UNIVERSITY OF NAIROBI
COLLEGE OF HEALTH SCIENCES
INSTITUTE OF TROPICAL AND INFECTIOUS DISEASES
(UNITID)

A COMPARISON OF COX AND POISSON REGRESSION IN THE ANALYSIS OF SURVIVAL DATA

MOSES MWANGI
W62/71080/2008

A Project submitted in partial fulfillment for the award of Masters of Science Degree in Medical Statistics at the University of Nairobi, Institute of Tropical and Infectious Diseases (UNITID)

On September 2010

DECLARATION

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

Signature....................................

Date...... $29/11/2010$ ....

Moses Mwangi   Reg. No. W62/71080/2008

Supervisors:

This project has been submitted for examination with our approval as University Supervisors.

1. Signature......................................   Date. $29/11/2010$ .....

    Mrs. Anne Wang'ombe

    Senior Lecturer

    University of Nairobi

2. Signature......................................   Date.... $29-11-2010$ ..

    Dr. Rosemary Nguti (Deceased)

    Senior Lecturer

    University of Nairobi

DEDICATION

This study is dedicated to my wonderful family: My mother Margaret W. Mwangi for always believing in me, my lovely wife Monica W. Mwangi and our bundle of joy son Ryan N. Mwangi for your understanding and sacrifice in the course of preparation of this work.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

ACRONYMS

AIDS	-	Acquired Immune Deficiency Syndrome
ART	-	Antiretroviral therapy
ARV	-	Antiretroviral drugs
BMI	-	Body Mass Index
CCC	-	Comprehensive Care Centre
CI	-	Confidence Interval
CPHR	-	Center for Public Health Research
F+NC	-	Food supplement and Nutrition counseling group
FANTA	-	Food and Nutrition Technical Assistance
FBP	-	Food By Prescription
HIV	-	Human Immunodeficiency Virus
HRQOL	-	Health-related Quality of Life
KDHS	-	Kenya Demographic and Health Survey
KEMRI	-	Kenya Medical Research Institute
KENQOL	-	Kenya Quality of Life
LTFU	-	Lost to follow up
MoH	-	Ministry of Health
NC	-	Nutrition counseling group
PLWHA	-	People Living with HIV/AIDS
QOL	-	Quality of Life
RUTF	-	Ready To Use Therapeutic Food
WHO	-	World Health Organization

ABSTRACT

When analyzing event data, one can decide to analyze either individual patient survival times or aggregated patient event rates. In this study two methods are used to analyze data arising from a study where the response variable is the length of time taken to change in nutritional status using body mass index (BMI). Patients whose BMI changed from <18.5 Kg/m^2 to $\geq$ 19.5 Kg/m^2 are considered to experience an event. The study adopted an interventional comparative design with 12 months follow-up after which, patients who improved within this period were considered to experience an event while those whose event could not be clearly established were censored. The Kaplan–Meier method, log rank test, Cox proportional hazards model and Poisson model are described.

Out of 330 adult PLWHA enrolled in the FBP program, 58.8% were females and 41.2% were males. Median age was 35 (IQR, 30 - 42) years with females generally younger than males (33 [IQR, 28 – 40] years vs. 37 [IQR, 33 – 44] years), (p<0.001). Median BMI of eligible clients was 17.35 (IQR, 16.40 - 17.96) kg/m^2.There was no significant difference in BMI between females and males. A total of 123 (37.3%) clients experienced nutritional improvement after 12 months of follow-up. There was no significant difference in distribution of nutritional outcome between the two treatment modalities (P=0.245). A higher proportion of clients on nutritional counseling alone (40.7%) experienced nutritional improvement compared to those on food and nutritional counseling (34.4%). Kaplan-Meier method revealed no significant difference in survival probabilities between the two treatment modalities (P=0.162). Median time to nutritional improvement among clients receiving food and nutritional counseling was 9 [95% CI= 8 – 10] months compared to 8 [95% CI= 7 – 9] months for clients place on nutrition counseling alone. Application of Cox and Poisson regression in bivariate and multivariate analysis generated similar results. Adjusting for treatment modality, CD4 change [1 to 100 counts (RR = 4.22; 95%CI 0.94-18.95); 101 – 200 counts (RR = 5.81; 95%CI 1.30-25.81); >200 counts (RR = 6.24; 95%CI 1.37-28.41); Deteriorated / No change in CD4 = Reference category] and source of socio-support [Medical professional and family members (PR = 5.04; 95%CI 1.05-1.24); Other= Reference category] were identified as significant factors associated with nutritional improvement.

Alluding to the outcome of the results, survival analysis is not limited by the nature of data presented, whether on rates or on survivorship. When presented with data on survivorship, Cox regression is the better option and when presented with data on rates, Poisson regression is the recommended option. Even though the input variables (dependent) are different in nature (Time-to-event for Cox and Count of events per time for Poisson), the output measurement is the same i.e. Relative risk. Both analysis yield to the same conclusion.

CHAPTER 1: INTRODUCTION

In analyzing event data, one can decide to analyze either aggregated patient event rates or individual patient survival times. In analyzing aggregated event rates, the response or outcome variable is the number of events that occur divided by the number of accumulated patient-time at exposure to event which can be referred to as the incidence of the event. The event rates can then be compared between treatment modalities.

A more common, alternative approach is to examine trends in events on the basis of individually determined patient survival times. Here, the response or outcome variable is the length of time until the event of interest takes place (e.g., nutritional improvement) or until some point in time where the patient is no longer followed (e.g., a patient is lost to follow-up or is still not nutritionally improved at the end of the study). When the latter occurs, the patient survival time is said to be censored. In particular, censoring occurs whenever the elapsed time to an event is known only partially. Censored data often arises in survival time data. This arises when the event of interest in not observed within the study time or follow up time.

1.1 Model for Survivorship

Survival analysis focuses on determining the distribution of survival times as well as determining the patient characteristics that may be associated with the events. The prototypical event in many situations may be death, from which the name 'survival

1

analysis' and much of its terminology derives, but the ambit of application of survival analysis is much broader. Essentially the same methods are employed in a variety of disciplines under various rubrics; an example is 'event-history analyses' in sociology. In this study, occurrence of an event will be defined as *'Change of BMI from < 18.5 Kg/m²* *to ≥ 19.5 Kg/m²'*. Therefore, the term *survival* is to be understood generically.

There are a number of statistical models that allow one to analyze subject's survival in the presence of censored data. Although there are well known methods for estimating unconditional survival distributions, most interesting survival modeling examines the relationship between survival and one or more predictors, usually termed *covariates*.

In his famous paper *"Regression Models and Life Tables"*, David R. Cox (1972) demonstrated that the length (time-to-event) may depend on conditional information. He therefore introduced regression methods subject to which Cox proportional-hazards regression model, a broadly applicable and the most widely used method in survival analysis was developed.

The Cox model assumes that the hazard function has a log-linear form, i.e.

$$\lambda(t, X) = \exp(\alpha(t) + X'\beta)$$

This hazard function can be approximated by dividing the period of follow-up into k intervals $(\tau_1, \tau_2](\tau_1, \tau_2], (\tau_2, \tau_3], \ldots, (\tau_k, \infty)$ and assuming the hazard is constant during each interval. Hence, the function $\alpha(t)$ is approximated by a step function $\alpha(t) = \alpha_k$ for

$\tau_k < \tau < \tau_{k+1}$. Such a model has been described by Holford (1976) and it gives rise to exponential survival during each interval.

1.2 Model for Rates

Models for rates are considered in which the underlying rate at which events occur can be represented by a regression function that describes the relation between the patient characteristics and the unknown rate of occurrence (Frome 1983). When the events of interest follow the Poisson distribution, Maximum Likelihood Estimation is used. Poisson regression models are generalized linear models with the logarithm as the (canonical) link function. To establish the relation between the dependent variable and the predictor variables a log-linear model is used.

As an alternative to the Cox model, one can carry out patient survival analysis using an interval Poisson model, also referred to a piecewise exponential model (Allison 1995). The interval Poisson model is similar to the Cox model in that both account for censored data and assume the event rates between any two groups of patients will be proportional to one another (Vonesh *et al.* 2000). Like the Cox model, the Poisson model also accommodates non-proportional event rates through the use of covariates. However, unlike the Cox model, the Poisson model is semi-parametric in that it assumes event rates are constant within specified intervals of time. In fact, this is the key difference between the two models. Specifically, in the Cox model, the reference population's event rate over a specified interval of time is left unspecified, while in the Poisson model it is

assumed constant. Both models assume the event rate for a comparative group of patients will be proportional to the event rate for the reference group within each specified interval of time. For short intervals of follow-up (e.g., every 3 months or every 6 months), it is entirely reasonable to assume that the event rates will be approximately constant. Consequently, by choosing appropriate intervals of follow-up, an interval Poisson model and an interval Cox model will give very nearly the same results with respect to relative risks.

1.3 Implications of Choice of Model

Assuming the underlying event rates for two groups of patients are roughly proportional to one another over time (i.e., the relative risk of event is constant over time), a Cox proportional hazards regression provides a robust method for estimating the relative risk of event. It also enables one to plot and compare adjusted patient survival curves between the two groups without making any unnecessary assumptions about the underlying event rates. By excluding any interaction between follow-up time and treatment modality, one can obtain a similar estimate of relative risk using interval Poisson regression. However, such estimates may vary slightly depending on one's choice of intervals (Allison 1995).

When the assumption of proportional event rates (i.e., constant relative risk) is violated, application of the standard Cox proportional hazards model yields an average relative risk. In some cases, this average risk may mislead investigators into thinking one therapy

4

is superior to another when in fact there are periods of time when the opposite is true. The use of interval Poisson regression avoids this by enabling the user to model the relative risk as a function of time. This is accomplished by including an interaction term between the interval follow-up times and treatment modality. Alternatively, one can apply an interval Cox proportional hazards model using the same set-up as for the interval Poisson model. The interval Cox model has the advantage of not assuming a constant event rate within each interval of follow-up and so it may be more robust to one's choice of intervals. However, it does not allow one to carry out formal hypothesis testing with respect to the shape of the event rates over time. In any case, one can achieve comparable results using either a Cox proportional hazards model or an interval Poisson model provided one has specified all other aspects of the model the same way (Vonesh *et al.* 2000).

1.4 Research question

The important question that this study attempts to answer is:

- Does survival analysis using Cox Proportional Hazards regression differ from Poisson Regression when fitted to the same data?

1.5 Objectives

The main objective of this study is to fit two models used in survival analysis using nutritional data of adults PLWHA and asses if they differ in the conclusion they lead to.

The specific objectives are;

- To compared survival probabilities in nutritional improvement between HIV clients on food supplements plus nutrition counseling ($S_1(t)$) and those on nutrition counseling alone ($S_2(t)$).

- To model time-to-nutritional improvement (Positive BMI change from <18.5 Kg/m^2 to $\geq$ 19.5 Kg/m^2) using Cox Proportional Hazards regression.

- To model rate of nutritional improvement (Positive BMI change from <18.5 Kg/m^2 to $\geq$ 19.5 Kg/m^2) using Poisson Regression.

1.6 Study justification

It has been a preferred notion to model and estimate *time-to-event* in place of *rate at which events occur* in longitudinal studies when the event of occurrence is rare. However, two types of data which arise from medical or epidemiological investigations are: data on rates and data on survivorship. The sources of data may be official vital statistics or disease registries which may involve a large number of individuals. Others may be obtained from Randomized Clinical Trials. One type of statistical summary is a set of rates [(number of events) / (total length of exposure) or (total number at risk)]

6

computed for the population broken down by several factors (e.g. age, race and sex). The second type of data may be the survival experience of the population, where the summary is based on the life table and individuals are again categorized by several variables.

When confronted with either of these situations it is critical to figure out on the suitable statistical methodology for analyzing the data. This study aims at using the two methodological approaches to survival analysis: Cox Proportional Hazards Regression and Poisson Regression. Two formats of the same data (Aggregated and non-aggregated) will be prepared and both methods applied.

CHAPTER 2: LITERATURE REVIEW

2.1 Models for Survivorship

A great deal of work has recently been done on the analysis of censored survival data, much of it inspired by the studies of Cox (1972). Typically, the analysis of individual patient survival times is carried out using a Cox proportional hazards regression model. Cox, DR and Oakes (1984), Kalbfleisch *et al.* (1980), and Allison (1995) are practical examples of Statistical Analysis of Survival Time Data.

The model got its name from the assumption that the hazard function or event rate for one group of patients will be proportional to the hazard function or event rate from another group. This is equivalent to assuming the relative risk of event between the two groups will be constant over time. This assumption does not require that the event rates themselves be constant in time; it merely requires that their ratio be constant. When the assumption of proportionality is violated, one can still use the Cox model by simply introducing an appropriate set of covariates into the regression and possibly interaction term with time. In addition to the above mentioned studies Holford (1976) and Aitkin *et al.* (1989) has described this type of modeling.

According to Vonesh *et al.* (2000), one of the chief advantages of the Cox model is that there are no assumptions regarding what the shape of the underlying hazard or event rate looks like. It is for this reason that estimates of relative risk are more robust under the

Cox model than what might otherwise be obtained using a fully parametric model. Allison (1995) points this as a disadvantage in that one can not formally test hypotheses about the shape of the hazard function although one can still estimate and describe its shape.

There have been a number of recent registry-based studies that examine the issue of mortality among patients with end-stage renal disease (ESRD), with particular emphasis placed on comparisons between patients receiving in-center hemodialysis (HD) versus those receiving home peritoneal dialysis (PD). On the surface, the results of some of these studies appear to contradict one another. The study done by Bloembergen *et al.* (1995) shows a survival advantage for Hemodialysis patients, the one done by Vonesh *et al.* (1999) shows no difference in survival between Hemodialysis and Peritoneal dialysis, while the study done by Fenton *et al.* (1997) shows a survival advantage for patients receiving peritoneal dialysis. Alluding to the two models afore described Vonesh *et al.* (2000) was able to demonstrated that differences in patient survival results as published in the literature are not due to differences in the statistical model used (i.e., a Cox regression model versus a Poisson regression model) but rather to the choice of analysis (Intention to treat versus As-treated) and type of the patients to be studied (Prevalence versus incidence). In this review, the Cox Proportional Hazards Regression method was demonstrated.

2.2 Models for Rates

Models for event rates have been used to model various diseases or outcomes of interest. The event rates can be compared between treatment modalities using Poisson regression. In the book 'Generalized Linear Models' McCullagh and Nelder (1989) have shown that closely to the Poisson model are models for the analysis of counted data in the form of proportions or ratio counts. They have explained that in medical and pharmaceutical trials, it is usually required to study not primarily the incidence of a particular disease but how the incidence is affected by factors such as age, social class, housing conditions, exposure to pollutants, and any treatment procedure under study. Generalized linear models permit us to study patterns of systematic variations in much the same way as ordinary linear models are used to study joint effects of treatment and covariates. Similarly, in his paper 'The analysis of rates using Poisson regression models' Frome (1983) has demonstrated that models for rates can be represented by a regression function that describes the relation between the patient characteristics and the unknown rate of occurrence.

Holford (1980) discusses the analysis of rates using log linear models. His discussion echoes what Bishop, Fienberg and Holland (1975) did on the application of log-linear models to rates where the rate for the subgroup i ($i = 1, \ldots, L$) of the population is $\lambda_i = {n_i}/{T_i}$. In this case, T_i is the total population size for subgroup i and n_i is the number in i who exhibited the event. However, in other contexts T_i may be a measure of total

length of time that the population was under observation. Berry (1983) was able to demonstrate this concept using data on observed mortality of a group of individuals allowing for age and period. Using subject-years method (Case and Lea 1955) Berry was able to determine expected deaths in which each person is assumed at risk up to the date of the analysis, the date of death or the date the person was LTFU, which ever come first.

One model which has been described by Armitage (1966) assumes that n_i has a Poisson distribution with mean $m_i = \lambda_i T_i$. λ_i, is assumed to have a log-linear relationship with the vector of J variables, X_i, thus $\lambda_i = \exp(\alpha + X_i'\beta)$. This yields a multiplicative model for the rates, which has also been postulated by Kilpatrick (1962), Bjarnason et al. (1974), Breslow and Day (1975), Osborn (1975) and Gail (1978). Haberman (1978) discusses the application of log-linear models to this case. To ensure a fair comparison, adjustments are often made for case-mix differences in covariates e.g. age, gender, level of education, and other exposure factors. The studies by Bloembergen et al. (1995), Vonesh et al. (1999) are examples of this kind of analysis.

CHAPTER 3: METHODOLOGY

3.1 Motivating study

3.1.1 HIV and Nutrition

Nutrition is acknowledged as an important factor in the management of HIV infection. It helps improve the health status of people living with HIV/AIDS (PLWHA) by providing nutrients that boost the immune system. It is an important component of comprehensive care for the HIV-infected individuals, and particularly so in resource-limited settings where malnutrition and food insecurity are endemic. There have been a number of studies on FBP comparing nutritional improvement rates between treatment groups among PLWHA. The motivating study to this review is one of the under takings of FANTA in addressing the existing gaps in comprehensive care for the HIV-infected individuals.

In addressing this gaps an experimental study entitled *'RANDOMISED CONTROLLED EVALUATION OF THE IMPACT OF FORTIFIED FOOD SUPPLEMENTS AMONG WASTED HIV INFECTED ADULTS ON ART & PRE-ART IN DIFFERENT SITES IN KENYA'*, **SSC 1023**, was initiated during the more recent roll out on Food by prescription in 2008 by a dedicated team of researchers from 1) Kenya Medical Research Institute - Kenya, 2) FANTA Project, AED - Uganda, 3) FANTA Project, AED - USA, 4) Kenyatta National Hospital - Kenya, and 5) Insta

Food Products Ltd - Kenya.

Data was collected from six CCC sites namely; Mbagathi, Muragua and Naivasha District Hospitals, Mathare and Riruta Health centres, and Nyeri Provincial General Hospital. The Study population comprised of HIV+ adults attending CCC in afore mentioned sites in Kenya.

3.1.2 Sampling and Analysis Design

Sampling Method

Two basic groups of clients were considered for evaluation in the bigger study. The first group (Group A) was wasted adult clients (BMI< 18.5) who were eligible for ART. The second group (Group B) was those adult clients who were wasted (BMI<18.5), symptomatic but not yet eligible for ART according to WHO guideline and Kenya national ART guidelines.

The impact of therapeutic and supplementary food supplements on ART clients was assessed in terms of nutritional status, effectiveness of and adherence to ART management of ART side effects, and overall QOL. This study concentrates on the nutritional component using BMI as the indicator of nutritional status.

Randomisation and treatment allocation: The study utilised a complete random block design in the allocation of subjects to various treatment intervention.

Clients from each arm were recruited into 60 blocks of 10 clients each (5 on nutritional counselling alone and 5 on nutritional counselling and Insta food supplement). Randomization was done within blocks to allocate the patients/clients to either treatment intervention within each block

A random numbered assignment of 10 numbers was generated and each patient/client was required to collect a card number which was matched to either of the two treatments available in opaque sealed envelope, which was given to the subject upon completing the informed consent process. The subject presented the card for food collection, at an adjacent room/site adjacent to the clinic, where nutritional counselling was offered to all for healthy eating with HIV and food supplements to 5 clients per block.

The food supplements were distributed monthly at the diet and nutrition centres within the study sites where nutritional counselling was available and anthropometric measurements were taken in all referred clients. Follow-up assessment and data collection was conducted during clients' monthly visits to the treatment centre. In cases of drop-outs, efforts were made through home visits and other approaches to determine whether death was the reason for drop-out.

Sample Size Determination

The primary outcome of the study was improvement on nutritional status measured using BMI change. The sample size needed to detect differences in BMI of 1.0, with 95% specificity and 90% power. Going by an average BMI increase of 1.0 kg/m^2, with an assumed standard deviation of 1.77 kg/m^2 the sample size required to detect such a difference was 66 per group. Using n: n ratio, minimum sample size requirement was **132 clients**.

The following formula was used for sample size computation:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 2\sigma^2}{(\mu_1 - \mu_2)^2}$$

Where:

α = significant level (0.05)

$1-\beta$ = the power of the study (90%)

$Z_{1-\alpha/2}$ = Z-value attributed to $\alpha/2$ (1.96)

$Z_{1-\beta}$ = Z-value attributed to $1-\beta$ (1.28)

$\mu_1 - \mu_2$ = the expected difference (after food supplementation and nutrition counseling) in BMI that can be detected by the sample population and that can be claimed to be the effect of supplementation (1.0)

Since the available records was more than the minimum required sample size, all the clients were considered for analysis, 180 on nutritional counseling and food supplement arm and 150 on nutritional counseling alone giving a total of **330 clients**.

Study Design

The study design for the bigger study was interventional comparative with 12 months follow-up recruiting both ART and pre-ART clients. Since clients on ART and pre-ART groups are at different stages of the illness and have different needs in terms of food and care, this study (my project) chose to take the ART clients under which two treatments modalities were administered.

```
┌─────────────────────────────────────────┐
│                 ART Arm                   │
│            WHO stage II or III            │
│     Started Cotrimoxazole (prophylaxis)   │
│              within 4 months              │
└─────────────────────────────────────────┘

┌──────────────────────────┐    ┌──────────────────────────┐
│ Food supplementation and │    │   Nutrition counseling    │
│    Nutrition counseling   │    │           only            │
└──────────────────────────┘    └──────────────────────────┘
```
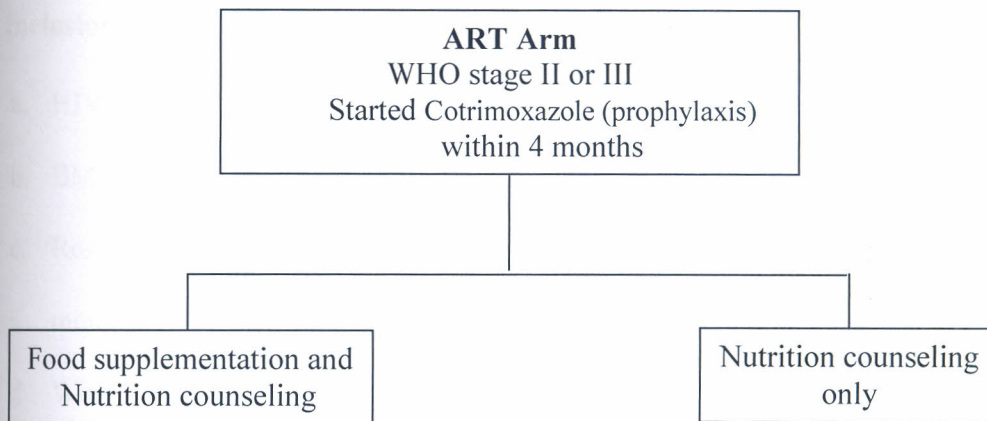
Subjects: A total of 330 subjects were randomly assigned to receive one of the two interventions:

Group	Intervention	Sample size
Group I	Nutritional counselling alone	150
Group II	Nutritional counselling and 300 g/d of Insta flour (blend of corn, Soya sugar, palm oil, and micronutrients)	180

Each of the recruited study subjects was required to attend the CCC once every month to undergo a nutrition assessment and a counseling session. During each visit, the patient's anthropometric measurements, i.e. Height, Weight, MUAC and Bioelectric measurements were taken. Those on the food arm were given supplies to last until the next visit.

Inclusion / Exclusion Criteria

Inclusion:

a. HIV-positive adults ($\geq$ 18 years) attending CCC

b. BMI < 18.5 kg/m^2 and > 14 kg/m^2

c. Residents within the site residence (or able to attend the clinic) for at least 12 months and not likely to move out

d. Willing to participate in the study voluntarily.

e. Not on ART but eligible to begin ART within one month (WHO stage II or III and/or CD4 count between 200 - 350 cells/mL)[20].

f. Beginning or have begun Cotrimoxazole prophylaxis within the past 4 months.

Exclusion:

a. HIV negative adults

b. BMI $\geq$ 18.5 kg/m^2 and $\leq$ 14kg/m^2

c. Pregnant and lactating women

d. Women who become pregnant during the study

e. Subjects already receiving another food supplement

f. Subject with previous contact with ART

3.1.3 Food by Prescription (Intervention)

Only one arm of the study received food supplement. The food supplement to be used was produced by Insta Foods Kenya Ltd and is known as Insta Food Foundation. It is a blend of maize, soya, sugar, palm oil, and micronutrient pre-mix. Insta Food Foundation is a pre-cooked RUTF and therefore needs to be hydrated using boiled hot water or milk before consumption. Nutritional counseling was provided to all referred clients. Follow-up assessment of adherence and the use of the food was conducted during client's monthly visits to the treatment centre.

3.1.4 Data collection

Data collection was undertaken by trained staff by use of questionnaires. The data collection tool was pre-tested on a trial run of 5 clients to ensure the procedures and logistic are working.

3.1.5 Nutrition assessment data

Body mass index (BMI) is a nutrition indicator that measures the body's weight relative to height. It is a better predictor of disease risk than body weight alone. BMI was used to determine the degree of wasting on the subjects. BMI was calculated by taking an individual's weight in kilograms and dividing it by subject's height in metres squared.

3.1.6 Ethical approval

Ethical permission was sought from the Scientific Steering Committee and Ethical Review Committee, KEMRI. Subjects were required to give consent to participate in the study by signing a consent form. Permission to undertake the study in the various health facilities was sought from the respective facility heads.

3.2 Background to this project

A Review of Kenya's Food by Prescription Program done in July 2009 was to examine service delivery under the FBP program with a view towards learning more about specific issues such as duration of food supplementation, LTFU, changes in client nutritional status, and the food delivery system. This study is aimed at shedding more light on duration of food supplementation in relation to changes in client nutritional status. BMI at recruitment (Entry into the program) was differenced from BMI at graduation (Exit from the program) therefore giving a positive, negative or no change in BMI. This study considers *'event'* to be nutritional improvement, i.e. *Positive BMI change from < 18.5 to $\geq$ 19.5 Kg/m^2*. Nutritional improvement (event of interest) is compared among individuals put on Food supplementation and nutritional counseling with those put on nutritional counseling alone.

3.3 Data Description

The bigger FANTA study collected information about adult HIV clients and this study selected the data needed for the achievement of its objectives. The data needed for the analysis of risk factors for improved nutritional status of adult PLWHA. Data on individual client characteristics that are likely to determine nutritional improvement and the length of time to nutritional improvement, for example sex of the client, education level, monthly family income, financial and social support, etc., were considered in the survival analysis. The unit of study was adult HIV client, this study analyzed data on

20

clients followed-up for 12 months. A sample of 330 adult HIV clients (136 males and 194 females) was created from the larger dataset with a total of 1380 records corresponding to both the ART and pre-ART clients enrolled into the study.

3.4 Analytical Methods

3.4.1 Survival analysis

In this section we describe briefly relevant techniques and methods of standard survival data analysis. In survival analysis the following are key variables;

Time variable (t_i): The variable measures duration to the event defined by the status variable. Time will be measured in months from enrollment into the study to improvement (Positive BMI change from <18.5 Kg/m^2 to $\geq$19.5 Kg/m^2), lost to follow-up, termination due to extraneous factors or end of the study whichever comes first.

Status variable (δ_i): It is also called the event or censoring variable. In this study the censoring variable is status of nutritional improvement. Those clients who improve are considered to experience an event while all the others are censored. Events are coded as 1 and censored as 0. In Cox regression, the outcome (dependent) variable is a combination of time and status variable (t_i, δ_i).

Covariates (X_i): These are predictors/independent variables which are assessed for their association with the event of interest. A number of covariates are tested for their association with time to improvement on nutritional status or to rate of improvement on nutritional status.

For survival analysis to be carried out the following are required.

- Well defined time of origin (Date of recruitment).
- Well defined event of interest (Improvement on nutritional status).
- Well defined scale of measurement (Duration in months from time of recruitment into the program to the time of exit from the program).

Describing the Distribution of Time to an Event

In routine data analysis, we may first present some summary statistics such as mean, standard error for the mean, etc. In analyzing survival data, however, because of possible censoring, the summary statistics may not have the desired statistical properties, such as unbiasedness. For example, the sample mean is no longer an unbiased estimator of the population mean (of survival time). So we elect to use other methods to present our data. One way is to estimate the underlying true distribution. When this distribution is estimated (either parametrically or non-parametrically), we then can estimate other quantities of interest such as mean, median, etc., of the survival times. The distribution

of the random variable T can be described in a number of equivalent ways. There is of course the usual (cumulative) distribution function.

Estimating the survival curve using the Kaplan–Meier method

In analyzing survival data, two functions that are dependent on time are of particular interest: the survival function and the hazard function. The survival function S (t) is defined as the probability of surviving at least to time t. The hazard function h (t) is the conditional probability of dying at time t having survived to that time. The graph of S (t) against t is called the survival curve. The Kaplan–Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution (Kaplan *et al.* 1958). The method is based on the basic idea that the probability of surviving k or more periods from entering the study is a product of the k observed survival rates for each period (i.e. the cumulative proportion surviving), given by the following:

$$S(k) = p_1 * p_2 * \ldots * p_k \qquad (3.1)$$

Here, p_1 is the proportion surviving the first period, p_2 is the proportion surviving beyond the second period conditional on having survived up to the second period, and so on. The proportion surviving period i having survived up to period i is given by:

$$p_i = \frac{n_i - d_i}{n_i} \qquad (3.2)$$

Where n_i is the number alive at the beginning of the period and d_i the number of deaths within the period.

The Hazard and Survival Functions

Let T be a non-negative random variable representing the waiting time until the occurrence of an event. For simplicity we will adopt the terminology of survival analysis, referring to the event of interest as `death' and to the waiting time as `survival' time, but the techniques to be studied have much wider applicability. They can be used, for example, to study age at marriage, the duration of marriage, the intervals between successive births to a woman, the duration of stay in a city (or in a job), and the length of life. In this study, occurrence of an event is defined by nutritional improvement i.e. Positive BMI change from $<18.5 \text{ Kg/m}^2$ to $\geq 19.5 \text{ Kg/m}^2$.

The Survival Function

We will assume for now that T is a continuous random variable with probability density function (p.d.f.) f(t) and cumulative distribution function (c.d.f.) $F(t) = \Pr[T \leq t]$, giving the probability that the event has occurred by duration t. It will often be convenient to work with the complement of the c.d.f, the survival function

$$S(t) = \Pr[T > t] = 1 - F(t) = \int f(x)dx, \qquad (3.4)$$

which gives the probability of being alive at duration t, or more generally, the probability that the event of interest has not occurred by duration t.

24

The Hazard Function

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event, defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t < T \le t + \Delta t \mid T > t\}}{\Delta t} \qquad (3.5)$$

The numerator of this expression is the conditional probability that the event will occur in the interval $(t, t + \Delta t)$ given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $(t, t + \Delta t)$ and $T > t$ (which is, of course, is the same as the probability that t is in the interval), to the probability of the condition $T > t$. The former may be written as $f(t)\Delta t$ for small Δt, while the latter is $S(t)$ by definition. Dividing by Δt and passing to the limit gives the useful result

$$h(t) = \frac{f(t)}{S(t)} \qquad (3.6)$$

where $f(t)$ is the density function and $S(t)$ is the survival function, Collet (2003).

A closely related function to the hazard function is the cumulative hazard function denoted by $H(t)$ and defined as,

$$H(t) = -\ln(S(t)) \qquad (3.7)$$

25

Comparing the survival of two groups

The simplest way of comparing survival times obtained from two groups of individual is to plot the corresponding estimates of the two survivor function on the same axes. The resulting plot can be quite informative. However, like in classical theory, the basic quantities or summary statistics obtained across the two groups can be compared.

In survival analysis, comparison of the two treatment modalities is done using a statistical hypothesis test called the *Log rank test,* Peto *et al.,* (1977). It is used to test the hypothesis that there is no difference between population survival curves (i.e. the probability of an event occurring at any time point is the same for each population).

Another is the Wilcoxon which is an example of a generalized Log-Rank test. The underlying null hypothesis is that there is no difference on the survival function of both group1 ($S(t)_1$) and group2 ($S(t)_2$).

H_0: $S(t)_1 = S(t)_2$ vs H_1: $S(t)_1 \neq S(t)_2$

Log-Rank Test

Let $t_{(1)} < t_{(2)} < t_{(3)} < \ldots \ldots < t_{(r)}$ be distinct ordered event times across the two groups.

Considering time $t_{(j)}$ then we have the following 2X2 at this time point.

Group	Event	Survivors	Total
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Where d_j and n_j are the response total number of events and those at risk at time $t_{(j)}$.

Further d_{ij} and n_{ij} are the respective numbers of events and those at risk at this time point in the *j-th* group; $j = 1, 2$.

Assuming that the null hypothesis is true and the margins are fixed, then this table can be solemnly determined by d1j – the number of events in group1. Thus d1j can be thought to be a variable that assumes integer values 0,…, min (n_{1j}, d_{1j}). Hence d_{1j} has a hypergeometric distribution whose probability density function is:

$$\frac{\binom{n_{1j}}{d_{1j}} \binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}} \qquad (3.8)$$

It thus follows that

$$E\,(d_j) = e_{1j} = \frac{n_j - d_j}{n_j} \tag{3.9}$$

e_{1j} is the expected number of events in group1 obtained by multiplying n_{1j} with the

probability of event $(d_j\,/\,n_j)$. Further,

$$Var[d_{1j}] = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{3.10}$$

The Log-Rank statistic is defined as;

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}) \tag{3.11}$$

This statistics belong to the (O – E) group of test statistics. U_L collects the (O – E) across

the 'r' event times.

Now,

$$E[U_L] = \sum_{j=1}^{r} E\!\left(d_{1j} - e_{1j}\right) = \sum_{j=1}^{r} E(d_{1j}) - e_{1j} \tag{3.12}$$

$$= \sum_{j=1}^{r}(e_{1j} - e_{1j}) = 0$$

Also,

$$Var[U_L] = \sum_{j=1}^{r} Var(d_{1j}) \tag{3.12}$$

Note: The underlying assumption is that there is independence across the event times.

28

$$Var[U_L] = \sum_{j=1}^{r} \left(\frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \right) = V_L \tag{3.13}$$

Asymptotically $U_L \sim N\ (0,\ V_L)$

Therefore,

$(U_L / V_L) \sim \chi^2\ (1)$, Collett, (1994).

Thus (U_L / V_L) is the Log-Rank Test Statistic.

At α – level of significance the decision rule for assessing the hypothesis is:

If $U_L > \chi_{1-\alpha}^2\ (1)$ we reject the null hypothesis.

If $U_L \leq \chi_{1-\alpha}^2\ (1)$ we do not reject the null hypothesis.

3.4.2 The Cox Proportional-Hazards Model

Cox Proportional Hazards models concerns with the analysis of data which have three main characteristics: (1) the dependent variable or response is the waiting time until the occurrence of a well-defined event, (2) observations are censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and (3) there are predictors or explanatory variables whose effect on the waiting time we wish to assess or control. We start with some basic definitions.

Fitting the Cox Proportional Hazard Model

This entails obtaining parameter estimates for the unknown beta (β) coefficients. The baseline hazard $h_0(t)$ may also be estimated. This two components can be estimated

separately by first estimating the beta *(β)* using the Maximum Likelihood Estimator methods and then $h_0(t)$ non-parametrically.

Sir David R Cox (1972) showed that one can obtain consistent highly efficient estimators of betas *(β)* by maximizing a Partial Likelihood independently of $h_0(t)$. Suppose that the data for *n* individuals consists of *r* distinct times assuming that there is only one event at each failure times (no ties).

Let $t_1 < t_2 < t_3 < \dots\dots\dots < t_m$ be the ordered event times and $R(t_{(j)})$ be the risk set at $t_{(j)}$, Then the partial likelihood Cox (1972) to be maximized for the betas (β) is

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(\underline{\beta}'\underline{X}_{(i)})}{\sum_{LR(t_{(j)})} \exp(\underline{\beta}'\underline{X}_L)} \quad ; \quad \text{i=1\dots n, j=1\dots m} \qquad (3.14)$$

Where $\underline{X}_{(i)}$ is the vector of covariates for the individuals whose event occurs at time $t_{(j)}$ and $R(t_{(j)})$ is the risk set at a time for *i-th* individual. The product is taken over all individuals. Those whose event occurs form the numerator while the denominator is formed by both the individuals whose event occur and those who are censored. This implies that the individuals who are censored make a contribution only in the risk set in the denominator.

$$\text{Let } \delta_i = \begin{cases} 1 \text{ if i}^{\text{th}} \text{ individual fails,} \\ \\ 0 \text{ if i}^{\text{th}} \text{ individual is censored} \end{cases}$$

Then the Partial Likelihood can also be given as,

$$\left[L(\underline{\beta}) = \prod_{i=1}^{n} \frac{\exp(\underline{\beta}'\underline{X}_{(i)})}{\sum_{LR(t_{(j)})} \exp(\underline{\beta}'\underline{X}_L)} \right]^{\delta_i} \quad (3.15)$$

The log-Partial Likelihood is,

$$\log L(\underline{\beta}) = \sum_{i=1}^{n} \delta_i \{ \underline{\beta}'\underline{X}_{(i)} - \log[\sum_{L \in R(t_{(j)})} \exp(\underline{\beta}'\underline{X}_L)]\} \quad (3.16)$$

By maximizing $L\ (\underline{\beta})$ (3.16), Maximum Likelihood Estimates for $\underline{\beta}$ in the proportional hazards model can be obtained. This is achieved by using numerical methods (iterative techniques) such as Newton Raphson procedure where an estimate of the vector of β parameters at the $(s+1)$-th cycle of the iterative procedure, $\hat{\beta}_{s+1}$, is given by

$$\hat{\beta}_{s+1} = \hat{\beta}_s + I^{-1}(\hat{\beta}_s)U(\hat{\beta}_s) \quad (3.17)$$

Here $U\hat{\beta}_{s+1}$ is the vector of efficient scores and $I^{-1}(\hat{\beta}_s)$ is the inverse of the information matrix,

$$I(\beta) = -\left[\frac{\partial^2}{\partial\beta_i \partial\beta_j} L(\beta) \right] \quad (3.18)$$

both evaluated at $\hat{\beta}_s$.

When the iterative procedure has converged, the variance-covariance matrix of the parameter estimates can be approximated by the inverse of the information

matrix, $I^{-1}(\hat{\beta}_s)$, evaluated at $\hat{\beta}_s$. The square of the diagonal elements of this matrix are then the standard errors of the estimated values of β parameters.

That is,

$$s.e(\hat{\beta}_j) = \sqrt{diag(I^{-1}(\hat{\beta}_j))}_j$$

Maximum Likelihood Estimates for $\underline{\beta}$ can be obtained using software (See R - syntax in the Appendix B).

Under the proportional hazards model for survival data, particularly in the construction of the partial likelihood function (3.15), it is assumed that the hazard function is continuous and tied survival times are not possible. But practically, survival times are usually recorded to the nearest weeks, days, months or years and so tied survival times can arise as a result of this rounding process or more than one than one censored time at an event time, Collet (2003). In such a case the likelihood function take a more complicated form and its computation can be time consuming. In presence of tied survival times, the simplest approximation to the likelihood function is Breslow (1974) and is computationally straightforward but only adequate when the number of tied observations at any one event time is not too large:

$$L(\underline{\beta}) = \prod_{j=1}^{m} \frac{\exp(\underline{\beta}'s_j)}{\left\{\sum_{L \in R(t_{(j)})} \exp(\underline{\beta}'\underline{X}_L)\right\}^{\delta_i}} \qquad (3.19)$$

Where s_j is the vector of sums of each of the p covariates for those individuals who experience the event at the jth event time $t_{(j)}$, $j = 1, \ldots, m$. The event d_j at time $t_{(j)}$ are considered to be distinct and to occur sequentially.

Efron (1977) proposed the following approximation which is closer to appropriate likelihood function than that due to Breslow although both often give similar results.

$$L(\underline{\beta}) = \prod_{j=1}^{m} \frac{\exp(\underline{\beta}'s_j)}{\prod_{k=1}^{d_i}\left\{ \sum_{L \in R(t_{(j)})} \exp(\underline{\beta}'s_L) - (k-1)d_i^{-1} \sum_{L \in D(t_{(j)})} \exp(\underline{\beta}'\underline{X}_L) \right\}} \qquad (3.20)$$

$D(t_{(j)})$ is the set of all individuals who experience the event at time $t_{(j)}$.

German Rodriguez (2006) compared exact, marginal, Breslow and Efron partial likelihood functions and showed that a good approximation is the Efron's. This is the default approximation in R package.

Inference for the beta (β) parameters

The Wald statistic for making inference about β is the quadratic form,

$$W = (\hat{\beta} - \beta)'I(\hat{\beta})(\hat{\beta} - \beta) \qquad (3.21)$$

and has a chi-square distribution with p degrees of freedom, where p is the dimension of the information matrix. This is useful for conducting a one-sided test by comparing W to $\chi^2_{1-\alpha}$ at α-level of significance. Asymptotically, $\hat{\beta}$ is a standard normal vector,

$$\hat{\beta} \sim N(\beta, I^{-1}(\hat{\beta}))$$

Taking the square root-like of W, we get a Z-standard normal statistic and use the ratio

$$Z_k = \frac{(\hat{\beta}_k - \beta_k)}{s.e(\hat{\beta}_k)} \sim N(0,1) \text{ with } s.e(\hat{\beta}_k) = \frac{1}{\sqrt{I(\hat{\beta}_k)}}$$

by conducting a two sided test of hypothesis about and constructing a confidence interval for a single parameter β_k at α-level of significance, Dobson (2002).

The statistic

$$Z_k = \frac{(\hat{\beta}_k)}{s.e(\hat{\beta}_k)} \tag{3.22}$$

is a Wald test statistic of the null hypothesis $H_0 : HR_k = 1 <=> \beta_k = 0$, where

$\beta_k = (\beta_2, ..., \beta_\alpha)$ are the $(\alpha - 1)$ coefficients corresponding to $Z_2, ..., Z_\alpha$ (or $Z_1, ..., Z_{\alpha-1}$, depending on the reference group).

Inference for the Hazard Ratio

Suppose that the model has one covariate X with two levels; $X=0$ and $X=1$,

Then

$$h_i(t) = h_0(t), \text{ if } X=0$$

$$h_i(t) = h_0(t)\exp(\beta), \text{ if } X=1$$

$$\frac{h_i(t)}{h_0(t)} = \exp(\beta) = \psi(HR) \tag{3.23}$$

Now,

$$\hat{\beta} \sim N[\beta, Var(\hat{\beta})]$$

Many software packages provide estimates of β, but the hazard ratio *(HR)* = exp *(β)* is usually the parameter of interest. We can use the delta method to get standard errors for exp *(β)*, Collett (2003).

Using Delta method we have that,

$$Var(\psi) = (\exp(\hat{\beta}))^2 Var(\hat{\beta})$$

$$= (\psi)^2 Var(\hat{\beta}) \qquad 3.24$$

Therefore,

$$s.e(\hat{\psi}) = \hat{\psi} \; s.e(\hat{\beta})$$

The (1-α) 100% Confidence Interval for ψ is,

$$\psi \pm Z_{\alpha/2} * s.e(\hat{\psi}) \qquad (3.25)$$

Constructing confidence intervals for exp *(β)* can be done using two options: (assuming that β is a scalar)

I. Using $s.e(\exp(\hat{\beta})$ obtained above via the delta method as

$$s.e(HR) = s.e(\exp(\hat{\beta})) = [Var(\exp(\hat{\beta}))] \qquad (3.26)$$

We calculate the endpoints as:

[Lower, Upper] = [HR $-$ 1.96 * s.e(HR), HR + 1.96 * s.e(HR)] $\qquad$ (3.27)

II. Form a confidence interval for $\hat{\beta}$, and then exponentiating the endpoints.

[Lower, Upper] = [exp ($\hat{\beta}$ $-$ 1.96 * s.e($\hat{\beta}$)), exp ($\hat{\beta}$ + 1.96 * s.e($\hat{\beta}$))] $\qquad$ (3.28)

Comparing Alternate Models

Suppose two models; 1 and 2 are fitted to the same data, and that model 1 contains a subset of explanatory variables in model 2. Then model 1 is said to be nested in model 2.

Specifically suppose model 1 have p variables $X_1, X_2, ..., X_p$ and model 2 has q additional variables $X_{p+1}, X_{p+2}, ..., X_{p+q}$. The problem is then to determine whether the additional q terms in model 2 are necessary indicating that model 1 will be more adequate for the data. Log likelihood Ratio test is used for such a comparison.

Log likelihood Ratio test

Suppose there are $(p + q)$ explanatory variables measured:

$X_1, X_2, ..., X_p, X_{p+1}, X_{p+2}, ..., X_{p+q}$ and proportional hazards are assumed.

Consider the following models:

• **Model 1:** (contains only the first p covariates)

$$\frac{\lambda_i(t,X)}{\lambda_0(t)} = \exp(\beta_1 X_1 + ... + \beta_p X_p)$$

• **Model 2:** (contains all $(p + q)$ covariates)

$$\frac{\lambda_i(t,X)}{\lambda_0(t)} = \exp(\beta_1 X_1 + ... + \beta_p X_p + \beta_{p+1} X_{p+1} + ... + \beta_{p+q} X_{p+q})$$

These are *nested* models. For such nested models, we can construct a **likelihood ratio** test of $H_0 : \beta_{p+1} = ... = \beta_{p+q} = 0$ as:

$$\chi^2_{LR} = -2[\log(L(1)) - \log(L(2))] \tag{3.29}$$

Under H_0, this test statistic is approximately distributed as χ^2 with q degrees of freedom.

Collet (2003), Venables and Ripley (1999) among others explain how comparison between a number of possible models, which need not necessarily be nested can also be made on the basis of the statistic known as *Akaike's Information Criterion (AIC)*,

$$AIC = -2\log \hat{L} + \alpha q \tag{3.30}$$

in which q is the number of unknown β parameters in the model and α is a predetermined constant. The *AIC* will tend to increase when unnecessary terms are added to the model, therefore the smaller the *AIC* the better the model.

3.4.3 Poisson Models for Count Data

Log-linear models for count data under the assumption of a Poisson error structure have many applications, not only to the analysis of counts of events, but also in the context of models for contingency tables and the analysis of survival data. The rationale for modeling the logarithm of the mean as a linear function of observed covariates results to a generalized linear model with Poisson response and link log.

The Poisson distribution

A random variable Y is said to have a Poisson distribution with parameter

μ if it takes integer values $y = 0,1,2,\ldots\ldots$ with probability

$$P\{Y = y\} = \frac{(e^{-\mu}\mu^{y})}{y!}$$
                                                                        3.31

for $\mu > 0$. The mean and variance of this distribution can be shown to be

$$E(Y) = Var(Y) = \mu$$

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data.

Poisson distribution in terms of a stochastic process can be described somewhat informally as follows; Suppose events occur randomly in time in such a way that the following conditions obtain:

- The probability of at least one occurrence of the event in a given time interval is proportional to the length of the interval.
- The probability of two or more occurrences of the event in a very small time interval is negligible.
- The numbers of occurrences of the event in disjoint time intervals are mutually independent.

Then the probability distribution of the number of occurrences of the event in a fixed time interval is Poisson with mean $\mu = \lambda t$, where λ is the rate of occurrence of the event per unit of time and t is the length of the time interval. A process satisfying the three assumptions listed above is called a Poisson process.

A useful property of the Poisson distribution is that the sum of independent Poisson random variables is also Poisson. Specifically, if Y_1 and Y_2 are independent with $Y_i \sim P(\mu_i)$ for $i = 1, 2$ then

$$Y_1 + Y_2 = P(\mu_1 + \mu_2)$$

This result generalizes in an obvious way to the sum of more than two Poisson observations. An important practical consequence of this result is that we can analyze individual or grouped data with equivalent results. Specifically, suppose we have a group of n_i individuals with identical covariate values. Let Y_{ij} denote the number of events experienced by the j-th unit in the i-th group, and let Y_i denote the total number of events in group i. Then, under the usual assumption of independence, if $Y_{ij} \sim P(\mu_i)$ for j = 1, 2.....n_i, then $Y_i \sim P(n_i\mu_i)$. In words, if the individual counts Y_{ij} are Poisson with mean μ_i, the group total Y_i is Poisson with mean $n_i\mu_i$. In terms of estimation, we obtain exactly the same likelihood function if we work with the individual counts Y_{ij} or the group counts Y_i.

Log-Linear Models

Suppose that we have a sample of n observations $y_1, y_2, \dots, y_n$ which can be treated as realizations of independent Poisson random variables, with $Y_i \sim P(\mu_i)$, and suppose that we want to let the mean μ_i (and therefore the variance!) depend on a vector of explanatory variables X_i. We could entertain a simple linear model of the form

$$\mu_i = X_i' \beta$$

This model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative.

A straightforward solution to this problem is to model instead the *logarithm* of the mean using a linear model. Thus, we take logs calculating $\eta_i = \log(\mu_i)$ and assume that the transformed mean follows a linear model $\eta_i = X_i' \beta$. Thus, we consider a generalized linear model with link log. Combining these two steps in one we can write the log-linear model as,

$$\log(\mu_i) = X_i' \beta \qquad\qquad 3.32$$

In this model the regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictor X_j. In other words increasing X_j by one unit is associated with an increase of β_j in the log of the mean. Exponentiating Equation 3.32 we obtain a multiplicative model for the mean itself:

$$\mu_i = \exp\{X_i^{'}\beta\}$$

In this model, an exponentiated regression coefficient $\exp\{\beta_j\}$ represents a multiplicative effect of the *j-th* predictor on the mean. Increasing X_j by one unit multiplies the mean by a factor $\exp\{\beta_j\}$.

A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to a much simpler model.

Estimation and Testing

The log-linear Poisson model is a generalized linear model with Poisson error and link log. Maximum likelihood estimation and testing is used for convergence in modeling Poisson distribution.

Maximum Likelihood Estimation

The likelihood function for *n* independent Poisson observations is a product of probabilities given by Equation 3.31. Taking logs and ignoring a constant involving $\log(y_i!)$, we find that the log-likelihood function is,

$$\log L(\beta) = \sum \{y_i \log(\mu_i) - \mu_i\},$$

Where μ_i depends on the covariates X_i and a vector of p parameters β through the log link of Equation 3.32. It is interesting to note that the log is the canonical link for the Poisson distribution. Taking derivatives of the log-likelihood function with respect to the elements of β, and setting the derivatives to zero, it can be shown that the maximum likelihood estimates in log-linear Poisson models satisfy the estimating equations

$$X'y = X'\hat{\mu}. \tag{3.33}$$

where X is the model matrix, with one row for each observation and one column for each predictor, including the constant (if any), y is the response vector, and $\hat{\mu}$ is a vector of fitted values, calculated from the m.l.e.'s $\hat{\beta}$ by exponentiating the linear predictor $\eta = X'\hat{\beta}$. This estimating equation arises not only in Poisson log-linear models, but more generally in any generalized linear model with canonical link, including linear models for normal data and logistic regression models for binomial counts. It is not satisfied, however, by estimates in probit models for binary data.

To understand equation 3.33 it helps to consider a couple of special cases. If the model includes a constant, then one of the columns of the model matrix X is a column of ones. Multiplying this column by the response vector y produces the sum of the observations. Similarly, multiplying this column by the fitted values $\hat{\mu}$ produces the sum of the fitted values. Thus, in models with a constant one of the estimating equations matches the sum of observed and fitted values.

As a second example suppose the model includes a discrete factor represented by a series of dummy variables taking the value one for observations at a given level of the factor and zero otherwise. Multiplying this dummy variable by the response vector y produces the sum of observations at that level of the factor. When this is done for all levels we obtain the so-called marginal total. Similarly, multiplying the dummy variable by the fitted values $\hat{\mu}$ produces the sum of the expected or fitted counts at that level. Thus, in models with a discrete factor the estimating equations match the observed and fitted marginals for the factor.

This result generalizes to higher order terms. Suppose we entertain models with two discrete factors, say A and B. The additive model $A+B$ would reproduce exactly the marginal totals by A or by B. The model with an interaction effect AB would, in addition, match the totals in each combination of categories of A and B, or the AB margin. This result is the basis of an estimation algorithm known as *iterative proportional fitting*.

In general, however, we will use the iteratively-reweighted least squares (IRLS) algorithm. For Poisson data with link log, the working dependent variable Z has elements

$$Z_i = \hat{\eta}_i + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}$$

and the diagonal matrix W of iterative weights has elements

43

$$w_{ii} = \hat{\mu}_i,$$

Where $\hat{\mu}_i$ denotes the fitted values based on the current parameter estimates.

Initial values can be obtained by applying the link to the data that is taking the log of the response, and regressing it on the predictors using OLS. To avoid problems with counts of 0, one can add a small constant to all responses. The procedure usually converges in a few iterations.

Goodness of Fit

A measure of discrepancy between observed and fitted values is the deviance. For Poisson responses the deviance takes the form

$$D = 2\sum \{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\}.$$

The first term is identical to the binomial deviance, representing 'twice a sum of observed times log of observed over fitted'. The second term, a sum of differences between observed and fitted values, is usually zero, because m.l.e.'s in Poisson models have the property of reproducing marginal totals.

For large samples the distribution of the deviance is approximately a chi-squared with $n-p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the

model. An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as,

$$\chi_i^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. The Pearson statistic has the same form for Poisson and binomial data, namely a `sum of squared observed minus expected over expected'.

In large samples the distribution of Pearson's statistic is also approximately chi-squared with $n - p$ d degrees of freedom. One advantage of the deviance over Pearson's chi-squared is that it can be used to compare nested models, as noted below.

Tests of Hypotheses

Likelihood ratio tests for log-linear models can easily be constructed in terms of deviances, just as earlier indicated in Cox regression models. In general, the difference in deviances between two nested models has approximately in large samples a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct. One can also construct Wald tests, based on the fact that the maximum likelihood estimator $\hat{\beta}$ has approximately in large samples a multivariate normal distribution with

mean equal to the true parameter value β and variance-covariance matrix $Var(\hat{\beta}) = X' WX$, where X is the model matrix and W is the diagonal matrix of estimation weights described earlier.

Grouped Data and the Offset

Let Y_{ijkl} denote the number of observed events in the *l-th* group of factor A *i-th* group of factor B, *j-th* group of factor C and *k-th* group of factor D, and let $Y_{ijk} = \sum_{l} Y_{ijkl}$ denote the group total. If each of the observations in this group is a realization of an independent Poisson variate with mean μ_{ijk}, then the group total will be a realization of a Poisson variate with mean $n_{ijk} \mu_{ijk}$, where n_{ijk} is the number of observations in the *(i, j, k)*-th cell. Suppose now that we postulate a log-linear model for the individual means, say

$$\log E(Y_{ijkl}) = \log(\mu_{ijk}) = X'_{ijk}\beta,$$

Where X_{ijk} is a vector of covariates. Then the log of the expected value of the group total is,

$$\log E(Y_{ijk}) = \log(n_{ijk} \mu_{ijk}) = \log(n_{ijk}) = X'_{ijk}\beta,$$

Thus, the group totals follow a log-linear model with exactly the same coefficients β as the individual means, except for the fact that the linear predictor includes the term $\log(n_{ijk})$. This term, which is known beforehand, is called an *offset*, and is a frequent

46

feature of log-linear models for counts of events. Often, when the response is a count of events the offset represents the log of some measure of exposure.

Thus, we can analyze the data by fitting log-linear models to the individual counts, or to the group totals. The parameter estimates and standard errors will be exactly the same. The deviances of course, will be different, because they measure goodness of fit to different sets of counts. Differences of deviances between nested models, however, are exactly the same whether one works with individual or grouped data.

3.5 Data Management and Analysis

Data managements and analysis was conducted using MS Access, SPSS, R statistical package and MS Excel applications. Exploratory data techniques were performed at the initial stage of analysis to reveal patterns in the population dataset and identify outlier or any unusual entered value.

Univariate analysis: All variables were subjected to descriptive data analysis. Descriptive statistics such as median and Interquratile range were used to summarize continuous variables while categorical variables were summarized using proportions. Kaplan-Meier method was used to estimate survival probabilities (proportion not experiencing nutritional improvement) at different time points. Survival probabilities were plotted against time points to come up with survival curve.

47

Bivariate Analysis: Comparison of survival probabilities between the two treatment modalities was done using the Log Rank Test. Cox Proportional Hazards regression and Poisson regression were used to model the hazard rates (incidence rates) of nutritional improvement at different time points. Hazard rates were compared across different levels within a single variable where one level was used as the reference category, adjusting for time interval. Hazard ratios (Relative risk) were used to measure the number of times nutritional improvement was experienced in one category of a single variable compared to the reference category.

Multivariate Analysis: Cox Proportional Hazards regression was then performed on all independent variables previously tested individually and confirmed to relate significantly with the outcome variable (nutritional improvement) at bivariate analysis. Model comparison was done for all possible combinations using Likelihood ratio test, Wald test, R-square and Score test, in order to identify the best fit model, there by developing a parsimonious model. This technique assisted in identification of potential confounders and effect modifiers; as a result establish independent predictors of nutritional improvement among the study participants. Similarly, Poisson regression was performed to model rates of nutritional improvement using significant variables.

The outcome results of the two models (Cox vs. Poisson) at bivariate and multivariate analysis were then compared. Level of significance was fixed at 0.05 ($p<0.05$) with a 95% Confidence interval.

CHAPTER 4: RESULTS AND FINDINGS

4.1 Baseline enrollment characteristics by study arm.

A total of 330 adult clients were enrolled, consisting of 58.8 percent females and 41.2 percent males. These proportions were different to those observed in the ART program in Kenya (66 percent females and 34 percent males), (LSTIK, 2007). The median age of all adult clients was 35 (IQR, 30 - 42) years. Females were generally younger than males (33 [IQR, 28 – 40] years vs. 37 [IQR, 33 – 44] years), (p<0.001). A profile of selected demographic, economic, behavioral and support characteristics is shown in **Table 4.1.1, Table 4.1.2** and **Table 4.1.3**.

Table 4.1.1: Selected demographic and economic characteristics.

Variables	NC (N=150)		F + NC (N=180)		Total (N=330)	
	n	%	n	%	n	%
Gender						
Male	57	38.0	79	43.9	136	41.2
Female	93	62.0	101	56.1	194	58.8
Age in years						
< 30	35	23.3	49	27.2	84	25.5
30 - 39	62	41.3	75	41.7	137	41.5
40 - 49	42	28.0	42	23.3	84	25.5
50+	11	7.3	14	7.8	25	7.6
Current marital status						
Single	29	19.3	47	26.1	76	23.0
Married	65	43.3	65	36.1	130	39.4
Once married	56	37.3	68	37.8	124	37.6
Education level						
None	10	6.7	9	5.0	19	5.8
Primary	98	65.3	111	61.7	209	63.3
Secondary	39	26.0	52	28.9	91	27.6
Tertiary	3	2.0	8	4.4	11	3.3
Main source of household income						
Formal employment	23	15.3	29	16.1	52	15.8
Small scale business	121	80.7	144	80.0	265	80.3
Welfare/NGO support	6	4.0	7	3.9	13	3.9
Number of people eating at the house						
0 to 1	41	27.3	37	20.6	78	23.6
2 to 3	67	44.7	78	43.3	145	43.9
4 to 5	26	17.3	43	23.9	69	20.9
>5	16	10.7	22	12.2	38	11.5
Amount spent on buying food per day in Kshs						
Nil	17	11.3	16	8.9	33	10.0
<50	35	23.3	40	22.2	75	22.7
50 - <100	64	42.7	80	44.4	144	43.6
100 - <200	26	17.3	35	19.4	61	18.5
>=200	8	5.3	9	5.0	17	5.2
Main source of domestic water						
Tap/rain water	78	52.0	114	63.3	192	58.2
Well/borehole	72	48.0	66	36.7	138	41.8
Distance to health facility in kilometers						
< 5	66	44.0	66	36.7	132	40.0
5 - 9.99	15	10.0	34	18.9	49	14.8
10+	69	46.0	80	44.4	149	45.2

Table 4.1.2: Behavioral and support characteristics.

Variables	NC (N=150)		F + NC (N=180)		Total (N=330)	
	n	%	n	%	n	%
Cigarette smoking						
Ever smoked	50	33.3	71	39.4	121	36.7
Never smoked	100	66.7	109	60.4	209	63.3
Drug abuse						
Ever abused	11	7.3	26	14.4	37	11.2
Never abused	139	92.7	154	85.6	293	88.8
Alcohol drinking						
Ever drank	81	54.0	102	56.7	183	55.5
Never drank	69	46.0	78	43.3	147	44.5
Physical activity/exercise twice weekly						
Performs	61	40.7	56	31.1	117	35.5
Does not perform	89	59.3	124	68.9	213	64.5
Social support received						
M & F	18	12.0	14	7.8	32	9.7
Others	132	88.0	166	92.2	298	90.3
One knowing HIV status						
M & F	21	14.0	19	10.6	40	12.1
Others	129	86.0	161	89.4	290	87.9
Support in the last one month						
Received	94	62.7	111	61.7	205	62.1
Not received	56	37.3	69	38.3	125	37.9

M & F - Medical professional(s) and family member(s)

Table 4.1.3: Baseline enrollment characteristics.

Characteristic	Female (N= 194)		Male (N= 136)	
	F + NC (N=101)	NC (N=93)	F + NC (N=79)	NC (N=57)
Age, Median (IQR)	32(27.5-39.5)	34(28-40)	36(32-42)	38(31-45)
Hb, Median (IQR)	9.9(8.4-11.1)	10.1(8.6-11.6)	11.1(9.3-13.1)	10.1(8.4-12.3)
RBC, Median (IQR), X10^6 UL	3.5(3.1-4.1)	4.0(3.3-4.4)	3.9(3.3-4.6)	3.7(3.1-4.5)
WBC, Median (IQR), X10^6 UL	3.9(2.8-5.4)	4.4(3.7-5.9)	4.5(3.5-6.2)	4.1(3.3-5.8)
CD4, Median, (IQR)	107(31-172)	89(51-189)	105(44-180)	110(47-169)
Weight , Median (IQR), kg	43(39-46)	44(41-48)	50(47-54)	49(45-52)
BMI, Median (IQR), kg/m^2	17.3(16.2-18)	17.4(16.4-18)	17.4(16.5-17.9)	17.4(16.6-17.9)

4.2 Survival Analysis

Out of 330 clients enrolled in the program, a total of 123 experienced nutritional improvement after 12 months of follow-up as shown in **Figure 4.2.1**.
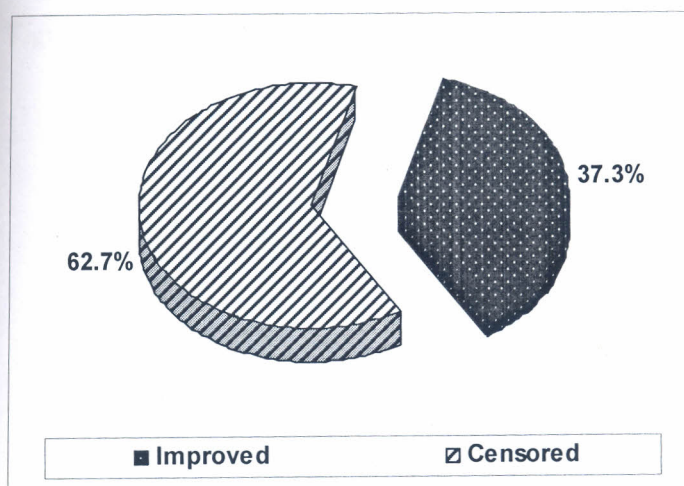
Figure 4.2.1: Nutritional outcome after 12 months follow-up.

There was no significant difference in distribution of nutritional outcome between treatment modalities (P=0.245). A slightly higher proportion of clients on nutritional counseling alone (40.7%) experienced nutritional improvement compared to those on food and nutritional counseling (34.4%), although not statistically significant.
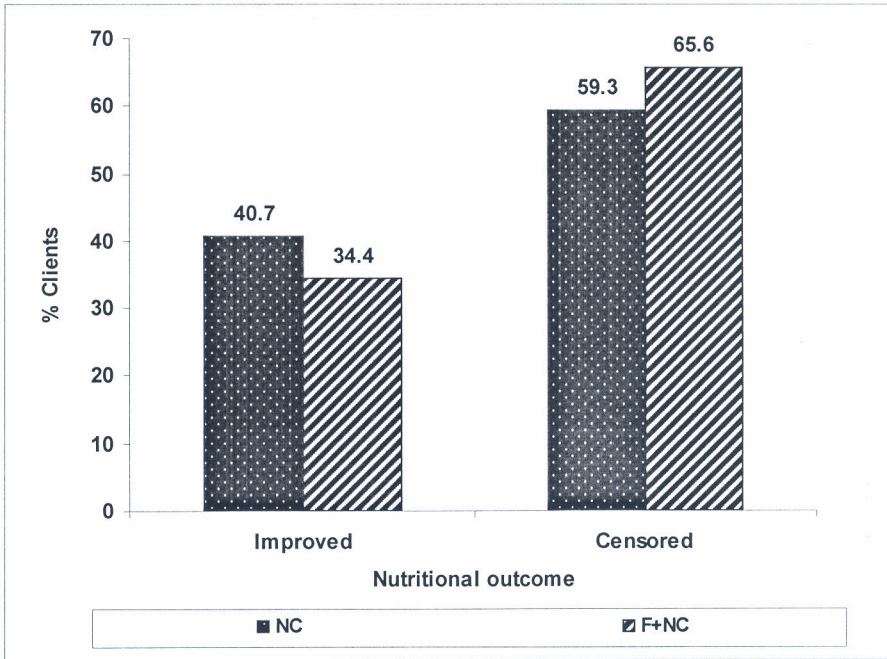
Figure 4.2.2: Nutritional outcome after 12 months follow-up by treatment groups.

4.2.1 Estimating the Survival Curve using the Kaplan–Meier method

The Hazard and Survival Functions

The Kaplan-Meir method was used to estimate survival probability (probability of non-improvement on nutritional status) at each time point during the 12 months follow-up. Using the results of **Table 4.2.1.1**, the cumulative proportion of non-improvement on nutrition status (Survival function) decreased from 1 to 0.101 in a span of 0 to 12 months as shown in **Figure 4.2.1.1**. The Median survival time-to-improved nutrition status was 8 [95% CI = 7 - 9] months.

Table 4.2.1.1: Cumulative proportion of non improvement on nutritional status during 12 months follow-up for all subjects.

Time (t)	N (at risk)	n (Event)	S (t)	std. err	95% C.I. for S (t) Lower	Upper
1	330	3	0.991	0.005	0.981	1.000
2	287	4	0.977	0.009	0.960	0.994
3	250	19	0.903	0.018	0.868	0.939
4	201	25	0.791	0.026	0.741	0.844
5	152	14	0.718	0.030	0.661	0.780
6	121	4	0.694	0.032	0.635	0.759
7	105	10	0.628	0.035	0.563	0.700
8	82	17	0.498	0.039	0.426	0.581
9	49	15	0.345	0.043	0.271	0.440
10	25	4	0.290	0.044	0.216	0.390
11	14	3	0.228	0.047	0.152	0.341
12	9	5	0.101	0.043	0.044	0.233
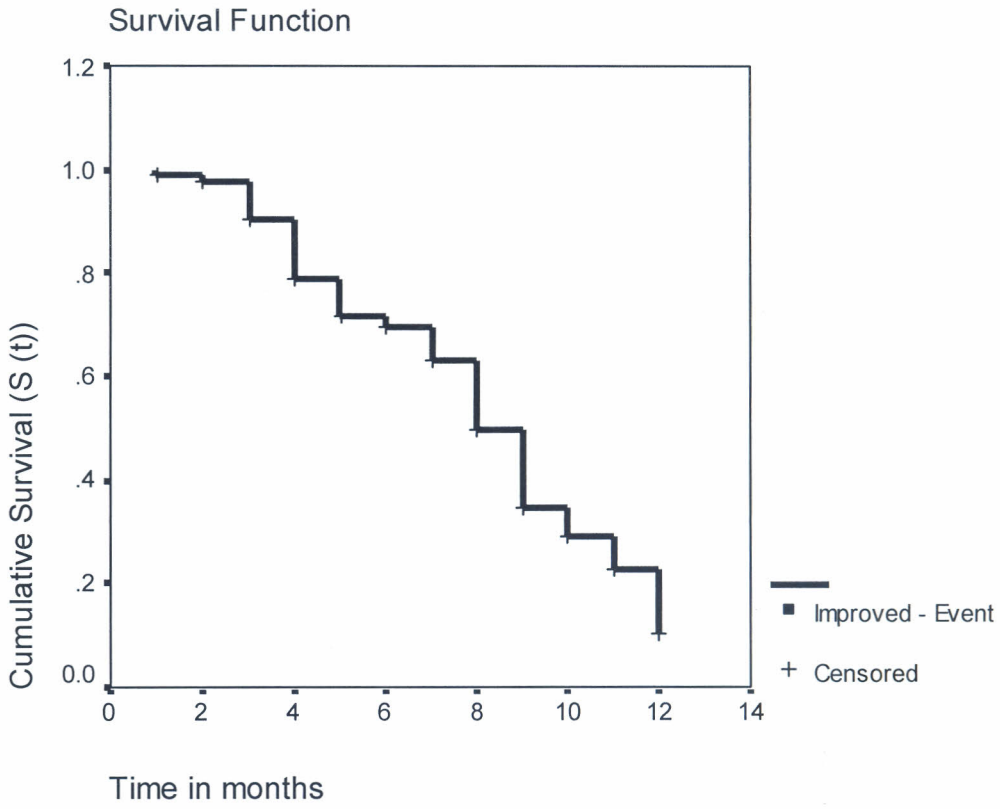
Figure 4.2.1.1: Survival function for non-improved nutritional status during 12 months follow-up for all subjects.

The cumulative proportion of improvement on nutrition status (Hazard function) is shown in **Figure 4.2.1.2.**
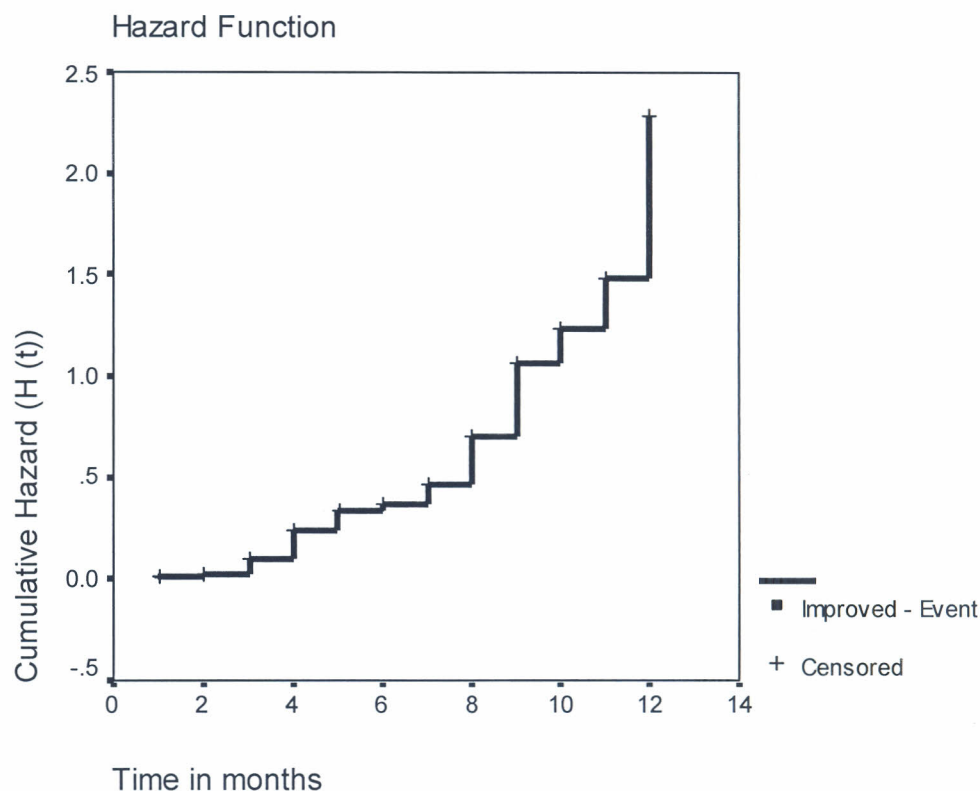
Figure 4.2.1.2: Hazard function for improved nutritional status during 12 months follow-up for all subjects.

Using the results of **Table 4.2.1.2** and **Table 4.2.1.3**, a plot of cumulative proportions of non-improvement on nutrition status (Survival function) by treatment groups is shown in **Figure 4.2.1.3**. There was no significant difference in survival experience between the two treatment modalities (P=0.162). Median time to improved nutritional status for F + NC group was 9 [95% CI= 8 – 10] while that for NC group was 8 [95% CI= 7 – 9].

Table 4.2.1.2: Cumulative proportion of non improvement on nutritional status during 12 months follow-up for NC group.

Time (t)	N (at risk)	n (Event)	$S_1(t)$	std. err	95% C.I. for S (t) Lower	Upper
1	150	3	0.980	0.011	0.958	1.000
2	127	4	0.949	0.019	0.913	0.987
3	106	10	0.860	0.032	0.799	0.924
4	90	10	0.764	0.040	0.689	0.847
5	71	6	0.700	0.045	0.617	0.793
6	56	2	0.675	0.046	0.590	0.772
7	48	4	0.618	0.050	0.527	0.725
8	36	10	0.447	0.059	0.345	0.578
9	21	6	0.319	0.061	0.220	0.463
10	12	3	0.239	0.061	0.146	0.393
11	5	1	0.191	0.065	0.099	0.371
12	3	2	0.064	0.056	0.011	0.361

Table 4.2.1.3: Cumulative proportion of non improvement on nutritional status during 12 months follow-up for F + NC group.

Time (t)	N (at risk)	n (Event)	$S_2(t)$	std. err	95% C.I. for S (t) Lower	Upper
1	180	0	1.000	-	-	-
2	160	0	1.000	-	-	-
3	144	9	0.938	0.020	0.899	0.978
4	111	15	0.811	0.035	0.745	0.883
5	81	8	0.731	0.042	0.654	0.817
6	65	2	0.708	0.043	0.629	0.798
7	57	6	0.634	0.048	0.546	0.735
8	46	7	0.537	0.053	0.443	0.652
9	28	9	0.365	0.060	0.265	0.502
10	13	1	0.337	0.061	0.236	0.480
11	9	2	0.262	0.067	0.159	0.431
12	6	3	0.131	0.063	0.051	0.336

Survival Functions
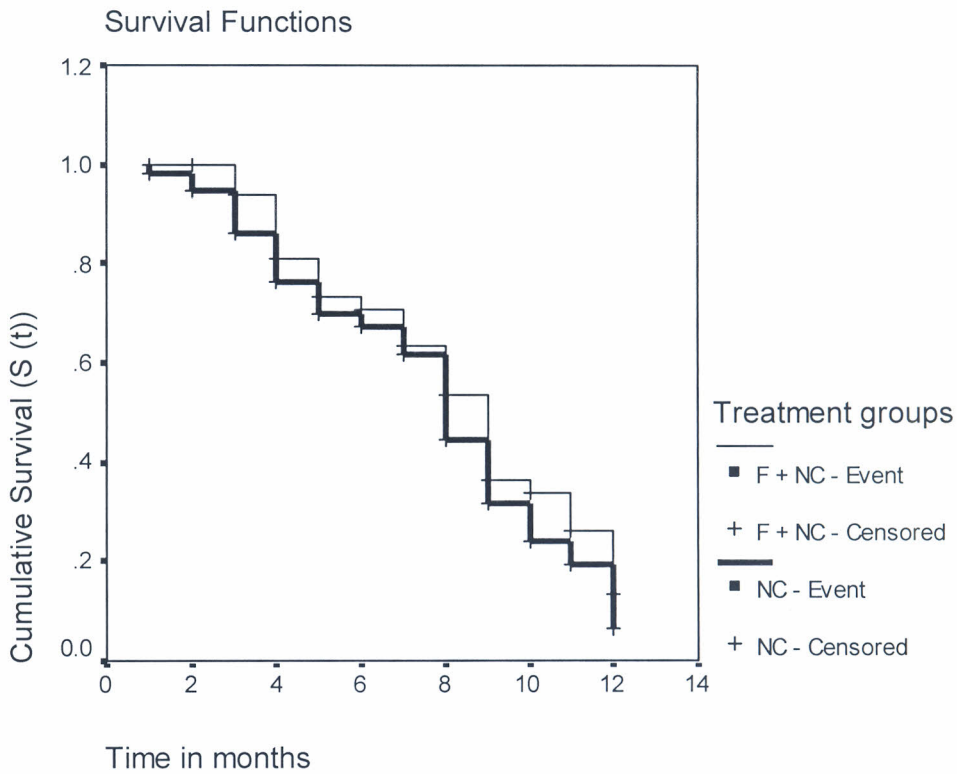
Figure 4.2.1.3: Survival function for non-improved nutritional status during 12 months follow-up by treatment group.

The cumulative proportion of improvement on nutrition status (Hazard function) by treatment groups is shown in **Figure 4.2.1.4**.
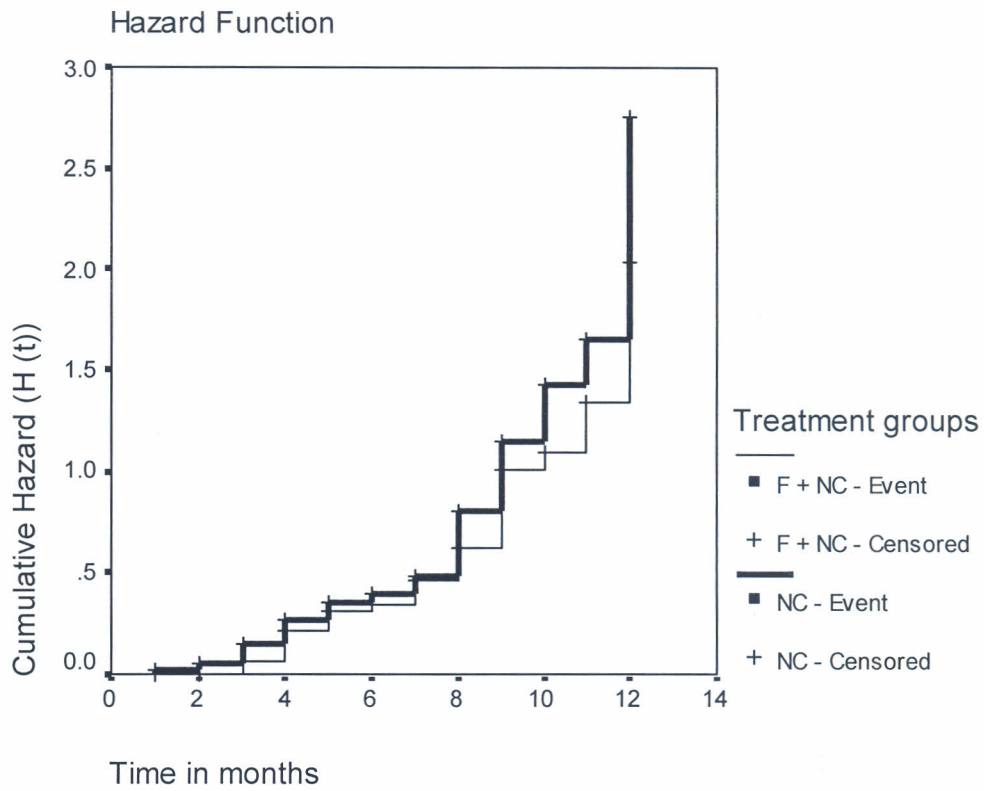
Figure 4.2.1.4: Hazard function for improved nutritional status during 12 months follow-up by treatment group.

4.3 Bivariate Analysis using Cox regression

An initial bivariate analysis of the selected socio-demographic, socio-economic, behavioral, socio-support, hematology, and treatment variables crossed with nutritional status was carried out to determine possible significant explanatory variables to be included in the model runs. bivariate analysis results are shown in **Table 4.3.1, Table 4.3.2,** and **Table 4.3.3.**

Table 4.3.1: Association between nutritional status and selected demographic and economic characteristics.

	β	S.E. (β)	z value	df	P value	Exp (β)	95% CI for Exp(β) Lower	95% CI for Exp(β) Upper
Gender;								
Male	-0.28	0.19	2.19	1	0.139	0.76	0.52	1.09
Female	Reference							
Age in years ;								
< 30	0.01	0.36	0.00	1	0.976	1.01	0.50	2.06
30 - 39	-0.23	0.35	0.46	1	0.498	0.79	0.40	1.56
40 - 49	-0.25	0.37	0.47	1	0.494	0.78	0.38	1.60
50+	Reference							
Current marital status ;								
Single	0.19	0.24	0.64	1	0.423	1.21	0.76	1.94
Married	-0.18	0.21	0.80	1	0.371	0.83	0.56	1.24
SDW	Reference							
Education level ;								
None	0.01	1.12	0.00	1	0.991	1.01	0.11	9.11
Primary	0.58	1.01	0.33	1	0.565	1.79	0.25	12.88
Secondary	0.48	1.02	0.23	1	0.635	1.62	0.22	11.87
Tertiary	Reference							
Main source of household income;								
Formal employment	-1.20	0.49	6.03	1	**0.014***	0.30	0.12	0.79
Small business	-1.08	0.43	6.28	1	**0.012***	0.34	0.15	0.79
Welfare/NGO	Reference							
Amount spent on buying food per day in Kshs;								
Nil	0.09	0.53	0.03	1	0.868	1.09	0.39	3.07
<50	0.18	0.44	0.16	1	0.687	1.20	0.50	2.85
50 - <100	-0.23	0.43	0.28	1	0.594	0.79	0.34	1.86
100 - <200	0.30	0.45	0.43	1	0.511	1.35	0.55	3.27
>=200	Reference							
Distance to health facility in kilometers;								
< 5	-0.08	0.20	0.16	1	0.693	0.92	0.63	1.37
5 - 9.99	0.45	0.27	2.83	1	0.092	1.57	0.93	2.64
10+	Reference							

SDW – Separated /divorced / widowed
* - Significant at $P<0.05$

Among the selected demographic and economic characteristics, main source of household income was significantly associated with nutritional status of HIV clients

(P<0.05). A client with a formal employment was 70% less likely to improve on their nutritional status as compared to one who relied on welfare/NGO support (P=0.014). Similarly, a client with a small scale business was 66% less likely to improve on their nutritional status as compared to one who relied on welfare/NGO support (P=0.012).

Table 4.3.2: Association between nutritional status and selected Behavioral and support characteristics.

	β	S.E. (β)	z value	df	P value	Exp (β)	95% CI for Exp(β) Lower	Upper
Cigarettes smoking;								
Ever smoked	-0.22	0.19	1.32	1	0.251	0.80	0.55	1.16
Never smoked	Reference							
Drug abuse;								
Ever abused	-0.24	0.32	0.59	1	0.442	0.78	0.42	1.45
Never abused	Reference							
Alcohol drinking;								
Ever drank	-0.25	0.18	1.86	1	0.172	0.78	0.54	1.11
Never drank	Reference							
Physical exercise twice weekly;								
Performs	-0.21	0.19	1.20	1	0.274	0.81	0.56	1.18
Does not perform	Reference							
Social support received;								
M & F	0.55	0.28	3.99	1	**0.046***	1.75	1.01	3.03
Others	Reference							
One knowing HIV status;								
M & F	0.27	0.26	1.09	1	0.297	1.30	0.79	2.17
Others	Reference							
Support in the last one month;								
Received	0.03	0.19	0.03	1	0.871	1.03	0.71	1.49
Not received	Reference							

M & F - Medical professional(s) and family member(s)
* - Significant at P<0.05

Among the selected behavioral and support characteristics, provision of socio-support was significantly associated with nutritional status of HIV clients (P=0.046). A client receiving socio-support from a medical professional (s) as well as family member (s) was 1.75 times more likely to improve on their nutritional status as compared to support from other sources.

Table 4.3.3: Association between nutritional status and selected hematology and nutrition variables.

	β	S.E. (β)	z value	df	P value	Exp (β)	95% CI for Exp(β) Lower	Upper
Treatment groups;								
F + NC	-0.23	0.18	1.67	1	0.197	0.79	0.56	1.12
NC	Reference							
CD4 change;								
1 to 100 counts	1.50	0.76	3.88	1	**0.049***	4.47	1.01	19.84
101 to 200 counts	1.85	0.76	6.01	1	**0.014***	6.38	1.45	28.08
> 200 counts	1.87	0.77	5.96	1	**0.015***	6.47	1.45	28.96
D / Nil	Reference							
WBC change;								
Improved	0.58	0.34	3.01	1	0.083	1.79	0.93	3.45
D / Nil	Reference							
RBC change;								
Improved	0.28	0.33	0.7	1	0.402	1.32	0.69	2.56
D / Nil	Reference							
HB change;								
Improved	0.24	0.48	0.25	1	0.614	1.28	0.50	3.33
D / Nil	Reference							
Baseline BMI;								
< 16 Kg/m^2	0.08	0.28	0.08	1	0.782	1.08	0.63	1.86
16 - 18.5 Kg/m^2	Reference							

D / Nil - Deteriorated / No change
* - Significant at $P<0.05$

Among the selected hematology and nutrition variables, CD4 change was a significant factor to nutritional improvement. Considering clients who CD4 deteriorated or did not change to be the reference category, clients whose CD4 count changed by 1 to 100 were 4.47 times more likely to improve on nutrition status (P=0.049). The likelihood

increased significantly to 6.38 (P=0.014) for those that changed by 101 to 200, to a significant high of 6.47 (P=0.015).

An exploratory model analysis was performed to explore the relations between the variables while simultaneously adjusting for all other variables that had significant association with nutritional status. After investigation of confounding, all variables with p-values of 0.05 or less were considered possible confounders and were retained for the model analysis.

4.4 Modeling Survivorship to nutritional improvement using Cox Proportional Hazards Regression

Comparison of F+NC and NC survival was carried out using Cox proportional hazards model adjusting for case-mix differences in two identified significant characteristics (CD4 count and Socio-support).

Dummy variables were created for reference cell coding of the categorical variables. These were necessary for the output of measures of association using the reference category of choice. *COXPH* was run using a manual backward stepwise model building approach. This created a final model with statistically significant effects of explanatory variables on survival times while controlling for possible confounding of exposure effects.

4.4.1 Comparing Alternative models

Different models were fitted using two popular methods ("efron" and "breslow") used for computation of the beta estimates. Method indicates how to handle observations that have tied (i.e., identical) survival times. **Table 4.4.1.1** and **Table 4.4.1.3** shows comparison of six models developed using the four variables identified to be candidate predictors of time to improved nutritional status. Although treatment variable was not significantly associated with time to improved nutritional status, it was considered as a key factor while the other three significant variables (CD4, Support, and Income) were used as stratification exposures.

The R-square, Likelihood test, Wald test, and score test were calculated to identify the best fit model for time to improved nutritional status. The comparison of output models developed using Breslow method is shown in **Table 4.4.1.1**.

Table 4.4.1.1: Comparison of models developed using Breslow method.

Model	R-square	Likelihood test	Wald test	Score test
Food + Support	0.015	4.98	5.49	5.60
Food + Income	0.020	6.64	8.25	8.95
Food + Support + Income	0.031	10.48	12.62	13.42
Food + CD4	0.102	10.87	6.88	8.48
Food + Income + CD4	0.121	13.01	9.76	11.99
Food + Support + CD4	0.128	13.85	11.31	14.43

Using R-square, Likelihood test, Wald test, and score test 'Food + Support + CD4' model was identified to be the best fit model. The estimated relative risks (RR) of

improved nutritional status are summarized in **Tables 4.4.1.2.** The analysis code is

shown in **Appendix B (c)**.

Table 4.4.1.2: The best fit model for time to improved nutritional status using Breslow method.

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
Treatment group: NC	Reference				
Treatment group: F + NC	-0.16	0.32	0.85	-0.497	0.619
CD4: Deteriorated / No change	Reference				
CD4: change by 1 to 100	1.44	0.77	4.22	1.879	0.060
CD4: change by 101 to 200	1.76	0.76	5.81	2.309	0.021 *
CD4: change by > 200	1.83	0.77	6.23	2.366	0.018 *
Support: Other	Reference				
Support: M & F	1.62	0.79	5.05	2.059	0.040 *

M & F - Medical professional(s) and family member(s)
* - Significant at P<0.05

According to the results shown in **Table 4.4.1.2**, there was no significant association

between treatment group and nutritional status. HIV clients on F + NC were 15%

unlikely to improve nutritionally as compared to those on NC. CD4 change was a

significant factor to nutritional improvement. Considering clients who CD4 deteriorated

or did not change to be the reference category, clients whose CD4 count changed by 1 to

100 were 4.22 times more likely to improve on nutrition status even though that was not

statistically significant (P=0.060). The likelihood increased significantly to 5.81

(P=0.021) for those that changed by 101 to 200, to a significant high of 6.23 (P=0.018).

Socio-support of the clients was significantly associated with their nutritional status

(P=0.040). A client receiving socio-support from a medical professional (s) as well as

family member (s) was 5.05 times more likely to improve on their nutritional status as compared to support from other sources. Similarly, comparison of output models developed using Efron method is shown in **Table 4.4.1.3**.

Table 4.4.1.3: Comparison of models developed using Efron method.

Model	R-square	Likelihood test	Wald test	Score test
Food + Support	0.017	5.68	6.34	6.49
Food + Income	0.023	7.76	9.88	10.91
Food + Support + Income	0.037	12.39	15.1	16.32
Food + CD4	0.12	12.86	8.04	10.08
Food + Income + CD4	0.141	15.41	11.5	14.42
Food + Support + CD4	0.149	16.32	13.26	17.38

Using R-square, Likelihood test, Wald test, and score test 'Food + Support + CD4' model was identified to be the best fit model. The estimated relative risks (RR) of improved nutritional status are summarized in **Tables 4.4.1.4.** The analysis code is shown in **Appendix B (d)**.

Table 4.4.1.4: The best fit model for time to improved nutritional status using Efron method.

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
Treatment group: NC	Reference				
Treatment group: F + NC	-0.20	0.33	0.82	-0.626	0.531
CD4: Deteriorated / No change	Reference				
CD4: change by 1 to 100	1.52	0.77	4.57	1.987	0.047 *
CD4: change by 101 to 200	1.91	0.77	6.75	2.50	0.013 *
CD4: change by > 200	1.96	0.78	7.10	2.53	0.012 *
Support: Other	Reference				
Support: M & F	1.77	0.79	5.87	2.25	0.024 *

M & F - Medical professional(s) and family member(s)
* - Significant at P<0.05

According to the results shown in **Table 4.4.1.4**, there was no significant association between treatment group and nutritional status. HIV clients on F + NC were 18% unlikely to improve nutritionally as compared to those on NC.

CD4 change was a significant factor to nutritional improvement. Considering clients who CD4 deteriorated or did not change to be the reference category, clients whose CD4 count changed by 1 to 100 were 4.57 times more likely to improve on nutrition status even though that was statistically significant (P=0.047). The likelihood increased significantly to 6.75 (P=0.013) for those that changed by 101 to 200, to a significant high of 7.10 (P=0.012).

Socio-support of the clients was significantly associated with their nutritional status (P=0.024). A client receiving socio-support from a medical professional (s) as well as family member (s) was 5.87 times more likely to improve on their nutritional status as compared to support from other sources.

4.4.2 Checking for Proportional Hazards

Tests for the proportional-hazards assumption are obtained from *cox.zph* in R software, which computes a test for each covariate, along with a global test for the model as a whole. A plot of *cox.zph* object gives the outcome for the two methods used as shown in **Table 4.4.2.1** and **Table 4.4.2.2**. There was no evidence of non-proportional hazards for all the covariates. **Figure 4.4.2.1** and **Figure 4.4.2.2** shows plots of scaled Schoenfeld

residuals against transformed time for each covariate in the best model fit to the data for "breslow" and "Efron" methods respectively. The solid line is a smoothing-spline fit to the plot, with the broken lines representing a ± 2-standard-error band around the fit.

Table 4.4.2.1: Test for the proportional-hazards assumption for the best fit model for time to improved nutritional status using Breslow method.

Variables	ρ	χ^2 value	p value
Treatment group: NC	Reference		
Treatment group: F + NC	-0.05937	0.15467	0.694
CD4: Deteriorated / No change	Reference		
CD4: change by 1 to 100	0.0923	0.34813	0.555
CD4: change by 101 to 200	0.17825	1.41986	0.233
CD4: change by > 200	0.12022	0.64912	0.420
Support: Other	Reference		
Support: M & F	-0.00549	0.00125	0.972
GLOBAL	NA	2.28004	0.809

M & F - Medical professional(s) and family member(s)

Figure 4.4.2.1: A plot of beta coefficients for the best fit model by time to improved nutritional status using Breslow method.
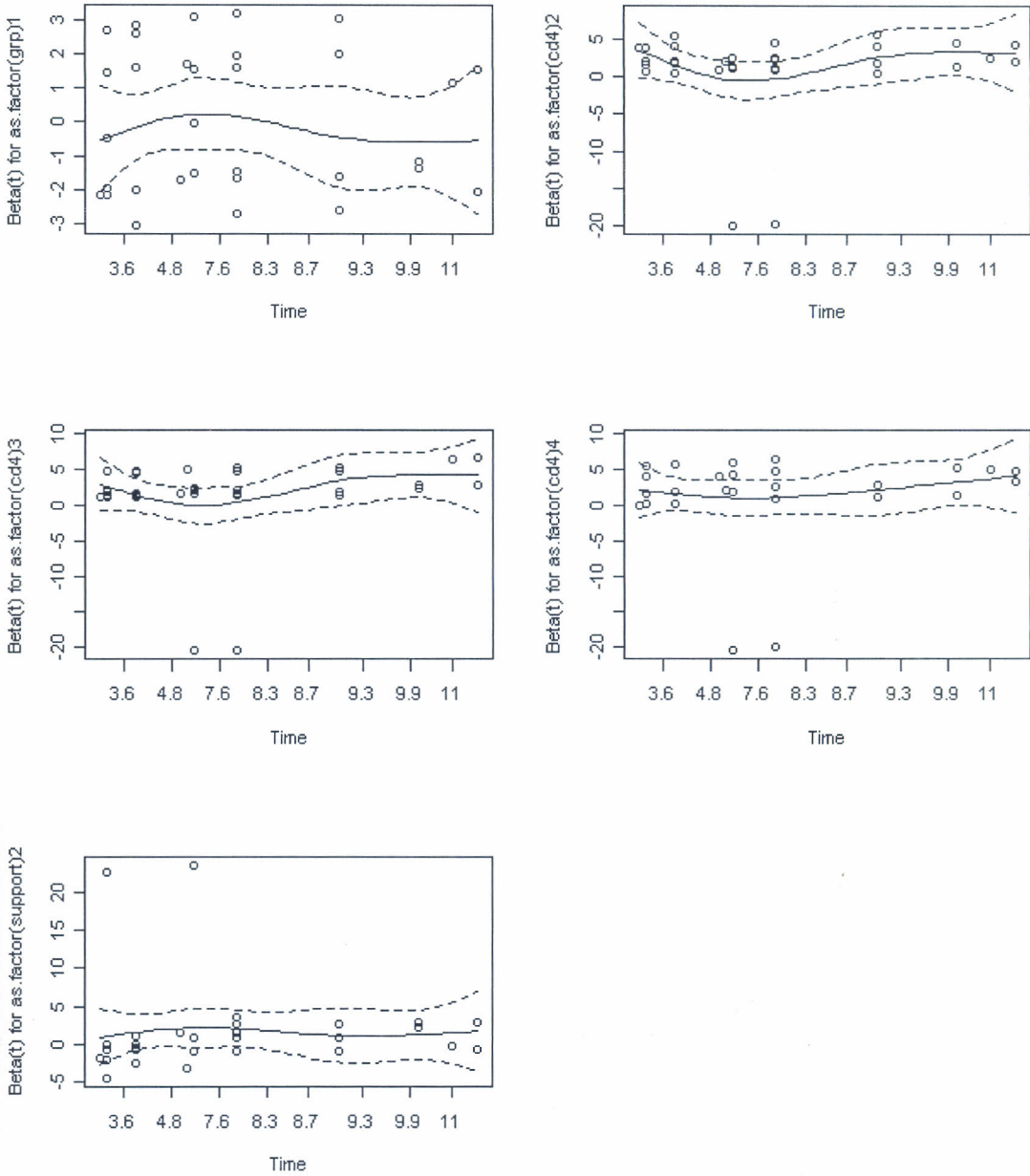
Table 4.4.2.2: Test for the proportional-hazards assumption for the best fit model for time to improved nutritional status using Efron method.

Variables	ρ	χ^2 value	p value
Treatment group: NC	Reference		
Treatment group: F + NC	-0.09117	0.379645	0.538
CD4: Deteriorated / No change	Reference		
CD4: change by 1 to 100	0.10953	0.500522	0.479
CD4: change by 101 to 200	0.20553	1.958618	0.162
CD4: change by > 200	0.13376	0.818629	0.366
Support: Other	Reference		
Support: M & F	0.00373	0.000595	0.981
GLOBAL	NA	3.194139	0.670

M & F - Medical professional(s) and family member(s)

Figure 4.4.2.2: A plot of beta coefficients for the best fit model by time to improved nutritional status using Efron method.
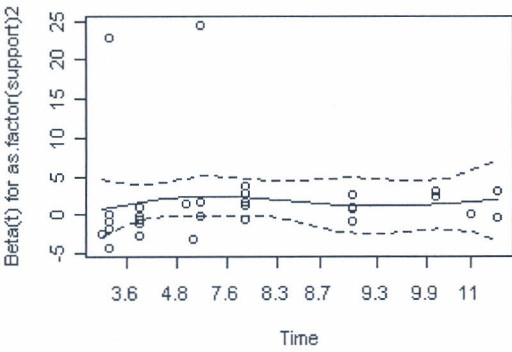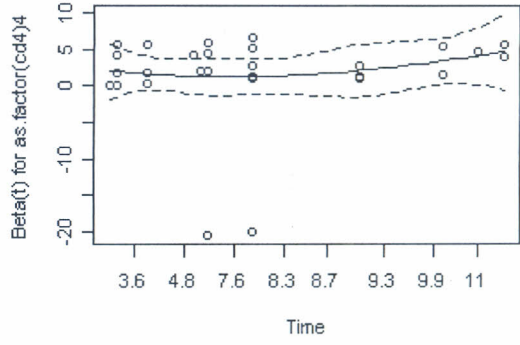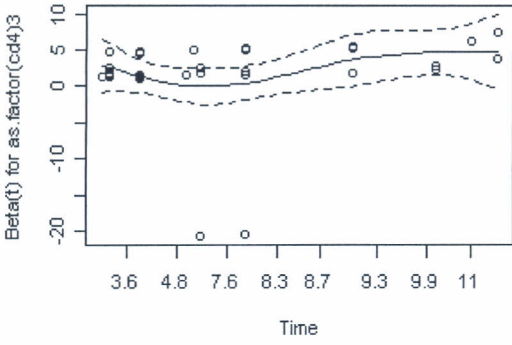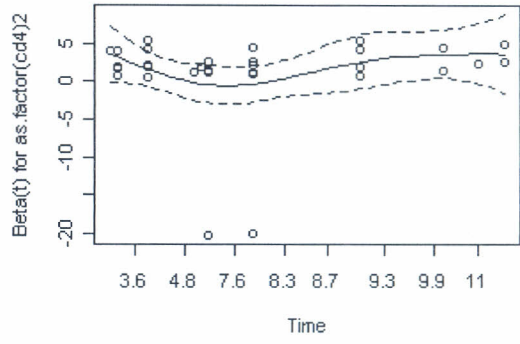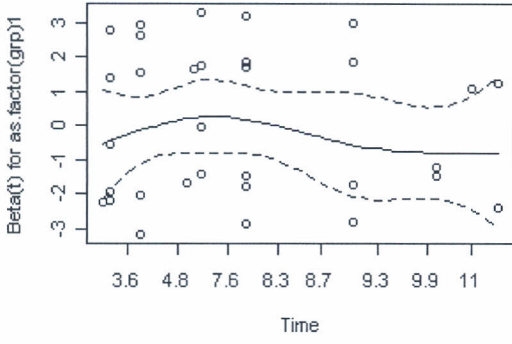
4.5 Rate of Nutritional Improvement

Rate of nutritional improvement during 12 months follow-up varied at every time point.

The rate increased from 0.009 to 0.556 as shown in **Table 4.5.1** and Figure **4.5.1**

Table 4.5.1: Rate of nutritional improvement during 12 months follow-up for all subjects.

Time (t)	N (at risk)	c (Censored)	d (Event)	n (c + d)	Rate (f (t))
1	330	40	3	43	0.009
2	287	33	4	37	0.014
3	250	30	19	49	0.076
4	201	24	25	49	0.124
5	152	17	14	31	0.092
6	121	12	4	16	0.033
7	105	13	10	23	0.095
8	82	16	17	33	0.207
9	49	9	15	24	0.306
10	25	7	4	11	0.160
11	14	2	3	5	0.214
12	9	4	5	9	0.556
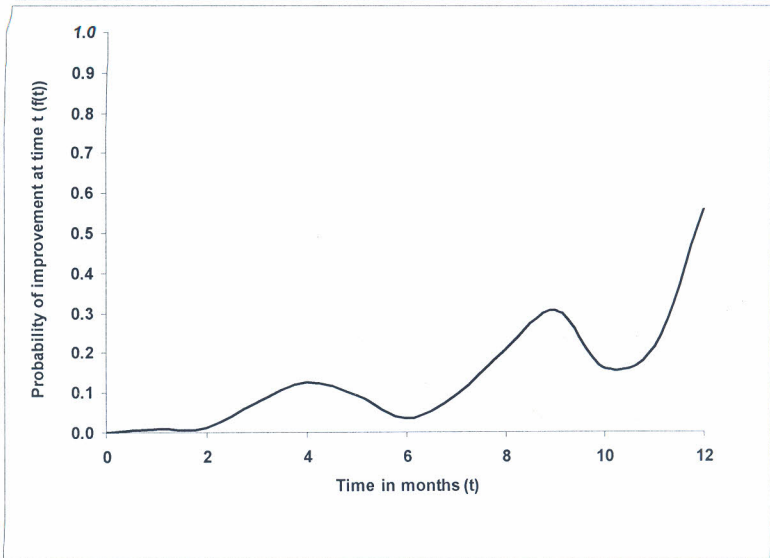
Figure 4.5.1: A plot of rate of nutritional improvement during 12 months follow-up for all subjects by time.

Rate of nutritional improvement between the treatment groups varied proportionally at every time point. The rate increased from 0 to 0.50 for clients on F+NC compared to 0.02 to 0.667 for those on NC as shown in **Table 4.5.2** and Figure **4.5.2**

Table 4.5.2: Rates of nutritional improvement during 12 months follow-up by treatment groups.

Time (t)	F + NC					NC				
	N	c	d	n (c + d)	Rate (f_1 (t))	N	c	d	n (c + d)	Rate (f_2 (t))
1	180	20	0	20	0.000	150	20	3	23	0.020
2	160	16	0	16	0.000	127	17	4	21	0.031
3	144	24	9	33	0.063	106	6	10	16	0.094
4	111	15	15	30	0.135	90	9	10	19	0.111
5	81	8	8	16	0.099	71	9	6	15	0.085
6	65	6	2	8	0.031	56	6	2	8	0.036
7	57	5	6	11	0.105	48	8	4	12	0.083
8	46	11	7	18	0.152	36	5	10	15	0.278
9	28	6	9	15	0.321	21	3	6	9	0.286
10	13	3	1	4	0.077	12	4	3	7	0.250
11	9	1	2	3	0.222	5	1	1	2	0.200
12	6	3	3	6	0.500	3	1	2	3	0.667
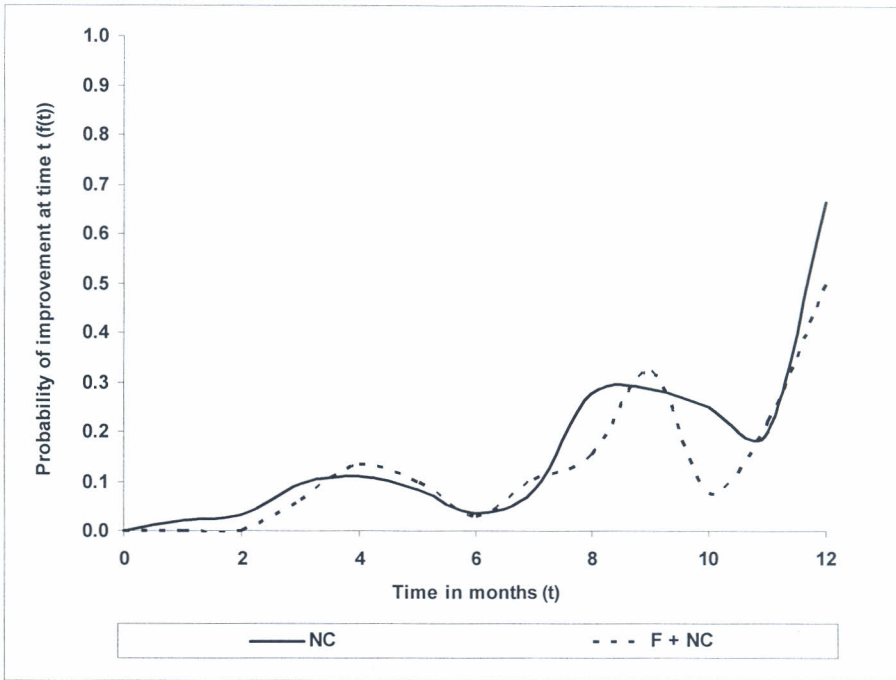
Figure 4.5.2: A plot of rate of nutritional improvement during 12 months follow-up for each treatment group by time.

4.5.1 Bivariate analysis using Poisson regression

Just like in bivariate analysis using Cox regression, an initial bivariate analysis of the selected socio-demographic, socio-economic, behavioral, socio-support, hematology, and treatment variables crossed with nutritional status was carried out using Poisson regression to determine possible significant explanatory variables to be included in the model runs. The results are shown in **Table A1** to **Table A20 (Appendix A).** Clients CD4 change, Main source of household income, Source of socio-support and Treatment group emerged to be the independent predictors of nutritional outcome.

An exploratory model analysis was performed to explore the relations between the variables while simultaneously adjusting for all other variables that had significant association with nutritional outcome. After investigation of confounding, all variables with p-values of 0.05 or less were considered possible confounders and were retained for the model analysis.

4.5.2 Modeling Rate of nutritional improvement using Poisson Regression

Data on client's time to improved nutritional status was reformatted to generate the number of event counts in aggregated format stratified by three covariates at every time interval. We have used covariates previously identified to develop the best fit model in modeling for survivorship. The resulting data is shown in **Table 4.5.2.1** and **Table 4.5.2.2**.

Table 4.5.2.1: Number of events (those who experienced nutritional improvement) at specific time interval during 12 months follow-up by CD4 change, Socio-support and Treatment group for adult PLWHA enrolled in the FBP program.

		(a) Number of events														
	CD4 change	1				2				3				4			
	Socio-support	1		2		1		2		1		2		1		2	
Interval	Treatment	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	-	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
3	-	0	0	0	0	1	0	0	0	0	1	0	1	1	1	0	0
4	-	0	0	0	0	3	1	0	0	1	1	0	0	0	1	0	0
5	-	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
6	-	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
7	-	0	1	0	0	0	0	0	0	0	0	0	1	1	2	0	0
8	-	0	1	0	0	1	0	0	0	2	1	0	0	4	2	0	0
9	-	0	0	0	0	1	2	0	0	3	1	0	0	0	0	0	0
10	-	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
11	-	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	-	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0

CD4 change; 1= Deteriorated / No change, 2= 1 to 100 count, 3= 101 to 200 count, 4=> 200 count

Socio-support; 1= others, 2= Medical professional(s) and family member(s)

Treatment; 1= NC, 2= F+NC

Table 4.5.2.2: Total number at risk at specific time interval during 12 months follow-up by CD4 change, Support and Treatment for adult PLWHA enrolled in the FBP program.

		(b) Total number at risk															
	CD4 change	1				2				3				4			
	Socio-support	1		2		1		2		1		2		1		2	
Interval	Treatment	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1	-	4	9	0	0	14	13	0	1	11	16	0	2	14	15	0	2
2	-	4	9	0	0	14	12	0	1	11	15	0	2	14	15	0	1
3	-	4	9	0	0	13	12	0	1	10	14	0	2	14	15	0	0
4	-	4	8	0	0	12	11	0	0	10	12	0	1	13	12	0	0
5	-	4	8	0	0	9	8	0	0	8	9	0	1	13	10	0	0
6	-	4	8	0	0	9	8	0	0	7	8	0	1	11	8	0	0
7	-	3	8	0	0	8	5	0	0	6	7	0	1	11	6	0	0
8	-	3	7	0	0	7	5	0	0	5	7	0	0	8	3	0	0
9	-	1	5	0	0	6	4	0	0	3	4	0	0	4	1	0	0
10	-	1	3	0	0	4	1	0	0	0	2	0	0	3	0	0	0
11	-	1	2	0	0	2	1	0	0	0	2	0	0	0	0	0	0
12	-	1	1	0	0	2	1	0	0	0	1	0	0	0	0	0	0

Interval Poisson Regression was used to model for Rates of nutritional improvement.

The analysis was performed using R software as shown in **Appendix B (e)**. The results are shown in **Table 4.5.2.3**.

Table 4.5.2.3: Model for rate of improved nutritional status using Interval Poisson Regression.

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-22.94	2478.27		-0.009	0.993
Interval: 1	Reference				
Interval: 2	16.70	2478.27		0.007	0.995
Interval: 3	18.39	2478.27		0.007	0.994
Interval: 4	18.92	2478.27		0.008	0.994
Interval: 5	17.15	2478.27		0.007	0.995
Interval: 6	17.26	2478.27		0.007	0.994
Interval: 7	19.00	2478.27		0.008	0.994
Interval: 8	20.11	2478.27		0.008	0.994
Interval: 9	20.16	2478.27		0.008	0.994
Interval: 10	19.67	2478.27		0.008	0.994
Interval: 11	19.73	2478.27		0.008	0.994
Interval: 12	21.10	2478.27		0.009	0.993
Treatment group: NC	Reference				
Treatment group: F + NC	-0.16	0.32	0.85	-0.497	0.619
CD4: Deteriorated / No change	Reference				
CD4: change by 1 to 100	1.44	0.77	4.22	1.879	0.060
CD4: change by 101 to 200	1.76	0.76	5.81	2.309	0.021 *
CD4: change by > 200	1.83	0.77	6.23	2.366	0.018 *
Support: Other	Reference				
Support: Medical professional and family members	1.62	0.79	5.05	2.059	0.040 *

* - Significant at P<0.05

The regression equation can be written as;

$$\text{Log Events} = \alpha + \beta_1 \text{Interval:2} + \beta_2 \text{Interval:3} + \beta_3 \text{Interval:4} + \beta_4 \text{Interval:5} + \beta_5 \text{Interval:6}$$

$$+ \beta_6 \text{Interval:7} + \beta_7 \text{Interval:8} + \beta_8 \text{Interval:9} + \beta_9 \text{Interval:10} + \beta_{10} \text{Interval:11} +$$

$$\beta_{11} \text{Interval:12} + \beta_{12} \text{Treatment group: (F + NC)} + \beta_{13} \text{CD4: change by 1 to 100} + \beta_{14} \text{CD4:}$$

$$\text{change by 101 to 200} + \beta_{15} \text{CD4: change by > 200} + \beta_{16} \text{Support: Medical professional}$$

and family

Log Events = -22.94 +16.70Interval:2 + 18.39Interval:3 + 18.92Interval:3 + 17.15Interval:4 + 17.15Interval:5 + 17.26Interval:6 + 19.00Interval:7 + 20.11Interval:8 + 20.16Interval:9 + 19.67Interval:10 + 19.73Interval:11 + 21.10Interval:12 - 0.16Treatment group: (F + NC) + 1.44CD4: change by 1 to 100 + 1.76CD4: change by 101 to 200 + 1.83CD4: change by >200 + 1.62Support: Medical professional and family members

$$\text{Events} = (e)^{-22.94} \times (e)^{16.70\text{Interval:2}} \times (e)^{18.39\text{Interval:3}} \times (e)^{17.15\text{Interval:4}} \times (e)^{17.15\text{Interval:5}} \times$$

$$(e)^{17.26\text{Interval:6}} \times (e)^{19.00\text{Interval:7}} \times (e)^{20.11\text{Interval:8}} \times (e)^{20.16\text{Interval:9}} \times (e)^{19.67\text{Interval:10}} \times$$

$$(e)^{19.73\text{Interval:11}} \times (e)^{21.10\text{Interval:12}} \times (e)^{-\,0.16\text{Treatment group: (F + NC)}} \times (e)^{1.44\text{CD4: change by 1 to 100}} \times$$

$$(e)^{1.76\text{CD4: change by 101 to 200}} \times (e)^{1.83\text{CD4: change by >200}} \times (e)^{1.62\text{Support: Medical professional and family}}$$

members

The equation is useful for estimating the relative risk of improved nutritional status using time interval, change in CD4, and source of socio-support.

For example, all factors held constant, the relative risk (RR) of improvement on nutritional status for change in CD4 by 101 to 200 as compared to change in CD4 by 1 to 100 can be estimated as;

$$RR = (e)^{1.76\text{CD4: change by 101 to 200}} \div (e)^{1.44\text{CD4: change by 1 to 100}} = (e)^{1.76*1} \div (e)^{1.44*1}$$

$RR = (e)^{1.76} \div (e)^{1.44} = 5.81 \div 4.22 = 1.38$. To be interpreted that clients whose CD4 count changed by 101 to 200 are 1.38 times more likely to improve on their nutritional status compared to those whose CD4 count changed by 1 to 100.

Similar estimates could be used for calculating the relative risk by different categories of change in CD4, Source of socio-support, Treatment group, or time interval. Interpretation of the results in **Table 4.5.2.3** is similar to that of **Table 4.4.1.2.**

To check the fitness of the model we carry out diagnostic plots of deviance against fits. A plot of residuals against fitted values is shown below:

Figure 4.5.2.1: A plot of residuals against fitted values.
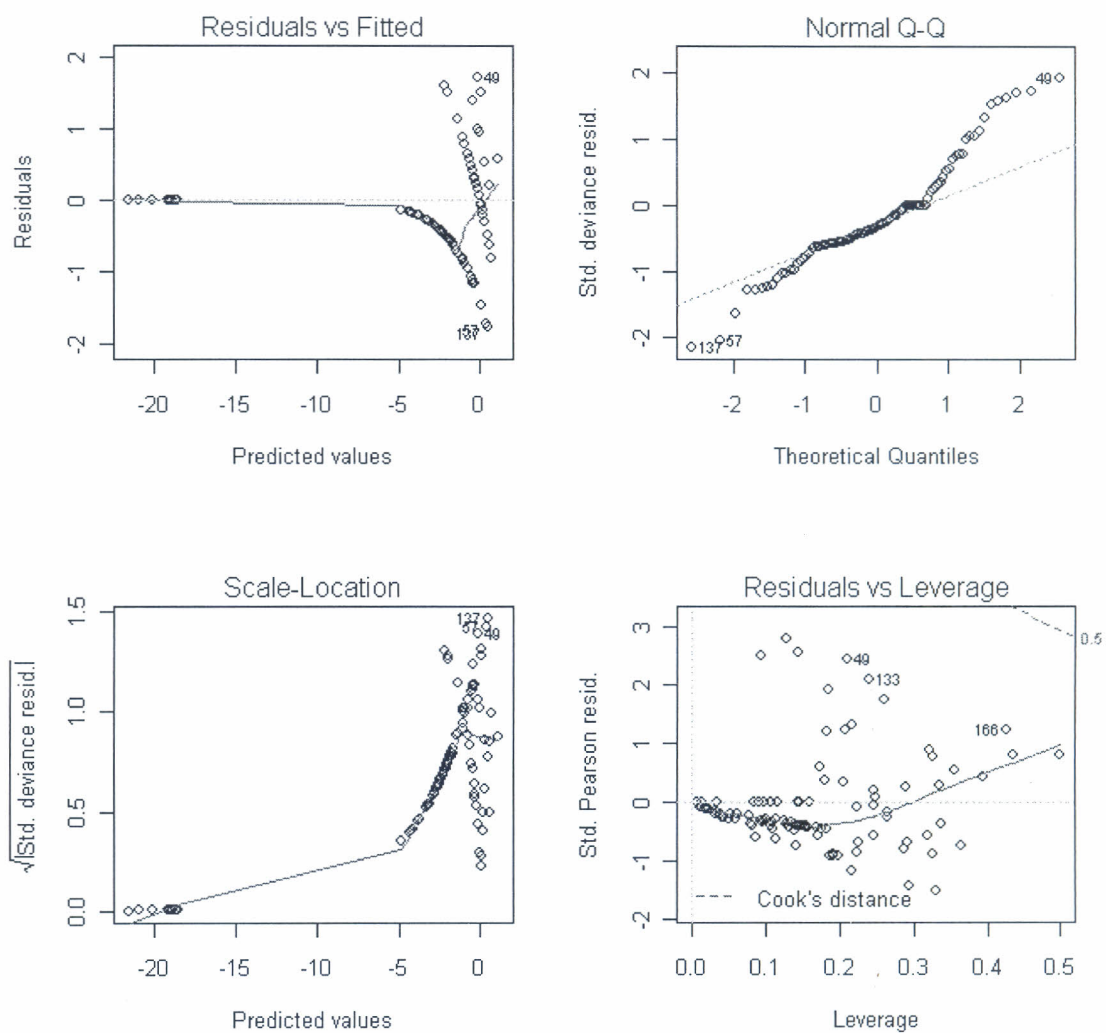
CHAPTER 5: DISCUSSION

By carefully summarizing results from this study and conducting some comparative analyses, some key results were identified. The Kaplan-Meir method was used to estimate survival probability adjusting for treatment modality. Survival probability was considered to mean the probability of non-improved nutritional status. Plotting the cumulative proportion surviving (non-improved nutritional status) against the survival times gives the stepped survival curve shown in **Figure 4.2.1.1**. This method is used in most statistical packages. The curves were generated using the Statistical Package for Social Sciences (SPSS) package. **Figure 4.2.1.3** shows output curve comparing the survival curves for the two treatment modalities using data on survivorship. Using the Log rank test, it was established that survival experience between the two treatment modalities was not significantly different (P=0.162). There was much less evidence on the impact of food intervention for malnutrition among PLWHA. This is in agreement with findings from a summary report done by Castleman *et al.* (2008). Insta foundation may not be optimal food for all groups. According to the report, adaptation and alternative formulations are underway.

The analysis was advanced farther to identify existence of other factors that could relate with nutritional outcome. Twenty factors were firstly analyzed using Cox regression and their results summarized in **Table 4.3.1** to **Table 4.3.3**. A repeat of the same analysis was done using Poisson regression (log-linear model). In relation to nutritional

improvement, the parameter estimates for each of the twenty variables upon adjusting for the intervals (**Tables A1** to **Table A20** in **Appendix A**) were exactly similar to those obtained using Cox regression. At bivariate analysis, Poisson regression procedure was able to identify the same variables identified by Cox regression.

In multivariate analysis, an exploratory model analysis was performed to explore the relations between the variables while simultaneously adjusting for all other variables that had significant association with nutritional outcome. After investigation of confounding, all variables with p-values of 0.05 or less were considered possible confounders and were retained for the model analysis.

Using Cox Proportional Hazards regression two factors; change in CD4 count, and client's socio-support, were found to significantly influence variability in survival (non-improvement) among the clients. Upon fitting this two into a model containing treatment modality as the co-model, the two variables did not modify the effect of treatment modality on client's nutritional outcome. The P value indicates that the difference between treatments was not statistically significant, where as there was strong evidence that CD4 change and source of client's socio-support was associated with length of survival. The analysis was done using R statistical package, where two methods used for handling ties ("Breslow" and "Efron") were applied (**Table 4.4.1.2** and **Table 4.4.1.4**).

In the R software, the default "efron" method is generally preferred to the once-popular "Breslow" method. Both are used to estimate the "Exact" method which is much more computationally intensive. "Efron" method gives the approximations to the "Exact" method without using the tremendous time it takes a CPU to run the "Exact" method. Both the "Efron" and the "Breslow" methods do reasonably well at approximating the "Exact" when there are not a lot of ties. If there are a lot of ties, then the "Breslow" approximation to the "Exact" will be very poor.

To fit a multivariate Poisson regression (Log-linear model) on the rate of nutritional improvement, time interval (12 levels) was used to adjust for the effect of treatment modality (2 levels), CD4 change (4 levels) and socio-support (2 levels). A unique combination of each level of every variable generated 192 permuted blocks (Contingency table). Nutritional status variable was used to determine the number of events and the number of censored. A total of both events and censored constituted the total number at risk at the start of each interval. The count of events (Nutritional improvement) was determined for each combination to form the counts variable. Upon regressing the count variable with the four variables; time interval, treatment modality, CD4 change and socio-support, the outcome were as shown in **Table 4.5.2.3.** The results are exact replica of those shown in **Table 4.4.1.2**, generated using Cox regression by specifying "method = breslow". Interpretation of these two results is the same since the outcome measurement is the relative risk. There was strong evidence of comparability of the results generated by Cox and Poisson regression. The results for **Table 4.4.1.4**,

generated using Cox regression by specifying "method = efron" are not different statistically speaking. The inference still remain as that for **Table 4.5.2.3** and **Table 4.4.1.2**.

CHAPTER 6: CONCLUSION

Conclusion

The outcome of the results confirms the initial intention i.e. to show that survival analysis is not limited by the nature of data presented, whether on rates or on survivorship. When presented with data on survivorship, Cox regression is the better option and when presented with data on rates, Poisson regression is the recommended option. Even though the input variables (dependent) are different in nature (Time-to-event for Cox and Count of events per time for Poisson), the output measurement is the same for all i.e. Relative risk. Both analysis yield to the same conclusion.

6.1 Application

This study forms a basis for more extended research work in applicability of other statistical methodologies (e.g. Logistic regression) in survival analysis; by comparing these methods with the standard statistical methodology of survival data analysis (i.e. Cox Proportion Hazards regression).

6.2 Study Limitations

- Duration of follow-up was based on months of visit since enrolment. The clients on Food and nutrition counseling arm were given food prescription enough for the whole month without monitoring to ensure they were the only consumers of

the food. There could be possibility of sharing the food at the household level, therefore dilution the impact of food in the treatment group.

- It was difficult to tell exactly when the event of interest (Positive BMI change from <18.5 Kg/m^2 to $\geq$ 19.5 Kg/m^2) was realized since the visits were monthly.

REFERENCES

Aitkin, M, Anderson, D A, Francis, B, Hinde, J P. Statistical Modelling in GLIM. 1989 Oxford, Clarendon Press.

Allison, P D. Survival Analysis Using the SAS® System: A Practical Guide. 1995 Cary, NC, SAS Institute Inc.

Armitage, P. The chi-square test for heterogeneity of proportions after adjustment for stratification. Journal of the Royal Statistical Society, 1966; **B28**: 150-163.

Berry, G. The analysis of mortality by the subject-years method. Biometrics 1983; **39**:173–180.

Bishop, Y M M, Fienberg, S F and Holland, P W. Discrete Multivariate Analysis. 1975 Cambridge, Massachusetts: M.I.T. Press.

Bjarnason, O, Day, N, Snaedal, J and Tulinius, H. The effect of year of birth on the breast cancer age-incidence curve in Ireland. International Journal of Cancer, 1974;**13**: 689-696.

Bloembergen, W E, Port, F K, Mauger, E A, Wolfe, R A. A comparison of mortality between patients treated with hemodialysis and peritoneal dialysis. J Am Soc Nephrol 1995; **6**:177–183.

Breslow N E. Covariance analysis of censored survival data. Biometrics 1974; **30**: 89-99.

Breslow, N E and Day, N E. Indirect standardization and multivariate models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. Journal of Chronic Diseases 1975; **28**: 289-303.

Case, R A M and Lea, A J. Mustard gas poisoning, chronic bronchitis and lung cancer. Brit. J. Prev. Soc. Med. 1955; **9**: 62-72.

Castelmen, T. Impacts of Specialized Food Products on HIV-infected Adults and Malnourished Children: Emerging Evidence from Randomized Trials. 2008 International Food Aid Conference.

Collett D. Modelling Survival Data in Medical Research. 1994 London: Chapman & Hall.

Collet, D. Modelling Survival Data in Medical Research, Second edition. 2003 Chapman & Hall / CRC, New York.

Cox, D R. Regression models and life tables (with discussion). J Royal Stat Society 1972; **B34**:187–220.

Cox, D R and Oakes, D: Analysis of Survival Data. 1984 London, Chapman & Hall.

Edward F Vonesh, Douglas E Schaubel, Wenli Hao, and Allan J Collins. Statistical methods for comparing mortality among ESRD patients: Examples of regional/international variations. Kidney International, 2000; **57**: S19–S27.

Efron, B. Bootstrap methods: Another look at the jackknife. Annals of Statistics 1977; **7**: 1 – 26.

Fenton, S S A, Schaubel, D E, Desmeules, M, Morrison, H I, Mao, Y, Copleston, P, Jeffery, J R, Kjellstrand, C M. Hemodialysis versus peritoneal dialysis: A comparison of adjusted mortality rates. Am J Kidney Dis 1997; **30**:334–342.

Frome E L. Analysis of Rates Using Poisson Regression Models:_Biometrics, 1983; **39(3)**:665-74.

Gail, M. The analysis of heterogeneity for indirect standardized mortality ratios. Journal of the Royal Statistical Society 1978; **A1419**: 224-234.

Haberman, S J. Analysis of Qualitative Data. Vol. I, Introductory Topics. 1978 New York: Academic Press.

Holford, T R. Life tables with concomitant information. Biometrics 1976; **32**: 587-597.

Holford, T R. The analysis of rates and of survivorship using log-linear models. Biometrics 1980; **36**:299–305.

Kalbfleisch, J D and Prentice, R L. The Statistical Analysis of Failure Time Data. 1980 New York, John Wiley & Sons, Inc.

Kaplan, EL and Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; **53**: 457–481.

Kilpatrick, S J. Occupational mortality indices. Population Studies 1962; **16**: 175-187.

McCullagh, P & Nelder, J A. Generalized Linear Models. 1989 London, Chapman and Hall.

Osborn, J. A multiplicative model for the analysis of vital statistical rates. Applied Statistics 1975; **24**: 75-84.

Peto, R, Pike, M C, Armitage, P, Breslow, N E, Cox, D R, Howard, S V, Mantel, N, McPherson, K, Peto, J, Smith, P G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: analysis and examples. Br J Cancer 1977; **35**: 1–39

Venables, W N and Ripley B D. Modern Applied Statistics with S-PLUS. 1999 Third Edition, Springer, ISBN 0-387-98825-4.

Vonesh, E F & Moran, J: Mortality in end-stage renal disease: A reassessment of differences between patients treated with hemodialysis and peritoneal dialysis. J Am Soc Nephrol 1999; **10**:354–365.

APPENDIX A: Poisson regression analysis tables

Table A1: Rate of nutritional improvement by gender adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.59	0.58		-7.91	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.56	0.575
Interval: 3	2.13	0.62		3.43	0.001
Interval: 4	2.62	0.61		4.29	<0.001
Interval: 5	2.31	0.64		3.63	<0.001
Interval: 6	1.29	0.76		1.69	0.090
Interval: 7	2.35	0.66		3.58	<0.001
Interval: 8	3.13	0.63		5.01	<0.001
Interval: 9	3.51	0.63		5.55	<0.001
Interval: 10	2.86	0.76		3.75	<0.001
Interval: 11	3.16	0.82		3.88	<0.001
Interval: 12	4.15	0.73		5.68	<0.001
Gender: Male	-0.28	0.19		-1.48	0.139
Gender: Female	Reference				

Table A2: Rate of nutritional improvement by age adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.55	0.65		-6.982	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.558	0.577
Interval: 3	2.13	0.62		3.421	<0.001
Interval: 4	2.62	0.61		4.294	<0.001
Interval: 5	2.33	0.64		3.658	<0.001
Interval: 6	1.31	0.76		1.720	0.086
Interval: 7	2.37	0.66		3.602	<0.001
Interval: 8	3.14	0.63		5.017	<0.001
Interval: 9	3.52	0.63		5.571	<0.001
Interval: 10	2.87	0.76		3.758	<0.001
Interval: 11	3.19	0.82		3.903	<0.001
Interval: 12	4.14	0.73		5.662	<0.001
Age in years: < 30	0.01	0.36		0.030	0.976
Age in years: 30 - 39	-0.23	0.35		-0.678	0.498
Age in years: 40 - 49	-0.25	0.37		-0.685	0.494
Age in years: 50+	Reference				

Table A3: Rate of nutritional improvement by current marital status adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.68	0.59		-7.955	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.558	0.577
Interval: 3	2.12	0.62		3.417	<0.001
Interval: 4	2.62	0.61		4.294	<0.001
Interval: 5	2.33	0.64		3.657	<0.001
Interval: 6	1.31	0.76		1.713	0.087
Interval: 7	2.37	0.66		3.601	<0.001
Interval: 8	3.15	0.63		5.022	<0.001
Interval: 9	3.55	0.63		5.602	<0.001
Interval: 10	2.91	0.76		3.803	<0.001
Interval: 11	3.23	0.82		3.946	<0.001
Interval: 12	4.19	0.73		5.721	<0.001
Marital status: Single	0.19	0.24		0.801	0.423
Marital status: Married	-0.18	0.21		-0.894	0.371
Marital status: Separated/divorced/ widowed	Reference				

Table A4: Rate of nutritional improvement by education level adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-5.21	1.14		-4.563	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.558	0.577
Interval: 3	2.12	0.62		3.405	<0.001
Interval: 4	2.61	0.61		4.266	<0.001
Interval: 5	2.30	0.64		3.621	<0.001
Interval: 6	1.28	0.76		1.676	0.094
Interval: 7	2.33	0.66		3.544	<0.001
Interval: 8	3.11	0.63		4.973	<0.001
Interval: 9	3.51	0.63		5.546	<0.001
Interval: 10	2.87	0.76		3.761	<0.001
Interval: 11	3.16	0.82		3.867	<0.001
Interval: 12	4.14	0.73		5.657	<0.001
Education: None	0.01	1.12		0.012	0.991
Education: Primary	0.58	1.01		0.576	0.565
Education: Secondary	0.48	1.02		0.475	0.635
Education: Tertiary	Reference				

Table A5: Rate of nutritional improvement by main source of household income adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-3.68	0.69		-5.320	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.561	0.575
Interval: 3	2.13	0.62		3.425	0.001
Interval: 4	2.62	0.61		4.293	<0.001
Interval: 5	2.36	0.64		3.710	<0.001
Interval: 6	1.35	0.76		1.766	0.077
Interval: 7	2.41	0.66		3.649	<0.001
Interval: 8	3.20	0.63		5.096	<0.001
Interval: 9	3.59	0.63		5.668	<0.001
Interval: 10	2.94	0.77		3.846	<0.001
Interval: 11	3.22	0.82		3.942	<0.001
Interval: 12	4.18	0.73		5.714	<0.001
Source of income: Formal employment	-1.20	0.49		-2.456	0.014
Source of income: Small scale business	-1.08	0.43		-2.506	0.012
Source of income: Welfare/NGO support	Reference				

Table A6: Rate of nutritional improvement by amount spent on buying food per day in Kenya shillings adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.73	0.70		-6.738	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.562	0.574
Interval: 3	2.13	0.62		3.428	0.001
Interval: 4	2.63	0.61		4.299	<0.001
Interval: 5	2.33	0.64		3.668	<0.001
Interval: 6	1.31	0.76		1.719	0.086
Interval: 7	2.38	0.66		3.616	<0.001
Interval: 8	3.15	0.63		5.028	<0.001
Interval: 9	3.54	0.63		5.601	<0.001
Interval: 10	2.89	0.76		3.782	<0.001
Interval: 11	3.23	0.82		3.948	<0.001
Interval: 12	4.13	0.73		5.650	<0.001
Amount spent per day: Nil	0.09	0.53		0.166	0.868
Amount spent per day: <50	0.18	0.44		0.403	0.687
Amount spent per day: 50 - <100	-0.23	0.43		-0.533	0.594
Amount spent per day: 100 - <200	0.30	0.45		0.657	0.511
Amount spent per day: >=200	Reference				

Table A7: Rate of nutritional improvement by distance to health facility in kilometers adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.75	0.59		-8.101	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.560	0.575
Interval: 3	2.12	0.62		3.414	0.001
Interval: 4	2.62	0.61		4.288	<0.001
Interval: 5	2.34	0.64		3.669	<0.001
Interval: 6	1.31	0.76		1.715	0.086
Interval: 7	2.38	0.66		3.613	<0.001
Interval: 8	3.16	0.63		5.041	<0.001
Interval: 9	3.58	0.63		5.652	<0.001
Interval: 10	2.94	0.76		3.840	<0.001
Interval: 11	3.21	0.82		3.929	<0.001
Interval: 12	4.22	0.73		5.749	<0.001
Distance to health facility: < 5	-0.08	0.20		-0.394	0.693
Distance to health facility: 5 - 9.99	0.45	0.27		1.683	0.092
Distance to health facility: 10+	Reference				

Table A8: Rate of nutritional improvement by whether the client has ever smoked a cigarette adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.63	0.58		-7.967	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.562	0.574
Interval: 3	2.13	0.62		3.425	0.001
Interval: 4	2.62	0.61		4.285	<0.001
Interval: 5	2.31	0.64		3.638	<0.001
Interval: 6	1.29	0.76		1.689	0.091
Interval: 7	2.35	0.66		3.573	<0.001
Interval: 8	3.13	0.63		5.001	<0.001
Interval: 9	3.52	0.63		5.561	<0.001
Interval: 10	2.87	0.76		3.753	<0.001
Interval: 11	3.19	0.82		3.904	<0.001
Interval: 12	4.18	0.73		5.704	<0.001
Cigarettes smoking: Ever smoked	-0.22	0.19		-1.148	0.251
Cigarettes smoking: Never smoked	Reference				

Table A9: Rate of nutritional improvement by whether the client has ever abuse a drug adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.68	0.58		-8.089	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.560	0.575
Interval: 3	2.13	0.62		3.422	0.001
Interval: 4	2.62	0.61		4.286	<0.001
Interval: 5	2.32	0.64		3.646	<0.001
Interval: 6	1.29	0.76		1.691	0.091
Interval: 7	2.35	0.66		3.572	<0.001
Interval: 8	3.13	0.63		4.993	<0.001
Interval: 9	3.51	0.63		5.542	<0.001
Interval: 10	2.86	0.76		3.745	<0.001
Interval: 11	3.15	0.82		3.859	<0.001
Interval: 12	4.11	0.73		5.631	<0.001
Drug abuse: Ever abused	-0.24	0.32		-0.768	0.442
Drug abuse: Never abused	Reference				

Table A10: Rate of nutritional improvement by whether the client has ever consumed alcohol adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.57	0.58		-7.821	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.557	0.578
Interval: 3	2.12	0.62		3.417	0.001
Interval: 4	2.61	0.61		4.279	<0.001
Interval: 5	2.31	0.64		3.626	<0.001
Interval: 6	1.28	0.76		1.679	0.093
Interval: 7	2.34	0.66		3.560	<0.001
Interval: 8	3.13	0.63		4.991	<0.001
Interval: 9	3.52	0.63		5.567	<0.001
Interval: 10	2.87	0.76		3.757	<0.001
Interval: 11	3.20	0.82		3.918	<0.001
Interval: 12	4.20	0.73		5.727	<0.001
Alcohol: Ever consumed	-0.25	0.18		-1.365	0.172
Alcohol: Never consumed	Reference				

Table A11: Rate of nutritional improvement by whether the client does physical exercise twice weekly adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.63	0.58		-7.979	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.562	0.574
Interval: 3	2.12	0.62		3.418	0.001
Interval: 4	2.61	0.61		4.278	<0.001
Interval: 5	2.32	0.64		3.640	<0.001
Interval: 6	1.29	0.76		1.695	0.091
Interval: 7	2.35	0.66		3.574	<0.001
Interval: 8	3.13	0.63		5.001	<0.001
Interval: 9	3.53	0.63		5.577	<0.001
Interval: 10	2.89	0.76		3.777	<0.001
Interval: 11	3.18	0.82		3.888	<0.001
Interval: 12	4.13	0.73		5.655	<0.001
Physical exercise: Performs	-0.21	0.19		-1.094	0.274
Physical exercise: Does not perform	Reference				

Table A12: Rate of nutritional improvement by source of social support received adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.77	0.58		-8.238	<0.001
Interval: 1	Reference				
Interval: 2	0.44	0.76		0.572	0.568
Interval: 3	2.13	0.62		3.432	0.001
Interval: 4	2.63	0.61		4.306	<0.001
Interval: 5	2.33	0.64		3.659	<0.001
Interval: 6	1.30	0.76		1.696	0.090
Interval: 7	2.35	0.66		3.570	<0.001
Interval: 8	3.14	0.63		5.019	<0.001
Interval: 9	3.53	0.63		5.577	<0.001
Interval: 10	2.91	0.76		3.805	<0.001
Interval: 11	3.23	0.82		3.949	<0.001
Interval: 12	4.18	0.73		5.716	<0.001
Socio-support: Medical professionals and family	0.55	0.28		1.998	0.046
Socio-support: Others	Reference				

Table A13: Rate of nutritional improvement by one knowing HIV status adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.74	0.58		-8.186	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.561	0.575
Interval: 3	2.13	0.62		3.422	0.001
Interval: 4	2.62	0.61		4.289	<0.001
Interval: 5	2.32	0.64		3.650	<0.001
Interval: 6	1.29	0.76		1.692	0.091
Interval: 7	2.35	0.66		3.572	<0.001
Interval: 8	3.13	0.63		4.999	<0.001
Interval: 9	3.50	0.63		5.540	<0.001
Interval: 10	2.86	0.76		3.740	<0.001
Interval: 11	3.17	0.82		3.888	<0.001
Interval: 12	4.12	0.73		5.636	<0.001
HIV status known by: Medical professionals and family	0.27	0.26		1.043	0.297
HIV status known by: Others	Reference				

Table A14: Rate of nutritional improvement by support received in the last one month adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.72	0.59		-8.010	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.559	0.576
Interval: 3	2.12	0.62		3.418	0.001
Interval: 4	2.62	0.61		4.282	<0.001
Interval: 5	2.32	0.64		3.641	<0.001
Interval: 6	1.29	0.76		1.691	0.091
Interval: 7	2.35	0.66		3.569	<0.001
Interval: 8	3.13	0.63		4.993	<0.001
Interval: 9	3.52	0.63		5.561	<0.001
Interval: 10	2.87	0.76		3.756	<0.001
Interval: 11	3.16	0.82		3.869	<0.001
Interval: 12	4.11	0.73		5.621	<0.001
Support: Received	0.03	0.19		0.163	0.871
Support: Not received	Reference				

Table A15: Rate of nutritional improvement by Treatment groups adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.58	0.58		-7.842	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.563	0.573
Interval: 3	2.13	0.62		3.429	0.001
Interval: 4	2.62	0.61		4.284	<0.001
Interval: 5	2.31	0.64		3.635	<0.001
Interval: 6	1.29	0.76		1.688	0.091
Interval: 7	2.35	0.66		3.568	<0.001
Interval: 8	3.13	0.63		4.999	<0.001
Interval: 9	3.52	0.63		5.570	<0.001
Interval: 10	2.86	0.76		3.747	<0.001
Interval: 11	3.18	0.82		3.898	<0.001
Interval: 12	4.14	0.73		5.668	<0.001
Treatment groups: F + NC	-0.23	0.18		-1.290	0.197
Treatment groups: NC	Reference				

Table A16: Rate of nutritional improvement by CD4 change adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-22.85	2425.03		-0.009	0.993
Interval: 1	Reference				
Interval: 2	16.62	2425.03		0.007	0.995
Interval: 3	18.28	2425.03		0.008	0.994
Interval: 4	18.75	2425.03		0.008	0.994
Interval: 5	16.99	2425.03		0.007	0.994
Interval: 6	17.10	2425.03		0.007	0.994
Interval: 7	18.87	2425.03		0.008	0.994
Interval: 8	19.89	2425.03		0.008	0.994
Interval: 9	19.94	2425.03		0.008	0.993
Interval: 10	19.47	2425.03		0.008	0.994
Interval: 11	19.47	2425.03		0.008	0.994
Interval: 12	20.87	2425.03		0.009	0.993
CD4 change: Changed by 1 to 100 counts	1.50	0.76		1.970	0.049
CD4 change: Changed by 101 to 200 counts	1.85	0.76		2.452	0.014
CD4 change: Changed by > 200 counts	1.87	0.76		2.441	0.015
CD4 change: Deteriorated /No change	Reference				

Table A17: Rate of nutritional improvement by WBC change adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-22.36	3975.80		-0.006	0.996
Interval: 1	Reference				
Interval: 2	17.61	3975.80		0.004	0.997
Interval: 3	19.04	3975.80		0.005	0.996
Interval: 4	19.39	3975.80		0.005	0.996
Interval: 5	17.95	3975.80		0.005	0.996
Interval: 6	18.08	3975.80		0.005	0.996
Interval: 7	19.83	3975.80		0.005	0.996
Interval: 8	20.87	3975.80		0.005	0.996
Interval: 9	20.67	3975.80		0.005	0.996
Interval: 10	19.91	3975.80		0.005	0.996
Interval: 11	20.33	3975.80		0.005	0.996
Interval: 12	21.33	3975.80		0.005	0.996
WBC change: Improved	0.58	0.34		1.734	0.083
WBC change: Deteriorated / No change	Reference				

Table A18: Rate of nutritional improvement by RBC change adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-21.93	4007.08		-0.005	0.996
Interval: 1	Reference				
Interval: 2	17.62	4007.08		0.004	0.996
Interval: 3	19.05	4007.08		0.005	0.996
Interval: 4	19.40	4007.08		0.005	0.996
Interval: 5	17.94	4007.08		0.004	0.996
Interval: 6	18.09	4007.08		0.005	0.996
Interval: 7	19.84	4007.08		0.005	0.996
Interval: 8	20.88	4007.08		0.005	0.996
Interval: 9	20.69	4007.08		0.005	0.996
Interval: 10	19.78	4007.08		0.005	0.996
Interval: 11	20.27	4007.08		0.005	0.996
Interval: 12	21.37	4007.08		0.005	0.996
RBC change: Improved	-0.28	0.33		-0.839	0.402
RBC change: Deteriorated / No change	Reference				

Table A19: Rate of nutritional improvement by Hb change adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-22.09	3593.29		-0.006	0.995
Interval: 1	Reference				
Interval: 2	17.40	3593.29		0.005	0.996
Interval: 3	18.83	3593.29		0.005	0.996
Interval: 4	19.37	3593.29		0.005	0.996
Interval: 5	17.73	3593.29		0.005	0.996
Interval: 6	17.86	3593.29		0.005	0.996
Interval: 7	19.62	3593.29		0.005	0.996
Interval: 8	20.67	3593.29		0.006	0.995
Interval: 9	20.46	3593.29		0.006	0.995
Interval: 10	19.47	3593.29		0.005	0.996
Interval: 11	19.93	3593.29		0.006	0.996
Interval: 12	21.34	3593.29		0.006	0.995
Hb change: Improved	0.24	0.48		0.505	0.614
Hb change: Deteriorated / No change	Reference				

Table A20: Rate of nutritional improvement by baseline BMI adjusting for intervals

Variables	β	S. E. (β)	Exp (β)	z value	Pr(>|z|)
(Intercept)	-4.76	0.62		-7.669	<0.001
Interval: 1	Reference				
Interval: 2	0.43	0.76		0.559	0.576
Interval: 3	2.12	0.62		3.417	0.001
Interval: 4	2.61	0.61		4.278	<0.001
Interval: 5	2.31	0.64		3.633	<0.001
Interval: 6	1.29	0.76		1.684	0.092
Interval: 7	2.34	0.66		3.557	<0.001
Interval: 8	3.12	0.63		4.981	<0.001
Interval: 9	3.51	0.63		5.547	<0.001
Interval: 10	2.86	0.76		3.742	<0.001
Interval: 11	3.15	0.82		3.846	<0.001
Interval: 12	4.10	0.73		5.600	<0.001
Baseline BMI: < 16 Kg/m^2	0.08	0.28		0.276	0.782
Baseline BMI: 16 - 18.5 Kg/m^2	Reference				

APPENDIX B: Analysis Codes

(a) Estimation of Survival function using Kaplan-Meir: R

```
dat1=read.csv("kaplan.csv",header=T)
attach(dat1)
dat1
names(dat1)
library(splines)
library(survival)
library(MASS)

fit1=survfit(Surv(time,cen)~1,data=dat1)
summary(fit1)
plot(fit1)

fit2=survfit(Surv(time,cen)~grp,data=dat1)
summary(fit2)
plot(fit2)
```

(b) Estimation of Survival function using Kaplan-Meir: SPSS

```
KM
  time  /STATUS=cen(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL HAZARD .

KM
  time  BY grp  /STATUS=cen(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL HAZARD
  /TEST LOGRANK BRESLOW
  /COMPARE OVERALL POOLED
```

(c) Survival analysis using Cox Proportional Hazards Regression – Estimation using Breslow method: R

```
dat1=read.csv("cox.reg.csv",header=T)
attach(dat1)
dat1
names(dat1)
library(splines)
library(survival)
library(MASS)

fit3 <- coxph(Surv(time,cen)~as.factor(grp)+as.factor(cd4)+as.factor(support),
method="breslow",data=dat1)
summary(fit3)
```

Call:
coxph(formula = Surv(time, cen) ~ as.factor(grp) + as.factor(cd4) + as.factor(support),
data = dat1, method = "breslow")

n=101 (229 observations deleted due to missingness)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
as.factor(grp)1	-0.1612	0.8511	0.3242	-0.497	0.6190
as.factor(cd4)2	1.4399	4.2201	0.7664	1.879	0.0603 .
as.factor(cd4)3	1.7583	5.8026	0.7615	2.309	0.0210 *
as.factor(cd4)4	1.8306	6.2374	0.7736	2.366	0.0180 *
as.factor(support)2	1.6176	5.0412	0.7858	2.059	0.0395 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
as.factor(grp)1	0.851	1.1750	0.4508	1.607
as.factor(cd4)2	4.220	0.2370	0.9397	18.953
as.factor(cd4)3	5.803	0.1723	1.3043	25.814
as.factor(cd4)4	6.237	0.1603	1.3692	28.414
as.factor(support)2	5.041	0.1984	1.0805	23.520

Rsquare = 0.128   (max possible= 0.962)
Likelihood ratio test = 13.85 on 5 df, p=0.01658
Wald test = 11.31 on 5 df, p=0.04563
Score (logrank) test = 14.43 on 5 df, p=0.01308

(d) Survival analysis using Cox Proportional Hazards Regression – Estimation using Efron method: R

```
dat1=read.csv("cox.reg.csv",header=T)
attach(dat1)
dat1
names(dat1)
library(splines)
library(survival)
library(MASS)


Fit4 <- coxph(Surv(time,cen)~as.factor(grp)+as.factor(cd4)+as.factor(support),
method="efron",data=dat1)
summary(fit4)


Call:
coxph(formula = Surv(time, cen) ~ as.factor(grp) + as.factor(cd4) +
    as.factor(support), data = dat1, method = "efron")
```

  n=101 (229 observations deleted due to missingness)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
as.factor(grp)1	-0.2043	0.8152	0.3262	-0.626	0.5312
as.factor(cd4)2	1.5215	4.5793	0.7658	1.987	0.0469 *
as.factor(cd4)3	1.9104	6.7561	0.7650	2.497	0.0125 *
as.factor(cd4)4	1.9641	7.1288	0.7774	2.527	0.0115 *
as.factor(support)2	1.7689	5.8645	0.7857	2.251	0.0244 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
as.factor(grp)1	0.8152	1.2267	0.4301	1.545
as.factor(cd4)2	4.5793	0.2184	1.0208	20.543
as.factor(cd4)3	6.7561	0.1480	1.5085	30.259
as.factor(cd4)4	7.1288	0.1403	1.5534	32.715
as.factor(support)2	5.8645	0.1705	1.2573	27.354

Rsquare = 0.149   (max possible= 0.959)
Likelihood ratio test= 16.32 on 5 df, p=0.005997
Wald test = 13.26 on 5 df, p=0.02107
Score (logrank) test = 17.38 on 5 df, p=0.003838

```
dat2=read.csv("pois.reg.csv",header=T)
attach(dat2)
dat2
fit5=glm(formula=event~as.factor(Interval)+as.factor(grp)+as.factor(cd4)+as.factor(sup
port),family=poisson,data=dat2,offset=logtotal)
summary(fit5)
Call:
glm(formula = event ~ as.factor(Interval) + as.factor(grp) + as.factor(cd4) +
    as.factor(support), family = poisson, data = dat2, offset = logtotal)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.763e+00	-5.384e-01	-3.147e-01	-3.806e-05	1.721e+00

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-22.9381	2478.2735	-0.009	0.9926	
as.factor(Interval)2	16.6971	2478.2735	0.007	0.9946	
as.factor(Interval)3	18.3913	2478.2734	0.007	0.9941	
as.factor(Interval)4	18.9232	2478.2734	0.008	0.9939	
as.factor(Interval)5	17.1467	2478.2735	0.007	0.9945	
as.factor(Interval)6	17.2558	2478.2735	0.007	0.9944	
as.factor(Interval)7	19.0013	2478.2734	0.008	0.9939	
as.factor(Interval)8	20.1110	2478.2734	0.008	0.9935	
as.factor(Interval)9	20.1623	2478.2734	0.008	0.9935	
as.factor(Interval)10	19.6683	2478.2735	0.008	0.9937	
as.factor(Interval)11	19.7349	2478.2736	0.008	0.9936	
as.factor(Interval)12	21.1017	2478.2734	0.009	0.9932	
as.factor(grp)2	-0.1612	0.3242	-0.497	0.6190	
as.factor(cd4)2	1.4399	0.7664	1.879	0.0603	.
as.factor(cd4)3	1.7583	0.7615	2.309	0.0210	*
as.factor(cd4)4	1.8306	0.7736	2.366	0.0180	*
as.factor(support)2	1.6176	0.7858	2.059	0.0395	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 123.049 on 99 degrees of freedom
Residual deviance: 50.519 on 83 degrees of freedom
(92 observations deleted due to missingness)
AIC: 154.84
Number of Fisher Scoring iterations: 17