_{i1}$ ~indep. Binary with mean μ_{i1j} and variance $\mu_{i1j}(1-\mu_{i1j})$,

$y_{i2j} \setminus b_{i2}$ ~indep. Gamma with mean μ_{i2j} and variance μ_{i2j}^2 / ν,

$logit(\mu_{i1j}) = \mathbf{x}'_{i1j}\beta + b_{i1}$,

$ln(\mu_{i2j}) = \mathbf{x}'_{i2j}\beta + b_{i2}$,

$$\mathbf{b}_i = (b_{i1}, b_{i2})' \sim MVN(\mathbf{0}, \mathbf{G}), \mathbf{G} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

This model shows how the GLMM approach can be extended for multivariate longitudinal data by assuming separate random effects for each outcome variable and then combining them by imposing a joint multivariate distribution on the random effects.

The data were analyzed with SAS PROC NLMIXED using the general log-likelihood option with Gaussian quadrature. The NLMIXED procedure using the general log-likelihood function allows one to impose a joint multivariate distribution on the random effects from two separate models. The results from fitting a joint model for the two response variables with uncorrelated random intercepts using the GLIMMIX procedure were used as initial parameter estimates. Gaussian quadrature with 20 quadrature points was used and it took roughly 6 hours to converge. The number of quadrature points was selected among several possibilities (10, 15, 20, 25) based on the stability of parameter estimates and standard errors. The results presented in Table 7.2 were the same within the required precision with a larger number of quadrature points, i.e. parameter estimates seemed to be stable from 15 quadrature points upwards.

Table 7. 2: Parameter estimates and standard errors of the joint model for the conditional independence random intercepts model with the adherence and duration outcomes

Effect	adherence Outcome			duration outcome		
	Estimates	**Std errors**	**p-value**	**Estimates**	**Std errors**	**p-value**
Intercept	1.590	0.299	<0.0001	9.066	0.039	<0.0001
Gender (ref=male)						
female	-1.471	0.516	0.007	-0.032	0.068	0.633
Education (ref=sec & higher)						
No schooling	1.664	0.706	0.019	0.099	0.093	0.290
Primary	-0.126	0.686	0.855	-0.011	0.089	0.897
Treatment site (ref=rural)						
Urban	1.222	0.181	<0.0001	0.006	0.023	0.811
Income (ref=not source)						
Source of income	-0.028	0.113	0.801	-0.015	0.014	0.299
Access to tapwater (ref=no)						
Yes	0.146	0.177	0.409	0.015	0.024	0.544
Househld with electricity (ref=no)						
Yes	0.070	0.153	0.648	0.013	0.020	0.526
Cell phone ownership (ref=no)						
Yes	0.244	0.102	0.017	0.004	0.013	0.772
WHO staging (ref=stage4)						
Stage1	-0.487	0.272	0.073	0.033	0.035	0.343
Stage2	-0.361	0.242	0.137	0.026	0.030	0.383
Stage 3	-0.298	0.225	0.187	0.023	0.028	0.408
Living with partner (ref=no)						
Yes	0.556	0.280	0.048	0.859	0.039	<0.0001
Reason for testing (ref=unwell)						
No specific reason	-0.090	0.194	0.641	-0.002	0.025	0.926
Exposed to the risk	-0.428	0.215	0.0476	-0.018	0.027	0.518
Baseline adherence (ref=no)						
yes	-0.113	0.121	0.352	0.026	0.016	0.096
Time (visit)	0.119	0.021	<0.0001	0.013	0.002	<0.0001
Age	-0.0197	0.013	0.142	0.001	0.002	0.779
Baseline CD4+	-0.001	0.001	0.326	0.0004	0.001	0.965
Baseline weight	-0.003	0.007	0.728	-0.001	0.001	0.191
Weight at follow-up visits	-0.003	0.006	0.618	0.001	0.001	0.191
Time*gender (ref=male)						
female	0.056	0.019	0.004	0.003	0.002	0.146
Time*site (ref=rural)						
urban	-0.0486	0.018	0.008	-0.008	0.002	0.0002
Time*reason (ref=unwell)						
No specific reason	-0.001	0.021	0.963	0.001	0.002	0.648
Exposed to the risk	0.0438	0.022	0.047	0.003	0.003	0.217
Age*gender						
Female	0.035	0.014	0.015	0.001	0.002	0.714
Age*education						
No schooling	-0.035	0.0175	0.047	-0.003	0.002	0.249
Primary	0.001	0.018	0.923	-0.001	0.002	0.824

Variance Components			
	estimate	**Standard error**	**Pvalue**
Var. R.I (adherence)	0.261	0.062	<0.0001
Var. R.I (duration)	0.005	0.001	<0.0001
Correlation between the R.I.	-0.684	0.139	<0.0001

The results show that cell phone ownership and living with a partner are significantly associated with HAART adherence. In addition, all the 2-way interaction terms (time*gender, time*site, time*reason to test, age*gender and age*education) were also significantly associated with HAART adherence. Even after accounting for duration between successive visits, these results reaffirm the results that have been obtained in the GLMM (Chapter 5) regarding the fixed effects.

The random effects for the two outcomes are significantly negatively associated (Table 7.2). This translates into a negative correlation between HAART adherence and duration. This means that increasing the number of days between clinic visits tends to decrease the chances of being adherent to medication. That is, the longer the duration between successive visits, adherence to medication is likely to be compromised. We are however aware that at times, the conditional independence assumption might be too restrictive and tests for checking the validity of this assumption are not well established in the statistical literature. We therefore attempted to relax the conditional independence assumption by re-fitting the joint random intercepts model but now allowing for correlated errors; the model failed to converge. Gueorguieva (2001) introduced conditional dependence by including one response in the linear predictor for the other response. In order to validate the observed correlation between the two outcomes emerging from the association of the random intercepts, we adopted Gueorguieva's (2001) approach. We fitted a generalized linear mixed model with HAART adherence as the outcome and included duration in the linear predictor. The results are presented in Table 7.3 and they show that HAART adherence is negatively associated with duration (p-value=<0.0001) after controlling for the other variables. That is, overall, if the number of days is increased between successive visits, the patients are less likely to adhere to medication.

Table 7. 3: Parameter estimates and standard errors of the HAART adherence outcome that include duration in the linear predictor

Effect			
	Estimates	**Std errors**	**p-value**
Intercept	3.556	0.551	<0.0001
Gender (ref=male)			
female	-1.491	0.535	0.006
Education (ref=sec & higher)			
No schooling	1.964	0.734	0.008
Primary	-0.202	0.712	0.777
Treatment site (ref=rural)			
Urban	-1.222	0.185	<0.0001
Income (ref=not source)			
Source of income	-0.046	0.117	0.694
Access to tapwater (ref=no)			
Yes	0.103	0.184	0.414
Househld with electricity (ref=no)			
Yes	0.094	0.159	0.552
Cell phone ownership (ref=no)			
Yes	0.224	0.105	0.034
WHO staging (ref=stage4)			
Stage1	-0.447	0.280	0.112
Stage2	-0.342	0.250	0.171
Stage 3	-0.262	0.232	0.258
Living with partner (ref=no)			
Yes	1.834	0.551	0.001
Reason for testing (ref=unwell)			
No specific reason	-0.1079	0.197	0.584
Exposed to the risk	-0.540	0.221	0.015
Baseline adherence (ref=no)			
yes	-0.092	0.125	0.463
Time (visit)	-0.042	0.044	0.337
Age	-0.052	0.027	0.054
Baseline CD4+	-0.001	0.001	0.282
Baseline weight	-0.007	0.008	0.358
Weight at follow-up visits	0.002	0.007	0.749
Duration	-0.021	0.004	<0.0001
Time*gender (ref=male)			
female	0.061	0.019	0.002
Time*site (ref=rural)			
urban	0.052	0.019	0.006
Time*reason (ref=unwell)			
No specific reason	0.003	0.021	0.876
Exposed to the risk	0.068	0.025	0.007
Age*gender			
Female	0.034	0.015	0.020
Age*education			
No schooling	-0.042	0.019	0.029
Primary	0.004	0.019	0.832

7.5 Summary

Joint modeling provides us with the ability to answer multivariate research questions, in addition to the possible gains in efficiency of parameter estimates. The results from fitting a joint model of adherence and time interval indicate that optimal HAART adherence is negatively associated with time interval between successive visits over time. That is, with shorter intervals between one visit to the next, the patients tend to maintain optimal adherence. Notwithstanding the negative association between optimal adherence and the time interval between successive visits, the results further revealed that if the patient is well adapted to the treatment program with a record of sustained optimal adherence, the gap between successive clinic visits might be longer depending on the treatment site the patient is enrolled with. The results reaffirm the significant determinants of optimal adherence over time reported in the previous chapters. That is, after accounting for the time interval measured in days between successive visits, cell phone ownership, living with a partner, and two-way interaction terms that involved time with gender, treatment site and reason for taking an HIV test, as well as age with gender and educational level were still associated with optimal adherence. Among other observations, an increasing linear effect on HAART adherence was noted, though it differed by gender, treatment site and reported reason for taking an HIV test. In the next chapter, to allow for more a flexible trajectory of the observed data, a semi-parametric approach was used to model the effect of time non-parametrically.

Chapter 8

Generalized additive mixed models

8.1 Introduction

In the previous chapters, we reviewed and fitted adherence data using parametric regression methods for longitudinal data. Specifically, the generalized estimating equations (GEEs), generalized linear mixed models (GLMMs) and transition models were used to fit the adherence data. These models provide a powerful tool for modeling the relationship between a response variable and covariates in longitudinal studies. Although these parametric mean models enjoy simplicity, they suffer from inflexibility in modeling complicated relationships between the response and covariates in various applications of practical longitudinal data. That is, for many applications, parametric models may be too restrictive or limited, and sometimes unavailable. This limitation has placed a strong demand in recent years on developing nonparametric regression methods for longitudinal data, where flexible functional forms can be estimated from the data to capture possibly complicated relationships between longitudinal outcomes and covariates (Lin and Carroll, 2008). The basic idea of the nonparametric approaches is to let the data determine the most suitable form of the functions. Wu and Zhang (2006) argue that nonparametric and parametric regression methods should not be regarded as competitors, instead they complement each other. At times, nonparametric techniques can be used to validate or suggest a parametric model. A combination of both nonparametric and parametric methods is more powerful than any single method in many applications.

There exist many nonparametric regression and smoothing methods for independent data in the literature. The most widely used methods include kernels and splines. A survey of these methods can be found in Hardle (1990),

Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996) and Ruppert, Wand and Carroll (2003) among others. Also the generalized additive models of Hastie and Tibshirani (1990) are widely used for nonparametric regression with independent data. However, very little effort has been made to develop nonparametric regression methods for longitudinal data analysis until recent years. The presence of the within-subject correlation among the repeated measures over time presents a major challenge in developing nonparametric techniques for longitudinal data analysis (Lin and Carroll, 2008). As a result, limited work has been done on nonparametric regression when the data are correlated, with the practical lagging behind even more so due to the lack of readily accessible statistical software.

The limited work that has been done on nonparametric regression methods for longitudinal data has been restricted to normally distributed outcomes, and their connection to mixed models (Zhang et al., 1998; Verbyla et al., 1999). For non-Gaussian longitudinal data, the developments in nonparametric regression methods have been mostly restricted to single covariate models, and these include local polynomial kernel GEEs (Lin and Carroll; 2000) and spline methods (Wang, 1998; Liang, Wu and Carroll, 2003). Extensions of single-covariate nonparametric models to multiple covariates pose a problem referred to as the 'curse of dimensionality', which means that the performance of nonparametric smoothing techniques deteriorates as the dimensionality increases (Hastie and Tibshirani, 1990; Fan, Heckman and Wand, 1995). One way of overcoming this problem is by using generalized additive models (Hastie and Tibshirani, 1990). To this end, for non-Gaussian longitudinal data, Berhane and Tibshirani (1998) extended generalized additive models to generalized estimating equations (Liang and Zeger, 1986). Moreover, modeling longitudinal data with non-Gaussian outcomes non-parametrically within the mixed effects model framework, Lin and Zhang (1999) proposed generalized additive mixed models (GAMMs), which are additive extensions of GLMMs in the spirit of Hastie (1990).

The aim of this chapter is to therefore review GAMMs and then fit them to adherence data. Specifically, we seek to model the effect of time on the patients' adherence status non-parametrically. In order to develop a better appreciation of GAMMS, a brief overview of nonparametric regression methods using generalized additive models (GAMs) for independent data is provided. This chapter is organized as follows. An overview of generalized additive models for independent data is presented in Sections 8.2. Section 8.3 reviews the generalized additive mixed models (GAMMs) for longitudinal data. The GAMM model is fitted to adherence data in Section 8.4.

8.2 Generalized Additive Models (GAMs)

Generalized additive models are presented by first introducing an additive model. An additive model is a generalization of the linear regression model and is defined by (Hastie and Tibshirani, 1990)

$$\mathbf{y} = \alpha + \sum_{j=1}^{p} f_j(\mathbf{x}_j) + \varepsilon, \qquad (8.1)$$

where $\mathbf{y}$ is the response variable, $\mathbf{x}_j$ are covariates and and the ε are i.i.d $N(0,\sigma^2)$. The f_j are arbitrary univariate and smooth functions, one for each predictor. Moreover, even though it is convenient to specify functions in additive models as univariate and smooth, functions of two or more dimensions as well as categorical variable terms and their interactions with continuous scale variables can still be specified. It is also assumed that $E\{f_j(\mathbf{x}_j)\} = 0$, since otherwise there will be free constants in each of the functions. An additive model retains an important interpretive feature of the linear model; that the variation of the fitted response surface holding all but one predictor fixed does not depend on the values of the other predictor. This means that once the additive model is fitted to the data, we can plot p coordinate functions separately to examine the roles of the predictors in modeling the response.

With additive models, it is necessary to represent the smooth functions in some way and to choose how smooth they should be. Hastie and Tibshirani (1990) suggest representing additive models using spline-like penalized regression smoothers. Spline smoothing can be used to represent smooth functions in such a way that model (8.1) becomes a linear model. This is done by specifying a set of basis functions, b_{ji} for each function so that the smooth function can be represented as (Wood, 2006)

$$f_j(x_j) = \sum_{j=1}^{q_j} \beta_{ji} b_{ji}(x_j), \tag{8.2}$$

where x_j maybe a vector quantity and β_{ji} are coefficients of the smooth, which will need to be estimated as part of model fitting. Examples of penalized regression smoothers include natural cubic splines, cubic smoothing splines, thin plate regression splines and tensor product bases (for more details, see Hastie and Tibshirani, 1990; Wood, 2006).

Controlling the roughness of the estimated function is a variant of the bias-variance tradeoff problem. Here, using too few basis functions may not allow the fitted curve to accurately represent the shape of the function, leading to biased estimation, while using too many will result in an overly close interpolation of the data. One way to control the model's smoothness is by adding a smoothness penalty to the least squares fitting objective. That is, rather than fitting a model by minimizing (Wood, 2006)

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta),$$

it could be fitted by minimizing

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \int_0^1 [f''(x)]^2 dx,$$

where the integrated square of second derivative penalizes models that are too smooth. The tradeoff between model fit and model smoothness is controlled by the smoothing parameter, λ. When $\lambda \to \infty$, this leads to a straight estimate of f, while when $\lambda = 0$, the result is an un-penalized regression spline estimate.

Because f is linear in the parameters, β_j, the penalty can always be written as a quadratic form in β,

$$\lambda\int_0^1 [f''(x)]^2 dx = \beta'\mathbf{S}\beta,$$

where $\mathbf{S}$ is a matrix of known coefficients. Therefore, the penalized regression spline fitting problem is to minimize

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\mathbf{S}\beta$$

with respect to β. Thus the penalized least squares estimator of β given λ is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}'\mathbf{y}.$$

The degree of smoothness for the model is obtained by estimating the smooth parameter, λ. Methods that include *ordinary cross validation* (OCV) and g*eneralized cross validation* (GCV) are used to estimate λ.

A generalized additive model is a generalized linear model (GLM) with a linear predictor involving a sum of smooth functions of covariates (Hastie and Tibshirani, 1990). The model has the following general structure

$$g(\mu_i) = \mathbf{x}_i^*\theta + \sum_{j=1}^{p} f_j(x_j), \qquad (8.3)$$

where $\mu_i = E(\mathbf{y}_i)$ and $\mathbf{y}_i$ is the response variable from the exponential family of distributions, g is a known, monotonic, twice differential link function, $\mathbf{x}_i^*$ is the ith row of a model matrix for any parametric components, θ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_j. The GAMs' representation is illustrated by using penalized regression smoothers based on splines in (8.2) and the derivation of the model parameters follows that of Wood (2006).

Whereas the additive model was estimated with penalized least squares, the GAM is estimated by penalized likelihood maximization. Furthermore, GAM can easily be represented as a GLM through re-parameterization of smooth terms

and as such the ordinary GLM likelihood, $l(\beta)$ can be used to estimate the model coefficients. There is however a good chance of substantially over-fitting the model with ordinary likelihood maximization and to guard against over-fitting, GAMs are estimated by penalized likelihood maximization, where the penalties are designed to suppress overly smooth estimates of the f_j terms. The penalized likelihood can be defined as

$$l_p(\beta) = l(\beta) - \frac{1}{2}\beta' \mathbf{S} \beta ,$$

where $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. λ_j are smoothing parameters and $\mathbf{S}_j$ is a matrix of known coefficients.

To fit the model in practice, the following penalized iteratively re-weighted least squares (P-IRLS) scheme is iterated until convergence (Wood, 2006):

1. Given the current linear predictor estimate, $\eta^{[k]}$, and corresponding estimated mean response vector, $\mu^{[k]}$, calculate:

   $$w_i \propto \frac{1}{V(\mu_i^{[k]}) g'(\mu_i^{[k]})^2} \text{ and } z_i = g'(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \mathbf{x}_i \beta^{[k]},$$

   where $var(\mathbf{y}_i) = V(\mu^{[k]})\phi$ and $\mathbf{x}_i$ is the ith row of $\mathbf{X}$.

2. Minimize

   $$\left(\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\right)' \left(\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\right) + \beta' \mathbf{S} \beta$$

   with respect to β to obtain $\beta^{[k+1]}$, and hence $\eta^{[k+1]} = \mathbf{X}\beta^{[k+1]}$. $\mathbf{W} = diag[\{\phi v(\mu_i) g'(\mu_i)^2\}^{-1}]$.

The generalized cross validation (GCV) method can be used to estimate λ in the generalized case (Wood, 2006). The GCV score for smoothing parameter selection can be obtained from the GAM fitting objective

$$\left(\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\right)' \left(\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\right) + \sum_{j=1}^{m} \lambda_j \beta' \mathbf{S}_j \beta ,$$

as

$$v_g = \frac{n\left(\sqrt{\mathbf{W}}(\mathbf{z}-\mathbf{X}\beta)\right)'\left(\sqrt{\mathbf{W}}(\mathbf{z}-\mathbf{X}\beta)\right)}{[n-tr(\mathbf{A})]^2},$$

where $\mathbf{A}$ is an influence matrix (Wood, 2006).

There are two possible numerical strategies for estimating smoothing parameters using v_g minimization:

- v_g can be minimized and smoothing parameters selected for each working penalized linear model of P-IRLS iteration. This is known as *performance iteration* and its algorithm details can be obtained in Gu and Wahba (1991) and Wood (2000, 2004).

- v_g can be minimized directly, which means that the P-IRLS scheme must be iterated to convergence for each trial set of smoothing parameters. This is usually referred to as *outer iteration* and the algorithm details can be found in Wood (2006).

8.3 Generalized additive mixed models

Generalized additive mixed models (Lin and Zhang, 1999) are an extension of generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) to allow the parametric fixed effects to be modeled non-parametrically using additive smooth functions in a similar spirit to Hastie and Tibshirani (1990). Suppose, there is an outcome variable $\mathbf{y}_i$, $i = 1,\ldots,n$ and p covariates $\mathbf{x}_i = (1, x_{i1},\ldots,x_{ip})'$ associated with fixed effects and a $q \times 1$ vector of covariates $\mathbf{z}_i$ associated with random effects. Given a $q \times 1$ vector of $\mathbf{b}$ of random effects, the observations $\mathbf{y}_i$ are assumed to be conditionally independent with means $E(\mathbf{y}_i \mid \mathbf{b}) = \mu_i$ and variances $var(\mathbf{y}_i \mid \mathbf{b}) = \phi\upsilon(\mu_i)$, where $\upsilon(\cdot)$ is a specified variance function and ϕ is a scale parameter. Then a generalized additive mixed model is given by (Lin and Zhang, 1999)

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \mathbf{z}_i\mathbf{b}, \qquad (8.4)$$

where $g(\cdot)$ is a monotonic differential link function, $f_j(\cdot)$ is a centred twice-differentiable smooth function, the random effects are assumed to be distributed as $N\{\mathbf{0}, \mathbf{G}(\gamma)\}$ and γ is a $c \times 1$ vector of variance components. The additive nonparametric functions are used to model covariate effects and random effects are used to model correlation between observations.

The integrated log quasi-likelihood of $\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\}$ is

$$\exp[l\{\mathbf{y}; \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\}] \propto |\mathbf{G}|^{-1/2} \int \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(\mathbf{y}; \mu_i) - \frac{1}{2}\mathbf{b}'\mathbf{G}^{-1}\mathbf{b}\right\} d\mathbf{b} \qquad (8.5)$$

where

$$\mathbf{y}_i = (y_1, \ldots, y_n)' \text{ and } d_i(\mathbf{y}; \mu) \propto -2\int_{\mathbf{y}_i}^{\mu_i} (\mathbf{y}_i - u)/v(u)du$$

defines the conditional deviance function of $\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot)\}$ given $\mathbf{b}$.

Statistical inference for GAMM involves inference on the nonparametric functions $f_j(\cdot)$, which requires the estimation of smoothing parameters, λ, and inference on variance components γ. Because of the close connection between the smoothing spline estimators and linear mixed models (Wang, 1998; Zhang et al., 1998; Verbyla et al., 1999), we will first discuss the construction of these estimators, specifically the natural cubic smoothing spline estimators of $f_j(\cdot)$ when λ and γ are known. The derivation of the natural cubic smoothing spline follows that of Lin and Zhang (1999). Given the values of λ and γ, the natural cubic smoothing spline estimators of $f_j(\cdot)$ maximize the penalized log quasi-likelihood

$$l\{\mathbf{y}; \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\} - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \int_{s_j}^{t_j} f_j''(x)^2 dx = l\{\mathbf{y}_i; \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\} - \frac{1}{2}\sum_{j=1}^{p} \lambda_j f_j' \mathbf{S}_j f_j \qquad (8.6)$$

where (s_j, t_j) defines the range of the jth covariate and $\lambda = (\lambda_1, \ldots, \lambda_p)'$ is a vector of smoothing parameters and controls the trade-off between goodness of

fit and the smoothness of the estimated functions. Here $f_j(\cdot)$ is an $r_j \times 1$ unknown vector of the values of $f_j(\cdot)$ evaluated at the r_j ordered distinct values of the x_{ij} $(i = 1,\ldots,n)$ and $\mathbf{S}_j$ is the smoothing matrix (Green and Silverman, 1994).

The GAMM model in (8.4) can be re-written in matrix notation:

$$g(\mu_i) = \mathbf{1}\beta_0 + \mathbf{N}_1 f_1 + \ldots + \mathbf{N}_p f_p + \mathbf{Zb}\,, \qquad (8.7)$$

where $g(\mu_i) = \{g(\mu_1),\ldots,g(\mu_n)\}'$, $\mathbf{1}$ is an $n \times 1$ vector of ones, $\mathbf{N}_j$ is an $n \times r_j$ matrix such that the ith component of $\mathbf{N}_j f_j$ is $f_j(x_{ij})$ and $\mathbf{Z} = (Z_1,\ldots,Z_n)'$.

The evaluation of expression (8.5) requires numerical integration (except for the Gaussian outcomes) and it is often difficult to calculate full natural cubic smoothing spline estimators of the f_j by directly maximizing (8.6). Lin and Zhang (1999) proposed an approximation using a double penalized quasi-likelihood (DPQL) within the framework of generalized linear mixed models (GLMM).

Double penalized quasi-likelihood estimation of nonparametric functions

The estimation of nonparametric functions f_j can be obtained by representing GAMM using the GLMM framework using double penalized quasi-likelihood (DPQL). The derivation of DPQL follows that of Lin and Zhang (1999). Since f_j is a centred parameter vector, it can be re-parameterized in terms of β_j (scalar) and $\mathbf{a}_j((r_j - 2)\times 1)$ through a one-to-one transformation as

$$f_j = \mathbf{x}_j^* \beta_j + \mathbf{B}_j \mathbf{a}_j\,, \qquad (8.8)$$

where $\mathbf{x}_j^*$ is an $r_j \times 1$ vector containing the r_j centred distinct values of the x_{ij} $(i = 1,\ldots,n)$, and $\mathbf{B}_j = \mathbf{L}_j(\mathbf{L}_j'\mathbf{L}_j)^{-1}$ and $\mathbf{L}_j$ is an $r_j \times (r_j - 2)$ full rank matrix

satisfying $\mathbf{S}_j = \mathbf{L}_j \mathbf{L}_j'$ and $\mathbf{L}_j' \mathbf{x}_j^* = \mathbf{0}$. Using the identity $f_j' \mathbf{S}_j f_j = \mathbf{a}_j' \mathbf{a}_j$, the double penalized quasi-likelihood becomes

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(\mathbf{y};\mu_i) - \frac{1}{2}\mathbf{b}'\mathbf{G}^{-1}\mathbf{b} - \frac{1}{2}\mathbf{a}'\Lambda^{-1}\mathbf{a}, \qquad (8.9)$$

where $\mathbf{a} = (a_1', \ldots, a_p')'$ and $\Lambda = diag(\tau_1 \mathbf{I}, \ldots, \tau_p \mathbf{I})$ with $\tau_j = 1/\lambda_j$. A small value of $\tau = (\tau_1, \ldots, \tau_p)'$ corresponds to over-smoothing. Plugging (8.8) into (8.7), expression (8.9) suggests that given γ and τ, the DPQL estimators $\hat{f}_j$ can be obtained by fitting the following GLMM using Breslow and Clayton's (1993) penalized quasi-likelihood approach:

$$g(\mu) = \mathbf{X}\beta + \mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b}, \qquad (8.10)$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{N}_1 \mathbf{x}_1^*, \ldots, \mathbf{N}_p \mathbf{x}_p^*)$, $\mathbf{B} = (\mathbf{N}_1 \mathbf{B}_1, \ldots, \mathbf{N}_p \mathbf{B}_p)$, $\beta = (\beta_0, \ldots, \beta_p)'$ is a $(p+1)\times 1$ vector of regression coefficients and $\mathbf{a}$ and $\mathbf{b}$ are independent random effects with distributions $\mathbf{a} \sim N(\mathbf{0}, \Lambda)$ and $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$. The DPQL estimator $\hat{f}_j$ is calculated as $\hat{f}_j = \mathbf{x}_j^* \hat{\beta}_j + \mathbf{B}_j \hat{\mathbf{a}}_j$, which is a linear combination of the Breslow and Clayton (1993) penalized quasi-likelihood estimators of the fixed effect $\hat{\beta}_j$ and the random effects $\hat{\mathbf{a}}_j$ in the working GLMM in (8.10).

Maximization of expression (8.9) with respect to $(\beta, \mathbf{a}, \mathbf{b})$ can proceed by using the Fisher scoring algorithm to solve

$$\begin{pmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{B} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{B}'\mathbf{W}\mathbf{X} & \mathbf{B}'\mathbf{W}\mathbf{B} + \Lambda^{-1} & \mathbf{B}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{B} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{W}\mathbf{y} \\ \mathbf{B}'\mathbf{W}\mathbf{y} \\ \mathbf{Z}'\mathbf{W}\mathbf{y} \end{pmatrix} \qquad (8.11)$$

where $\mathbf{y}$ is a working vector defined as

$$\mathbf{y} = \beta_0 \mathbf{1} + \sum_{j=1}^{p} \mathbf{N}_j f_j + \mathbf{Z}\mathbf{b} + \Delta(\mathbf{y} - \mu)$$

and $\Delta = diag[g'(\mu_i)]$, $\mathbf{W} = diag[\{\phi v(\mu_i) g'(\mu_i)^2\}^{-1}]$. Expression (8.11) shows that it corresponds to the normal equation of best linear unbiased predictors (BLUPs) of β and $(\mathbf{a}, \mathbf{b})$ under the linear mixed model

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b} + \varepsilon, \tag{8.12}$$

where $\mathbf{a}$ and $\mathbf{b}$ are independent random effects with $\mathbf{a} \sim N(\mathbf{0}, \Lambda)$ and $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$ and $\varepsilon \sim N(\mathbf{0}, \mathbf{W}^{-1})$. This suggests that the DPQL estimators $\hat{f}_j$ and random effect estimators $\hat{\mathbf{b}}$ can be easily obtained using the BLUPs by iteratively fitting model (8.12) to the working vector $\mathbf{y}$.

To compute the covariance matrix of $\hat{f}_j$, it is more convenient to calculate β and $\mathbf{a}$ by using

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{B} \\ \mathbf{B}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{B}'\mathbf{R}^{-1}\mathbf{B} + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{a} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{B}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{8.13}$$

where $\mathbf{R} = \mathbf{W}^{-1} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$. Denoting by $\mathbf{H}$ the coefficient matrix on the left-hand side of equation (8.13) and $\mathbf{H}_0 = (\mathbf{X}, \mathbf{B})'\mathbf{R}^{-1}(\mathbf{X}, \mathbf{B})$, the approximate covariance matrix of $\hat{\beta}$ and $\hat{\mathbf{a}}$ is

$$cov(\hat{\beta}, \hat{\mathbf{a}}) = \mathbf{H}^{-1}\mathbf{H}_0\mathbf{H}^{-1}.$$

It follows that the approximate covariance matrix of $\hat{f}_j$ is $(\mathbf{x}_j^*, \mathbf{B}_j) cov(\hat{\beta}_j, \hat{\mathbf{a}}_j)(\mathbf{X}_j, \mathbf{B}_j)'$, where $cov(\hat{\beta}_j, \hat{\mathbf{a}}_j)$ can be easily obtained from the corresponding blocks of $\mathbf{H}^{-1}\mathbf{H}_0\mathbf{H}^{-1}$. Here we assume that the $\hat{f}_j(\cdot)$ are smooth functions in calculating the co-variances of the $\hat{f}_j$.

Estimation of smoothing parameters and variance components

In the previous section, we had assumed that the smoothing parameters λ and the variance components γ are known when estimation was made on the

nonparametric functions f_j. However, they are usually estimated from the data. Under the classical nonparametric regression model

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon, \tag{8.14}$$

where the ε are independent random errors following $N(0,\sigma^2)$, Wahba (1985) and Kohn, Ansley and Tharm (1991) proposed to estimate the smoothing parameter λ by maximizing a marginal likelihood. The marginal likelihood of $\tau = 1/\lambda$ is constructed by assuming that $f(\mathbf{X})$ has been prior specified in the form of (8.8) with $\mathbf{a} \sim N(\mathbf{0}, \tau\mathbf{I})$ and a flat prior for β and integrating out $\mathbf{a}$ and β as follows:

$$\exp\{l_M(\mathbf{y};\tau,\sigma^2)\} \propto \tau^{-1/2} \int \exp\left\{ l(\mathbf{y};\beta,\mathbf{a},\sigma^2) - \frac{1}{2\tau}\mathbf{a}'\mathbf{a} \right\} d\mathbf{a}\, d\beta \tag{8.15}$$

where $l(\mathbf{y};\beta,\mathbf{a},\sigma^2)$ is the log-likelihood (normal) of f under model (8.14). Speed (1991) and Thompson (1985) pointed out that the marginal likelihood (8.15) of τ is in fact the REML under the linear mixed model

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\beta_1 + \mathbf{B}\mathbf{a} + \varepsilon$$

where $\mathbf{a} \sim N(\mathbf{0}, \tau\mathbf{I})$ and $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{B}$ was defined earlier; τ is regarded as a variance component. Hence the marginal estimator of τ is a REML estimator. It has been shown that the maximum marginal likelihood estimator of τ has similar and often better performance compared with the GCV estimator in estimating the nonparametric function (Kohn et al., 1991).

Zhang et al., (1998) extended these result to estimate the smoothing parameter λ and variance component γ jointly using REML for longitudinal data with normally distributed outcomes and a nonparametric mean function. Their model can be written as

$$\mathbf{y} = f(\mathbf{X}) + \mathbf{Z}\mathbf{b} + \varepsilon \tag{8.16}$$

where $f(\mathbf{X})$ denotes the values of the nonparametric function $f(\cdot)$ evaluated at the design points of $\mathbf{X}(n\times 1)$, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G}(\gamma))$ and $\varepsilon \sim N(\mathbf{0}, \mathbf{V}(\gamma))$. If $f(\cdot)$ is

estimated using cubic smoothing spline (8.8), Zhang et al., (1998) rewrote model (8.16) as a linear mixed model

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\beta_1 + \mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b} + \varepsilon \tag{8.17}$$

where $\mathbf{a} \sim N(\mathbf{0}, \tau\mathbf{I})$ and distribution of $\mathbf{b}$ and ε are the same as those in model (8.16). They hence proposed to treat τ as an extra variance component in addition to γ in model (8.17) and to estimate τ and γ jointly by using REML. This REML corresponds to the marginal likelihood of (τ, γ) constructed by assuming that f takes the form of (8.8) with $\mathbf{a} \sim N(\mathbf{0}, \tau\mathbf{I})$ and a flat prior for β and integrating out $\mathbf{a}$ and β as follows:

$$\exp\{l_M(\mathbf{y};\tau,\gamma)\} \propto |\mathbf{G}|^{-1/2}\ \tau^{-1/2} \int \exp\left\{ l(\mathbf{y};\beta,\mathbf{a},\mathbf{b}) - \frac{1}{2}\mathbf{b}'\mathbf{G}^{-1}\mathbf{b} - \frac{1}{2\tau}\mathbf{a}'\mathbf{a} \right\} d\mathbf{b}\, d\mathbf{a}\, d\beta\,, \tag{8.18}$$

where $l(\mathbf{y};\beta,\mathbf{a},\mathbf{b}) = l(\mathbf{y};f,\mathbf{b})$ is the conditional log-likelihood (normal) of f given the random effects $\mathbf{b}$ under model (8.16). The marginal log-likelihood $l_M(\mathbf{y};\tau,\gamma)$ in (8.18) has a closed form.

Lin and Zhang (1999) proposed to extend the marginal likelihood approach to GAMM in (8.4) and to estimate τ and γ jointly by maximizing a marginal quasi-likelihood. Specifically, the GLMM representation of GAMM in (8.10) suggests that τ may be treated as extra variance components in addition to γ. Similar to the REML (8.18), the marginal quasi-likelihood of (τ, γ) can be constructed under the GAMM in (8.4) by assuming that f_j takes the form (8.8) with $\mathbf{a}_j \sim N(\mathbf{0}, \tau_j\mathbf{I})$ $(j = 1, \ldots, p)$ and integrating $\mathbf{a}_j$ and β out as follows:

$$\begin{aligned}
\exp\{l_M(\mathbf{y};\tau,\gamma)\} &\propto |\Lambda|^{-1/2} \int \exp\left\{ l(\mathbf{y};\beta,\mathbf{a},\gamma) - \frac{1}{2}\mathbf{a}'\Lambda^{-1}\mathbf{a} \right\} d\mathbf{a} d\beta \\
&\propto |\mathbf{G}|^{-1/2} |\Lambda|^{-1/2} \int \exp\left\{ \sum_{i=1}^{n} -\frac{1}{2\phi} d_i(\mathbf{y};\mu_i) - \frac{1}{2}\mathbf{b}'\mathbf{G}\mathbf{b} - \frac{1}{2}\mathbf{a}'\Lambda\mathbf{a} \right\} d\mathbf{b} d\mathbf{a} d\beta,
\end{aligned} \tag{8.19}$$

where $l(\mathbf{y};\beta,\mathbf{a},\gamma) = l(\mathbf{y};\beta_0, f_1, \ldots, f_p, \gamma)$ was defined in (8.5). Under the Gaussian nonparametric mixed model (8.16), the marginal quasi-likelihood reduces to the

REML (8.18). Unlike in the Gaussian situation, an evaluation of the marginal quasi-likelihood (8.19) for non-Gaussian outcomes is hampered by often intractable numerical integration. It is thus approximated using Laplace's method. Lin and Zhang (1999) showed that taking the quadratic expansion of the exponent of the integrand of (8.19) about its mode before integration and approximating the deviance statistic $d_i(\mathbf{y};\mu_i)$ by the Pearson's χ^2-statistic (Breslow and Clayton, 1993), the approximate marginal log-quasi-likelihood is given as

$$l_M(\mathbf{y};\tau,\gamma) \approx -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}), \qquad (8.20)$$

where $\mathbf{V} = \mathbf{B}\Lambda\mathbf{B}' + \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{W}^{-1}$. Expression (8.20) corresponds to the REML log-likelihood of the working vector $\mathbf{y}$ under the linear mixed model (8.12) with both $\mathbf{a}$ and $\mathbf{b}$ as random effects and τ and γ as variance components. It follows that τ and γ can be estimated by iteratively fitting model (8.12) using REML.

Specifically, Lin and Zhang (1999) showed that differentiating (8.20) with respect to $\vartheta = (\tau,\gamma)$, some calculation gives the estimating equations for τ and γ, which are the REML equations under the working linear mixed model (8.12)

$$-\frac{1}{2}tr\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\vartheta_k}\right) + \frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\beta})\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\vartheta_k}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}) = \mathbf{0},$$

which can be equivalently written in terms of the $\hat{f}_j$ as

$$-\frac{1}{2}tr(\mathbf{P}\mathbf{N}_j\mathbf{B}_j\mathbf{B}_j'\mathbf{N}_j') + \frac{1}{2}\left(\mathbf{y}-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}\mathbf{N}_k\hat{f}_k\right)'\mathbf{R}^{-1}\mathbf{N}_j\mathbf{B}_j\mathbf{B}_j'\mathbf{N}_j'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}\mathbf{N}_k\hat{f}_k\right) = 0$$

$$-\frac{1}{2}tr\left(\mathbf{P}\frac{\partial\mathbf{R}}{\partial\gamma_l}\right) + \frac{1}{2}\left(\mathbf{y}-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}\mathbf{N}_k\hat{f}_k\right)'\mathbf{R}^{-1}\frac{\partial\mathbf{R}}{\partial\gamma_i}\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{1}\hat{\beta}_0-\sum_{k=1}^{p}\mathbf{N}_k\hat{f}_k\right) = 0$$

where calculations of $\partial\mathbf{V}/\partial\vartheta_k$ and $\partial\mathbf{R}/\partial\gamma_j$ ignore the dependence of $\mathbf{W}$ and ϑ and

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{X},\mathbf{B})'\mathbf{R}^{-1}$$

is the projection matrix under the linear mixed model (8.12) and $(\hat{\beta}_0, \hat{f}_1, \ldots, \hat{f}_p)$ is DPQL estimator.

The Fisher information matrix of the approximate marginal quasi-likelihood estimators $\hat{\vartheta} = (\hat{\tau}, \hat{\gamma})$ can be approximated by

$$\Gamma(\vartheta) = \begin{pmatrix} \Gamma_{\tau\tau} & \Gamma_{\tau\gamma} \\ \Gamma'_{\tau\gamma} & \Gamma_{\gamma\gamma} \end{pmatrix}, \qquad (8.21)$$

where the $(j,k)th$ element of $\Gamma(\vartheta)$ is $\Gamma_{\vartheta_j\vartheta_k} = 0.5tr(\mathbf{P}\partial\mathbf{V}/\partial\vartheta_j\mathbf{P}\partial\mathbf{V}/\partial\vartheta_k)$. It should be noted that expression (8.21) is mainly used to construct an approximate covariance matrix of $\hat{\gamma}$.

8.4 Fitting adherence data using GAMMs

We had earlier (Chapter 5) fitted a parametric logistic-normal model (GLMM) and assumed a linear time effect. Now our aim is to model the effect of time non-parametrically while the other covariates remain parametric using GAMM. Recall that the final GLMM model was

$$\begin{aligned} g(\mu_{ij}) = {} & \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\ & + \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\ & + \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\ & + \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} time_{ij} + \beta_{21} time_{ij} gender_i \\ & + \beta_{22} time_{ij} site_i + \beta_{23} time_{ji} reason\_no_i + \beta_{24} time_{ji} reason\_risk_i + \beta_{25} age_i educ\_no_i \\ & + \beta_{26} age_i educ\_p_i + \beta_{27} age_i gender_i + b_{0i} + b_{1i} time_{ij} . \end{aligned} \qquad (8.22)$$

We now seek to fit a semi-parametric logistic regression model by re-fitting model (8.22) but now with a nonparametric time effect including all interactions that involved time as follows:

$$
\begin{aligned}
g(\mu_{ij}) = {} & \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\
& + \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\
& + \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\
& + \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + \beta_{20} age_i educ\_no_i + \beta_{21} age_i educ\_p_i \\
& + \beta_{22} age_i gender_i + f_1(time_{ij}) + f_2(time_{ij}) gender_i + f_3(time_{ij}) site_i + f_4(time_{ij}) reason\_no_i \\
& + f_5(time_{ij}) reason\_risk_i + b_{0i} + b_{1i} time_{ij}
\end{aligned}
\tag{8.23}
$$

where $g(\cdot)$ is the logit link function, β are parametric regression coefficients, f_j are centred smooth functions and the random effects, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}(\gamma))$. It should be noted that the semi-parametric logistic regression model (8.23) differs slightly from the GAMM in (8.4) in the extra parametric part $\mathbf{x}'_{ij}\beta$, which can be incorporated in the fixed effects part of the working GLMM in (8.10). Therefore, the estimation procedures discussed for fitting GAMMs in Section 8.3 can be used to fit model (8.23) with trivial modifications. The R package *mgcv* was used to fit the data; it has a number of options available for controlling the model smoothness using splines.

Regarding model fitting, attempts were made to fit model (8.23) with several different penalized regression smoothers but because of the computational intensity entailed in fitting such a huge model with a large data set, the model failed to converge. Then model (8.23) was reduced by removing the random slope and the parametric interactions and the resulting model was as follows:

$$
\begin{aligned}
g(\mu_i) = {} & \beta_0 + \beta_1 gender_i + \beta_2 educ\_no_i + \beta_3 educ\_p_i + \beta_4 site_i + \beta_5 income_i + \beta_6 tapwater_i \\
& + \beta_7 electricity_i + \beta_8 cellphone_i + \beta_9 WHOstage1_i + \beta_{10} WHOstage2_i + \beta_{11} WHOstage3_i \\
& + \beta_{12} partner_i + \beta_{13} reason\_no_i + \beta_{14} reason\_risk_i + \beta_{15} adh\_month1_i + \beta_{16} age_i \\
& + \beta_{17} base\_cd4_i + \beta_{18} baseweight_i + \beta_{19} weight_{ij} + f_1(time_{ij}) + f_2(time_{ij}) gender_i \\
& + f_3(time_{ij}) site_i + f_4(time_{ij}) reason\_no_i + f_5(time_{ij}) reason\_risk_i + b_{0i}
\end{aligned}
\tag{8.24}
$$

When the thin plate shrinkage smoothers were used to fit model (8.24), convergence was reached. These smoothers have the advantage of avoiding

the knot placement and can be constructed for smooths of any number of predictor variables. In addition, the shrinkage smoothers are constructed in such a way that smooth terms can be penalized away altogether making no contribution to the model (Wood, 2006). The results of the parametric coefficients and the approximate significance of the smooth terms are presented in Table 8.1. From the parametric part model, the results indicate that gender, treatment site, educational attainment, living with a partner and cell phone ownership has a significant effect on optimal HAART adherence (Table 8.1). Since gender and treatment site interact significantly with the nonparametric effect of time, they will be interpreted with the smooth terms of the model.

Moreover, the results showed that patients who owned a cell phone were 1.30 ($e^{0.26}$) times more likely to adhere to medication than those without a cell phone. Also, patients who stayed with a partner were 1.32 ($e^{0.28}$) times more likely to be adherent to medication that those who were not staying with a partner. Contrary to our expectation, the results reveal that patients with primary and secondary levels of education were respectively 0.69 ($e^{-0.37}$) and 0.65 ($e^{-0.43}$) times less likely to adhere to medication than those with no schooling.

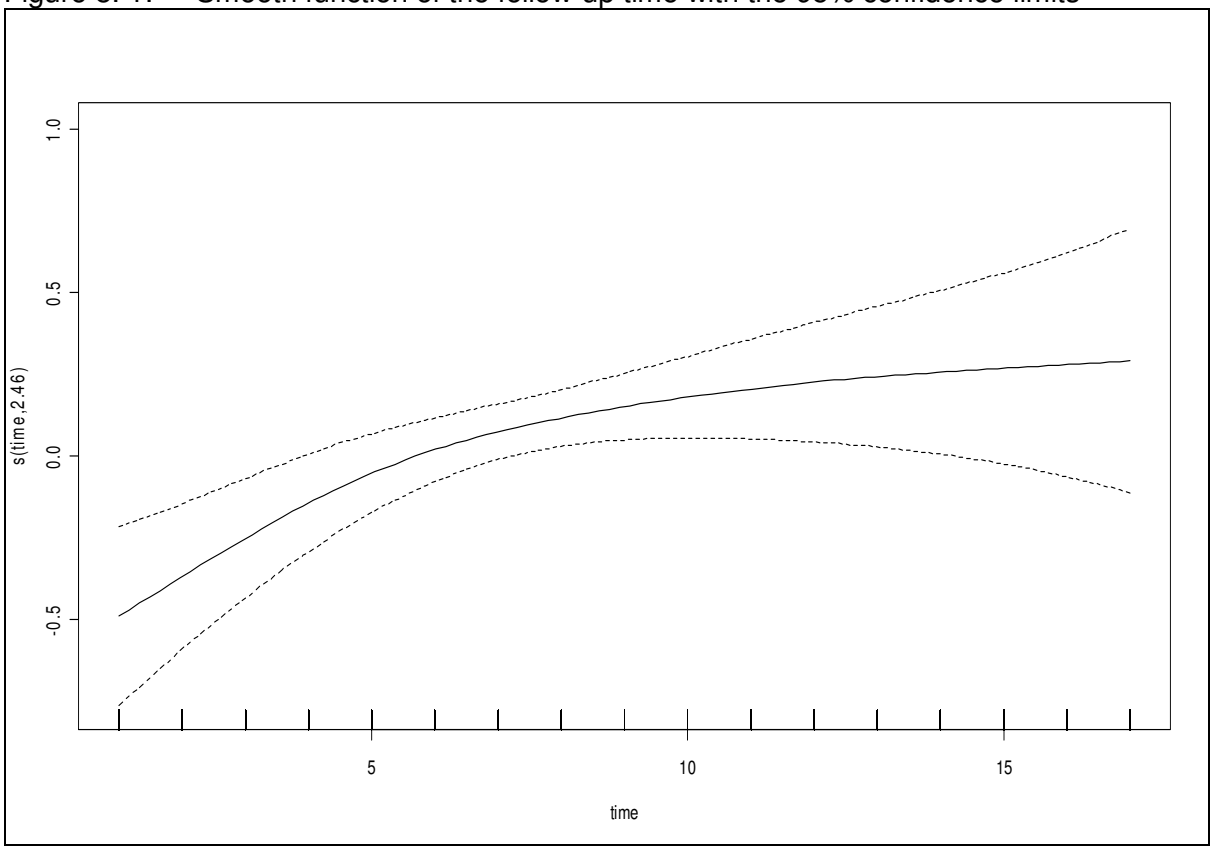
Follow up time has been fitted as a smooth function on its own and at different levels of gender, treatment site and reason for taking an HIV test. The results in Table 8.1 show that the follow-up time had a significant effect on optimal HAART adherence. The smooth term for the follow-up time has been depicted in Figure 8.1 and it shows that on the overall, HAART adherence has been increasing over the follow-up visits but the rate of increase was highest in the first five follow-up visits after which the rate of increase declined.

Table 8. 1: The parameter estimates of the fixed part of the GAMM model comprised of the parametric coefficients and the approximate significance of the smooth terms

Parametric coefficients	Estimate	Standard error	tvalue	pvalue
Intercept	1.9590	0.4346	4.508	<.0001*
Gender (ref = female)				
Male	-0.2584	0.0940	-2.749	0.0060*
Education (ref = secondary and higher)				
No schooling	-0.3742	0.1692	-2.212	0.0270*
Primary	-0.4289	0.1565	-2.741	0.0062*
Treatment site (ref = rural)				
Urban	-0.6986	0.1021	-6.845	<.0001*
Source of household income (ref = not source)				
Source of income	0.0311	0.0961	0.324	0.7459
Access to tap water (ref = no)				
Yes	0.1065	0.1516	0.702	0.4825
Household with electricity (ref = no)				
Yes	0.0448	0.1375	0.326	0.7446
Cell phone ownership (ref = no)				
Yes	0.2596	0.0874	2.971	0.0030*
WHO staging of the disease (ref = stage4)				
Stage1	0.0635	0.1600	0.397	0.6911
Stage2	0.2237	0.1497	1.494	0.1351
Stage3	0.3397	0.2251	1.509	0.1313
Living with a partner (ref = no)				
Yes	0.2832	0.0100	2.846	0.0044*
Reason for testing (ref = unwell)				
No specific reason	-0.0822	0.1330	-0.618	0.5368
Exposed to the risk	0.0074	0.1170	0.063	0.9498
Baseline adherence (ref = adherent)				
No adherent	0.1137	0.1059	1.074	0.2830
Age	-0.0027	0.0061	-0.442	0.6588
Baseline CD4+ cell count	-0.0006	0.0007	-0.789	0.4303
Baseline weight	-0.0024	0.0064	-0.370	0.7113
Weight at follow-up visits	-0.0002	0.0056	-0.039	0.9688
Approximate significance of smooth terms				
Smooth terms		**edf#**	**Fvalue**	**Pvalue**
S(time)		2.461	4.873	0.0042*
S(time): gender (female)		1.043	11.09	0.0007*
S(time): gender (male)		0.044	0.037	-
S(time): treatment site (urban)		0.002	0.001	-
S(time): treatment site (rural)		1.019	14.99	<.0001*
S(time): reason for taking HIV test (exposed to the risk)		0.922	8.017	0.0057*
S(time): reason for taking HIV test (unwell)		0.000	0.000	-
S(time): reason for taking HIV test (no specific reason)		0.000	0.000	-

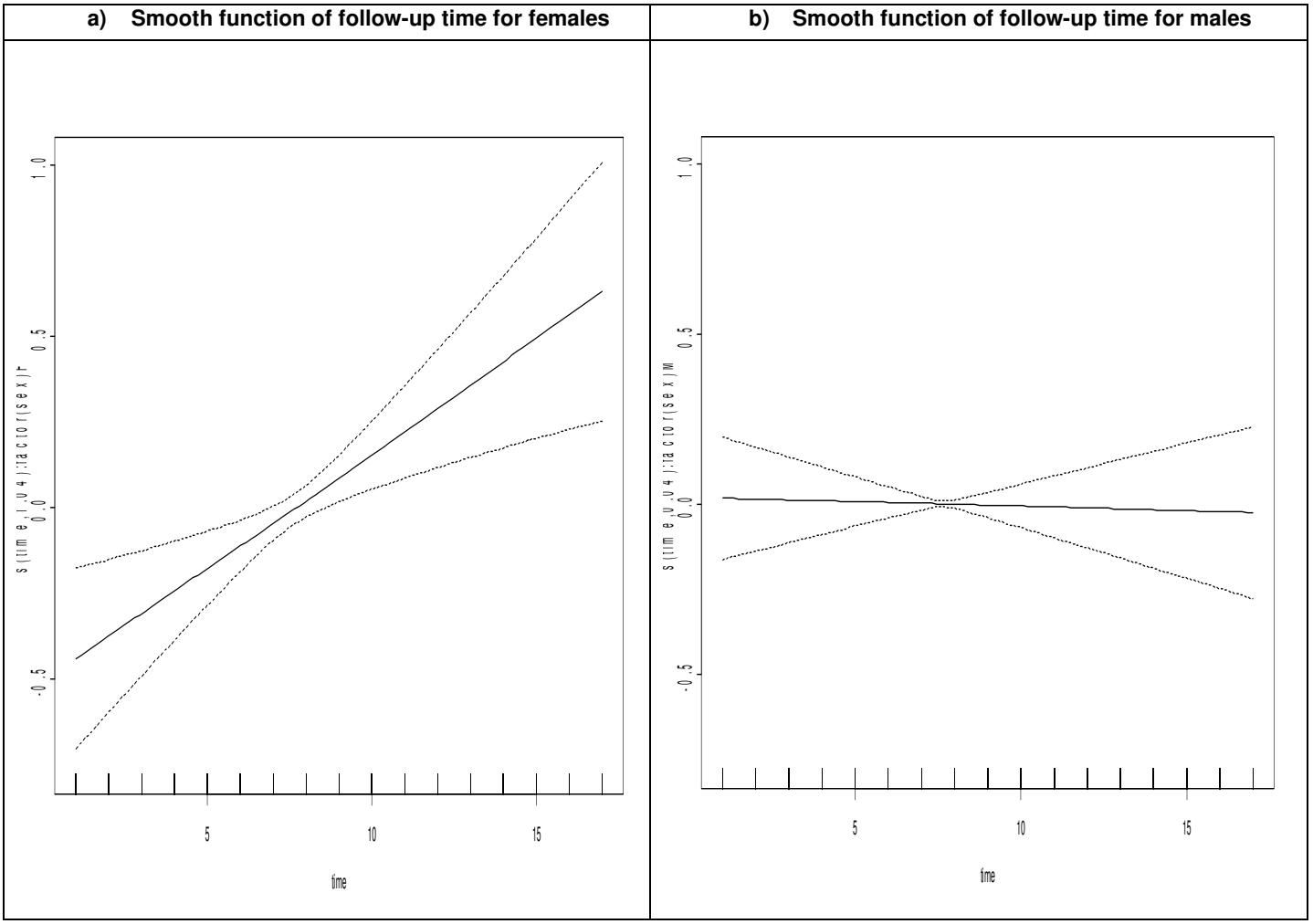
#estimated degrees of freedom
*significant at 5% level

Figure 8. 1: Smooth function of the follow-up time with the 95% confidence limits



Moreover, the results further revealed that the effect of the follow-up time differed by gender, treatment site and the reason for taking an HIV test. It is shown (Table 8.1) that the follow-up time had a significant effect on HAART adherence for females whereas the follow-up time had no significant effect on adherence for males. The smooth function for the follow-up time by gender is depicted in Figure 8.2. Figure 8.2(a) shows that HAART adherence increased over the follow-up visits for female patients whereas Figure 8.2(b) highlights the result that optimal adherence has remained almost constant for the follow-up period.

Figure 8. 2: Smooth functions of the follow-up time by gender with the 95% confidence limits



Also, the follow-up time had a significant effect on HAART adherence for patients in the rural treatment site while the follow-up time had no significant effect on adherence for patients in the urban treatment site. Figure 8.3 presents the smooth function of the follow-up time for the urban and rural patients.

Figure 8. 3: Smooth functions of the follow-up time by treatment site with the 95% confidence limits

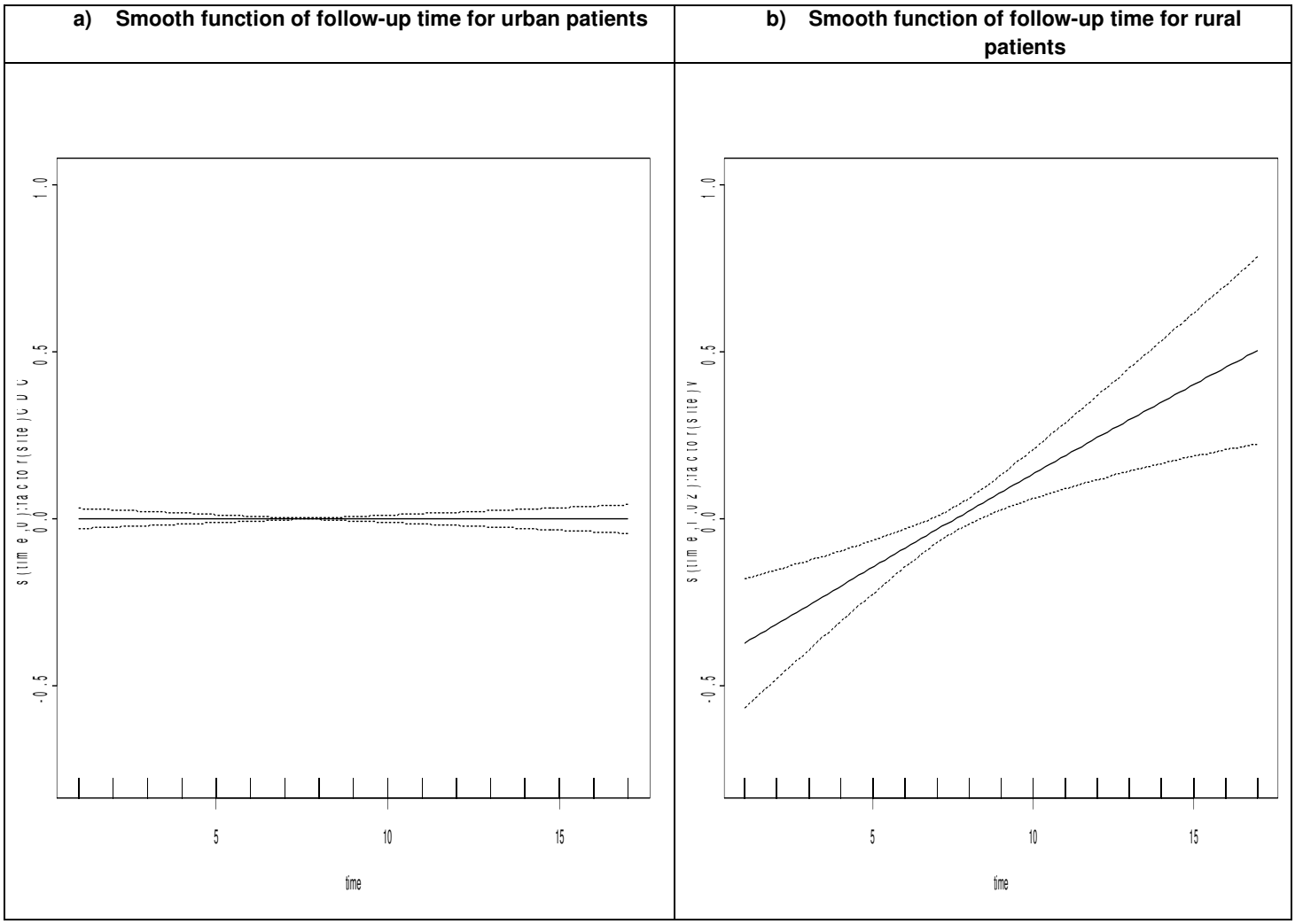
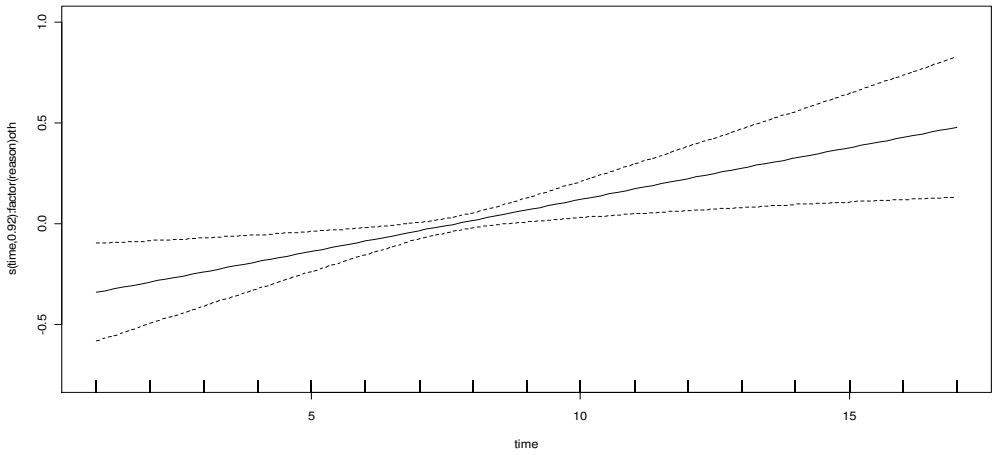


Figure 8.3a shows that HAART adherence remained constant for the urban patients over the follow-up period while the rural patients experienced an increase in HAART adherence over the same period (Figure 8.3b).

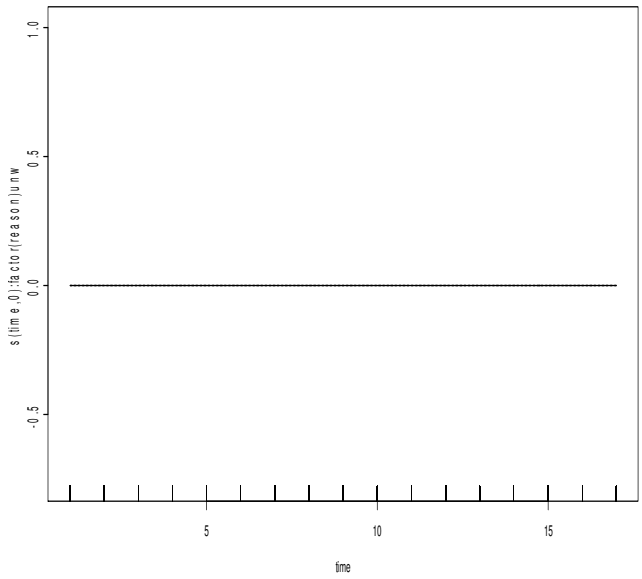
Regarding the reason that patients reported for taking an HIV test, the follow-up time had a significant effect on adherence for the patients who reported to have been exposed to the risk of contracting HIV while there was no significant effect for patients who reported being tested for HIV for no specific reason or because they were not well. The smooth functions are depicted in Figure 8.4.

Figure 8. 4 Smooth functions of the follow-up time by reason for taking an HIV test with the 95% confidence limits

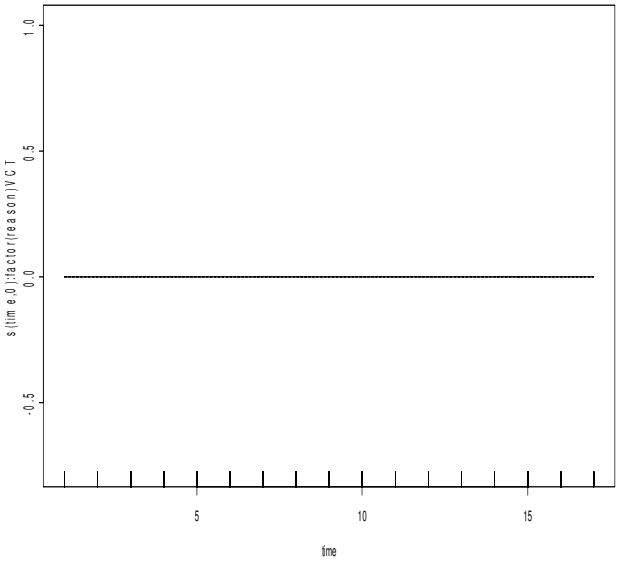
a) Smooth function of follow-up time for patients who felt exposed to the risk of contracting HIV



b) Smooth function of follow-up time for unwell patients



c) Smooth function of time for patients who tested for no specific reason



It is shown In Figure 8.4a that HAART adherence increased over the follow-up for patients who felt exposed to the risk of contracting the HIV disease and remained constant for those who were unwell (Figure 8.4b) and those who took an HIV test for no specific reason (Figure 8.4c).

8.5 Summary

The results from the GAMM model with a nonparametric follow-up time validate the results from previous models and generally gave more insight regarding the trend of optimal HAART adherence. The results from the parametric part of the model confirm that cell phone ownership and living with a partner tends to enhance one's adherence to medication. Moreover, the smooth term of follow-up time confirmed that overall, HAART adherence is increasing over time but it further revealed that the rate of increase has not been constant over the entire follow-up period. In addition, it is revealed that HAART adherence increased over the follow-up period for females while the adherence trend remained constant for males. Likewise, the rural patients showed an increase in HAART adherence over the follow-up period whereas adherence remained constant for the urban patients. Moreover, adherence remained constant over the follow-up time for patients who took the HIV test for no specific reason and those who took the test because they were unwell. However, adherence increased for those who tested because they felt they have been exposed to the risk of contracting the HIV disease.

Chapter 9

Discussions and Conclusions

This work has focused on statistical methods aimed at modeling the data that are correlated. More specifically, we have been concerned with statistical methods for longitudinal binary data, which is frequently encountered in applied statistics. Although many approaches to the analysis of longitudinal data have been studied, most are restricted to the setting in which the response variable is normally distributed. While the development of methods for analysis of longitudinal data with discrete outcomes has received substantially less attention in the past, this has more recently become an important area of research. Still, the methodology is not nearly as well-developed as for normally distributed outcomes. The practical application of methods for repeated discrete outcomes also lags behind that for normal-theory methods because among other reasons, the development of readily accessible software lags behind. Even with recent software developments, data analysts still face the challenge of choosing the appropriate method of analysis to address their research questions adequately. The other challenge relates to the estimation procedure; a method will often have more than one estimation procedure to choose from, and in choosing an estimation procedure, one has to be aware of the fact that the chosen estimation method will have a direct bearing on which tests will be available for inference. In this thesis, we attempt to give more insight into the different longitudinal approaches when one has a binary outcome. These methodologies have been demonstrated with in-depth analyses of a practical data set with a binary outcome. The data relates to HAART adherence over time among HIV positive adults who are on treatment.

With the advent of HAART, morbidity and mortality among HIV-infected individuals have been dramatically reduced. However, HAART requires strict adherence, thus identifying and overcoming factors that reduce adherence is

critical if optimal clinical and survival benefits are to be attained and maintained over time. Determining the predictors of optimal long-term HAART adherence status of patients, and whether factors affecting initial (first month) adherence status also influence long-term HAART adherence status was the focus for the application of longitudinal methods. Patients were considered optimally adherent if they took at least 95% of their prescribed medication; otherwise they were classified as non-adherent. Optimal HAART adherence status therefore was considered as a binary response variable (adherent and non adherent) that follows a Bernoulli distribution. Consequently, generalized linear models (GLMs) and its extensions to correlated data become relevant models for analysis.

Because of the importance of HAART adherence at the onset of treatment, the study began by identifying factors affecting initial (one month) adherence status of patients. The appropriate methods for analysis would be those for independent data as we have one datum point for each patient and a logistic regression model was used. The results revealed that baseline CD4+ cell count and two-way interaction terms between gender and reason for taking an HIV test, age and cell phone ownership as well as treatment site and income were important. That is, optimal HAART adherence was negatively associated with a higher baseline CD4+ cell count. Female patients had better adherence if they voluntarily attended testing and counseling or if they had taken an HIV test because they were unwell, while male patients had higher adherence if they were tested due to perceived risk of HIV infection. Optimal HAART adherence was positively associated with higher age amongst patients who possessed cell phones and amongst patients who provided a source of income in the urban setting, but not in the rural setting.

Although factors associated with optimal adherence at the initial stages of therapy give important information, factors that affect long-term optimal adherence may be more relevant, as lifetime adherence to HAART is required. Consequently, evaluation of determinants of long-term optimal HAART

adherence entailed data with a longitudinal structure. Longitudinal data tend to be correlated and this poses a challenge in the analysis because the correlation has to be accounted for to obtain valid inference. It was therefore important to explore the dependence among outcomes in order to inform further analysis. With a binary outcome, a correlation-based approach for exploring the association structure is not feasible as the range of correlation is constrained by the means. Thus a Lorelogram, which measures the association between the repeated binary outcomes with odds ratios (Heagerty and Zeger, 1998), was used in this study. The correlation appeared to decrease with increasing lag between repeated responses indicating that an AR-1 correlation structure may be appropriate for describing the relationship between follow-up visits with regard to optimal HAART adherence.

Missing data are almost always present in the analysis of longitudinal data and as a result, the data becomes unbalanced over time. Therefore, methods that require balanced data cannot be used. Because at times missing data can introduce bias, thereby leading to misleading inferences about changes in the mean, one must consider the reasons for missingness (missing data mechanisms). These are missing completely at random (MCAR), which refers to missingness if the missing values are independent of both unobserved and observed data. In this case the observed values of the responses are a random subsample of all values of the responses and no bias will arise with almost any method of analysis. There is also missing at random (MAR) and occurs if conditional on observed data, the missingness is independent of the unobserved measurements. In this case, analysis restricted to data from completers will yield biased estimates because observed values are not necessarily a random subsample of all responses. The other mechanism of missingness is referred to as missing not at random (MNAR). In this case the missing data process is neither MCAR nor MAR but is non-random.

The adherence data was mainly characterized by dropout. We began by testing the most restrictive assumption of missingness, i.e. MCAR, with adherence data and logistic regression was used to carry out the test. The results revealed that dropout did not depend on the previous outcome. This implied that there was no evidence against MCAR in favour of MAR. It was therefore concluded that MCAR assumption holds for the adherence data and as a result, methods of longitudinal data analysis will yield valid estimates.

For identifying long-term predictors of optimal HAART adherence, generalized linear models for longitudinal data were employed. These include several broad model families that include marginal, subject-specific, transition, joint and semi-parametric additive models. While these models can be viewed as direct extensions of generalized linear models for independent observations to the context of correlated data, there are differences in the way they address the dependency in the data. All the models were used to assess the determinants of long-term optimal adherence and each one of the methods has its own strengths and weaknesses.

We began by fitting a marginal model to adherence data. With marginal models, the marginal probabilities of the response are directly modeled. In general the specification of a joint distribution for discrete longitudinal data is difficult, even in cases where it might be possible to specify the joint distribution, the likelihood can be too complicated to evaluate. Consequently, marginal models were fitted using the method of generalized estimating equations (GEE) (Liang and Zeger, 1986). The GEE method is an extension of GLM to accommodate correlated data using a quasi-likelihood approach. It requires the correct specification of mean and variance, but not the specification of a distribution for the response. The GEE method is appealing because of its computational simplicity compared to the maximum likelihood approaches for discrete data. In addition, it is robust to misspecification of the correlation structure for large samples. However, since, it does not completely specify the joint distribution, it does not have a likelihood

function, thus, likelihood-based methods are not available for testing fit, comparing models and conducting inference about parameters. Thus, the generalized Wald test was used for model comparison for fixed effects while quasi-likelihood under the independence model criterion (QIC) was used for the selection of the best fitting correlation structure. Note that although the GEE estimates are consistent with misspecification of the covariance structures; care must be taken in choosing such a structure because a poor choice of the working correlation may hamper the efficiency of the estimates.

With adherence data, the QIC results indicated AR-1 as the best fitting correlation structure among the competing ones. Moreover, the results showed that cell phone ownership and living with a partner enhanced optimal HAART adherence. Further, five 2-way interaction terms, namely, time*gender, time*site, time*reported reason, age*gender and age*education were significantly associated with optimal HAART adherence. That is, with the two-way interactions that involve time, the results showed that optimal HAART adherence increased on average over time, however, the rate of increase differed by (a) gender in favour of females; (b) treatment site in favour of the rural treatment site and (c) reason for taking an HIV test where the rate of increase in optimal adherence was higher for patients who tested due to possible exposure to HIV, than for patients who tested because they were unwell. Further analysis revealed that the rate of increase in optimal adherence was higher for patients who tested due to possible exposure to HIV, than for those who reported no specific reason for taking an HIV test. Age also interacted significantly with gender and education. As the age of patients increased, females tended to adhere better to HAART than males. Among older patients, those with no schooling were less likely to achieve optimal HAART adherence than those with secondary and higher education. To this end, analysis revealed that as patients get older, those with primary education were more likely to achieve optimal adherence than those with no schooling.

In addition to exploring the population effects associated with optimal HAART adherence through the GEE method, we further explored the subject's specific effects that are unique to a particular individual by fitting a generalized linear mixed model (GLMM). GLMMs are considered as straight forward extensions of generalized linear models by appending the random effects in the linear predictor, which gives them the ability to model subject-specific evolutions in addition to determining the association structure of the data. However, the estimation procedures are more complex with GLMMs. For maximum likelihood, different approximation methods are used for parameter estimation with each method having its own strengths and limitations. These include approximation methods based on numerical integration techniques (e.g. quadrature methods) and model approximation (linearization).

Integral approximation methods approximate the log likelihood function of the GLMM and the approximated function is numerically optimized. These methods provide an actual objective function for optimization and thus likelihood-based methods can be used for inferential procedures such as likelihood ratio tests. Because of the computational complexity entailed in these methods, they are not able to accommodate a large number of random effects as well as models that contain both random effects and serial correlation. With model approximation methods, algorithms are expressed in terms of Taylor series to linearize the response and the resulting model can be viewed as a linear mixed model for pseudo-data. Then the algorithm for fitting the linear mixed model is employed and the resulting estimates are used to update the pseudo-data and this whole scheme is iterated until convergence is reached. The linearization estimation technique can also incorporate models that have both the random effects and the different residual covariance structures. However, likelihood based methods cannot be used because the likelihood is not based on the observed data but rather the likelihood corresponds to the linear mixed models for pseudo-data. This is because when one changes the error structure, then the linearized response also changes and likelihood values can only be compared across

models when the response variable for each model is the same. Approximate Wald tests can be used to test hypotheses about fixed effects as well as variance components. For both methods, care has been taken when the null-hypothesis is on the boundary of the parameter space.

Consequently, fitting data using GLMM was quite challenging. Regardless of the embedded differences between the approximation methods, the results obtained from all methods largely agree, which gives us more confidence in all the approximation methods. The results showed that there was considerable variability from patient to patient; however, all the factors that were important at the population level were still important even at the subject-specific level. In addition, the association structure revealed a strong negative correlation between the random intercepts and slopes indicating that there was a decline over time in adherence rates of patients who had high levels of adherence at the beginning of follow-up visits and vice versa.

To assess whether the current optimal adherence status of a patient depended on the previous adherence measurements in addition to the explanatory variables, a transition model was fitted. That is, the transition model was used to predict optimal HAART adherence status on the basis of explanatory variables and all available information of the previous adherence status. In this model, the correlation of repeated observations is dealt with by treating previous outcomes as additional explanatory variables. Fitting transition models is relatively straightforward because subsequent measurements, given their previous history, are considered to be independent; as a result, standard GLM procedures can be employed. Also transition models are likelihood-based methods, therefore inferential procedures such as likelihood ratio tests can be used.

The results revealed that the current adherence status had a very strong dependence on the two previous measurements of adherence status. In addition there were three significant interactions between the previous responses and

explanatory variables. These were lag1*time, lag2*cell phone ownership and lag2*baseline adherence where lag1 and lag2 represents adherence measurements lagged by 1 and 2 follow-up visits respectively. Patients who were adherent in lag1 follow-up visits were more likely to be adherent to medication than those who were not adherent in the same lag follow-up visits. Moreover, patients who owned a cell phone were more likely to adhere to medication than those without a cell phone provided they were adherent in lag2 follow-up visits. Furthermore, patients who were adherent at baseline were less likely to adhere to medication than those who were not adherent given that they failed to adhere to medication in two consecutive visits (both lag2 and lag1 follow-up visits).

With all the analyses using the different models for longitudinal data, we assumed that the time interval (duration) between one clinic visit to the next is the same across all follow-up visits. In reality, however, this assumption of equal interval between successive visits might not necessarily hold because at times patients are unable to keep their scheduled clinic appointments for various reasons/circumstances; they either make an early or a late visit to the clinic. A joint modeling approach was used to further investigate the joint effect of the predictor variables on both optimal HAART adherence status of patients and duration between successive visits. More specifically, we assessed whether the explanatory variables that were found to be significantly related with optimal HAART adherence in the marginal and random effects models would still have a significant effect on long-term optimal adherence even when duration between successive visits was accounted for. Also assessing the association between the two outcomes (optimal adherence and duration) was of interest.

Advantages of joint over separate fitting of models include better control over type I error rates in multiple tests, possible gains in efficiency in the parameter estimates and the ability to answer multivariate questions (Gueorguieva, 2001). The difficulties in analyzing longitudinal data arise because of correlations usually

present between observations on the same subject. In case of multiple outcomes, two types of correlations must be taken into account: correlations between measurements on different variables and correlations between measurements on the same variable within a subject. While joint models for longitudinal and time-to-event data are widely used, joint models for two longitudinal outcomes of similar or dissimilar nature are less widespread. For models that jointly analyze discrete and continuous outcomes, the challenge has been the lack of multivariate distributions for combining both types of outcomes; as a result, specification of a joint distribution of the responses is not straightforward.

Broadly, there are two approaches adapted to joint modeling. A first approach avoids direct specification of a joint distribution; it is based on a conditioning argument that allows joint distribution to be factored in a marginal component and a conditional component, where the conditioning can be done either on the discrete or on the continuous outcome (Catalano and Ryan, 1992; Faes et al., 2004). A disadvantage of mixed outcome models based on conditional models is that they do not directly lead to marginal inferences (Verbeke and Davidian, 2008). Also the correlation among the two outcomes cannot be directly estimated (Faes et al., 2008). A second approach directly formulates a joint model for both outcomes. The latent variable idea has been used to directly specify the joint distribution of the discrete and continuous outcomes based on two methods (Regan and Catalano, 2002). Instead of using a latent variable approach, one can directly specify the joint distribution for both outcomes through a mixed model, by specification of the marginal distribution, conditional on the correlated random effect (Faes et al., 2008). The latter approach was adopted for fitting adherence data.

A joint marginal model and a joint random effects model were fitted to assess the joint effects of the predictor variables at the population level and at the subject specific level respectively. The joint random effects model was also used to

assess the correlation between the two outcomes, which is induced by the association between the random effects of the two outcomes. The results from the joint models showed that even after accounting for the interval between successive visits, the GEE and GLMM results remained the same. Moreover, the results indicated a negative correlation between optimal HAART adherence and duration between successive visits. That is, on the overall, the association between optimal HAART adherence and duration was negative, however, the joint effect of the time and treatment site interaction on both outcomes revealed that as the number of follow-up visits increased, the interval between successive visits also increased while at the same time high levels of optimal adherence were maintained in the rural treatment site

Since the parametric statistical models employed for the analysis of adherence data may not have been flexible enough to capture the main features of the longitudinal profiles, a semi-parametric approach was adopted. Specifically, generalized additive mixed models were used to model the effect of time as well as interactions associated with time non-parametrically. The results showed that the smooth term of follow-up time confirmed that overall, HAART adherence is increasing over time but it further revealed that the rate of increase has not been constant over the entire follow-up period. In addition, it was revealed that HAART adherence increased over the follow-up period for females while the adherence trend remained constant for males. Likewise, the rural patients showed an increase in HAART adherence over the follow-up period whereas adherence remained constant for urban patients. Moreover, adherence remained constant over the follow-up time for patients who took the HIV test for no specific reason and those who took the test because they were unwell while adherence increased for those who tested because they felt they had been exposed to the risk of contracting the HIV disease.

This work demonstrates that with appropriate statistical modeling of real life data, taking into account the nature of such data and scientific settings, a useful

contribution to the knowledge/literature in areas of specific application can be made. For instance, the findings from this study identified specific groups of high risk patients for whom adherence counseling should be targeted and tailored. For example first month HAART adherence can be improved by targeting patients initiated on treatment with high CD4+ cell counts. With the long-term adherence, the findings helped identify sub-populations, such as the urban and male population, that required vigorous ongoing adherence counseling. This work highlighted the reality that the longer the duration between successive clinic visits, the more compromised optimal HAART adherence is likely to be. However, with the scaling-up of HIV-treatment, given the human resource constraints in Sub-Saharan Africa, the gap between successive visits might be increased for patients well adapted to the treatment programme and with a good record of sustained optimal adherence to reduce the burden in the health facilities.

There are avenues for further work in this research. One of the most important steps in the process of data modeling is to check various features of the fitted model. This usually involves checking goodness-of-fit of the model, checking model assumptions and detecting possibly influential observations. Little work has been done on model checking and diagnostics in GEE, GLMM and multivariate generalized joint modeling and this deserves further research. In addition, the difficulty in evaluating the likelihood for models with discrete correlated data has been a limiting feature and was noticeable in the computational aspects of fitting GLMM where intensive computing times were experienced with the very large adherence data set. The computing intensity became even worse with multivariate generalized random effects models (joint modeling). For instance, with joint modeling we were unable to fit models that contained both random intercepts and slopes because that required powerful computational resources that we did not have. This therefore highlights the need to further investigate the performance of simplified computational methods which would not require powerful computational resources. Moreover, the alternative

method of estimation that uses the pseudo-likelihood from the pseudo-data is not computationally intense but still requires further investigation in many respects. Since pseudo-likelihood is essentially composed of likelihoods, we might be inclined to believe that it is more robust to missing data mechanism than other methods based purely on estimating equations (such as GEE). It would however be useful to evaluate the robustness of this method to missing data. Further, model selection tools are lacking with this estimation method; a lack most keenly felt when a best fitting residual covariance structure has to be selected. This is because the usual likelihood based methods such as AIC cannot be used, therefore it is also a fertile area for further research.

There exist many nonparametric regression and smoothing methods for independent data in the literature. The presence of the within-subject correlation among the repeated measures over time presents a major challenge in developing nonparametric techniques for longitudinal data analysis. Most of these developments focus on nonparametric population mean models and mixed-effects models for normal longitudinal data. Limited work has been done on developing these nonparametric models for non-normal longitudinal data. Moreover, the practical application of nonparametric methods for longitudinal outcomes (normal and non-normal) is lagging behind due to a lack of readily accessible statistical software. Further research is needed in this area.

With evaluation of the determinants of optimal HAART adherence, multilevel modeling can be considered for future research to address questions about individual effects while adjusting for facility location (urban/rural), staffing issues, and equipment availability among other things. Furthermore, future research should aim at exploring determinants of optimal HAART adherence with an optimal adherence outcome measured using patient's self report in addition to the pill count method which were used in this study.

In summary, five families of models were reviewed and applied to HAART adherence data. All the analyses demonstrated that these families of models are useful in the study of binary longitudinal responses. Furthermore the thesis highlighted the direction and development of longitudinal binary data analyses. Highly complicated longitudinal binary data arising in practice are challenging, but they also provide for great opportunities and the advancement of this important area of research. One of the future directions of this thesis is to compare the different families of methods using simulations. The other possible direction of the study is to develop diagnostic tools for each of the models.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.

Aerts, M. Geys, H. Molenberghs, G. and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman and Hall.

Agresti, A., Booth, J. G., Hobert, J. P, Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology* **30**, 27-80.

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B* **46**, 118-119.

Ammassari, A., Trotta, M. P., Murri, R., Castelli, F., Narciso, P., Noto, P., Vecchiet, J., D'Arminio-Monforte, A., Wu, A. W. and Antinori, A. (2002). Correlates and predictors of adherence to highly active antiretroviral therapy: overview of published literature. *Journal of Acquired Immune Deficiency Syndromes* **31** (Suppl 3), S123-127.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley

Antonio, K. and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics* **40**, 58-76.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods* **7**, 127-150.

Berg, K. M., Demas, P. A., Howard, A. A., Schoenbaum, E. E., Gourevitch, M. N. and Arnsten, J. H. (2004). Gender differences in factors associated with adherence to antiretroviral therapy. *Journal of General Internal Medicine* **19**, 1111-1117.

Berhane, K. and Tibshirani, R. J. (1998). Generalized additive models for longitudinal data. *Canadian Journal of Statistics* **26**, 517-535.

Birnbaum, A. (1954). Statistical methods for Poisson processes and exponential populations. *Journal of the American Statistical Association* **49**, 254-266.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Brown, H. and Prescott, R. (2006). *Applied Mixed Models in Medicine*, 2nd ed. New York: John Wiley & Sons.

Cantoni, E., Flemming, J. M. and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Statistics in Medicine* **61**, 507-514.

Carey, V. C., Zeger, S. L. and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517-526.

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87**, 651-658.

Chesney, M. A. (2000). Factors affecting adherence to antiretroviral therapy. *Clinical Infectious Diseases* **30** (Suppl 2), S171-176.

Collett, D. (2003). *Modelling Binary Data*, 2nd ed. New York: Chapman & Hall/CRC.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113-120.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of Repeated Measures*. London: Chapman and Hall/CRC Press.

Cui, J. and Qian, G. (2007). Selection of the working correlation structure and best model in GEE analyses of longitudinal data. *Communication in Statistics – Simulation and Computation* **36**, 1987-1996.

Cui, J. and Feng, L. (2009). Correlation structure and model selection for negative Binomial distribution in GEE. *Communication in Statistics – Simulation and Computation* **38**, 190-197.

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Demidenko, E. (2004). *Mixed Models: Theory and Application.* New York: John Wiley and Sons.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics* **45**, 1255 – 1258

Diggle, P. J., Heagerty, P., Liang, K. - Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models.* 2nd ed. London: Chapman & Hall/CRC Press.

Dunlop, D. D. (1994). Regression for longitudinal data: a bridge from least squares regression. *The American Statistician* **48**, 299-303.

Faes, C., Geys, H., Aerts, M., Molenberghs, G. and Catalano, P. (2004). Modelling combined continuous and ordinary outcomes in a clustered setting. *Journal of Agricultural Biological and Environment Statistics* **9**, 515-530.

Faes, C., Geys, H., and Catalono, P. (2008). Joint models for continuous and discrete longitudinal data. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds). London: Chapman & Hall/CRC.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Mixed Models.* New York: Springer-Verlag.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141-150.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Ferguson, T. F., Stewart, K. E., Funkhouser, E., Tolson, J., Westfall, A. O. and Saag, M. S. (2002). Patient-perceived barriers to antiretroviral adherence: associations with race. *AIDS Care* **14**, 607-617.

Fieuws, S. and Verbeke, G. (2004). Joint modeling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in Medicine* **23**, 3093-3104.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association* **90**, 340-355.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. New York: John Wiley & Sons.

Fitzmaurice, G. M. and Molenberghs, G. (2008). Generalized linear mixed-effects models. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds). London: Chapman & Hall/CRC.

Gill, C. J., Hamer, D. H., Simon, J. L., Thea, D. M. and Sabin, L. L. (2005). No room for complacency about adherence to antiretroviral therapy in sub-Saharan Africa. *AIDS* **19**, 1243-1249.

Godin, G., Cote, J., Naccache, H., Lambert, L. D. and Trottier, S. (2005). Prediction of adherence to antiretroviral therapy: a one-year longitudinal study. *AIDS Care* **17**, 493-504.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Gu, C. and Wahba, G. (1991). Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing* **12**, 383-398.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modeling of clustered outcomes in the exponential family. *Statistical Modelling* **1**, 177-193.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. A*merican Statistician* **58**, 16-24.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations*. New York: Chapman and Hall.

Hardle, W. (1990). *Applied Non-parametric Regression*. Cambridge: Cambridge University Press.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Heagerty, P. J. and Zeger, S. L. (1998). Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. *Journal of the American Statistical Association* **93**, 150-162.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58**, 342-351.

Hedeker, D. and Gibbons D. R. (2006). *Longitudinal Data Analysis*. New York: John Wiley & Sons.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Von Krosig, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192-218.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modeling of measurements and event time data. *Biostatistics* **1**, 465-480.

Henderson, R., Diggle, P. and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3**, 33-50.

Horton, N. J. and Lipsitz, S. R. (1999). Review of software to fit generalized estimating equations (GEE) regression models. *The American Statistician* **55**, 24-254.

Hosmer D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.

Hwang, H. and Takane, Y. (2005). Estimation of growth curve models with structured error covariances by generalized estimating equations. *Behaviourmetrika* **32**, 141-153.

Islam, M. A. and Crowdhury, R. I. (2006). A higher order Markov model for analyzing covariate dependence. *Applied Mathematical Modelling* **30**, 477-488.

Jiang, J. (2001). A non-standard χ^2-test with application to generalized linear model diagnostics. *Statistics and Probability Letters* **53**, 101-109.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

Kalichman, S. C., Ramachandran, B. and Catz, S. (1999). Adherence to combination antiretroviral therapies in HIV patients of low health literacy. *Journal of General Internal Medicine* **14**, 267-273.

Kingman, J. F. C. (1993). Poisson Processes. New York: Oxford Science Publications.

Kleeberger, C. A., Phair, J. P., Strathdee, S. A., Detels, R., Kingsley, L. and Jacobson, L. P. (2001). Determinants of heterogeneous adherence to HIV-antiretroviral therapies in the Multicenter AIDS Cohort Study. *Journal of Acquired Immune Deficiency Syndrome* **26**, 82-92.

Korn, E. L. and Whittemore, A. S. (1979). Methods of analyzing panel studies of acute health effects of air pollution. *Biometrics* **35**, 795-802.

Korn, R., Ansley, C. F. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association* **86**, 1042-1050.

Kuchibhatla, M. and Fillenbaum, G. G. (2003). Comparison of methods for analyzing longitudinal binary outcomes: cognitive status as an example. *Aging and Mental Health* **7**, 462-468.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill Irwin.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Laniece, I., Ciss, M., Desclaux, A., Diop, K., Mbodj, F., Ndiaye, B., Sylla, O., Delaporte, E. and Ndoye, I. (2003). Adherence to HAART and its principal determinants in a cohort of Senegalese adults. *AIDS* **17** (Suppl 3), S103-108.

Liang, H., Wu, H. and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics* **4**, 297-312.

Liang, K. - Y. and Zeger, S. .L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K. - Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3-40.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society, Series B* **61**, 381-400.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor variable is measured with/without error. *Journal of the American Statistical Association* **95**, 520-534.

Lin, X. and Carroll, R. J. (2008). Non-parametric and semi-parametric regression methods for longitudinal data. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds). London: Chapman & Hall/CRC.

Lin, Y. - L. and Chen, Y. – W. (2009). Effects of ad types, positions, animation, lengths, and exposure times on the click-through rate animated online advertisings. *Computers and Industrial Engineering* **57**, 580-591.

Lindsey, K. L. (1997). *Applying Generalized Linear Models*. New York: Springer-Verlag.

Lindsey, K. L. (1999). *Models for Repeated measurements*, 2nd ed. New York: Oxford University Press.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, D. R. and Schabenberger, O. (2006). *SAS System for Mixed Models*, 2nd ed. Cary, NC: SAS Institute Inc.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**, 1198-1202.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley and Sons.

Liu, Q. and Pierce, D. (1994). A note on gauss-hermite quadrature. *Biometrika* **81**, 624-629.

Liu, X., Daniels, M. J. and Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to smoking cessation trial. *Journal of the American Statistical Association* **104**, 429-438.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford: Oxford University Press.

Maqutu, D., Zewotir, T., North, D., Naidoo, K. and Grobler, A. (2010a). Factors affecting first-month adherence to antiretroviral therapy among HIV-positive adults in South Africa. *African Journal of AIS Research* **9**, 117-124.

Maqutu D., Zewotir T., North D., Naidoo K., & Grobler A. (2010a). Determinants of optimal adherence over time to antiretroviral therapy amongst HIV positive adults in South Africa: a longitudinal study. *AIDS and Behavior*. Advance online publication. doi: 10.1007/s10461-010-9688-x.

Martus, P., Stroux, A., Junemann, A. M., Korth, M., Jones, J. B., Horn. F. K. and Ziegler, A. (2004). GEE approaches to marginal regression models for medical diagnostics tests. *Statistics in Medicine* **23**, 1377-1398.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

McCulloch, C. E. (2003). *Generalized Linear Mixed Models*. Beachwood, OH: Institute of Mathematical Statistics.

Mehta, S., Moore, R. D. and Graham, N. M. H. (1997). Potential factors affecting adherence with HIV therapy. *AIDS* **11**, 1665-1670.

Milliken, G. A. and Johnson, D. E. (2002). *Analysis of Messy Data, Volume III: Analysis of Covariance*. London: Chapman and Hall/CRC.

Moeti, A. (2007). Factors affecting the health status of the people of Lesotho. *M.Sc. Dissertation*: University of KwaZulu-Natal.

Mohammed, H., Kieltyka, L., Richardson-Alston, G., Magnus, M., Fawal, H., Vermund, S. H., Rice, J. and Kissinger, P. (2004). Adherence to HAART

among HIV-infected persons in rural Louisiana. *AIDS Patient Care STDS* **18**, 289-296.

Molenberghs, G. and Ryan, L. M. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics* **10**, 279-300.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Chichester, UK: John Wiley and Sons.

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score and Wald tests in a constrained parameter space. *The American Statistician* **61**, 22-27.

Muenz, L. R. and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics* **41**, 91-101.

Murphy, D. A., Marelich, W. D., Hoffman, D. and Steers, W. N. (2004). Predictors of antiretroviral adherence. *AIDS Care* **16**, 471-484.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370-384.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991). A Comparison of cluster-specific and population-averaged approaches for analysing correlated binary data. *International Statistical Review* **59**, 25-35.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Paterson, D. L., Swindells, S., Mohr, J., Brester, M., Vergis, E. N., Squier C. Wagener, M. M. and Singh, N. (2000). Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. *Annals of Internal Medicine* **133**, 21-30.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**, 89-118.

Penedo, F. J., Gonzalez, J. S., Dahn, J. R., Antoni, M., Malow, R., Costa, P. and Schneiderman, N. (2003). Personality, quality of life and HAART adherence

among men and women living with HIV/AIDS. *Journal of Psychosomatic Research* **54**, 271-278.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825-839.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical recipes in FORTRAN*, 2nd ed. Cambridge: Cambridge University Press.

Rabe-Hesketh, S. and Skrondal, A. (2008). Generalized linear mixed-effects models. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds). London: Chapman & Hall/CRC.

Ramroop, S. (2008). Analysis of longitudinal binary data: An application to a disease process. *PhD Thesis*: University of KwaZulu-Natal

Regan, M. and Catalano, P. (2002). Combined continuous and discrete outcomes. In: *Topics in Modelling of Clustered Data*. Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M. (Eds). London: Chapman and Hall.

Ridout, M. S. (1991). Reader Reaction: Testing for random dropouts in repeated measurement data. *Biometrics* **47**, 1617-1619.

Robins. J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106-121

Roca, B., Lapuebla, C. and Vidal-Tegedor, B. (2005). HAART with didanosine once versus twice daily: adherence and efficacy. *International Journal Infectious Diseases* **9**, 195-200.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2006). *The GLIMMIX Procedure*. Cary, NC: SAS Institute Inc.

Schabenberger, O. and Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Science*. New York: CRC Press.

Schall, R. (1991). Estimation of Generalized Linear Models with Random Effects. *Biometrika* **78**, 719-727.

Searle, S. R., Casella, G., and McCulloch, C. E. (2006). *Variance Components*. Hoboken, New Jersey: John Wiley & Sons.

Self, S. G. and Liang, K. - Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non standard conditions. *Journal of the American Statistical Association* **82**, 605-610.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association* **90**, 342-349.

Skogvoll, E. and Linqvist, B. H. (1999). Modeling the occurrence of cardiac arrest as a Poisson process. *Annals of Emergency Medicine* **33**, 409-417.

South African National Department of Health. (2004). *National antiretroviral treatment guidelines*, 1st ed. Available at

http://www.hst.org.za/uploads/files/sa_ART_gudelines1.pdf/. Accessed May 7, 2009.

Skovgaard, M. and Ritz, C. (2007). Influence on tests with focus on linear models. *Journal of Statistical Planning and Inference* **137**, 1979-1991.

Speed, T. (1991). Discussion on BLUP is a good thing: the estimation of random effects, by Robinson, G.K. *Statistical Science* **6**, 15-51.

Stokes, M. E., Davis, C. S. and Koch, G. G. (2000). *Categorical data analysis using the SAS system*, 2nd ed. Cary, NC: SAS Institute, Inc.

Stram, D. O. and Lee, J. W. (1994). Variance Components testing in the longitudinal mixed effects models. *Biometrics* **50**, 1171-1177.

Stram, D. O. and Lee, J. W. (1995). Correction to: Variance components testing in the longitudinal mixed effects model. *Biometrics* **51**, 1196.

Thompson, R. (1985). Discussion on some aspects of the spline smoothing approach to non-parametric regression curve fitting (by B.W. Silverman). *Journal of the Royal Statistical Society, Series B* **47**, 43-44.

Tsay, R. (1984). Regression models with time series errors. *Journal of the American Statistical Association* **79**, 118-124.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

Tuerlinckx, F., Rijimen, F., Verbeke, G., and De Boeck, P. (2006). Statistical Inference in generalized linear mixed models: a review. *British Journal* of *Mathematical and Statistical Psychology* **59**, 225-255.

Twisk, J. W. R. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge: Cambridge University Press.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika* **92**, 351-370.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G. (2003). The use of Score tests for inference of variance components. *Biometrics* **59**, 254-262.

Verbeke, G. and Davidian, M. (2008). Joint models for longitudinal data: Introduction and overview. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Eds). London: Chapman and Hall/CRC.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics* **48**, 269-311.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C. and McCulloch, C. E. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*. New York: Springer.

Wahba, G. (1985). A comparison of of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* **13**, 1378-1402.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B* **60**, 159-174.

Ware, J. H., Lipsitz, S. and Speizer, F. E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine* **7**, 95-107.

Wedderburn, R. W. M. (1974). Qausi-likelihood functions, generalized linear models, and Gauss-Newton method. *Biometrika* **61**, 439-447.

Weiss, R. E. (2005). *Modeling Longitudinal data*. New York: Springer.

West, B. T., Welch, K. B. and Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* New York: Chapman and Hall/CRC.

Williams, D. A. (1987). Generalized linear models diagnostics using the deviance and single-case deletions. *Applied Statistics* **36**, 181-191.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of statistical computation and simulation* **48**, 233-243.

Wolfinger, R. D. (1998). Towards practical application of generalized linear mixed models, in B. Marx and H. Friedl, editors, *Proceedings of 13th International Workshop on Statistical Modeling*, 388-395, New Orleans, Louisiana, USA.

Wood, S. N. (2000). Modeling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B* **62**, 413-428.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99**, 673-686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. New York: Chapman and Hall/CRC.

Wu, H. and Zhang, J. - T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

Zeger, S. L. and Liang, K. - Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Zeger, S. L., Liang, K. - Y. and Albert, P. A. (1988). Model for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019-1031.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of American Statistical Association* **93**, 710-719.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642-648

APPENDIX A: A published paper based on the application of GLM to HAART adherence data

African Journal of AIDS Research 2010, 9(2): 117–124
Printed in South Africa — All rights reserved

AJAR
ISSN 1608–5906 EISSN 1727–9445
doi:

Factors affecting first-month adherence to antiretroviral therapy among HIV-positive adults in South Africa

Dikokole Maqutu,[1,2]* Temesgen Zewotir[1], Delia North[1], Kogieleum Naidoo[2] and Anneke Grobler[2]

[1]School of Statistics and Actuarial Science, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, South Africa
[2]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Private Bag X7, Congella 4013, South Africa
**Corresponding author, e-mail address: 206521995@ukzn.ac.za*

This study explores the influence of baseline factors on first-month adherence to highly active antiretroviral therapy (HAART) among adults. The study design involved a review of routinely collected patient information in the CAPRISA AIDS Treatment (CAT) programme, at a rural and an urban clinic in KwaZulu-Natal Province, South Africa. The records of 688 patients enrolled in the CAT programme between June 2004 and September 2006 were analysed. Adherence was calculated from pharmacy records (pill counts) and patients were considered adherent if they had taken at least 95% of their prescribed drugs. Logistic regression was used to analyse the data and account for confounding factors. During the first month of therapy, 79% of the patients were adherent to HAART. HAART adherence was negatively associated with a higher baseline CD4 count. Women had better adherence if they attended voluntarily testing and counselling or if they had taken an HIV test because they were unwell, while men had higher adherence if they were tested due to perceived risk of HIV infection. HAART adherence was positively associated with higher age among patients who possessed cell phones and among patients who provided a source of income in the urban setting, but not in the rural setting. Though long-term data from this cohort is required to fully evaluate the impact of non-adherence in the first month of treatment, this study identifies specific groups of patients at higher risk for whom adherence counselling should be targeted and tailored. For example, first-month HAART adherence can be improved by targeting patients initiated on treatment with a high CD4 count.

Keywords: baseline survey, CD4 count, compliance, HAART, health information, pill counts, statistical analysis

Introduction

Advances in highly active antiretroviral therapy (HAART) have dramatically reduced morbidity and mortality due to AIDS (Kalichman, Ramachandran & Catz, 1999; Berg, Demas, Howard, Schoenbaum, Gourevitch & Arnsten, 2004; Gill, Hamer, Simon, Thea & Sabin, 2005). However, HAART includes complex drug regimens, which require strict adherence to complicated schedules for the attainment of optimal long-term clinical and survival benefits. At least 95% adherence is required for adequate virological and immunological response (Paterson, Swindells, Mohr, Brester, Vergis, Squier *et al.*, 2000). Strict adherence to all the antiretroviral (ARV) drugs prescribed is required for optimal management of HIV infection (Chen, Hoy & Lewin, 2007). This strict adherence is one of the most critical behavioural challenges in the treatment of HIV infections (Ferguson, Stewart, Funkhouser, Tolson, Westfall & Saag, 2002). Adherence appears to be the strongest predictor of both the durability of ARV medication (Esch, Klem, Kuhman, Hewitt & Morse, 2002) and the rate of cycling of the HAART regimen. Hence, it is imperative to evaluate the effect of adherence comprehensively.

Recent studies on the impact of HAART adherence on treatment outcomes in sub-Saharan Africa (in an era of treatment rollout) indicate improved immunologic response and clinical outcomes among patients with optimal adherence (Abaasa, Todd, Ekoru, Kalyango, Levin, Odeke & Karamagi, 2008; Chi, Cantrell, Zulu, Mulenga, Levy, Tambatamba *et al.*, 2009). Even after one month of therapy, patients with optimal adherence demonstrate a significant increase in CD4 count, viral suppression and significant weight gain (Brechtl, Breitbart, Galietta, Krivo & Rosenfeld, 2001). Moreover, it is shown that optimal initial adherence is positively associated with improved long-term treatment outcomes (Carrieri, Raffi, Lewden, Sobel, Michelet, Cailleton *et al.*, 2003). These studies highlight the need for strict adherence from the beginning of treatment in order to maintain prolonged clinical benefits. However, clinicians are not able to accurately predict individuals at risk of suboptimal HAART adherence at the beginning of treatment (Bangsberg, Hetch, Clague, Charlebois, Ciccarone, Chesney & Moss, 2002). Since these individuals do not have a prior adherence track record, patient-related factors (such as such as age, gender, income) and factors related to the disease characteristics at baseline might prove to be informative predictors of initial and future HAART adherence (Chesney, 2000; Reynolds, Testa, Marc, Chesney, Neidig, Smith *et al.*, 2004).

There is limited data regarding factors that predict initial optimal HAART adherence to ARV medication. In particular, there are no studies that examine how patient-related factors relate to each other (interact) and their subsequent influence on initial optimal HAART adherence. The purpose of this study is to identify whether specific clinical and socio-demographic factors present at baseline (and their respective possible interactions) influenced first-month optimal adherence to HAART among HIV-positive adults. The knowledge and understanding of such factors is particularly important given the increasing number of patients enrolled on HAART who are to be maintained in therapy. Improved first-month adherence could also help to avoid switching patients to more costly second-line regimens. The findings will be useful in developing tools to assist clinicians in the identification of factors related to poor adherence prior to initiating therapy. The results can be further used to shape communication and counselling strategies, prior to treatment initiation.

Materials and methods

Study design

The Centre for the AIDS Programme of Research in South Africa's (CAPRISA) AIDS Treatment (CAT) programme is an ongoing rollout programme for ART services that was started in 2004. The objective is to describe the profile of patients presenting at the CAT programme with respect to their social characteristics, clinical status, and clinical course during care, including HAART regime, HAART adherence and clinical outcomes. Adult patients who met HAART eligibility criteria were enrolled into the CAT programme. Eligibility criteria included a CD4 count below 200 cells/μL or patients with World Health Organization (WHO) stage 4 of HIV disease, regardless of CD4 count. The accrual of patients had been evenly distributed over time since initiation of the programme. All patients received three sessions of adherence education and motivation and preparedness training prior to HAART initiation. These adherence counselling sessions were held one week prior to starting treatment. The first two sessions were held the day the decision was taken to start the patient on treatment, and the third session was held two or three days subsequent to the first two sessions. All the sessions lasted 15 to 60 minutes, depending on the patient who is counselled. The CAT programme offers HIV-care services at two different sites in KwaZulu-Natal Province, namely the eThekwini Clinical Research Site located in an urban area and the Vulindlela Clinical Research Site located in a rural area (see <http://www.caprisa.org>). The programme started providing free HAART through a grant from the US President's Emergency Plan for AIDS Relief (PEPFAR) at a time when access to HAART was limited.

All patients were on a regimen containing two nucleoside reverse transcriptase inhibitors and one non-nucleoside reverse transcriptase inhibitor. Patients in the urban clinic received efavirenz (EFV), lamivudine (3TC) and didanosine (ddI or ddI-EC). These regimens were chosen because they can be co-administered with anti-tuberculosis (TB) medication. The regimen dispensed at the rural clinic consisted of EFV, 3TC and stavudine (d_4T), which is the standard government regimen in South Africa. A few patients (3.8%) received nevirapine (NVP) rather than EFV because they were pregnant.

All information for the CAT programme was recorded on data-collection sheets (administered at the treatment sites) which underwent two levels of quality control and were subsequently faxed to a data-management centre. The data used for analysis in this study consisted of a retrospective review of the records of patients in the CAT programme between June 2004 and September 2006. Only patient records with data on pill-count information for the defined study period were included in the analysis. Analysis of this data was approved by the University of KwaZulu-Natal's biomedical research ethics committee (ref. E 248/05).

Variables of interest

Response variable — The outcome of interest is optimal adherence to HAART. Pharmacy records that contained the detailed pill-count information were used to calculate adherence rates. Patients were then classified as optimally adherent if they took at least 95% of their prescribed drugs in a given regimen; otherwise, they were considered non-adherent. Thus, the response variable is binary, indicating whether a patient was optimally adherent or not adherent.

Independent variables — Baseline socio-demographic variables included in the analysis were age (years), gender, educational status, treatment site, whether or not a patient lived with a partner, whether or not the patient was the source of household income, the patient's access to tap water and electricity, as well as whether or not a patient owned a cell phone. The variables used to characterise the health status of the patients at baseline included World Health Organization (WHO) stage of HIV disease, CD4 count (cells/μL), and weight (kg). Information about why the HIV test was done was sought from the patients, and their responses included being unwell, attending voluntary counselling and testing (VCT), and various 'other' reasons. Other reasons included a partner having died of the disease, being ill or being unfaithful; thus they were classified as at risk of having been exposed to HIV.

Data analysis

The socio-demographic and clinical characteristics of the study population were summarised using the median with the inter-quartile range (IQR) for continuous variables and using proportions for the categorical variables.

The data were analysed by fitting a multivariate logistic regression model. The deviance was used to compare alternative models during model selection. Change in the deviance was used to measure the extent to which the fit of the model improved when additional variables were included. The main effects and possible combinations of two-way interaction terms were fitted, while attention was given to the hierarchic principle in statistical modelling. The selected model was the one with the smallest deviance.

The selected model was assessed for goodness of fit using the Hosmer-Lemeshow statistic (see Hosmer & Lemeshow, 1989; Collett, 2002). Influential observations were identified by plotting the Cook's distance statistic

against the observations (see Collett, 2002). The appropriateness of the link function was assessed by regressing the linear predictor and its square on the dependent variable. The link function is appropriate if the linear predictor is significant and its square is insignificant Vittinghoff, Gliden, Shiboski & McCulloch, 2005).

Results

Study population

The CAT programme enrolled 1 184 patients between June 2004 and September 2006. Pill-count records were only available for 688 of these patients and all 688 were included in this study. We studied relatively equal numbers of patients from each of the treatment sites, 54% from the urban treatment site and 46% from the rural treatment site. There were no differences in the baseline characteristics between those included in the study and those excluded with regard to age (means: included = 34.1 years, excluded = 34 years; $p = 0.90$), gender (men: included = 30%, excluded = 31.8%; $p = 0.51$) and baseline CD4 count (means: included = 107.6 cells/μL, excluded = 111.5 cells/μL; $p = 0.47$).

Baseline socio-demographic and clinical characteristics of the patients

The baseline socio-demographic and clinical characteristics of the patients included in the analysis are presented in Table 1. The median age of the patients was 32.5 years (inter-quartile range [IQR]: 28–38 years), 70% were women, and 69% had attained a secondary-school or higher level of education. A large portion of the patients were not living with a partner (75%). Only 28% were classified as a source of their household's income. At the time of enrolment, 64% of the patients were classified as WHO stage 3 of HIV disease; the sample's median weight was 60 kg (IQR: 53–69 kg) and the median CD4 count was 108 cells/μL (IQR: 52–159 cells/μL). Over half (56%) reported that they had decided to take an HIV test because they were not well, while 26% reported having done so because they attended VCT. The remaining 18% said they had taken an HIV test because they felt exposed to the risk of contracting HIV. Over 90% of the patients stayed in households that had access to tap water and electricity, while only 42% of the patients owned cell phones. In the first month post-HAART initiation, 79% of the patients were at least 95% adherent to HAART.

Results of multivariate logistic regression

From a set of alternative models, a model with all the main effects and three interaction terms had the smallest deviance and thus was selected. Goodness-of-fit of the model was found to be satisfactory (Hosmer-Lemeshow statistic = 6.45; $p = 0.60$). The index plot of the Cook's distance statistic indicated that there were no influential observations. The link function was appropriate since the linear predictor was significant ($p < 0.001$) while the square of it was insignificant ($p = 0.50$). The adjusted odds ratios (AOR) and their corresponding 95% confidence intervals (95% CI) for the selected model are presented in Table 2.

Gender, treatment site, contribution to household income, cell phone ownership, and baseline CD4 count were all found to be significant main effects (Table 2). There were three significant interaction terms: between age and cell phone ownership; between gender and reported reason for taking an HIV test; and between treatment site and source of household income. All significant main effects, except baseline CD4 count, were involved in significant interaction terms (Table 2). For a unit increase in CD4 count (cells/μL), the odds of HAART adherence decreased by 5% (AOR = 0.995 [0.992–0.999]; $p = 0.02$) (Table 2). The interaction effects are presented in the next sections. It should be noted that for the interaction terms that involved two categorical variables, the meaningful odds ratios for comparison needed to be further calculated from the table of results (i.e. Table 2). The post-hoc effects of the interactions between gender and reported reason for taking an HIV test as well as between treatment site and the patient's contribution to household income are reported in Tables 3 and 4, respectively.

Interactions between patient's age and cell phone ownership

As age increased, optimal HAART adherence was less likely for patients without cell phone ownership than those with cell phone ownership (AOR = 0.927 [0.869–0.987]; $p = 0.019$) (Table 2). More specifically, the rate of change in optimal HAART adherence increased with age for patients with cell phones, whereas it decreased as age increased for patients without cell phones (Figure 1). Figure 1 shows that the gap in optimal HAART adherence between groups of patients with and without cell phone ownership widened with increasing age.

Interactions between gender and patient's reason for taking an HIV test

Optimal HAART adherence was significantly higher for women than for men among patients who reported having attended VCT as a reason for taking an HIV test (AOR = 4.911 [1.892–12.75]; $p < 0.001$) as well as those who reported having tested because they were not well (AOR = 2.039 [1.066–3.900]; $p = 0.031$) (Table 3). There was, however, no significant difference in optimal HAART adherence between women and men who reported having tested for HIV because they felt exposed to the risk of contracting HIV (AOR = 0.299 [0.059–1.519]; $p = 0.145$) (Table 3). It is also shown that for men, HAART adherence was significantly lower for patients who reported VCT as a reason for taking an HIV test than for those who tested because they felt exposed to the risk of contracting HIV [AOR = 0.185 [0.035–0.993]; $p = 0.049$) (Table 3). These results confirm the observed proportions of optimal HAART adherence for gender classified by reported reason for taking an HIV test as depicted in Figure 2.

Interactions between treatment site and patient's contribution to household income

For patients who were not sources of their household's income, optimal HAART adherence was significantly higher for patients at the urban treatment site than at the rural

Table 1: Baseline socio-demographic and clinical characteristics of the HAART patients (n = 688)

Characteristics	Median (inter-quartile range)	n (%)
Age (years)	32.5 (28–38)	
Gender:		
Men		206 (30)
Women		482 (70)
Education:		
No schooling		74 (12)
Primary school		116 (19)
Secondary school or higher		429 (69)
Treatment site:		
Urban		369 (54)
Rural		319 (46)
Living with or without a partner:		
Living with a partner		168 (25)
Living without a partner		510 (75)
Contribution to household income:		
Source of income		186 (28)
Not source of income		489 (72)
WHO stage of HIV disease:		
Stage 1		71 (10)
Stage 2		121 (16)
Stage 3		438 (64)
Stage 4		58 (8)
Baseline CD4 count (cells/μL)	108 (52–159)	
Baseline weight (kg)	60 (53–69)	
Reason for taking HIV test:		
Unwell		374 (56)
Attended VCT		170 (26)
Risk of exposure to HIV		121 (18)
Household access to tap water:		
Yes		611 (91)
No		59 (9)
Household access to electricity:		
Yes		607 (91)
No		63 (9)
Cell phone ownership:		
Yes		281 (42)
No		389 (58)
First-month optimal HAART adherence:		
Optimally adherent		546 (79)
Not optimally adherent		142 (21)

treatment site [AOR = 4.347 [2.258–8.369]; $p < 0.001$) (Table 4), whereas there was no difference between treatment sites in regard to patients who were sources of household income [AOR = 0.751 [0.237–2.385]; $p = 0.628$) (Table 4). For the rural treatment site, optimal HAART adherence was significantly higher for patients who were a source of household income than those who were not a source of household income (AOR = 3.828 [1.311–11.17]; $p = 0.014$) (Table 4; Figure 3).

Discussion and conclusions

First-month adherence to HAART decreased with increasing baseline CD4 count. It has been previously established that healthier patients tend to have significantly greater rates of missing appointments than immunocompromised patients (Esch *et al.*, 2002). Patients with a higher CD4 count may not have experienced debilitating opportunistic infections (despite laboratory evidence of the need to start HAART) at the time they start HAART, and this seems to have an adverse effect on adherence (Sarna, Pujari, Sengar, Garg, Gupta & Dam, 2008). Lack of experience of severe opportunistic infections may influence the patient's perceptions about the severity of the disease and the need to maintain a high level of adherence.

Forgetfulness has been found to be the most frequently mentioned reason for missed doses among patients on HAART (e.g. Chesney, 2000; Bartlett, 2002; Barfod, Sorensen, Nielsen, Rodkjaer & Obel, 2006). As a result, many interventions consist of providing memory aids for dosing times. These include the use of new technologies such as reminders through cell phones (Bartlett, 2002; Ickovics & Meade, 2002; Abel & Painter, 2003; Osterberg & Blaschke, 2005). Our results seem to indicate that older patients who have cell phones use them effectively

Table 2: Adjusted odds ratios (AOR) (95% confidence interval [CI]) for the multivariate logistic regression model on optimal HAART adherence (ref. = reference category)

Parameter	AOR (95% CI)	*p*-value
Intercept	0.422 (0.044–4.024)	0.454
Age (years)	1.047 (0.992–1.106)	0.093
Gender (ref. = men)		
Women	2.039 (1.066–3.900)	0.031
Education (ref. = secondary school and above)		
No schooling	0.594 (0.277–1.276)	0.182
Primary school	1.483 (0.725–3.033)	0.280
Treatment site (ref. = rural site)		
Urban site	4.347 (2.258–8.369)	<0.001
Contribution to household income (ref. = not source of income)		
Source of income	3.828 (1.311–11.17)	0.014
Partner (ref. = living with a partner)		
Living without a partner	0.908 (0.530–1.557)	0.727
WHO stage of HIV disease (ref. = stage 4)		
Stage 1	1.575 (0.437–5.679)	0.488
Stage 2	0.829 (0.275–2.497)	0.739
Stage 3	0.860 (0.319–2.320)	0.766
Baseline CD4 count (cells μL)	0.995 (0.992–0.999)	0.020
Baseline weight (kg)	1.002 (0.983–1.022)	0.834
Reason for taking HIV test (ref. = unwell)		
Attended VCT	0.614 (0.271–1.389)	0.242
Risk of exposure to HIV	3.319 (0.653–16.88)	0.143
Household access to tap water (ref. = yes)		
No	1.153 (0.473–2.811)	0.755
Household has electricity (ref. = yes)		
No	0.981 (0.441–2.184)	0.963
Cell phone ownership (ref. = yes)		
No	15.55 (1.773–136.4)	0.013
Age * cell phone ownership (ref. = own cell phone)		
Age * no cell phone ownership	0.927 (0.869–0.987)	0.019
Gender * reason for testing (ref. = men * unwell)		
Women * attended VCT	2.409 (0.767–7.566)	0.132
Women * risk of exposure to HIV	0.147 (0.025–0.846)	0.032
Site * household income (ref. = rural site * not source of income)		
Urban site * source of income	0.173 (0.050–0.602)	0.006

Table 3: Post-hoc effects of the interaction between gender and reported reason for taking an HIV test (adjusted odds ratio [AOR] with 95% confidence interval [CI])

Interaction effect	AOR (95% CI)	*p*-value
Women versus men:		
Attended VCT	4.911 (1.892–12.75)	<0.001
Risk of exposure to HIV	0.299 (0.059–1.519)	0.145
Unwell	2.039 (1.066–3.900)	0.031
Men:		
Attended VCT versus risk of exposure to HIV	0.185 (0.035–0.993)	0.049
Attended VCT versus unwell	0.614 (0.271–1.389)	0.242
Risk of exposure of HIV versus unwell	3.319 (0.653–16.88)	0.148
Women:		
Attended VCT versus risk of exposure to HIV	0.446 (0.073–2.739)	0.383
Attended VCT versus unwell	1.479 (0.657–3.330)	0.345
Risk of exposure to HIV versus unwell	3.319 (0.653–16.88)	0.148

as reminders to pre-empt forgetfulness and thus will have higher adherence than older patients without cell phones.

Women who had sought VCT were more likely to have adhered to their medication than were their male counterparts. The expectation is that patients in a VCT setting are ready for behaviour change, therefore better compliance can be expected. Lower adherence levels for men than women who attended VCT might be attributed to the suggestion that men are less likely to adopt positive behaviour change (Laforge, Velicer & Owen, 1999).

Table 4: Post-hoc effects of the interaction between HAART treatment site and patient's contribution to household income (adjusted odds ratio [AOR] with 95% confidence interval [CI])

Interaction effect	AOR (95% CI)	*p*-value
Urban versus rural treatment site:		
Source of household income	0.751 (0.237–2.385)	0.628
Not source of household income	4.347 (2.258–8.369)	<0.001
Source versus not source of household income:		
Urban site	0.662 (0.342–1.278)	0.219
Rural site	3.828 (1.311–11.17)	0.014

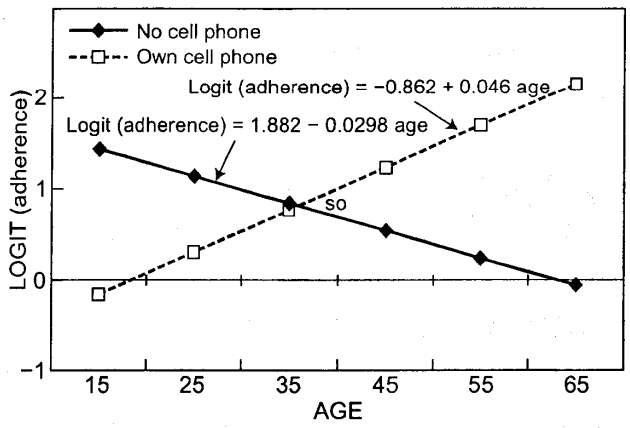


Figure 1: Log odds ratio associated with optimal HAART adherence and age for patients with and without cell phones (based on estimates of the coefficients from the fitted multivariate logistic regression model)

Men who tested for HIV because there were compelling reasons to take an HIV test tended to adhere to medication significantly better than those who had attended VCT. In view of the general reluctance of men to seek healthcare (Macintyre, Hunt & Sweeting, 1996; Laforge *et al.*, 1999), the attitudes and behaviour of the men who admitted that they had been exposed to the risk of HIV and consequently sought healthcare might have led to higher adherence than would be expected generally.

A major challenge facing rural communities is food insecurity, which has a negative relationship with income (Laforge *et al.*, 1999; Nord & Winicki, 2000). In the rural setting we found that adherence was significantly lower among patients who were not sources of their household's income as compared to patients who were sources of income. Lack of food and hunger following HAART introduction and has been a regularly cited reason for non-adherence to HAART among patients (Marston & De Cock, 2004).

It has been shown that demographic factors are not consistently associated with adherence to HAART (e.g. Haubrich, Little, Currier, Forthal, Kemper, Beall *et al.*, 1999; Fong, Ho, Fung, Lee, Tse, Yuen *et al.*, 2003). The results from this study demonstrate that age, gender, baseline CD4 count, and contribution to household income have an effect on first-month HAART adherence through significantly interacting with other variables, which include cell phone ownership, reasons given by patients for taking an HIV test, and whether a patient resides in an urban or a rural setting. We found urban versus rural differences with regard to some of the factors that might affect first-month adherence. Due attention should be paid to address the specific needs in each setting. It has been established that non-adherence to treatment is associated with faster disease progression, even for those who start HAART at a relatively high baseline CD4 count (Wood & Hogg, 2003). This, combined with lower adherence among patients with higher CD4 counts, implies that pre-treatment counselling interventions should also be targeted at people initiated on HAART with a relatively high CD4 count, and should be tailored to the specific needs of these patients.

Owing to the considerable amount of information collected from the patients at the two treatment sites for the CAT programme, as well as the quality-control measures undertaken before the data was faxed to the data management centre, the pill-count information was available for 688 out of 1 184 patients enrolled at the time of the analysis. However, we established that there were no major demographic differences between those included and those excluded in the analysis, thus not biasing the results.

One limitation of this study is that the interactions between variables were identified using the data and model fit techniques. The interactions were not pre-specified or expected during data collection. Therefore, detailed information on why these interactions influenced HAART adherence were not collected and the reason for some of these findings cannot be explained. For example, we can only speculate about why older, but not younger, patients who claimed cell phone ownership were found to adhere better.

Furthermore, this study focused on first-month HAART adherence and there was no evidence that first-month HAART adherence was indicative of longer-term adherence in this cohort. Early non-adherence to HAART has been associated with initial attrition of patients from HIV-treatment programmes (Mocroft, Youle, Moore, Sabin, Madge, Lepri *et al.*, 2001; O'Brien, Clark, Besch, Myers & Kissinger, 2003) and is therefore important to understand. An understanding of underlying factors that may contribute to non-adherence has the potential to improve the effective scaling up of HAART programmes. The researchers intend to further study the factors influencing long-term adherence to HAART. It would also be interesting to see whether factors that influence first-month adherence would also

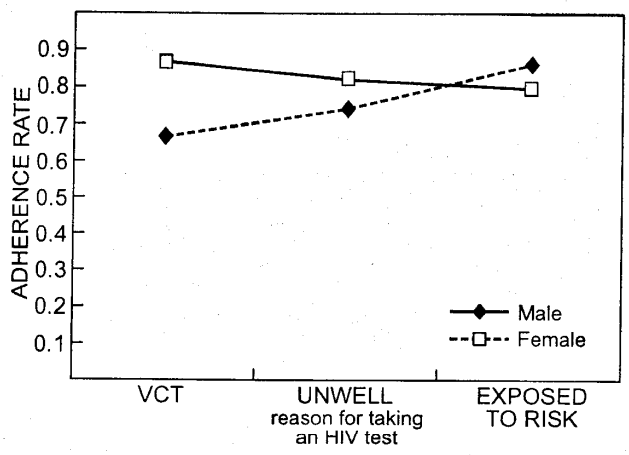


Figure 2: Percentage of HAART adherence associated with gender and reported reason for taking an HIV test, namely: attended voluntary counseling and testing (VCT), felt unwell or exposed to risk of HIV infection (based on observed proportions of optimal HAART adherence)

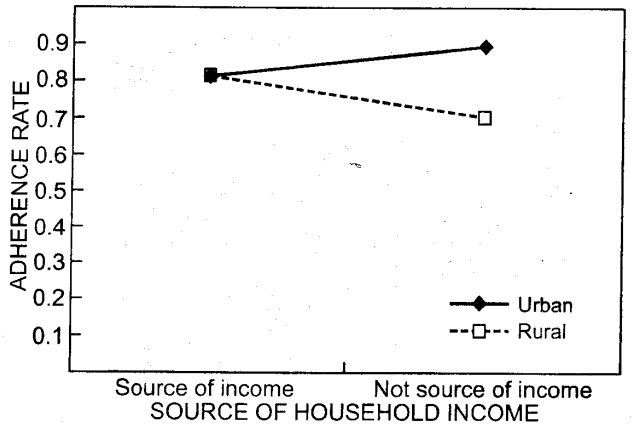


Figure 3: Percentage of HAART adherence associated with urban or rural treatment site and whether or not a patient was a source of household income (based on observed proportions of optimal adherence)

influence adherence over a longer period of time using the same dataset.

The power of this study lies in the fact that it suggests possible interactions between certain characteristics of the patients, which merit further research. In addition, it identifies specific groups of patients at higher risk for whom ARV adherence counselling should be targeted and tailored.

Acknowledgements — CAPRISA was established in 2002 through a Comprehensive International Program of Research on AIDS (CIPRA) grant (AI51794) from the American National Institutes of Health (NIH), as a multi-institutional collaboration, incorporated as an independent non-profit AIDS research organisation. The NIH funded the development of the research infrastructure, including the data management, laboratory and pharmacy cores established through the CIPRA grant. A PEPFAR grant (1U2GPS001350) funded the care of the patients in the CAT programme. Dikokole Maqutu was supported by the Columbia University-Southern African Fogarty AIDS International Training and Research Programme (AITRP), funded by the Fogarty International Center (grant D43TW00231). We gratefully acknowledge all the patients in the CAT programme. We also thank all the staff who worked on treating patients in the programme and who helped with the data collection. Special thanks are extended to the pharmacists for collection of the pill-count data.

References

Abaasa, M.A., Todd, J., Ekoru, K., Kalyango, J.N., Levin, J., Odeke, E. & Karamagi, C.A.S. (2008) Good adherence to HAART and improved survival in a community HIV/AIDS treatment and care programme: the experience of The AIDS Support Organisation (TASO), Kampala, Uganda. *BMC Health Service Research* 8, p. 241.

Abel, E. & Painter, L. (2003) Factors that influence adherence to HIV medications: perceptions of women and health care providers. *Journal of the Association of Nurses in AIDS Care* 14(4), pp. 61–69.

Bangsberg, D., Hetch, F.M, Clague, H., Charlebois, E.D., Ciccarone, D., Chesney, M. & Moss, A. (2002) Provider assessment of adherence to HIV antiretroviral therapy. *Journal of Acquired Immune Deficiency Syndromes* 26(5), pp. 435–442.

Barfod, T.S., Sorensen, H.T., Nielsen, H., Rodkjaer, L. & Obel, N. (2006) 'Simply forgot' is the most frequently stated reason for missed doses of HAART irrespective of degree of adherence. *HIV Medicine* 7(5), pp. 285–290.

Bartlett, J.A. (2002) Addressing the challenges of adherence. *Journal of Acquired Immune Deficiency Syndromes* 29(supplement 1), pp. S2–S10.

Berg, K.M., Demas, P.A., Howard, A.A., Schoenbaum, E.E., Gourevitch, M.N. & Arnsten, J.H. (2004) Gender differences in factors associated with adherence to antiretroviral therapy. *Journal of General Internal Medicine* 19(11), pp. 1111–1117.

Brechtl, J.R., Breitbart, W., Galietta, M., Krivo, S. & Rosenfeld, B. (2001) The use of highly active antiretroviral therapy (HAART) in patients with advanced HIV infections: impact on medical, palliative care and quality of life outcomes. *Journal of Pain and Symptom Management* 21(1), pp. 41–51.

Carrieri, M.P., Raffi, F., Lewden, C., Sobel, A., Michelet, C., Cailleton, V., Chene, G., Leport, C., Moatti, J.P., Spire, B. & APROCO Study Group (2003) Impact of early versus late adherence to highly active antiretroviral therapy on immunovirological response: a 3-year follow-up study. *Antiviral Therapy* 8(6), pp. 585–594.

Chen, L.F., Hoy, J. & Lewin, S.R. (2007) Ten years of highly active antiretroviral therapy for HIV infection. *Medical Journal of Australia* 186(3), pp.146–151.

Chesney, M.A. (2000) Factors affecting adherence to antiretroviral therapy. *Clinical Infectious Diseases* 30(supplement 2), pp. S171–S176.

Chi, H.B., Cantrell, R.A., Zulu, I., Mulenga, L.B., Levy, J.W., Tambatamba, B.C., Reid, S., Mwango, A., Mwinga, A., Bulterys, M., Saag, M.S. & Stringer, J.S.A. (2009) Adherence to first-line antiretroviral therapy effects non-virologic outcomes among patients on treatment for more than 12 months in Lusaka, Zambia. *International Journal of Epidemiology* 38(3), pp. 746–756.

Collett, D. (2002) *Modelling Binary Data* (Second edition). New York, Chapman & Hall/CRC Press.

Esch, L., Klem, K., Kuhman, L., Hewitt, R. & Morse, G. (2002) *Intensive adherence interventions improve virologic response to antiretroviral therapy (ART) in treatment-naive patients.* Poster

[#MoPeB3301] at the 14th International AIDS Conference, Barcelona, Spain, 7–12 July 2002.

Ferguson, T.F., Stewart, K.E., Funkhouser, E., Tolson, J., Westfall, A.O. & Saag, M.S. (2002) Patient-perceived barriers to antiretroviral adherence: associations with race. *AIDS Care* 14(5), pp. 607–617.

Fong, O.W., Ho, C.F., Fung, L.Y., Lee, F.K., Tse, W.H., Yuen, C.Y., Sin, K.P. & Wong, K.H. (2003) Determinants of adherence to highly active antiretroviral therapy (HAART) in Chinese HIV/AIDS patients. *HIV Medicine* 4(2), pp. 133–138.

Gill, C.J., Hamer, D.H., Simon, J.L., Thea, D.M. & Sabin, L.L. (2005) No room for complacency about adherence to antiretroviral therapy in sub-Saharan Africa. *AIDS* 19(12), pp. 1243–1249.

Haubrich, R.H., Little, S.J., Currier, J.S., Forthal, D.N., Kemper, C.A., Beall, G., Johnson, N.D., Dube, M.P., Hwang, J.Y. & McCutchan, J.A. (1999) The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response: California Collaborative Treatment Group. *AIDS* 13(9), pp. 1099–1107.

Hosmer, D. & Lemeshow, S. (1989) *Applied Logistic Regression.* New York, Wiley.

Ickovics, J.R. & Meade, C.S. (2002) Adherence to HAART among patients with HIV: breakthroughs and barriers. *AIDS Care* 14(3), pp. 309–318.

Kalichman, S.C., Ramachandran, B. & Catz, S. (1999) Adherence to combination antiretroviral therapies in HIV patients of low health literacy. *Journal of General Internal Medicine* 14(5), pp. 267–273.

Laforge, R.G., Velicer, W.F. & Owen, N. (1999) Stage distributions for five health behaviors in the United States and Australia. *Preventive Medicine* 28(1), pp. 61–74.

Macintyre, S., Hunt, K. & Sweeting, H. (1996) Gender differences in health: Are things really as simple as they seem? *Social Science and Medicine* 42(4), pp. 617–624.

Marston, B. & De Cock, K. (2004) Multivitamins, nutrition and antiretroviral therapy for HIV disease in Africa. *New England Journal of Medicine* 351(1), pp. 78–80.

Mocroft, A., Youle, M., Moore, A., Sabin, C.A., Madge, S., Lepri, A.C., Tyrer, M., Chaloner, C., Wilson, D., Loveday, C., Johnson, M.A. & Phillips, A.N. (2001) Reasons for modification and discontinuation of antiretrovirals: results from a single-treatment centre. *AIDS* 15(2), pp. 185–194.

Nord, M. & Winicki, J. (2000) Prevalence of household hunger declines in rural households. *Rural Conditions and Trends* 11(5), pp. 80–86.

O'Brien, M., Clark, R.A., Besch, C.L., Myers, L. & Kissinger, P. (2003) Patterns and correlates of discontinuation of the initial HAART regimen in an urban outpatient cohort. *Journal of Acquired Immune Deficiency Syndromes* 34(4), pp. 407–414.

Osterberg, L. & Blaschke, T. (2005) Adherence to medication. *New England Journal of Medicine* 353(5), pp. 487–497.

Paterson, D.L., Swindells, S., Mohr, J., Brester, M., Vergis, E.N., Squier, C., Wagener, M.M. & Singh, N. (2000) Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. *Annals of Internal Medicine* 133(1), pp. 21–30.

Reynolds, N.R., Testa, M.A., Marc, L.G., Chesney, M.A., Neidig, J.L., Smith, S.R., Vella, S. & Robbins, G.K. (2004) Factors influencing medication adherence beliefs and self-efficacy in persons naive to antiretroviral therapy: a multicenter, cross-sectional study. *AIDS and Behaviour* 18(2), pp. 141–150.

Sarna, A., Pujari, S., Sengar, A.K., Garg, R., Gupta, I. & Dam, J. (2008) Adherence to antiretroviral therapy and its determinants amongst HIV patients in India. *Indian Journal of Medical Research* 127(1), pp. 28–36.

Vittinghoff, E., Gliden, D.V., Shiboski, S.C. & McCulloch, S.E. (2005) *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated-Measures Models.* New York, Springer.

Wood, E. & Hogg, R.S. (2003) Effect of medication adherence on survival of HIV-infected adults who start highly active antiretroviral therapy when the CD4 cell count is 0.200 to 0.350 x 10^9 cells/L. *Annals of Internal Medicine* 139(10), pp. 810–816.

APPENDIX B: A published paper based on the application of GEE to HAART adherence data

AIDS Behav
DOI 10.1007/s10461-010-9688-x

ORIGINAL PAPER

Determinants of Optimal Adherence Over Time to Antiretroviral Therapy Amongst HIV Positive Adults in South Africa: A Longitudinal Study

Dikokole Maqutu · Temesgen Zewotir · Delia North · Kogieleum Naidoo · Anneke Grobler

Author Proof

Abstract Highly active antiretroviral therapy (HAART) requires strict adherence to achieve optimal clinical and survival benefits. A study was done to explore the factors affecting HAART adherence among HIV positive adults by reviewing routinely collected patient information in the Centre for the AIDS Programme of Research in South Africa's (CAPRISA) AIDS Treatment Programme. Records of 688 patients enrolled between 2004 and 2006 were analysed. Patients were considered adherent if they had taken at least 95% of their prescribed drugs. Generalized estimating equations were used to analyse the data. The results showed that HAART adherence increased over time, however, the rate of increase differed by some of the socio-demographic and behavioural characteristics of the patients. For instance, HAART adherence increased in both urban and rural treatment sites over time, but the rate of increase was higher in the rural site. This helped identify sub-populations, such as the urban population, that required ongoing adherence counseling.

Keywords Adherence HAART · Generalized estimating equations (GEE) · Longitudinal study · Pill count approach

Introduction

Highly active antiretroviral therapy (HAART) has dramatically reduced morbidity and mortality among HIV-infected individuals [1–3], and requires strict adherence to attain optimal clinical and survival benefits [4–6]. Patients who take 95% or more of their prescribed medication benefit more from treatment than those who take less than 95% [4]. Identifying and overcoming factors that reduce adherence to HAART is therefore critical if optimal clinical and survival benefits are to be attained. Optimal adherence to HAART is often influenced by a variety of factors, including social, demographic, economic and behavioural [7, 8]. There have, however, been inconsistent findings regarding the association between adherence, demographic and economic factors. For instance, some studies have shown an association between adherence and age [9, 10], while others did not find such a relationship [11]. Income has also been associated with HAART adherence in some studies [12, 13], while others have concluded that there was no link between the two [14]. Social factors that include family support have been consistently associated with adherence [15]. The fact that associations between adherence and demographic and economic factors are observed inconsistently highlights the need for an evaluation and understanding of how these factors interact among themselves, and how they interact with other social, clinical and behavioural factors.

In Sub-Saharan African, studies have been conducted in countries such as Ethiopia [16], Uganda [17, 18], Uganda and Zimbabwe [19], South Africa [3, 20], Senegal [12, 21, 22], Nigeria [23] and Botswana [24], to identify factors affecting adherence to HAART. Some studies were cross-sectional and did not assess adherence over time. A limitation in some studies was the lack of multivariate analyses

D. Maqutu · T. Zewotir · D. North
School of Statistics and Actuarial Science,
Private Bag X01, Scottsville 3209, South Africa

D. Maqutu (✉) · K. Naidoo · A. Grobler
CAPRISA, University of KwaZulu-Natal, Pietermaritzburg,
Private Bag X01, Scottsville 3209, South Africa
e-mail: 206521995@ukzn.ac.za

Springer

Journal : Large 10461 | Dispatch : 22-3-2010 | Pages : 10
Article No. : 9688 | LE | TYPESET
MS Code : AIBE-09-476 | CP | DISK

that included an evaluation of interaction terms between demographic, social and economic factors to control for potential confounding variables.

The cross-sectional work on the analysis of social, demographic, behavioural, economic and clinical factors influencing initial adherence for this cohort revealed that two-way interaction terms between age and cell phone ownership, gender and reason for taking an HIV test, as well as treatment site and income were important [25]. Although factors associated with optimal adherence at the initial stages of therapy give important information, factors that affect long-term optimal adherence may be more relevant, as lifetime adherence to HAART is required. It is therefore important to identify factors influencing adherence over time, as well as the relationships (interactions) between these factors. Studies that evaluate relationships between factors influencing adherence are limited in the literature, hence the motivation for this study. The aim of this study was to determine the predictors of optimal long-term adherence, and whether factors affecting initial adherence also influence long-term HAART adherence.

Methods

Study Design

The Centre for the AIDS Programme of Research in South Africa's (CAPRISA) started a HAART rollout program in 2004. The CAPRISA AIDS Treatment (CAT) Programme offers HIV care services at two sites in KwaZulu-Natal, South Africa, namely the eThekwini Clinical Research site located adjacent to the Prince Cyril Zulu Communicable Disease Clinic in the center of Durban, and the Vulindlela Clinical Research site, located in a rural area outside the town of Howick, approximately 95 km from Durban. The programme started providing free HAART through a President's Emergency Plan for AIDS Relief (PEPfAR) grant at a time when access to HAART in the public sector was limited. Adult patients with a CD4+ count below 200 cells/μl, or patients with World Health Organisation (WHO) stage 4 of the HIV disease, were eligible for HAART initiation.

During the first month, patients visited the clinic once a week for the first 2 weeks and again 2 weeks later for intensive clinical monitoring. Thereafter, patients visited the treatment sites monthly to collect their treatment and to undergo a clinical examination. Prior to HAART initiation, all patients received three sessions of adherence education, motivation and preparedness training. All patients were on regimens containing two nucleoside reverse transcriptase inhibitors and one non-nucleoside reverse transcriptase inhibitor. Patients in the urban clinic received Efavirenz (EFV), Lamivudine (3TC) and Didanosine (ddI or ddI-EC). This regimen was chosen as it can be co-administered with anti-tuberculosis (TB) medication. The regimen in the rural clinic consisted of EFV, 3TC and Stavudine (d_4T), which is recommended according to the South African HIV treatment guidelines [26]. A few pregnant patients (3.8%) received Nevirapine (NVP) rather than EFV.

Patient information was recorded on data collection sheets at the clinics; it underwent two levels of quality control, and was faxed to a central data management centre. The data analysed in this study consisted of a retrospective review of patients' records in the CAT programme between June 2004 and September 2006. Only patients with pill count data for the initial visit, and at least one other clinic visit for the defined study period, were included in the analysis. The number of follow-up visits differed per patient, as some patients started treatment earlier and therefore had more visits, while some patients dropped out of the treatment programme prematurely. Approval for the data collection and analysis was obtained from the University of KwaZulu-Natal Biomedical Research Ethics Committee.

Model Variables and Measurements

The data routinely collected at the treatment sites included the patients' demographic details, medical history, CD4+ cell count and viral load, pill counts, clinical status and a laboratory safety assessment. For this analysis, the outcome was measured using the pill counts data, and the covariates were selected from the various aspects of the collected data. The covariates can be classified as time-independent and time-varying. Measurement and coding of the outcome and covariates is described below.

Adherence Measurement

Adherence to HAART was measured using pill counts conducted by pharmacists at the treatment sites. Patients were provided with more medication than required, i.e. tablets were usually dispensed in multiples of 30, whereas visits were booked in multiples of 28 days. Patients were asked to bring all medication bottles and unused pills to each clinic visit, but were not told that the returns were to be counted. Adherence at every visit (monthly, i.e. 28 days) for all the drugs was calculated as the total number of drugs dispensed, minus the total number of drugs returned, divided by the total number of days between clinic visits, times the daily dose. This method takes into account the date of the clinic visits and adjusts for late return to the clinic (meaning pills would have been missed) and early return to clinic (meaning pills would still be available for return).

Journal : Large 10461 | Dispatch : 22-3-2010 | Pages : 10
Article No. : 9688 | LE | TYPESET
MS Code : AIBE-09-476 | CP | DISK

Outcome Variable

The outcome was optimal adherence to HAART, and patients were classified as optimally adherent if they took at least 95% of the prescribed drugs in a given regimen [4], otherwise, they were considered non-adherent. At each visit, the response variable was binary, indicating whether a patient was optimally adherent or not (1 = adherence score of 95% and above, i.e. the patient was optimally adherent; 0 = adherence score less than 95%, i.e. the patient was not optimally adherent).

Time Independent Covariates

As adherence was monitored very closely for the first month after initiation of HAART (three visits), compared to subsequent visits, the adherence for the first month (initial adherence) was treated as a covariate in the analysis. It has been argued that the use of a baseline response as a covariate in a longitudinal study permits the use of each subject as their own control to assess the effect of treatment over time [27, 28]. Initial adherence was calculated as the sum of drugs dispensed in the first month, minus the sum of drugs returned in the first month (over the three visits), divided by the number of days between the first and fourth visit multiplied by the daily dose. Those who took 95% of the prescribed dose were considered adherent (1 = optimal initial adherence; 0 = not optimal initial adherence).

Baseline demographic and socio-economic variables included in this analysis were age (in years); gender (1 = female, 0 = male); educational status (2 = no schooling, 1 = primary and 0 = secondary and higher); treatment site (1 = urban, 0 = rural); whether or not a patient lived with a partner (1 = living with a partner, 0 = not living with a partner); whether or not the patient was the source of household income (1 = source of household income, 0 = not a source of household income); access to tap water (1 = yes, 0 = no) and electricity (1 = yes, 0 = no), and whether a patient owned a cell phone (1 = yes, 0 = no). Other variables recorded at baseline and included were WHO HIV stage (3 = stage 1, 2 = stage 2, 1 = stage 3 and 0 = stage 4), CD4+ cell count (cells/μl), weight (in kilograms). Patients were asked why they did an HIV test and their responses included being unwell, testing for no specific reason, testing because a partner died of HIV, being ill and unfaithfulness. Reason for testing was therefore classified as follows: (2 = possible exposure to HIV, 1 = no specific reason and 0 = unwell).

Time Varying Covariates

Time was measured as a continuous variable representing monthly follow-up visits to the treatment site. The variable time starts with the value 1 for the first follow-up visit, 2 for the second visit, up to 17 for the seventeenth follow-up visit. Weight was measured at every follow-up visit and was modeled as a time-varying covariate.

Data Analysis

Differences in the baseline characteristics of CAT patients excluded and included in the analysis were compared using the chi-squared test for categorical variables, and the *t*-test for the continuous variables. Socio-demographic and clinical characteristics of the study population were summarised using the median and inter-quartile range (IQR) for continuous variables and proportions for categorical variables. The trend of optimal adherence over time was tested using the Cochran-Armitage test for trend [29].

To account for multiple measurements of each patient (correlated data), and to allow all patients, regardless of the number of visits, to be included in the analysis, generalized estimating equations (GEEs) [30] were used to assess factors associated with optimal adherence over time. Quasi-likelihood function value obtained under the independence model criterion (QIC) [31] was used to select the model with the best fitting correlation structure. Model selection was done by first including all predictor variables in the model and then evaluating whether any interaction terms needed to be incorporated into the model. This was determined by fitting each of the product terms formed from the predictor variables, one at a time, and keeping the significant terms in the model. As GEE parameters are estimated using quasi-likelihood procedures, there is no associated likelihood underlying the model. Therefore, the usual likelihood ratio tests could not be applied to compare models. However, the generalized Wald test was used for model comparison [32].

The GEE method accommodates missing data, however, it yields valid estimates if missing data can be assumed to be 'missing completely at random' (MCAR). This means that missing values do not depend on the observed (previous) or unobserved outcomes [33, 34]. A logistic regression model was used to assess whether dropout in this study is independent of the previous outcome, i.e. whether the missing data mechanism is MCAR. All statistical tests were conducted at a 5% level of significance and analyses were done using SAS (version 9.1.3).

Results

Study Population

Between June 2004 and September 2006, 1,184 patients were enrolled in the CAT programme, 411 (35%) at the

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

urban site, and 773 (65%) at the rural site. A total of 688 patients, 369 (54%) from the urban site and 319 (46%) from the rural site were included in the analysis. Only patients with pill count data for the initial and at least one other clinic visit were included in the analysis. There were no differences between those included in the study and those excluded with regard to age (mean: included = 34.1 years, excluded = 34.0 years; t-value = 0.13, $P = 0.90$), gender (males: included = 30.0%, excluded = 31.8%; chi-square-value = 0.43, P=0.51) and baseline CD4+ cell count (mean: included = 107.6 cells/μl, excluded = 111.5 cells/μl; t-value = 0.72, $P = 0.47$). Furthermore, power calculations were performed for the available sample size (688 patients) to detect a difference in proportion of adherent patients between the first follow-up visit and the final follow-up visit of 0.27. With a sample size of 600, this gave more than 90% power when a test of proportions was done.

Baseline Socio-Demographic and Clinical Characteristics of Patients

The baseline socio-demographic and clinical characteristics of patients included in the analysis are presented in Table 1. The median age of patients was 32.5 years (IQR: 28–38 years), 70% were male and 75% were not living with a partner. Over two-thirds of the patients had attained secondary or higher level of education (69%), and 28% of patients were classified as breadwinners. Over 90% of the patients stayed in households that had access to tap water and electricity, while 42% of the households had cell phones. At enrolment, the median weight was 60 kg (IQR: 53–69 kg), median CD4+ cell count was 108 cells/μl (IQR: 52–159 cells/μl) and 64% of patients were classified as WHO stage 3. Over half of the patients (56%) reported to have taken an HIV test as they were not well, while 26% reported no specific reason for testing and 18% took an HIV test as they were concerned that they had been exposed to HIV. In the initial month of treatment, 79% of the patients were at least 95% adherent to HAART.

Table 1 Baseline socio-demographic and clinical characteristics of the patients

Characteristics	Value
Age (years)	
Median (IQR)	32.5 (28, 38)
Gender, *n* (%)	
Male	206 (30)
Female	482 (70)
Education, *n* (%)	
No schooling	74 (12)
Primary	116 (19)
Secondary and higher	429 (69)
Treatment site, *n* (%)	
Urban site	369 (54)
Rural site	319 (46)
Source of household income, *n* (%)	
Source of income	186 (28)
Not source of income	489 (72)
Living with or without a partner, *n* (%)	
Living with a partner	168 (25)
Living without a partner	510 (75)
Household with or without electricity, *n* (%)	
Yes	607 (91)
No	63 (9)
Access to tap water, *n* (%)	
Yes	611 (91)
No	59 (9)
Cell phone ownership, *n* (%)	
Yes	281 (42)
No	389 (58)
WHO stage of HIV disease, *n* (%)	
Stage 1	71 (10)
Stage 2	121 (16)
Stage 3	438 (64)
Stage 4	58 (8)
Reason for taking HIV test, *n* (%)	
Unwell	374 (56)
Possible exposure to HIV	170 (26)
No specific reason	121 (18)
Initial (baseline) optimal adherence, *n* (%)	
Optimally adherent	546 (79)
Not optimally adherent	142 (21)

Overall Optimal Adherence and Follow-up Visits

The number of follow-up visits ranged between 2 and 17 per patient, with the median of 8 (IQR: 5–12). Figure 1 presents the total number of patients expected at each visit, the number of patients who actually attended the visit and those who dropped out at each visit. The proportion of patients who dropped out gradually increased over the follow-up period.

The proportion of patients who were at least 95% adherent (optimally adherent) to HAART increased from 58% at the first follow-up visit to 86% at the last follow-up visit (Fig. 2). The Cochran-Armitage statistic [29] ($Z = 17.52$ and $P < 0.0001$) provides strong evidence of an increasing adherence rate over time.

Selection of the Multivariate Model

Using the QIC [31], the correlation structure that fitted the data well was First-Order Autoregressive (AR1) structure,

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

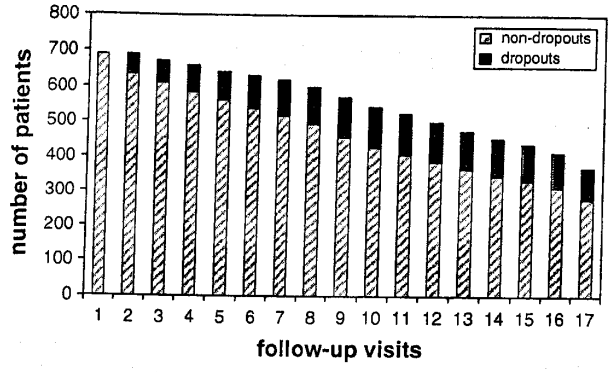


Fig. 1 Total number of patients expected at every visit classified by the number of patients who actually attended the clinic (non-dropouts) and those who dropped out

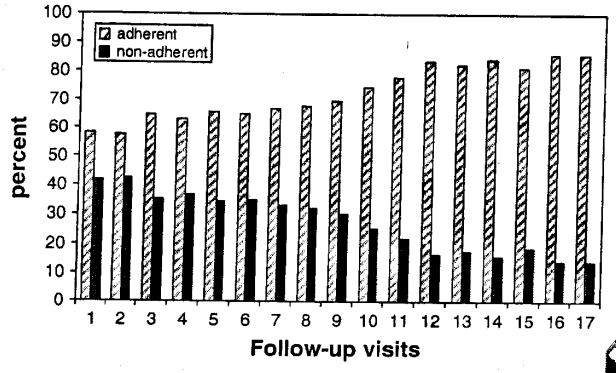


Fig. 2 Optimal adherence and non-adherence rates over the follow-up visits

which reinforces the assumption of equally spaced measurement occasions.

Cross sectional analysis of predictors of initial optimal adherence showed that two-way interaction terms between age and cell phone ownership, gender and reason for taking an HIV test, and treatment site and income were significant [25]. Using the same data set, we evaluated whether these predictors of initial adherence also predicted long-term adherence using the GEE method. The results were not significant (Wald statistic = 4.78 with four degrees of freedom, $P = 0.31$), indicating no association between long-term optimal adherence and the interaction terms associated with initial adherence.

To further examine the relationship between long-term optimal adherence and other variables, a model was built by fitting all the predictor variables (listed in Table 1), followed by an assessment of whether any interaction terms need to be incorporated into the model. Consequently, the final model contained five two-way interaction terms and all the predictor variables. The two-way interactions were between gender and time, treatment site and time, reported reason for taking an HIV test and time, age and gender, as well as age and educational attainment (Table 2).

Table 2 Adjusted odds ratio (aOR) from the GEE model with associated 95% confidence intervals (CI)

Parameter	aOR (95% CI)	*P*-value
Intercept	1.802 (0.832, 3.903)	0.146
Age	0.988 (0.972, 1.004)	0.147
Gender (ref = male)		
Female	0.030 (0.155, 0.591)	0.001
Education (ref = secondary and above)		
No schooling	5.038 (1.837, 13.82)	0.002
Primary	1.104 (0.378, 3.223)	0.856
Treatment site (ref = rural site)		
Urban site	3.[illegible] (2.355, 4.438)	<0.001
Source of household income (ref = source)		
Not source of income	1.008 (0.828, 1.229)	0.934
Access to tap water (ref = no)		
Yes	1.114 (0.831, 1.494)	0.469
Household with electricity (ref = no)		
Yes	1.004 (0.786, 1.283)	0.975
Cell phone ownership (ref = no)		
Yes	1.260 (1.056, 1.503)	0.010
Reason for taking an HIV test (ref = unwell)		
No specific reason	0.973 (0.685, 1.381)	0.877
Possible exposure to HIV	0.615 (0.431, 0.877)	0.007
Staying with a partner (ref = no)		
Yes	1.335 (1.097, 1.625)	0.004
WHO staging of HIV disease (ref = stage 4)		
Stage 1	0.716 (0.486, 1.054)	0.091
Stage 2	0.778 (0.554, 1.092)	0.147
Stage 3	0.898 (0.660, 1.221)	0.492
Baseline CD4+ cell count	1.000 (0.998, 1.001)	0.532
Initial optimal adherence (ref = not adherent)		
Adherent	0.869 (0.704, 1.072)	0.190
Time (visit)	1.105 (1.059, 1.153)	<0.001
Baseline weight	0.996 (0.981, 1.010)	0.557
Weight at follow-up visits	1.001 (0.987, 1.015)	0.890
Time × gender (ref = male)		
Female	1.074 (1.034, 1.116)	<0.001
Time × treatment site (ref = rural site)		
Urban site	0.945 (0.908, 0.982)	0.004
Time × reason for taking the test (ref = unwell)		
No specific reason	0.989 (0.947, 1.033)	0.624
Possible exposure to HIV	1.058 (1.010, 1.108)	0.018
Age × gender (ref = male)		
Female	1.024 (1.006, 1.043)	0.010
Age × education (ref = secondary and above)		
No schooling	0.968 (0.943, 0.993)	0.012
Primary	0.998 (0.972, 1.025)	0.875

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

Predictors of Optimal Adherence

After controlling for other variables in the model, optimal adherence was significantly higher when patients had cell phones than when they did not have cell phones [adjusted odds ratio (aOR) = 1.260, 95% confidence interval (CI): (1.056, 1.503), $P = 0.010$] and when they lived with a partner compared to when they did not live with a partner [aOR = 1.335, 95% CI: (1.097, 1.625), $P = 0.004$] (Table 2).

Optimal HAART adherence increased on average over time, however, since interactions with time were significant, the rate at which optimal adherence increased differed by treatment site, gender and the patient's reported reason for taking an HIV test. Age interacted significantly with gender and education. The interaction effects are presented below.

Interaction Between Gender and Time

Optimal HAART adherence increased over time for both males and females. However, since there was a significant interaction between gender and time, the rate of increase was not the same for males and females after controlling for other covariates in the model. The rate of increase was 7.4% higher for females than for males [aOR = 1.074, 95% CI: (1.034, 1.116); $P < 0.001$] (Table 2). Estimated probabilities of optimal adherence were higher for males at the beginning of the follow-up visits, but by the end of the study period (17th follow-up visit), they were similar for both groups (Fig. 3a).

Interaction Between Treatment Site and Time

The rate at which optimal adherence increased over time differed in the urban and rural treatment sites. After controlling for other variables in the model, the rate of increase in optimal adherence was 6% higher in the rural treatment site than in the urban treatment site [aOR = 1.06, 95% CI: (1.002, 1.100); $P = 0.004$] (Table 2). The estimated probability of optimal adherence at the first follow-up visit was 66% at the rural site and 86% at the urban site (Fig. 3b). Since the rate of increase was higher in the rural site relative to the rate of increase in the urban site, the gap in adherence between the treatment sites gradually decreased over time until, by the end of the study, the estimated optimal adherence probabilities were similar (at 91 and 92%, respectively).

Interaction Between Reason for Taking an HIV Test and Time

Optimal adherence increased over time, but the rate at which it increased differed with the patient's reported reason for taking an HIV test. The rate of increase in optimal adherence was 5.8% higher over the study period for patients who tested due to possible exposure to HIV, than for patients who tested as they were unwell [aOR = 1.058, 95% CI: (1.010, 1.108); $P = 0.016$] (Table 2). There was, however, no significant difference in the rate of change of optimal adherence between patients who tested because they were unwell and those who reported no specific reason for taking an HIV test [$Z = -0.43$, $P = 0.666$] (Table 2). Further analysis revealed that the rate of increase in optimal adherence over the study period was 7% higher for patients who tested due to possible exposure to HIV, than for those who reported no specific reason for taking an HIV test [aOR = 1.069, 95% CI: (1.016, 1.126), $P = 0.0107$].

Figure 3c indicates that the estimated probabilities of optimal adherence for patients who tested as they were unwell, and those who reported no specific reason for taking an HIV test, were similar throughout the study.

It is again shown in Fig. 3c that at the beginning of the follow-up period, estimated probabilities of optimal adherence for patients who tested due to possible exposure to HIV were less than the estimated probabilities of patients who reported no specific reason for taking an HIV test, as well as those who tested as they were unwell. However, during the middle of the follow-up visits, estimated probabilities were similar for all the reported reasons. Towards the end of the study, the probabilities of optimal adherence were higher for patients who tested due to possible exposure to HIV than those who reported no specific reason for testing for HIV, or those who tested as they were unwell.

Interaction Between Age and Gender

Optimal HAART adherence differed by age for males and females. As the age of patients increased, females tend to adhere better to HAART than males [aOR: 1.024; 95% CI: (1.006, 1.043), $P = 0.010$]. It is shown in Fig. 3d that the estimated probabilities of optimal adherence were higher with younger males than with younger females, whereas with older patients, estimated probabilities were higher with females than males.

Interaction Between Age and Education

Optimal HAART adherence differed by age at different education levels. Among older patients, those with no schooling were less likely to achieve optimal HAART adherence than those with secondary and higher education [OR = 0.97; 95% CI: (0.94, 0.99); $P = 0.012$] (Table 2). There was, however, no significant difference in optimal HAART adherence between patients with secondary education and patients with primary education, regardless of

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

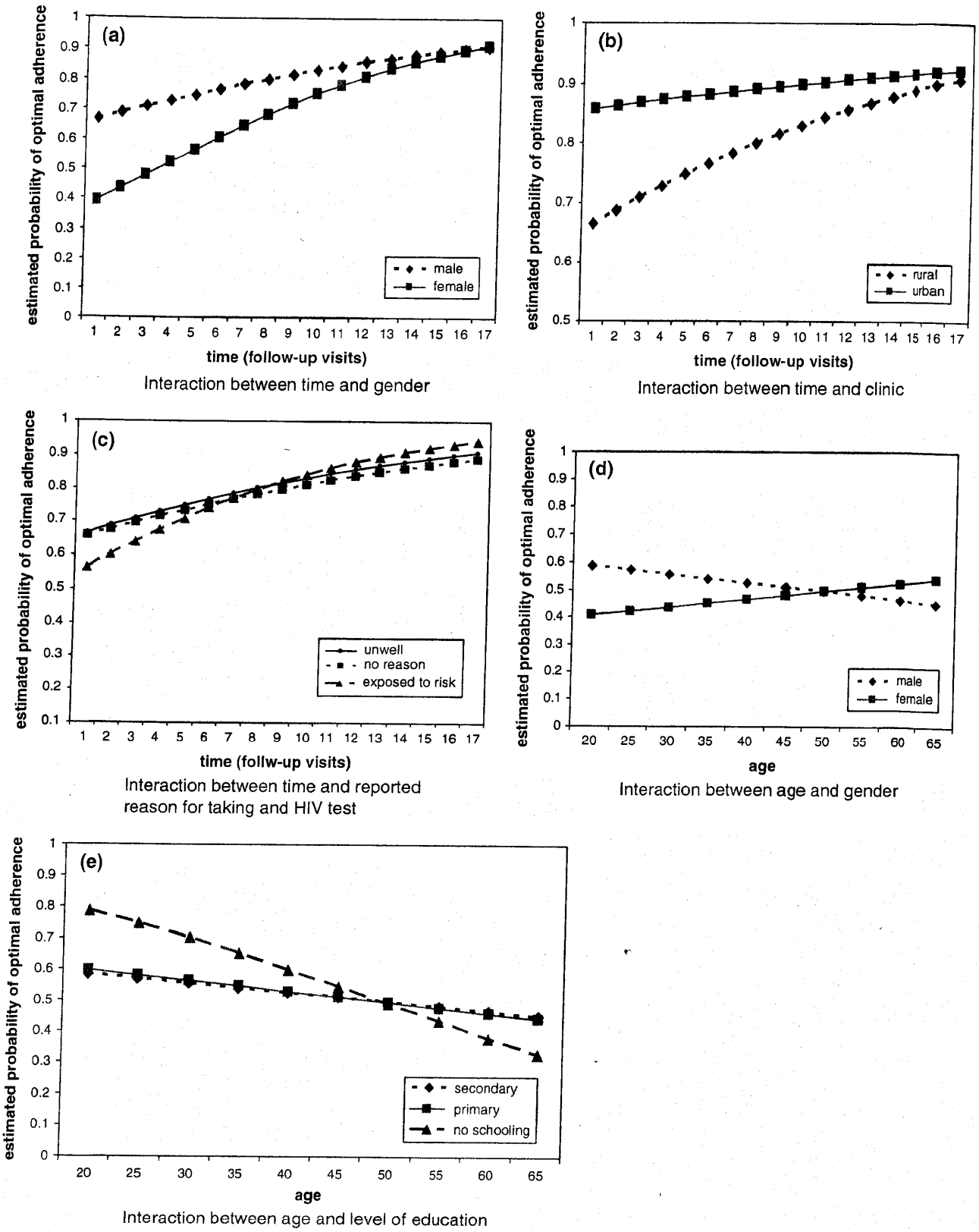


Fig. 3 All the two-way interaction terms in the final model of optimal adherence

Springer

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

age [$Z = -0.16$, $P = 0.875$)] (Table 2). Further analysis revealed that as patients got older, those with primary education were more likely to achieve optimal adherence than those with no schooling [aOR = 1.03, 95% CI: (1.002, 1.070); $P = 0.048$]. Figure 3e indicates that the estimated probabilities of optimal adherence by patients decreased with age. More specifically, the probability of optimal adherence for patients with secondary and primary education was similar for all ages, whereas probabilities of optimal adherence for those with no schooling were higher for younger ages and lower for older ages, relative to those with primary and secondary schooling.

Impact of Dropout on the Analysis

The analysis revealed that dropout did not depend on previous outcome (Chi-squared-value = 0.103 with one degree of freedom; $P = 0.748$). It can therefore be concluded that MCAR holds and as a result, GEE method yielded valid estimates.

Discussion

Consistent with results from recent longitudinal studies in Sub-Saharan Africa [16, 19], the findings showed that optimal HAART adherence increased with time, indicating that programmatic interventions conducted during the CAT programme may have had a positive impact on adherence. These interventions included patients being exposed to ongoing peer education when queuing for services within the clinic, and through one-on-one session with a peer educator at the end of the clinic visit after the pills were collected. Every 6 months, patients received individual education on adherence maintenance. This highlights the importance of prioritising adherence counseling when setting up new HAART clinics. These results reinforce the message that it is possible to achieve and sustain high levels of HAART adherence in Sub-Saharan Africa, given strong support measures.

Studies have consistently shown that demographic characteristics are generally poor predictors of HAART adherence [35, 36]. However, the results from this study demonstrate that demographic factors predict HAART adherence through interactions among themselves or with other variables. For instance, an interaction between age and educational attainment predict HAART adherence. More specifically, as patients get older, adherence decreases over time, and the decrease is more pronounced with patients who have no schooling and in male patients.

HAART adherence increased over time but at different rates in males and females. Males had higher adherence at the beginning of the follow-up period but by the end of the study period, adherence was similar between males and females. This might be explained by social and behavioural factors associated with HAART adherence being different in males and females [37]. For instance, women are generally responsible for care-giving duties which might hamper adherence at the beginning of treatment, until they get used to the new routine of taking their own medication [37, 38].

HAART adherence has been higher in urban clinics than in rural clinics, which has been attributed to the latter being less well resourced [39]. However, the rate at which adherence increases with time is higher in the rural clinic than in the urban clinic. Studies have shown that in rural settings, optimal adherence tends to increase over time [40]. Since rural populations live in a more communal setting, the positive impact of HAART on HIV positive-patients might be more visible, which may positively influence optimal adherence to HAART.

Patients who lived with a partner adhered better to HAART than those who lived with no partner. Lack of social support has been associated with a decrease in adherence [41] and living with a partner has been associated with increased social support and optimal adherence [42]. Furthermore, cell phone ownership enhanced long-term optimal HAART adherence. This reinforces proposed interventions of providing memory aids for dosing times that include the use of new technologies such as reminders through cell phones [43–47], and verbal or text messaging for check-up reminders and general adherence messages [48].

Three shortcomings are acknowledged in our research. Firstly, adherence was assessed only through pill counts. Pill counts method is attractive due to its simplicity and empirical nature; however, it has disadvantages that include patients switching medicines between bottles or discarding pills before visits [47]. Despite these problems, it has been shown that pill counts method has a strong linear relationship with viral load [49].

Secondly, interactions between variables were identified using the data and model fit techniques. The interactions were not pre-specified or expected during data collection. Detailed information on why these interactions influenced adherence was therefore not collected, and the reasons for some of these findings cannot be explained.

Thirdly, an equal interval (monthly) between successive follow-up visits by all patients was assumed. In reality, some patients visited the clinic a number of days earlier or later than scheduled appointments. However, the impact of this assumption on the analysis is negligible for a number of reasons. The patients in the CAT programme have shown excellent clinic attendance with scheduled appointment. In addition, patients were given 2-day buffer stock with each supply to cover appointment delays and

Journal : Large 10461 Dispatch : 22-3-2010 Pages : 10
Article No. : 9688 LE TYPESET
MS Code : AIBE-09-476 CP DISK

this was accounted for in calculating the adherence rate. Generally, when we say monthly data, we do not expect an equal number of days in every month. Furthermore, the effect of unequally spaced interval correlation structures was tested using QIC and was found to be insignificant, which reinforces the assumption of equal interval between clinic visits.

The study showed that HAART adherence increased over time on average, however, the rate at which it increased differed by treatment site, gender and the patient's reported reason for taking an HIV test. Consequently, due attention should be paid to address the specific needs of each group of patients, specifically with respect to urban–rural and gender differentials. HAART adherence over time increased for patients with cell phones and patients living with a partner, and decreased as age of patients' with no schooling increased. HAART programs need to take these factors into account in the design and implementation of long-term adherence strategies. Specific groups of patients at risk for less than optimal adherence should be targeted with long-term adherence boosting sessions tailored to their specific needs.

Acknowledgments CAPRISA was established in 2002 through a Comprehensive International Program of Research on AIDS (CIPRA) grant (AI51794) from the US National Institutes of Health (NIH), as a multi-institutional collaboration, incorporated as an independent non-profit AIDS Research Organization. The NIH funded the development of the research infrastructure, including the data management, laboratory and pharmacy cores established through the CIPRA grant. The US President's Emergency Plan for AIDS Relief (PEPfAR) grant (1U2GPS001350) funded the care of all the patients in the CAT Programme. Dikokole Maqutu was supported by and Anneke Grobler received career development support from the Columbia University-Southern African Fogarty AIDS International Training and Research Programme (AITRP) funded by the Fogarty International Center, National Institutes of Health (Grant # D43TW00231). We gratefully acknowledge the patients in the CAT Programme. We also thank all the staff who worked on the CAT Programme, treating patients and helped in the data collection. Special thanks to the pharmacists for collection of pill count data.

References

1. Hogg RS, Heath KV, Yip B, et al. Improved survival among HIV-infected individuals following initiation of antiretroviral therapy. JAMA. 1998;279(6):450–4.
2. Mocroft A, Ledergerber B, Katlama C, et al. Decline in the AIDS and death rates in the EuroSIDA study: an observational study. Lancet. 2003;362(9377):22–9.
3. Nachega JB, Stein DM, Lehman DA, et al. Adherence to antiretroviral therapy in HIV-infected adults in Soweto, South Africa. AIDS Res Hum Retroviruses. 2004;20(10):1053–6.
4. Paterson DL, Swindells S, Mohr J, et al. Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Ann Intern Med. 2000;133(1):21–30.
5. Bangsberg DR, Perry S, Charlebois ED, et al. Non-adherence to highly active antiretroviral therapy predicts progression to AIDS. AIDS. 2001;15(9):1181–3.
6. Knobel H, Guelar A, Valldecillo G, et al. Response to highly active antiretroviral therapy in HIV-infected patients aged 60 years or older after 24 months follow-up. AIDS. 2001;15(12):1591–3.
7. Chesney MA. Factors affecting adherence to antiretroviral therapy. Clin Infect Dis. 2000;30(Suppl 2):S171–6.
8. Ferguson TF, Stewart KE, Funkhouser E, Tolson J, Westfall AO, Saag MS. Patient-perceived barriers to antiretroviral adherence: associations with race. AIDS Care. 2002;14(5):607–17.
9. Murphy DA, Marelich WD, Hoffman D, Steers WN. Predictors of antiretroviral adherence. AIDS Care. 2004;16(4):471–84.
10. Penedo FJ, Gonzalez JS, Dahn JR, et al. Personality, quality of life and HAART adherence among men and women living with HIV/AIDS. J Psychosom Res. 2003;54(3):271–8.
11. Roca B, Lapuebla C, Vidal-Tegedor B. HAART with didanosine once versus twice daily: adherence and efficacy. Int J Infect Dis. 2005;9(4):195–200.
12. Laniece I, Ciss M, Desclaux A, et al. Adherence to HAART and its principal determinants in a cohort of Senegalese adults. AIDS. 2003;17(Suppl 3):S103–8.
13. Kleeberger CA, Phair JP, Strathdee SA, Detels R, Kingsley L, Jacobson LP. Determinants of heterogeneous adherence to HIV-antiretroviral therapies in the Multicenter AIDS Cohort Study. J Acquir Immune Defic Syndr. 2001;26(1):82–92.
14. Mohammed H, Kieltyka L, Richardson-Alston G, et al. Adherence to HAART among HIV-infected persons in rural Louisiana. AIDS Patient Care STDS. 2004;18(5):289–96.
15. Ammassari A, Trotta MP, Murri R, et al. Correlates and predictors of adherence to highly active antiretroviral therapy: overview of published literature. J Acquir Immune Defic Syndr. 2002;31(Suppl 3):S123–7.
16. Amberbir A, Woldemichael K, Getachew S, Girma B, Deribe K. Predictors of adherence to antiretroviral therapy among HIV-infected persons: a prospective study in Southwest Ethiopia. BMC Public Health. 2008;8:265.
17. Bangsberg D, Emenyonu N, Andia I, et al. No decline in high levels of electronic pill cap, unannounced home pill count, and patient-reported adherence to free ARV therapy over 12 months in rural Uganda. In: 15th conference on retroviruses and opportunistic infections. Boston, Massachusetts, 3–6 Feb 2008 [abstract no. 801]. http://www.retroconference.org/2008/index.asp. Accessed 4 Feb 2010.
18. Byakika-Tusiime J, Oyugi JH, Tumwikirize WA, Katabira ET, Mugyenyi PN, Bangsberg DR. Adherence to HIV antiretroviral therapy in HIV+ Ugandan patients purchasing therapy. Int J STD AIDS. 2005;16(1):38–41.
19. Muyingo SK, Walker AS, Reid A, et al. Patterns of individual and population-level adherence to antiretroviral therapy and risk factors for poor adherence in the first year of the DART trial in Uganda and Zimbabwe. J Acquir Immune Defic Syndr. 2008;48(4):468–75.
20. Orrell C, Bangsberg DR, Badri M, Wood R. Adherence is not a barrier to successful antiretroviral therapy in South Africa. AIDS. 2003;17(9):1369–75.
21. Etard JF, Laniece I, Fall MB, et al. A 84-month follow up of adherence to HAART in a cohort of adult Senegalese patients. Trop Med Int Health. 2007;12(10):1191–8.
22. Laurent C, Diakhate N, Gueye NF, et al. The Senegalese government's highly active antiretroviral therapy initiative: an 18-month follow-up study. AIDS. 2002;16(10):1363–70.
23. Iliyasu Z, Kabir M, Abubakar IS, Babashani M, Zubair ZA. Compliance to antiretroviral therapy among AIDS patients in Aminu Kano Teaching Hospital, Kano, Nigeria. Niger J Med. 2005;14(3):290–4.
24. Weiser S, Wolfe W, Bangsberg D, et al. Barriers to antiretroviral adherence for patients living with HIV infection and AIDS in Botswana. J Acquir Immune Defic Syndr. 2003;34(3):281–8.

Journal : Large 10461 | Dispatch : 22-3-2010 | Pages : 10
Article No. : 9688 | LE | TYPESET
MS Code : AIBE-09-476 | CP | DISK

25. Maqutu D, Zewotir T, North D, Naidoo K, Grobler A. Factors affecting early adherence to antiretroviral therapy amongst the HIV positive adults in South Africa. In: 50th annual conference of the South African statistical association, 29 Oct–2 Nov 2007. Johannesburg, South Africa.
26. South African National Department of Health. National antiretroviral treatment guidelines. 1st ed. (2004). http://www.hst.org.za/uploads/files/sa_ART_gudelines1.pdf/. Accessed 7 May 2009.
27. Littell RC, Milliken GA, Stroup WW, Wolfinger DR, Schabenberger O. SAS system for mixed models. 2nd ed. Cary, NC: SAS Institute Inc; 1999.
28. Milliken GA, Johnson E. Analysis of messy data. New York: Chapman & Hall/CRC; 2002.
29. Agresti A. Categorical data analysis. 2nd ed. New York: Wiley; 2002.
30. Liang K-Y, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
31. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001;57:120–5.
32. Hedeker D, Gibbons R. Longitudinal data analysis. New York: Wiley; 2006.
33. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
34. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.
35. Fong OW, Ho CF, Fung LY, et al. Determinants of adherence to highly active antiretroviral therapy (HAART) in Chinese HIV/AIDS patients. HIV Med. 2003;4(2):133–8.
36. Haubrich RH, Little SJ, Currier JS, et al. The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response California Collaborative Treatment Group. AIDS. 1999;13(9):1099–107.
37. Berg KM, Demas PA, Howard AA, Schoenbaum EE, Gourevitch MN, Arnsten JH. Gender differences in factors associated with adherence to antiretroviral therapy. J Gen Intern Med. 2004;19(11):1111–7.
38. Skhosana N, Struthers H, Gray G, McIntyre J. HIV disclosure and other factors that impact on adherence to antiretroviral therapy: the case of Soweto, South Africa. AJAR. 2006;5(1):17–26.
39. Harvey KM, Carrington D, Duncan J, et al. Evaluation of adherence to highly active antiretroviral therapy in adults in Jamaica. West Indian Med J. 2008;57(3):293–7.
40. Birbeck GL, Chomba E, Kvalsund M, et al. Antiretroviral adherence in rural Zambia: the first year of treatment availability. Am J Trop Med Hyg. 2009;80(4):669–74.
41. Williams A, Friedland G. Adherence, compliance, and HAART. AIDS Clin Care. 1997;9(7):51–4, 8.
42. Godin G, Cote J, Naccache H, Lambert LD, Trottier S. Prediction of adherence to antiretroviral therapy: a one-year longitudinal study. AIDS Care. 2005;17(4):493–504.
43. Bartlett JA. Addressing the challenges of adherence. J Acquir Immune Defic Syndr. 2002;29(Suppl 1):S2–10.
44. Ickovics JR, Meade CS. Adherence to HAART among patients with HIV: breakthroughs and barriers. AIDS Care. 2002;14(3):309–18.
45. Abel E, Painter L. Factors that influence adherence to HIV medications: perceptions of women and health care providers. J Assoc Nurses AIDS Care. 2003;14(4):61–9.
46. Nachega JB, Hislop M, Dowdy DW, Chaisson RE, Regensberg L, Maartens G. Adherence to nonnucleoside reverse transcriptase inhibitor-based HIV therapy and virologic outcomes. Ann Intern Med. 2007;146(8):564–73.
47. Osterberg L, Blaschke T. Adherence to medication. N Engl J Med. 2005;353(5):487–97.
48. Chang LW, Kagaayi J, Nakigozi G, et al. Telecommunications and health care: an HIV/AIDS warmline for communication and consultation in Rakai, Uganda. J Int Assoc Physicians AIDS Care (Chic Ill). 2008;7(3):130–2.
49. Bangsberg DR, Hecht FM, Charlebois ED, et al. Adherence to protease inhibitors, HIV-1 viral load, and development of drug resistance in an indigent population. AIDS. 2000;14(4):357–66.

Journal : Large 10461 | Dispatch : 22-3-2010 | Pages : 10
Article No. : 9688 | LE | TYPESET
MS Code : AIBE-09-476 | CP | DISK