LONGITUDINAL ANALYSIS OF INCIDENCE, RISK FACTORS AND OUTCOMES AMONG TB/HIV COINFECTED CHILDREN

Odhiambo Collins Ojwang

Reg. No: I56/77141/2009

University of Nairobi

College of Biological and Physical Sciences

School of Mathematics

In Partial Fullfillment of the Requirement for the award of the degree

of

Master of Science in Biometry

Declaration

This dissertation is my original work and has not been presented for a degree in any other university .

Odhiambo, Collins Ojwang'

I56/77141/2009

Signature _____ Date 01/08/2011

Declaration by Supervisor

This dissertation has been submitted for examination with our approval as supervisors

Mr. James Ngugi Mwangi

Signature _____ Date 02/08/2011

Dr. Nelson Owuor

Signature _____ Date 02/08/2011

Contents

List of Tables

List of Figures

Abbreviations and Acronyms

TB⇒ Tuberculosis

HAART⇒ Highly Active Antiretroviral Therapy

CD4⇒ cluster of differentiation 4

CDC⇒ Center for Desease Control

PEPFAR⇒ President's Emergency Plan for AIDS Relief

PMTCT⇒ preventing mother-to-child transmission

VCT⇒voluntary counseling and testing

DNA PCR⇒ Deoxyribonucleic acid polymerase chain reaction

BMI⇒ Body Mass Index

EMR⇒ Electronic medical record

ART⇒ Antiretroviral Therapy

PYFU⇒ Person-years of follow-up

GEE⇒ Generalized Estimation Equation

PEP⇒Post Exposure Prophylaxis

PCR -VE⇒ polymerase chain reaction Negative

WHO⇒ World Health Organisation

AFBS⇒Acid-Fast Bacillus (AFB) Smear

PTB⇒ Extra-Pulmonary Tuberculosis

EPTB⇒Pulmonary Tuberculosis

CI⇒Confidence Interval

NVP⇒Nevirapine

EFV⇒ Efavirenz

LPV/r⇒Lopinavir/ritonavir

NRTI⇒ Nucleotide Reverse Transcriptase Inhibitors

CxR⇒Chest X-ray

Acknowledgement

I am very grateful to my Employer University of Nairobi, School of Medicine PACT-Coe for providing data for analysis for my Masters of Science degree. They also granted me study leave in order to concentrate on my study. I would also like to thank all people and institutions that have contributed to my success of my Masters degree in Biometry. May I thank the teaching and non-teaching staff in The School of Mathematics, University of Nairobi for their support. I would like to recognize the contribution of all my M.Sc. Lecturers (The late Prof. J.O. Owino, Prof. M. Manene, Prof. J.A.M Otieno, Dr. P. Weke, Dr. T. Achia, Mr. J.N. Mwangi, and Mr. E. Obudho)
May I single out my supervisors (Mr. J.N. Mwangi and Dr. Nelson Owuor) for their guidance right from the start of my project formulation.
I appreciate the support from my family (my wife Sarah Ojwang') Above all, I thank The Almighty God for good health throughout the study period.

Dedication

I dedicate this degree to my wife Sarah Ojwang for her outstanding support, prayers, encouragement and nurturing conducive environment during my study period.

Abstract

Tuberculosis (TB) being the most common opportunistic infection in human immunodeficiency virus (HIV)-infected people worldwide has attracted a lot of interest in research. TB manifestations are more severe in HIV-positive children and progression to death is more rapid than in HIV-negative children. TB also hastens the progression of HIV disease by increasing viral replication and reducing CD4 counts further. We performed a retrospective cohort study at a large HIV initiation site-Kenyatta National Hospital CCC, and use Generalized Estimating Equations to longitudinally identify predictors of incidence and risk factors among children initiated HAART, and in particular to determine the impact of TB co treatment on the outcomes among children with confirmed TB/HIV. Results showed that, for fifty one children who had information on CD4 before TB treatment and after completing TB treatment. The median CD4 percent before was 10% (IQR: 7: 15.6) and 29 (IQR: 20.5: 42.7) after completing TB treatment. Before starting TB treatment, 62.8% of children were immune-suppressed and this proportion reduced to 19.6% after treatment. The difference in the proportion of immune-suppressed children was statistically significant using McNemar test (p-value =0.001)

Chapter 1

Introduction

Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured. The main advantage of longitudinal studies is that they can distinguish changes over time within each individual (longitudinal effects) from differences between people in their baseline values (cross-sectional effects). Further, longitudinal data are characterized by the fact that repeated observations tend to be correlated. Statistical techniques which assume independent observations, such as linear regression analysis and logistic regression analysis, cannot directly be used in longitudinal studies due to correlation. Repeated observations made over time on HIV/TB co-infected children, is a classical case of this nature where certain parameters such as weights, CD4 percentages, Hb, Creatinine, viral load, height, adherence and other clinical outcomes are repeatedly measured. Models utilized to analyze such data i.e. generalized linear models enable us to account for various types of heterogeneity, including between and within subjects of interest. We narrow our interest to Generalized Estimating Equations (GEE) which provides a regression methodology to analyze the correlated data that often result from a longitudinal study.

Chapter 2

Literature

2.1 Literature Review

Long term outcome of TB/HIV in children under HAART in a large typical pediatric hospital remains poorly documented in resource-limited settings. Though previous studies have examined the effect of TB on mortality in HIV-infected individuals within specific cohorts for example, a recent study of TB patients in Nigeria (2011 Au-Yeung et al) which found that 15.5% of HIV-positive individuals on treatment died, compared with 3.1% of HIV-negative individuals; much previous work has primarily focused on epidemiological distribution.

2011 Au-Yeung et al(2010) outlined TB and HIV as two devastating global infectious diseases. TB is one of the deadliest diseases of the 20th century and continues to claim millions of lives. The global HIV epidemic also continues to claim many lives, despite the global scale-up of antiretroviral (ARV) therapy. In 2008, an estimated 33.4 million people worldwide were living with HIV, and 2.7 million were newly infected. More than 25 million people worldwide have died from HIV since the epidemic began. Immune-suppressed HIV-positive individuals are more likely to become co-infected with TB, which is a leading cause of death in this population, especially among those who reside in Sub-Saharan Africa. The dual burden of TB and HIV infection increases the likelihood of dying compared with having either disease separately. There remains a dearth of research on global TB mortality in HIV-positive individuals.

Further, (011 Au-Yeung et al(2011) the study found that Initiating antiretroviral therapy (ART) together with tuberculosis (TB) therapy could reduce mortality risk among coinfected patients by 55%, according to interim findings from a large clinical trial in

South Africa. Historically, ART has been delayed in many coinfected patients until after the completion of TB therapy because of concerns about drug-drug interactions, high pill burden, overlapping drug toxicities, and immune reconstitution inflammatory syndrome.(2010, Joshi YP, et al) illustrates Tuberculosis and HIV co-infection as complex disease where there are hurdles to cross at each stage. Diagnosis and treatment is complex and involve a clear understanding of innovative laboratory methods as well as complex drug-drug interactions. The epidemiology clearly shows that the HIV and TB epidemics go hand in hand and indeed, fuel each other. Thus, prevention can not be underplayed. This review brings us up to date on the current thoughts on this co-infection and hopefully encourages more research in this area.

On the other hand analysis of longitudinal data remains an exciting topic. Interest in this topic has grown steadily since the watershed Harris volume (1963) and has been reflected in other books such as Collins & Horn (1991), Collins & Sayer (2001), Gottman (1995), Nagin (2005), Nesselroade &Baltes (1979), Singer &Willett (2003a), and von Eye (1990a,b), as well as in countless journal articles. Computing resources are becoming increasingly powerful, and readily available specialized software enables scientists to apply new and highly sophisticated statistical analyses in their work. Some rich longitudinal data sets have been archived and are available to the scientific public for statistical analysis. Further, creative use of new technology, such as handheld computers and other devices, is making collection of longitudinal data characterized by more frequent and intense measurement increasingly feasible (e.g., Shiffman & Stone 2006, Walls & Schafer 2006). As more longitudinal studies have been undertaken, and the length and intensity of longitudinal studies have increased, a fundamental tension has emerged between what Molenaar (2004) has termed "attention to interindividual variation," that is, variation between individuals, and "attention to intraindividual variation," that is, variation within individuals. Approaches focusing on interindividual variation emphasize establishment of general developmental principles that apply to all individuals. In contrast, approaches focusing on intraindividual variation emphasize understanding change within the individual, with establishment of general principles a secondary goal. This article takes the perspective that growth is a phenomenon that occurs within the individual, and therefore intraindividual variability is a primary interest in statistical modeling of longitudinal data. The importance of modeling intraindividual variability has emerged forcefully in longitudinal research (e.g., Nesselroade 1991, Rogosa et al. 1982, Rogosa & Willett 1985) and has been the impetus

behind the development of many of the statistical procedures discussed here. At the same time, in science we are always engaged in inductive reasoning and so attempting to abstract general principles about interindividual variability in intraindividual change is the ultimate goal (Curran & Wirth 2004).

Longitudinal research is most likely to approach characteristics by the seamless integration of three elements:

- a well articulated theoretical model of change observed using
- a temporal design that affords a clear and detailed view of the process, with the resulting data analyzed by means of
- a statistical model that is an operationalization of the theoretical model.

The integration of these three elements is necessary to ideal longitudinal research but not sufficient.

Integration of the Three Elements

The seamless integration of theoretical model, temporal design, and statistical model is an ideal that rarely, if ever, is met in social and behavioral research. For some theoretical models of change, particularly those that are very sophisticated, a tailor-made statistical model may not yet be available, forcing the investigator to use a statistical model that is not completely appropriate. In many empirical settings, the degree of correspondence between the statistical model and the theoretical model is limited by the temporal design. To take a simple illustration, if a theoretical model of growth in a continuous outcome is highly complex, involving many peaks and valleys, and the temporal design provides data collected at only two points in time, the only choice of statistical model available to the investigator is linear growth, which clearly does not correspond to the theoretical model. Resource limitations, logistical considerations, and concerns such as the prospect of overburdening study participants frequently may mean that the best temporal design scientists reasonably can use will allow statistical modeling to provide only a rough approximation to a complex and nuanced theoretical model of human development. Nevertheless, this conceptual framework is useful in interpreting the results of longitudinal research. Consideration of the ways in which a particular study approaches or fails to approach the ideal of integration of theoretical model, temporal design, and statistical model may help to identify the strengths and limitations of the study, the generalizability and likely replicability of the conclusions,

and directions for future research.

The most commonly used approach to modeling change in continuous variables is growth curve models. Growth curve models, such as hierarchical linear models (Raudenbush 2000), fit growth trajectories for individuals and relate characteristics of these individual growth trajectories (e.g., slope) to covariates. The individual growth trajectory can be expressed as

$$Y_{ti} = \beta_{0i} + \beta_{1i}x_{ti} + e_{ti}$$

for a linear model of growth.Y_{ti} represents individual i's outcome score at time t, where $t = 1, \cdots T$;x_{ti} represents the measure of time for individual i; and β_{0i} and β_{1i} represent the intercept and slope, respectively, of linear growth for individual i. This is often referred to as the level-1 equation. The intercept and slope parameters are random effects; in other words, they may vary across individuals, as reflected in the need for the i subscript denoting individual. This leads directly to the level-2 equations:

$$\beta_{0i} = \gamma_{00} + \mu_{0i}$$

$$\beta_{1i} = \gamma_{10} + \mu_{1i}$$

Much longitudinal research is at the intersection of a theoretical model concerning change in continuous variables and a temporal design that is some version of a longitudinal panel design. Under these circumstances, the theoretical model that is fit in data will usually need to be limited to fairly simple polynomial models of growth. Even with this limitation, such models can be very sophisticated, particularly if several measurement occasions are available. Raudenbush (2000) and McArdle & Nesselroade (2003) provide excellent overviews of growth modeling for longitudinal panel designs. Some theoretical models postulate that there is a discontinuity in continuous change; in other words, there is a distinct change point at which growth accelerates, decelerates, or levels off. This may simply be a change in acceleration, or it may represent a qualitative shift in some underlying process, so that different covariates are expected to predict different phases of the process. In piecewise growth models, the growth curve is divided into segments that are fit simultaneously with separate growth parameters. Cumsille et al. (2000) and Li et al. (2001) demonstrated how to fit piecewise growth curve models in which the change point is known and is the same for all subjects. Cudeck & Klebe (2002) presented an approach to piecewise growth models in which the change point may be unknown and estimated as a random effect. Even when discontinuity in

continuous change is not expected, the piecewise approach may offer a straightforward and intuitively appealing alternative for fitting nonlinear models of growth.

There are times when the investigator poses the question, what distinct patterns of growth characterize this sample? In other words, are there subgroups of individuals undergoing similar growth? Growth mixture models (Muthen 2001; Muthen & Muthen 2000; Muthen & Shedden 1999; Nagin 1999, 2005; Nagin & Tremblay 2001) identify subgroups corresponding to distinct patterns of growth in longitudinal data and can be used to estimate the prevalence of the patterns and to relate the patterns to covariates.

Some interesting developmental research questions involve continuous change unfolding across a period of time that may be years in duration. One alternative to a temporal design that involves observing the process over the entire period of development in a single cohort is an accelerated longitudinal design (Bell 1953, McArdle & Hamagami 2001). In an accelerated longitudinal design, multiple cohorts of different ages are observed longitudinally for a shorter period of time. The cohorts must be overlapping, so that for each cohort, there is at least one age at which at least one of the other cohorts is also observed. Then a statistical approach is used to combine the cohorts and estimate a single growth trajectory, extending from the youngest age observed to the oldest. The accelerated longitudinal design can save a significant amount of time, but it makes the assumption that there is no age-by-cohort interaction affecting development; in other words, it assumes that a single growth trajectory can reasonably represent all the cohorts. Duncan et al. (1996) combined data from four different overlapping age cohorts, each of which was measured at one-year intervals over a three-year period. This enabled them to estimate a growth trajectory for adolescent alcohol use extending from age 12 to age 17. They compared this growth trajectory with one based on a smaller single-cohort sample over the same age range, and found them to be essentially similar, suggesting that the assumption of no age-by-cohort interaction was met. Miyazaki & Raudenbush (2000) showed how to test this assumption empirically and presented a general framework for analysis of data from accelerated longitudinal designs.

Standard growth models extend directly to intensive longitudinal data, and are currently one of the most frequently chosen statistical models in this context. One very common implementation of intensive longitudinal designs in psychology is in controlled laboratory studies involving animals. In these studies, it is typical to have a nearly continuous record of important outcome behaviors, such as lever presses to obtain rewards,

over several days or weeks. Donny et al. (2004) and Lanza et al. (2004) illustrated how to fit growth models to individual animal data and how to use a growth curve framework to address experimental hypotheses like the ones typically motivating laboratory studies. Growth models also can fit theoretical models of increased complexity. The complexity of the theoretical models that motivate collecting intensive longitudinal data often goes beyond periodicity. Change over time may not be easily characterized by a polynomial. Instead, change may have many irregular ups and downs. There may be time-varying covariates, and some of these time-varying covariates may even have time-varying effects. A timevarying effect occurs when the strength and/or direction of a covariate changes as a function of time. It is even possible for an intervention to operate on an effect. For example, a drug abuse prevention program aimed at adolescents could reduce the influence of peer substance use on adolescent substance use; a therapy intervention aimed at depression could reduce the relation between everyday stressful events and anxiety. Functional data analysis (Fan & Gijbels 1996, Fok & Ramsay 2006, Li et al. 2006) provides a statistical model that is well suited both to these kinds of theoretical models and to intensive longitudinal data. Fok & Ramsay (2006) and Li et al. (2006) have illustrated how to apply functional data analysis in intensive longitudinal data using a growth modeling approach. Li et al. analyzed intensive longitudinal data on affect and urge to smoke collected in a sample of smokers enrolled in a cessation program. They found that urge to smoke was associated with negative affect, and that this relation grew stronger immediately after quitting smoking. The standard longitudinal panel design involves relatively long intervals, often months or years, between occasions of measurement. Data collected using this kind of temporal design may reveal during which interval a transition, often called an event, occurred, but beyond this cannot be used to determine when the event took place. For example, consider a standard longitudinal panel study in which adolescents are measured once per year. At each time, the adolescents are asked if they have ever tried smoking. Such data may indicate that an adolescent tried smoking for the first time between the previous observation and the current one, but exactly when in that interval the encounter with tobacco occurred is unknown. Often a research question such as the following is of interest: Given that at time interval t an individual has never tried smoking, what is the probability that the individual will try smoking during a particular subsequent time interval? This question and related questions can be addressed by means of discrete-time survival analysis (e.g., Cox 1972). As the name suggests,

survival analysis was originally developed in the biostatistics literature to model time until the occurrence of death and other medical events such as relapse of disease. Today its application in psychology and other areas in the behavioral sciences is growing (Singer & Willett 2003a,b). Fundamental to survival analysis are two closely related functions, the hazard function and the survival function. Singer & Willett (2003a) presented a survival analysis of data collected yearly on a cohort of special education teachers newly hired in the Michigan public schools. The target event was leaving teaching. Let T_i represent the time interval j when individual i experienced the event; in this case, left teaching. Then the hazard for a particular time interval is

$$h(t_{ij}) = P[T_i = j | T_i \geq j]$$

, or in words, the probability that individual i will experience the event during time interval j, given that individual i has not experienced the event during a previous time interval. There is a hazard associated with each time interval. For example, Singer &Willett (2003a) found that given that an individual did not leave during the first year of teaching, the probability was about 11% of leaving during the second year; given that an individual remained for at least six years, the probability was about 6% of leaving during the seventh year. Closely related is the survival function $S(t_{ij}) = P[T_i > j]$, which is the probability that individual i will not experience the event (will "survive"),conditional on not already having experienced the event in a previous time interval. Often of particular interest is the median lifetime, which is time interval during which the survival function reaches 0.5; in other words, the time interval by which half of the sample has experienced the event. Singer & Willett (2003a) found that sometime during year seven of teaching, 50% of the teachers in their sample had left. It is possible to introduce time-invariant and time-varying covariates into a survival analysis, in order to address research questions such as whether the hazard function differs across groups or whether elevation in the hazard function corresponds to an elevation or decrease in another variable. For example, if it is of interest to examine whether more experienced teachers are more likely to remain in teaching longer, a covariate representing years of prior experience before beginning the current teaching position can be included. The temporal design of a study has an impact on the conclusions that can be drawn from data. This is very evident in censoring, a topic that is of great concern in survival analysis. Suppose a study is evaluating a new approach to psychotherapy for depressed inpatients. It is expected that patients receiving the

new psychotherapy will go longer before experiencing a recurrence of depression. In this case, the event of interest is recurrence of depression. Suppose the new therapy is delivered, and then the patients are followed for two years. For each patient, it is recorded when the individual has a first recurrence of depression. Of course, at the end of two years not every patient will have had a recurrence of depression. At the conclusion of the two-year period, all that is known is that some patients have not had a recurrence; it is not known whether these patients will have a recurrence in the future. This is called right-censoring. Right-censoring is present in most survival analyses because rarely is it practical to conduct a study long enough for the event in question to occur for all subjects-and in some cases it is expected that some subjects will not experience a reoccurrence of the event. Somewhat less common, and more problematic, is left-censoring. Left-censoring is present when for some individuals the event in question occurred at some indeterminate time before the start of the study.

This dissertation emphasized situations where there is a clear theoretical model guiding research. But many studies, of necessity, are not guided by theory. Some results are breaking new ground in areas where theory is incomplete or even nonexistent. Sometimes the original theory guiding a study has been soundly refuted by the empirical data collected to test it, leaving the investigators to conduct secondary analyses with the hope of beginning to build a newtheory. Where theory cannot inform choice of temporal design and statistical model, these choices must be made in a more exploratory manner. Where there is little theory to guide choices, it may be wise for investigators to do their best to keep options open. For example, consider the choice of number and spacing of observations in a temporal design. If there are too many occasions of measurement spaced too close together, it is always possible to base analyses on a subset of measurement occasions or to aggregate (e.g., sum) occasions that are close together. However, if there are too few occasions or they are spaced too far apart, there may be little the investigator can do to recover information about what happened between observations. Thus, erring on the side of more, and more frequent, observation may be prudent. In cases where secondary analyses are being performed on existing data in order to build a theory, it may be helpful to interpret the results in the light of the temporal design used to collect the data. This interpretation could involve considering what kinds of change processes the temporal design might reasonably have been able to reveal and what kinds the temporal design would tend to hide from view. For example, consider a secondary analysis based on a three-wave longitudinal panel study, with

yearly data collection. A linear model of growth might simultaneously be an excellent representation of the empirical data and a poor representation of the underlying growth process. There may be considerable curvilinearity, even repeated peaks and valleys, in the growth process, but the temporal design obscures these features of growth, preventing them from being observed. Exploratory studies and secondary data analyses will always have a place in social and behavioral research. The argument here is not that such studies and analyses should be avoided. Rather, the argument is that even when theory is absent, the impact of the relation between temporal design and statistical model is an important consideration in longitudinal research. Careful thought about these matters can be tremendously helpful in interpreting empirical results and building theories to be tested in future studies.

We argued that ideal longitudinal research is characterized by the seamless integration of a well-articulated theoretical model of change, an appropriate temporal design, and a statistical model that is an operationalization of the theoretical model. Although no research study is perfect, it is useful to articulate an ideal as a standard for evaluation. Using the ideal as a conceptual framework, this article has surveyed a number of familiar and less familiar approaches to analyzing longitudinal data. Of the three elements discussed here, the theoretical model is perhaps under the most direct control of the investigator. A clear and detailed theoretical model is a necessary foundation for all longitudinal research. This model then provides the basis for choosing a temporal design. In many cases, the temporal design is not completely under the control of the investigator, due to resource limitations and other considerations. Nevertheless, an articulation of the ideal temporal design will provide a useful perspective when difficult tradeoffs have to be made and compromises reached with respect to the design that is chosen for implementation. The most appropriate statistical model is as close to a direct operationalization of the theoretical model as possible, subject to the limitations imposed by the temporal design. Even when the ideal of seamless integration of the three elements is met, unplanned missing data and measurement reliability and validity are important issues that can have a major impact. We are entering a new era of longitudinal data analysis. Increasingly elegant statistical models and new technology supporting more intensive longitudinal data collection are enabling data analysis and design to catch up with sophisticated and nuanced psychological theories of human development and change. It will be interesting to see what the next decade brings!

We performed a retrospective cohort study at a large HIV initiation sites-Kenyatta

National Hospital CCC, to identify predictors of incidence and risk factors among children initiating HAART, and in particular to determine the impact of TB co treatment on the outcomes among children with confirmed TB/HIV's mortality in co-infected patients

2.2 Problem statement

Given a resource -limited setting, what is the incidence and risk factors for TB/HIV co-infection in children under HAART in a large typical pediatric hospital and what are the outcomes of treatment and factors influencing the outcomes. We apply generalised estimating equations which account for correlates.

2.3 Broad Objective

The objective of the study is to longitudinally describe the observed incidence of and risk factors for TB diagnosis among HIV-infected children enrolled in a large pediatric hospital.

2.4 Specific Objectives

- To describe the clinical burden of pediatric TB diagnosed in an HIV-infected population at KNH CCC
- To describe the social and clinical factors associated with a TB diagnosis in HIV-infected children
- To describe the effect of combination antiretroviral treatment on the probability of being diagnosed with TB, among these HIV-infected children.

2.5 Significance

Long term outcome of TB/HIV in Children under HAART in a large typical pediatric hospital remains poorly documented in a resource -limited settings. In this retrospective study carried out at Kenyatta National Hospital's (KNH) Comprehensive Care

Centre (CCC) in Nairobi, Kenya we wish to longitudinally describe the observed incidence of and risk factors for TB diagnosis among HIV-infected children enrolled in a large pediatric hospital.

Chapter 3

Methodology

3.1 Study area

This is a retrospective study carried out at Kenyatta National Hospital's (KNH) Comprehensive Care Centre (CCC) in Nairobi, Kenya. KNH is a national public referral hospital that also serves as the teaching facility for the University of Nairobi Medical School. The CCC provides care for HIV infected adults and children providing antiretroviral treatment as well as treatment of opportunistic infections. The study subjects are children seen at the CCC between April 2005 and November 2010.

3.2 Study population

The study population consisted of HIV-infected children aged 0-19 years who were receiving HAART at the CCC.

3.3 Study procedure

Clients to the CCC are usually referred from various sources including KNH wards, KNH outpatient/consultant clinics, PMTCT, KNH VCT and other VCTs in and out of Nairobi. Patients are then referred to a nurse who assesses and determines whether the patients need to open a file or whether the patient can be referred to another HIV care clinic in an area near their residence. Once the patient is registered, they are then referred to a nurse for triage. At the triage desk, the nurse receives the patient file and examines the patient recording the temperature, height, weight and blood

pressure. This information is recorded in a standard clinical data form which is put in the patients file. New patients are then sent to the counselor and clinician.

1. Counseling: Counselors take a detailed social history and record this in the adherence counseling checklist session.

2. Clinician Assessment: Clinicians completes standardized clinical questionnaires and submits it to the data clerk for data entry in to the electronic medical records. Patients are then reviewed in two weeks with the baseline results to determine eligibility of HAART. The clinician monitors adherence at every visit and refers patients with poor adherence to the counselor.

3. Pharmacy: The pharmacist keeps electronic records of all drugs dispensed and the expected dates of refill as well as patient details such as weight, height, age and gender.

4. Laboratory services: Baseline tests including CD4count, Full blood count (FBC), alanine transaminase (ALT) and creatinine (CREAT) are taken. These tests are repeated at least every 6 months during follow-up or any other time that the clinician finds it necessary. Other tests may include DNA PCR for early infant diagnosis, RNA PCR for viral load, HIV ELISA, fasting lipid profile, liver function tests, lactate levels when indicated.

5. Nutritional assessment: The nutritionist carries out an assessment of the nutritional status of the patient and provides nutritional counseling.

6. Social worker: collection of exit data (mortality, transfer outs, lost to follow-up, defaulters) and track follow-ups

3.4 Eligibility criteria

Inclusion criteria: Patients on HAART and had TB at enrolment or developed TB during follow-up

3.5 Ethical consideration

The study was approved by the University of Nairobi/KNH Ethical Review Committee.

3.6 Study design

We define longitudinal data design as follows

i=1,..., N subjects.

For a balanced design in which all subjects have complete data, and are measured on the same occasion, we index the measurements occasion as

j=1,...,n observations

or in the unbalanced case of unequal number of measurements or different time points for different subjects

$j = 1, \ldots, n_i$ observations for subject i.

The total numbers of observations are given by

1. $\sum_{i=1}^{N} = n_i$
   The repeated response or outcomes or dependent measures for subject i are denoted as vector

2. $Y_i = n_i \times 1$
   The values of the p predictors, or covariates, or independent variables for subject I on occasion j are denoted as

3. $X_{ij} = p \times 1$ For time-invariant predictors (between subject, e.g. sex), the values of

4. $_{ij}$ are constant for

5. $j = 1, , n_i$. For time-varying predictors (within-subject, e.g. age), the

6. x_{ij} can take on subject and time point-specific values. To describe entire matrix of predictors for subject i, we use notation

7. $X_i = n_i \times p$.

3.7 Data Layout

In this univariate design ni varies by subject and so the number of data lines per subject can vary. In terms of covariates, if X_r is time-invariant (i.e. between subject variable) then, for a given subject i, the covariate values are the same across time, namely,

Table 3.1: Data illustration

Subject	Observation	Response	Covariates
1	1	y_{11}	$x_{11_1} \cdots x_{11_p}$
1	2	y_{12}	$x_{12_1} \cdots x_{12_p}$
⋮	⋮	⋮	⋮
1	n_i	y_{12_1}	$x_{1_n 1_1} \cdots x_{1_n 1_p}$
⋮	⋮	⋮	⋮
N	1	y_{N_1}	$x_{N1_1} \cdots x_{N1_p}$
N	2	y_{N_2}	$x_{N2_1} \cdots x_{N2_p}$
⋮	⋮	⋮	⋮
N	n_N	y_{Nn_N}	$x_{N_n N_1} \cdots x_{N_n N_p}$

$X_{i1_r} = x_{i2_r} = x_{i3_r} = \ldots = x_{i_n i_r}$

The above layout depicts what is called a 2-level in the multilevel and hierarchical linear modeling literatures

3.8 Analysis Considerations

3.8.1 Generalized Linear Models and Estimating Equations

Generalized linear models are the generalization of certain general linear models. These are namely ANOVA, ANCOVA, MANOVA and MANCOVA, as well as the regression models. Generalized Linear Model supports non-normal distributions for dependent or criterion variables. Thus, it can be said that the generalized linear model involves logistic models for binary dependent variables, log linear analysis, Poisson regression, etc.

3.8.2 Generalized Estimating Equations

GEE extends Generalized Linear Models further by involving dependent data for repeated measures, logistic regression and various other models for the time series or correlated data.

The dependent variable in the implementation of Generalized Estimating Equations

and Generalized Linear Models are distributed in the following distributions:

- The dependent variable during the implementation of Generalized Estimating Equations and Generalized Linear Models takes the form of Normal distribution when the dependent variable is continuous (numeric)
- The dependent variable during the implementation of Generalized Estimating Equations and Generalized Linear Models takes the form of Multinomial distribution when the dependent variable is ordinal (numeric or string)
- The dependent variable during the implementation of Generalized Estimating Equations and Generalized Linear Models takes the form of Binomial distribution when the dependent variable is binary
- The dependent variable during the implementation of Generalized Estimating Equations and Generalized Linear Models takes the form of Poisson distribution when the dependent variable is count in nature or when the events are rare in nature
- The dependent or criterion variables in Generalized Estimating Equations and Generalized Linear Models are not distributed as free variables
- The link function in Generalized Linear Models relates the Generalized Linear Models in a specified design matrix

Generalized Estimating Equations assumptions

- Generalized Estimating Equations do not assume that the dependent/independent variables are not normally distributed
- Generalized Estimating Equations neither assume linearity between the predictors and the dependent variables, nor homogeneity of variance for the range of the dependent variable
- There must be linearity in the link function as assumed by the Generalized Estimating Equations
- It is assumed in Generalized Estimating Equations that the multicollinearity is absent

- The data in Generalized Estimating Equations must be centered in order to reduce multicollinearity
- The dependent data in Generalized Estimating Equations are either interval or ordinal, and they are sometimes binary or count type

3.9 Data and Data collection

Records were analyzed for patients followed up for the period 2005-2010. Data was obtained from the existing EMR as well as abstraction from medical charts. Details of demographics, clinical presentation, investigations, management and outcome were extracted. Data abstraction was done by clinicians who completed forms with information on demographic and clinical outcomes form patient's records. The abstracted data was then entered in Microsoft database. The information on ART use and outcome was extracted from KNH drug management system developed by Management Sciences for Health (MSH). Identifying information was stripped of for all analytical files and individual records were identified using patient identification number that was provided by the hospital staff. The study included 711 patients aged 0 to 19 years who were enrolled in the program with at least one follow-up visit between 2005 and 2010. Nutritional status was assessed using weight-for-age and BMI-for-age based on WHO reference 2007. The indicator of weight-for-age is determined for the age range of 0-10 years while BMI-for-age is determined for age range 0-19 years. The BMI for the children aged below 24 months was based length and height otherwise.

3.10 Statistical Analysis

The description of the sample was presented in form of descriptive statistics mean, median, inter-quartile range (IQR) and proportions. Because the primary aim of the analysis was to describe the incidence of new cases of TB in this cohort from time of enrolment, children with a prior diagnosis of TB before enrolled were censored in estimating the incidence. We considered the incidence of TB during the 6 years after enrolled and also considered the incidence of TB after starting ART. Person-years of follow-up (PYFU) were calculated from the date of enrolment or starting ART until the date of death, the date of last follow-up visit, or October 2010; whichever occurred

first. The following baseline factors were considered: sex, age group, period (2005-2006, 2007-2008 and 2009-2010), WHO stage at enrolment and at start of ART. The analysis of TB incidence used the entire sample of 711 children and the rest of analysis was performed using only children who were diagnosed with TB. The proportions of different methods of HIV diagnosis (rapid test, Elisa and PCR) was presented by age groups (¡18 mths, 18 mths - 59 mths, 5 yrs - ¡10 yrs, 10 yrs - 19 yrs). The response or outcome of TB treatment was assessed using McNemar test for paired binary outcomes. We employ Generalized Estimation Equation (GEE) to assess adherence rate before and after TB treatment while controlling for possible correlates. The analysis for change used only children who had outcome both before and after TB treatment. Nutrition status before TB diagnosis was restricted to 1 month before or the same month TB was diagnosed while nutritional status after was considered for 1 to 2 months after TB was diagnosed. CD4 count was also restricted to first available CD4 12 months before to 1 month after as before TB diagnosis while 12 months to 18 months after TB diagnosis.

Chapter 4

Exploratory Data Analysis

4.1 Introduction

This is an approach to analyzing data for the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses. It is the approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set
2. uncover underlying structure
3. extract important variables
4. detect outliers and anomalies
5. test underlying assumptions
6. develop parsimonious models and
7. determine optimal factor settings

EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.
We describe our data as

4.2 Descripive statistics

This section narrates the breakdown of abstructed data to the number of patients who were eventually incorporated to the analysis.The study included 711 patients aged 0 to 19 years who were enrolled in the program with at least one follow-up visit between 2005 and 2010 [2mm]

and Settings/data manager/Desktop/desktop/project.JPG

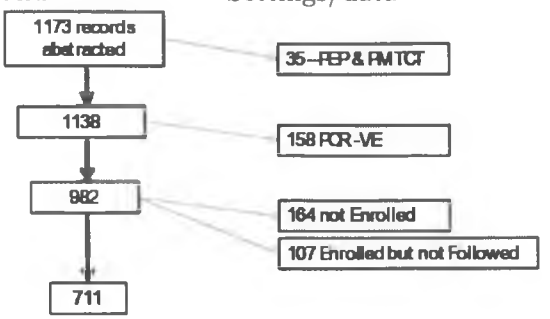


Figure 4.1: Description of Data abstraction

4.2.1 Child characteristics

In total, 213(30%) of 711 children had been diagnosed with TB and 150 (70.4%) of 213 had TB diagnosed before they were enrolled. The age of children diagnosed with TB ranged from 0 to 18 years with average age of 6.7 years (median =6). About 53% of the children were female, 33.3% were orphan with 18.6% of them being double orphan. At the time of data abstraction 139 (65.3%) were still active i.e. still in the program. Further observation are illustrated at the table below

Table 4.1: Child age distribution at enrolment for 213 children diagnosed with TB

Mean	6.7
Median	6.0
Lower IQR	3.0
Upper IQR	10.0
Minimum	0.1
Maximum	18.0

Table 4.2: Child Characteristics for TB sample

	N=213	%
4 age groups		
Less than 18 months	23	10.8
18-59 months	63	29.6
5-10 years	71	33.3
10-19 years	56	26.3
The sex of the patient		
Male	101	47.4
Female	112	52.6
Is The child an orphan?		
Yes	71	33.3
No	142	66.7
Which Parent is deseased?		
Mother	37	52.9
Father	20	28.6
Both Parents	13	18.6
Time of TB diagnosis		
TB before enrolment	150	70.4
TB between enrolment and ART	23	10.8
TB afetr ART	30	14.1
TB and never on ART	10	4.7
18-59 months	63	29.6
Current client's status		
Active	139	65.3
Deceased	5	2.3
Lost to Follow-up	46	21.6
Transfer out	23	10.8
Total	213	100

4.3 Method of HIV diagnosis

A total of 152 (71.4%) had information on the method of HIV diagnosis and the distribution of method of diagnosis is presented by age group in table ( ). Overall, a half of these children had HIV rapid test, 45.4% had Long HIV ELISA test and only 14.5% had HIV diagnosed through DNA PCR. HIV DNA PCR was common (57.1%) method of HIV diagnosis among children below 18 month compare to older age groups. The table is interpreted as for example 12 out of 21 children aged less than 18 months

Table 4.3: Type of HIV diagnosis by age group

	HIV rapid test	Long HIV ELSA	HIV DNA PCR	Total
less than 18 months	11	5	12	21
18-59 months	27	21	7	46
5-10 years	21	27	2	49
10-19 years	18	16	3	36
Total	77	69	22	152

had DNA PCR done

4.4 TB diagnosis

Among the children diagnosed with TB, 76.5% (165 out of 213) reported to have had contact with person with confirmed TB and most common symptom was cough for more than 2 weeks with fatigue as the least common symptom. About 48% had none of the symptoms (cough, fever, poor weight gain/weight loss and fatigue) or was not documented in the patient files and only 4.2% had all the four symptoms. Considering TB signs, 16.9% had un-resolving pneumonia and 4.7% had TB lymphadenopathy. About 26% of the children diagnosed with TB had CxR suggestive for TB while only 2.3% had sputum positive for AFBS, 10% had Mantoux test positive and 1.9% had tissue positive for AFBS. Twelve children (5.6%) had a combination of both sputum positive and Mantoux test positive. Majority (59.6%) of children had PTB type of TB, 12.7% had EPTB type and 5.2% had both two types of TB.

Table 4.4: The symptom, lab findings and type of TB distribution

Symptoms	**n**	**%**
Contact with person with confirmed TB	50	23.5
Cough $\geq$ 2 weeks	102	47.9
Fever or night sweats	51	23.9
Poor weight gain	31	14.6
Weight loss	35	16.4
Had fatigue	14	6.6
Number of symptoms		
None	103	48.4
One	40	18.8
Two	41	19.2
Three	20	9.4
Four	9	4.2
Signs		
Had Un-resolving pneumonia	36	16.9
TB lymphadenopathy	10	4.7
Lab findings		
Sputum positive for AFBS	5	2.3
Tissue positive for AFBS	4	1.9
CXR Suggestive for TB	56	26.3
Mantoux test positive	21	9.9
Spatum and mantoux positive	12	5.6
Type of TB		
PTB type of TB	127	59.6
EPTB type of TB	27	12.7
Both PTB and EPTB types	11	5.2
Total	213	100

4.5 Hospitalization

We check descriptively the the number of days the HIV/TB confirmed cohort in our study has been hospitalised. On the negative of x-axis, we represents the number of days for hospitalisation for non-TB HIV positive patients while the positive represents description of the number of days for hospitalisation for patients with confirmed TB/HIV. 0 means no record of hospitalisation. Most of the children (64%) were hospi-

and Settings/data manager/Desktop/desktop/hosp.JPG

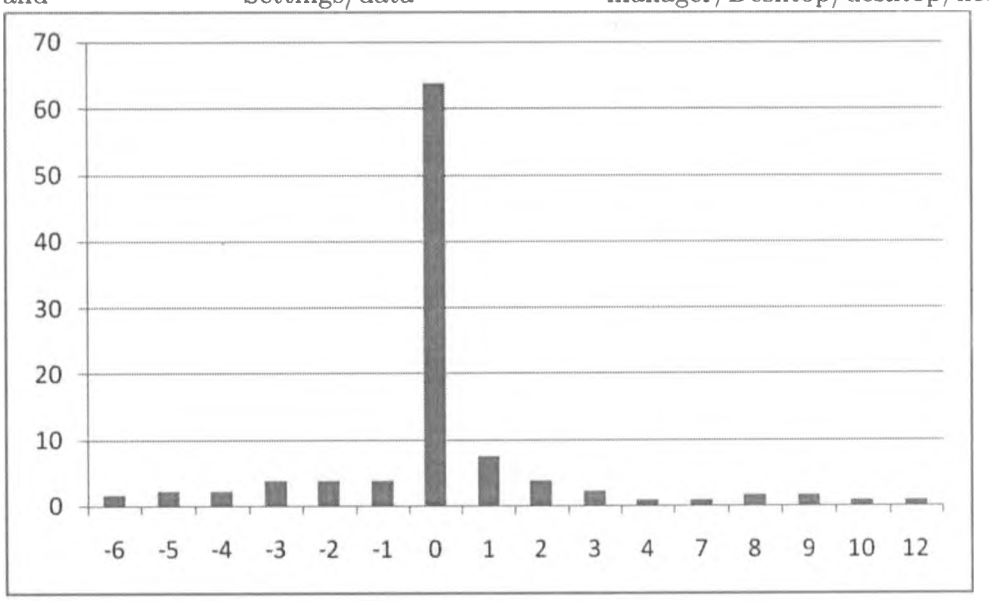


Figure 4.2: Hospitalisation

talized by the time TB was diagnosed and most of them (53%) were referred from KNH wards. The proportion hospitalized one month after starting TB treatment dropped to 7.4%.

Chapter 5

Confirmatory Data Analysis

5.1 Introduction

The purpose of Confirmatory Data Analysis are Provide precise information in the right circumstances and Well-established theory and methods.The analysis involves the following

- Heavy reliance on probability models
- Must accept untestable assumptions
- Look for definite answers to specific questions
- Emphasis on numerical calculations
- Hypotheses determined at outset
- Hypothesis tests and formal confidence interval estimation

5.1.1 Generalized Estimating Equation

We use GEE to fit the parameters of a generalized linear model where unknown correlation is present. The GEE allows for correlation without explicitly defining a model for the origin of the dependency, hence it is most suitable when the random effects and their variances are not of direct interest. The focus is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual. GEEs are usually used in conjunction with Huber-White

standard errors. GEEs belong to a class of semi parametric regression techniques as they rely on specification of only the first two moments. Under mild regularity conditions, parameter estimates from GEEs are consistent. They are a popular alternative to the likelihood-based generalized linear mixed model which is more sensitive to variance structure specification. They are frequently applied in multi-site cohort studies as they can handle many types of unmeasured dependence.

Formulation

Given a mean model, u_{ij}, and variance structure, V_i, the estimating equation is formed via

$$U(\beta) = \sum_{i=1}^{N} \frac{\partial u_{ij}}{\partial \beta_k} V_i^{-1} \{Y_i - \mu_i(\beta)\}$$

The parameter estimates solve $U(\beta) = 0$ and are typically obtained via the Newton-Raphson algorithm. The variance structure is chosen to improve efficiency. The Hessian of the solution to the GEEs in the parameter space can be used to construct robust standard errors. The most popular form of inference on GEE regression parameters is the Wald test using naive or robust standard errors, though the Score test is also valid and preferable when it is difficult to obtain estimates of information under the alternative hypothesis. The likelihood ratio test is not valid in this setting because the estimating equations are not likelihood equations.

5.1.2 Estimating Equations for Longitudinal Data

The Generalised Linear Models estimating equations were extended by Liang and Zeger(1986) to obtain a regression methodology that accounts for correlated measurements from longitudinal data.

Let individuals labeled $i =, \cdots, I$ be observed at times $t = 1, \cdots, T_i$ which results into a total of $N = \sum_{i=1}^{I}, T_i$,observations.

The outcome vector of individual i is $\underline{Y}_i = [Y_{i1}, \cdots, Y_{iT_i}]'$, and the associated covariate information is $X_i = \left[\underline{X_{i1}}, \cdots, \underline{X_{iT_i}}\right]'$.

Assume that outcomes for the same individual are correlated, but outcomes for different individuals are independent. The covariance matrix of $\underline{Y_i}$ has the form $\sigma^2 V_i$ where

$$V_i = \{diag[V(\mu_{i1}), \cdots, V(\mu_{iT_i})]\}^{1/2} \times R_i \{diag[V(\mu_{i1}), \cdots, V(\mu_{iT_i})]\}^{1/2}$$

and R_i is a correlation matrix among the outcomes measured at different times on the same individuals. The covariance matrix of all N outcomes $\underline{Y} = \left[\underline{Y_1'}, \cdots, \underline{Y_I'}\right]$ is a block diagonal covariance structure, the estimating equations can be written in the form

$$\sum_{i=1}^{I} D_i' V_i^{-1} \left(\underline{Y_i} - \mu_i\right) = 0$$

where $D_i = \dfrac{\partial \mu_i}{\partial \underline{\beta}}$.

These are generalised estimating equations (GEE). The extra term *generalised* distinguishes theses as estimating equations that accomodate the correlation structure. The GEE are identical in form to the GLM estimaing equations except that the matrix V contains nonzero off-diagonal terms. The solution to GG gives a consistent estimate of $\underline{\beta}$ that is asymptotically multivariate normal with covariance matrix

$$\sigma^2 \left[D' V^{-1} D\right]^{-1} = \sigma^2 \left\{\sum_{i=1}^{I} \left[D_i' V_i^{-1} D_i\right]\right\}^{-1}$$

5.2 TB

5.2.1 Incident and Prevalent TB

Table 5.1: Prevalent TB by two year cohorts

Two year cohorts	No TB	Prevalent TB	TB within 4 weeks of Enrolment	Total
2005 - 2006	153	58	6	217
2007 - 2008	298	77	17	392
2009 - 2010	83	15	4	102
Total	534	150	27	711

The tables uses all 711 children and interpreting this table is by cohort for example, 217 children were enrolled during 2005-2006 and prevalent TB was 26.7% among these children.

The proportion of children enrolled after TB was diagnosed was 21.2% (150 out of 711) and it differed by different enrollment cohorts. Prevalent TB was high (26.7%) among those who were enrolled in the year 2005-2006 which reduced to 19.6% and 14.7% for those enrolled in the periods 2007-2008 and 2009-2010 respectively. The proportion

enrolled with severe WHO HIV stage IV decreased by period from 16.1% in 2005-2006 to 14.9% and 11.1% for 2007-2008 and 2009-2010 respectively.

5.2.2 TB incidence

Table 5.2: TB incidence from time of enrolment and from time of starting ART per 100 person years

	Enrolment			start of ART		
	Cases	Rate	95% CI	Cases	Rate	95% CI
Overall	63	5.5	4.3-7.0	30	3.0	2.1-4.3
Male	29	5.1	3.5- 7.3	12	2.4	1.4 -4.2
Female	34	5.9	4.2- 8.3	18	3.6	2.3- 5.7
Period						
2005 - 2006	9	7.3	3.8-14.0	2	2.0	0.5- 7.9
2007 - 2008	38	7.6	5.5- 10.5	17	4.0	2.5- 6.4
2009 - 2010	16	3.0	1.9 -5.0	11	2.3	1.3- 4.2
WHO stage(initiation)						
WHO stage I	5	4.0	1.7- 9.5	1	1.5	0.2- 10.9
WHO stage II	13	5.2	3.0- 8.9	10	3.0	1.6 -5.5
WHO stage III	24	8.6	5.7 -12.8	9	3.1	1.6 -5.9
WHO stage IV	10	9.6	5.2- 17.8	6	5.4	2.4- 12.0
Age group						
Less than 18 months	7	4.3	2.1 -9.1	2	1.4	0.3- 5.5
5-10 years	20	5.8	3.7- 8.9	9	3.0	1.6 -5.8
10-19 years	15	4.9	2.9- 8.1	8	3.0	1.5- 6.0

The incidence of TB was determined for both after enrolment and after starting ART. During the 5 year period (1147.74 person years), 63 children enrolled developed TB resulting in the overall TB incidence after enrolment of 5.1 per 100 person years. The median time from enrolment to TB diagnosis was 1.87 months (IQR 0.49: 10.77 months). The incidence among female children was slightly higher compared to males and was low in the period of 2009-2010 compared to earlier years of 2005-2008. There was no clear pattern of TB incidence and the age of the children though incidence was

higher among children aged 18-59 months. Incidence of TB was related to WHO stage at the time of enrolment of the child. Incidence was higher among children who were enrolled with severe HIV stage i.e. the incidence was 4.0, 5.2, 8.6 and 9.6 per 100 person years for stage I to stage IV respectively.

5.2.3 Recurrence or Relapse of TB

TB relapse is defined as recurrence of TB diagnosis after 12 months of the first diagnosis. Based on follow-up period, a child was considered to have completed TB treatment if they were followed for more than 6 months. Twenty six children (12.2%) did not complete TB treatment and are expected to have had more than 6 months of follow up. There was under reporting on the date of ending TB treatment as only 74 children had this information recorded. There was TB relapse among 17 (8%) children with median time to TB relapse of 21.7 months with interquartile range (IQR) of [16.9; 30.6]. [3mm]

Table 5.3: Completing TB treatment and relapse

	n	%
TB treatment complete		
No	26	12.21
Yes	187	87.79
TB Relapse		
No	196	92.02
Yes	17	7.98

5.2.4 Response to TB analysis

To determine incidence TB from the start of ART, a total of 999.94 person years was observed with 30 children who were diagnosed with TB after starting ART treatment. The median time from the start of ART treatment to diagnosis of TB was 6.0 months (IQR 0.7: 22.6). Fifty percent of 30 children were diagnosed with TB within 6 months of starting ART treatment. The overall incidence after starting ART treatment was 4.6 per 100 person years. Similar pattern of incidence for period and age group is observed as that of incidence after enrolment. TB incidence after start of ART was

Table 5.4: overall TB Response

	Duration to End of TB treatment (n=74)	Duration to TB relapse (n=
Total	74	1
Mean	5.8	26.5
Median	6.0	21.7
Lower IQ	5.0	16.9
Upper IQ	6.5	30.6
Min	0.0	12.8
Max	18.1	61.4

dependent on the WHO stage at the time of HAART initiation. Incidence was hig among children who were in severe stage at the start of ART (1.5 per 100 person y for those who were in stage I and 7.7 per 100 person years for those who were in st IV).

5.3 Antiretroviral Therapy (ART) analysis

Among the children who were enrolled with TB but had not started TB, 95.7% of th were initiated on ART within first week of starting TB treatment. Twelve children were enrolled with TB and were on ART already, they all initiated ART in first week starting TB treatment. All 23 children who were diagnosed with TB after enrollm and starting ART, TB diagnosis happened within the first week of initiating A Among 30 children who were diagnosed with TB after enrollment and were not ART, 53% were initiated on ART after 6 months of TB treatment and none of th was initiated on ART within the first week of starting TB treatment. Among children, 10 were not on ART by the last time of analysis and 23.8% of children on NVP based regimen, 51.8% on EFV based, 4.2% on LPV/r based, 17.5% on t NRTI based with remaining 2.8% being on other regimen. Among 30 children who on ART before TB diagnosis, 6 (20%) had their ART regimen changed. Three chi had the ART regimen changed with the first month of starting TB treatment the remaining three the regimen was changed during first, second and third mo Majority (67.3%) of children aged more than 3 years were EFV regimen and m children aged below 3 years (52.8%) were on triple NRTI.

and Settings/data manager/Desktop/desktop/treat tb.JPG

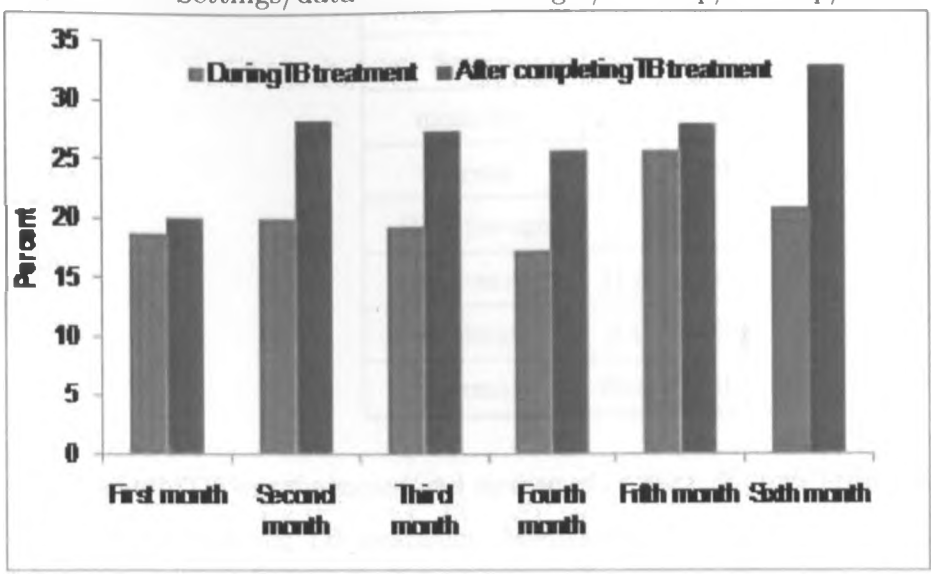


Figure 5.1: Response to TB

5.4 Outcome

5.4.1 Weight-for-age

Table 5.5: Weight-for-age

	Before		After	
	zwfa (42)	zbfa (61)	zwfa	zbfa
median	-1.16	-0.56	-1.295	-0.67
lower IQ	-2.47	-1.44	-2.45	-1.8
Upper IQ	-0.26	0.5	-0.17	0.53

[3mm]

5.4.2 BMI-for-age and Weight-for-age

The number of children who had information on nutritional status before and about two months after starting TB treatment was 42 for weight-for-age and 61 for BMI-for-age. The median weight-for-age score was -1.16 (IQR -2.47:-0.26) one month before or

Table 5.6: BMI-for-age and Weight-for-age

Weight-for -age	%	%
Severe	17.8	16.7
moderate	11.1	21.4
normal	71.1	61.9
BMI-for-age		
Severe	11.3	9.8
moderate	8.1	9.8
normal	80.6	80.3

same month TB was diagnosed and median of -1.30 (IQR -1.45: -0.17) after about two months after starting TB treatment. Median BMI-for-age was -0.56 (IQR -1.44:0.50) before and -0.67 (IQR: -1.80: 0.53) after starting TB treatment. See tables 5.5 and 5.6 above.

5.4.3 Immunosuppression

Table 5.7: CD4 count

	Before		After	
	cd4count	percent	cd4count	percent
median	495	10	928	29
Lower IQ	216	7	507	20.5
Upper IQ	671	15.5581	1367	42.7

Table 5.8: Immunosuppression

	Not Immune-suppressed	Immune-suppressed	Total
Before	19	32	51
After	41	10	51
Total	60	42	102
Odds Ratio	0.12	P-value	0.001

Fifty one children had information on CD4 before TB treatment and after completing TB treatment. The median CD4 percent before was 10% (IQR: 7: 15.6) and 29 (IQR: 20.5: 42.7) after completing TB treatment. Median CD4 count was 495 (IQR

216: 671) before and 928 (IQR: 505: 1367) after completing TB treatment. Before starting TB treatment, 62.8% of children were immune-suppressed and this proportion reduced to 19.6% after treatment. The difference in the proportion of immune-suppressed children was statistically significant using McNemar test (p-value =0.001). See tables 5.7 and 5.8 above.

5.4.4 Adherence

Adherence was determined from the pharmacy data by comparing number of days taken to come back for drug replacement and the number of days the drugs were supposed to last. While on TB treatment, 23 (18.7%) missed at least one dose and 36 (26.9%) after completing TB treatment. The figure () below shows the distribution of the proportion missed at least one dose by month. There seem to be poor adherence during the fifth month of TB treatment and it increases from the second month after completing TB treatment. The difference in adherence level before and after TB treatment was determined controlling for orphan status. The odds of missing at least one dose after completing TB treatment increased by 54% as compared to time while on TB treatment. Orphan children were 46% less likely to miss at least one dose.

Chapter 6

Survival analysis

6.1 Introduction

6.1.1 Definition

This is a branch of statistics which deals with exits i.e. death in biological organisms. More generally, survival analysis involves the modeling of time to event data; in this context, death or failure is considered an event in the survival analysis literature.
Survival analysis attempts to answer questions such as: what is the fraction of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the odds of survival?

6.1.2 Survival function

The object of primary interest is the survival function, conventionally denoted S, which is defined as

$$S(t) = Pr(T > t)$$

Where t is some time, T is a random variable denoting the time of death, and "Pr" stands for probability. That is, the survival function is the probability that the time of death is later than some specified time t. The survival function is also called the *survivor function* or *survivorship function* in problems of biological survival. Usually one assumes S(0) = 1, although it could be less than 1 if there is the possibility of immediate death or failure.

The survival function is non-increasing: $S(u) \leq S(t)$ if $u \geq t$. This property follows directly from F(t) = 1 - S (t) being the integral of a non-negative function. This reflects the notion that survival to a later age is only possible if all younger ages are attained. Given this property, the lifetime distribution function and event density (F and f below) are well-defined. The survival function is assumed to approach zero as age increases without bound, i.e., $S(t) \to 0$ as $t \to \infty$, although the limit could be greater than zero if eternal life is possible. For instance, we could apply survival analysis to a mixture of stable and unstable carbon isotopes; unstable isotopes would decay sooner or later, but the stable isotopes would last indefinitely. Lifetime distribution function and event density

The **lifetime distribution function**, conventionally denoted F, is defined as the complement of the survival function,

$$F(t) = Pr(T) = 1 - S(t)$$

and the derivative of F, which is the density function of the lifetime distribution, is conventionally denoted *f*,

$$f(t) = F'(t) = \frac{d}{dt}F(t)$$

The function f is sometimes called the **event density**; it is the rate of death or failure events per unit time. The survival function is often defined in terms of distribution and density functions

$$S(t) = Pr(T > t) = \int_t^{\infty} f(u)\,du = 1 - F(t)$$

Similarly, a survival event density function can be defined as

$$s(t) = S'(t) = \frac{d}{dt}S(t) = \frac{d}{dt}\int_t^{\infty} f(u)\,du = \frac{d}{dt}[1 - F(t =)] = -f(t)$$

6.2 Hazard function and cumulative hazard function

6.2.1 hazard function

The function conventionally denoted λ, is defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$),

$$\lambda(t)\,dt = Pr(t \leq T < t + dt | T \geq t) = \frac{f(t)\,dt}{s(t)} = -\frac{S'(t)\,dt}{S(t)}$$

Force of mortality is a synonym of hazard function which is used particularly in demography and actuarial science, where it is denoted by μ. The term hazard rate is another synonym.

The hazard function must be non-negative, $\lambda(t) \geq 0$, and its integral over $[0, \infty]$ must be infinite, but is not otherwise constrained; it may be increasing or decreasing, non-monotonic, or discontinuous. An example is the bathtub curve hazard function, which is large for small values of t, decreasing to some minimum, and thereafter increasing again; this can model the property of some mechanical systems to either fail soon after operation, or much later, as the system ages.

The hazard function can alternatively be represented in terms of the **cumulative hazard function**, conventionally denoted Λ:

$$\Lambda(t) = -logS(t)$$

so transposing signs and exponentiating

$$S(t) = exp(-\Lambda(t))$$

or differentiating (with the chain rule)

$$\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t).$$

6.2.2 cumulative hazard function

The function is derived from the fact that

$$\Lambda(t) = \int_0^t \lambda(\mu)\, du$$

which is the "accumulation" of the hazard over time.

From the definition of $\Lambda(t)$, we see that it increases without bound as t tends to infinity (assuming that S(t) tends to zero). This implies that $\lambda(t)$ must not decrease too quickly, since, by definition, the cumulative hazard has to diverge. For example, $exp(-t)$ is not the hazard function of any survival distribution, because its integral converges to 1.

6.3 Kaplan-Meier estimator

The **Kaplan-Meier estimator**, also known as the **product limit estimator**, estimates the survival function from life-time data.

In medical research, it is used to measure the fraction of patients living for a certain amount of time after treatment.

A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant.

An important advantage of the Kaplan-Meier curve is that the method can take into account some types of censored data, particularly *right-censoring*, which occurs if a patient withdraws from a study, i.e. is lost from the sample before the final outcome is observed. On the plot, small vertical tick-marks indicate losses, where a patient's survival time has been right-censored. When no truncation or censoring occurs, the Kaplan-Meier curve is equivalent to the empirical distribution.

In medical statistics, a typical application involves grouping patients into categories, for instance, in our case

we group TB and No-TB cases and assess their exposures. In the graph, patients with TB die much more quickly than those without TB.

6.3.1 Formulation

Let S(t) be the probability that an item from a given population will have a lifetime exceeding t. For a sample from this population of size N let the observed times until death of N sample members be

$$t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_N.$$

Corresponding to each t_i is n_i, the number at risk just prior to time t_i, and d_i, the number of deaths at time t_i.

Note that the intervals between each time typically will not be uniform. The Kaplan-Meier estimator is the nonparametric maximum likelihood estimate of S(t). It is a product of the form

$$\widehat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}.$$

When there is no censoring, n_i is just the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). It is only those surviving cases that are still being observed (have not yet been censored) that are at risk of an (observed) death. There is an alternative definition

that is sometimes used, namely

$$\widehat{S}(t) = \prod \frac{n_i - d_i}{n_i}.$$

The two definitions differ only at the observed event times. The latter definition is right-continuous whereas the former definition is left-continuous.

Let T be the random variable that measures the time of failure and let F(t) be its cumulative distribution function. Note that

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t).$$

Consequently, the right-continuous definition of may be preferred in order to make the estimate compatible with a right-continuous estimate of F(t).

6.3.2 Statistical considerations

The Kaplan-Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common such estimators are Greenwood's formula:

$$\widehat{Var}\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}$$

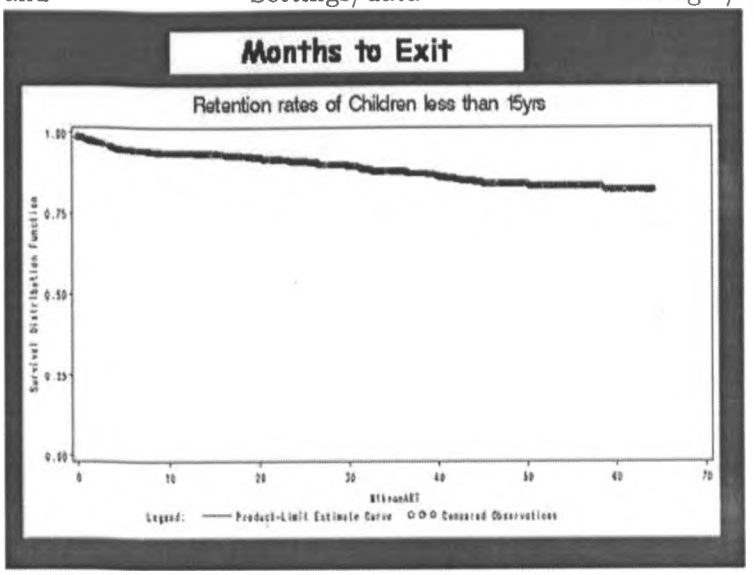
In our case, we wish to compare different Kaplan-Meier curves- TB verses non-TB cases

6.3.3 Month to exit

A bit of care is necessary to construct accurate confidence interval bounds from a product-limit estimated survival probability and its estimated variance. The expression for a 95% confidence interval based directly on an estimated parameter and the normal distribution is estimate $1.960\sqrt{variance(estimate)}$. However, the accuracy of a confidence interval is frequently improved by calculating the confidence bounds from a function of a parameter and using these limits to derive the confidence interval bounds for the parameter. The confidence interval for the survival probability P_k is such a case. The function is $\widehat{S}_k = log[-log(\widehat{P}_k)]$. The estimate $\widehat{s}_k$ has a more normal-like distribution than the distribution of $\widehat{P}_k$.

Figure 6.1: Month to Exit

and Settings/data manager/Desktop/desktop/exit.JPG

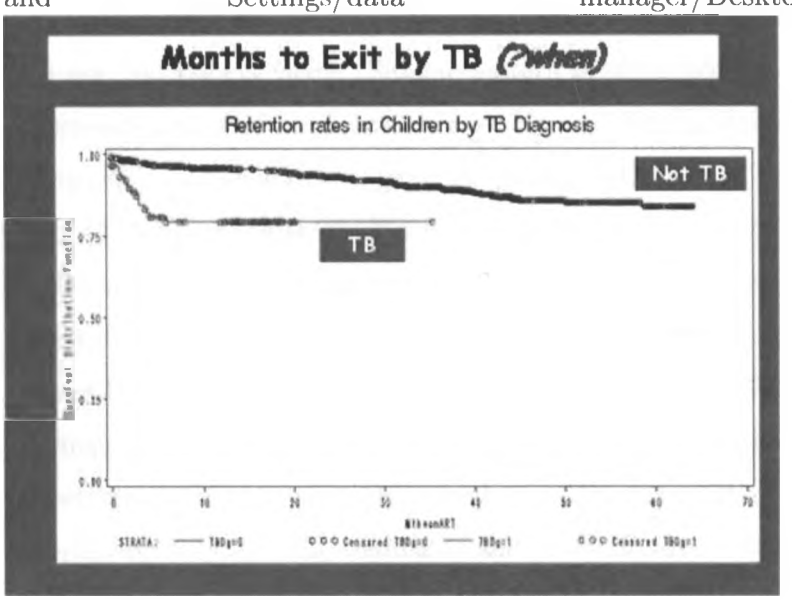


6.3.4 Month to Exit by TB

Comparison of survival distributions for TB and no-TB patients. The no-TB patients were followed-up for a median of 32 months (IQR 1945 months), whereas the elderly patients were followed-up for a median of 31 months (IQR 1945 months). The estimated probability of survival (for no-TB and TB, respectively) after 3 months on treatment was 96.3% (95% CI 96.096.5%) and 95.7% (95% CI 94.996.5%); at 6 months was 94.4% (95% CI 94.194.8%) and 93.5% (95% CI 92.594.5%); at 1 year of follow-up was 92.5% (95% CI 92.192.8%) and 91.3 (95% CI 90.292.4%); and at 4 years of follow-up was 89.4% (95% CI 88.989.8%) and 86.5% (95% CI 85.288.1%). Estimated probabilities of survival at other time points are provided in Fig. 1, which also shows KaplanMeier product limit estimates of the survival distribution for the TB and no-TB patients, and show that the no-TB group had better survival than the TB group (log-rank P ¡ 0.001).

Figure 6.2: Retention

and Settings/data manager/Desktop/desktop/retention.JPG



Chapter 7

Conclusions and Recommendations

Our study has several implications. This longitudinal analysis highlights the importance of considering the dynamic features of HIV/TB pediatric treatment and its long-term outcomes. Fifty percent of 30 children were diagnosed with TB within 6 months of starting ART treatment. The overall incidence after starting ART treatment was 4.6 per 100 person years. Similar pattern of incidence for period and age group is observed as that of incidence after enrolment. TB incidence after start of ART was dependent on the WHO stage at the time of HAART initiation. Prevalent TB was high (26.7%) among those who were enrolled in the year 2005-2006 which reduced to 19.6% and 14.7% for those enrolled in the periods 2007-2008 and 2009-2010 respectively. TB is one of the most common infections seen in children in areas of the world where HIV is prevalent. Clinical features overlap with HIV disease and radiographic findings may be non-specific. Although treatment using standard anti-TB regimens is effective, outcome is poor due to high mortality rates. Wider access to anti-retroviral therapy is likely to reduce mortality and morbidity due to TB in HIV-infected children and adults. Research is required to identify better

References

Elston JW, Thaker HK.(2008). Co-infection with human immunodeficiency virus and tuberculosis. *Indian J Dermatol Venereol Leprol* 74:194-9

Neal D. Hoffman, Colleen Kelly and Donna Futterman Young (1996). A New York City Cohort Tuberculosis Infection in Human Immunodeficiency Virus-Positive Adolescents and Positive Pediatrics .*Paediatrics Journal 97*;198-203

Soumya Swaminathan (2010).Tuberculosis in HIV-infected children. *HIV-associated Lung diseases Journal* 6(8):672-678

Pilheu JA.(1998): Tuberculosis and Lung Diseases .*International Journal- Tuberculosis* 2: 696-703.

Seth V, Narain S, Simpson L., Seth V, Kabra SK (2001). *Essentials of Tuberculosis in Children.* New Delhi: Jaypee Brothers, ; pp. 199-208.

Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. (1999).*Global burden of tuberculosis: estimated incidence, prevalence and mortality by country.* WHO Global Surveillance and Monitoring Project. JAMA ; 282: 677-686.

A. Odhiambo at al. (1996).Increasing Annual Risk of Tuberculosis Infection in Kenya. *Tuberculosis and the HIV Epidemic Journal:* 1986-1996J.

Harries AD. Tuberculosis and human immunodeficiency Virus infection in developing countries. Lancet. 1990; 335: 387-390.

WHO. Towards universal access: scaling-up priority HIV/AIDS interventions in the health sector. Progress report. Geneva: World Health Organization, 2007.

Impact of tuberculosis cotreatment on viral suppression rates among HIV-positive children initiating HAART Brian C. Zanonia

P. M. Jeena, P. Pillay, T. Pillay, H. M. Coovadia (2009).Impact of HIV-1 co-infection on presentation and hospital-related mortality in children with culture proven pulmonary tuberculosis in Durban, South Africa. *International Journal Tuberculosis and Lung Diseases* 6(8):672-678

Petros Isaakidis 1Mdecins Sans Frontires, Phnom Penh, Cambodia Isaakidis et al(2010).High survival and treatment success sustained after two and three years of first-line ART for children in Cambodia *Journal of the International AIDS Society* 13:11

Braitstein et al(2010). The Clinical Burden of Tuberculosis Among Human Immunodeficiency Virus-Infected Children in Western Kenya and the Impact of Combination Antiretroviral Treatment Infectious Disease Journal

Petra Bukov and Thomas Lumley (2007).Longitudinal Data Analysis for Generalized Linear Models with Follow-up Dependent on Outcome-Related Variables.*The Canadian Journal of Statistics* ,Vol. 35, No. 4, pp. 485-500

Christopher Au-Yeung et al(2011).Tuberculosis mortality in HIV-infected individuals: a cross-national systematic assessment: *Clinical Epidemiology.* 3: 21-29. 10.2147/CLEP.S15574 [3mm] Joshi YP, et al.(2001).Co-infection of Mycobacterium tuberculosis and Human Immunodeficiency *Co-infection of TB and HIV* ViruVol. 27 No. 1 45 :- 45

Swaminathan S . (2004). Tuberculosis in HIV-infected children. *Pediatric Respiration Review* 5(3):225-30

James A. Hanley on (2002) *Statistical Analysis of Correlated Data Using Generalized Estimating Equations*: An Orientation.

Bell RQ. 1953. Convergence: an accelerated longitudinal approach. Child Dev. 24:145-52

Boker SM, Graham J. (1998). A dynamical systems analysis of adolescent substance abuse. *Multivar. Behav. Res.* 33:479-507

Boker SM, Laurenceau J-P. (2006). Dynamical systems modeling: an application to the regulation of intimacy and disclosure in marriage. See Walls & Schafer 2006, pp. 195-218

Boker SM, Nesselroade JR. (2002).Amethod for modeling the intrinsic dynamics of intraindividual variability: recovering the parameters of simulated oscillators in multi-

wave panel data. *Multivar. Behav. Res.* 37:127-60

Bollen KA, Curran PJ. (2004). Autoregressive Latent Trajectory (ALT) models: a synthesis of two traditions. *Sociol. Methods Res.* 32:336-83

Chung H, Park Y, Lanza ST. (2005). Latent transition analysis with covariates: pubertal timingand substance use behaviors in adolescent females. *Stat. Med.* 24:2895-910

Cohen P. (1991). A source of bias in longitudinal investigations of change. Collins & Horn 1991, pp. 18-25

Collins LM, CliffN. (1990). Using the Longitudinal Guttman Simplex as a basis for measuring growth. *Psychol. Bull.* 108:128-34

Collins LM, Graham JW. (1991). Comment on "A source of bias in longitudinal investigationsof change: a problem posed to attenders at the 'Best Methods for the Analysis of Change' conference." See Collins & Horn 1991, pp. 26-30

Collins LM, Graham JW. (2002). The effect of the timing and temporal spacing of observations in longitudinal studies of tobacco and other drug use: temporal design considerations. *Drug Alcohol Depend.* 68:S85- 96

Collins LM, Horn JL, eds. (1991).*Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: Am. Psychol. Assoc.

Collins LM, Sayer AG, eds. (2001). *New Methods for the Analysis of Change.* Washington, DC: Am. Psychol. Assoc.

Collins LM,Schafer JL,KamCK. (2001).Acomparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychol. Methods* 6:330-51

Cox DR. (1972). Regression models and life tables.*J. Roy. Stat. Soc. Ser.* B 34:187-202

Cox DR, Lewis PAW. (1966). *The Statistical Analysis of Series of Events.* London: Chapman & Hall

Cressie N. (1991). *Statistics for Spatial Data.* New York: Wiley

Csikszentmihalyi M, Larson R. (1987). Validity and reliability of the experience-

sampling method. J. Nerv. Ment. Disord. 175:526-36

Cudeck R, Klebe KJ. (2002). Multiphase mixedeffects models for repeated measures data. *Psychol. Methods* 7:41-63

Cumsille PE, SayerAG, Graham JW. (2000). Perceived exposure to peer and adult drinking as predictors of growth in positive alcohol expectancies during adolescence. *J. Consult. Clin. Psychol.* 68:531-36

Duncan SC, Duncan TE, Hops H. (1996). Analysis of longitudinal data within accelerated longitudinal designs. *Psychol. Methods* 1: 236-48

Von Eye A, ed. (1990a.) *Statistical Methods in Longitudinal Research, Volume 2: Time Series and Categorical Longitudinal Data. Boston,* MA: Academic

Linda M. Collins . (2006).Analysis of Longitudinal Data: The Integration of Theoretical Model, Temporal Design, and Statistical Model.*Annu. Rev. Psychol.* 57:505-28

Appendices

MacNemar Definition

The test is applied to a 2 2 contingency table, which tabulates the outcomes of two tests on a sample of n subjects, as follows. The null hypothesis of marginal homogeneity

Table 7.1: Type of HIV diagnosis by age group

	Test 2 positive	Test 2 negative	Row total
Test 1 positive	a	b	a + b
Test 1 negative	c	d	c + d
Column total	a + c	b + d	n

states that the two marginal probabilities for each outcome are the same, i.e. $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$. Thus the null hypothesis is[1]

$$p_b = p_c$$

Here p_a, etc., denote the theoretical probability of occurrences in cells with the corresponding label. The **McNemar test statistic** with **Yates' correction for continuity** is given by:

$$\chi^2 = \frac{(|b - c| - c)}{b + c}$$