Survival Analysis and Generalized Estimating Equations for Repeated Measures in Mosquito Dose-Response

Gabriel Otieno Okello

A dissertation submitted in partial fulfillment of the requirements for the award of the degree of Master of Science in Research Methods of Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

This dissertation is my original work and it has not been presented for a degree in any other University. The work of others used in this study has been dully acknowledged.

Signature. . . . . Date . . . .

Gabriel Otieno Okello

This dissertation submitted with our knowledge and evaluated under our guidance as Jomo Kenyatta University of Agriculture and Technology and African Insect Science for Food and Health (ICIPE)

We declare that, this dissertation is from the student's own work and effort and where he has used other sources of information, it has been acknowledged.

Signature. . . . . Date . . . .

Dr. Gichuhi .A. Waititu,

JKUAT, Kenya

Signature. . . . . Date . . . .

Dr. Daisy Salifu,

ICIPE, Kenya

DEDICATION

I dedicate this dissertation to my family as a whole, for their constant love, support and inspiration

ACKNOWLEDGEMENT

My sincere appreciation goes to the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) for facilitating my studies through sponsorship.

I would like to thank Dr Waititu and Dr Salifu for supporting my work and the helpful advices they gave me to finish this challenging dissertation. Despite their numerous obligations they always found time to provide their advices when the problems arose.

Sincere thanks goes to African Insect Science for Food and Health (ICIPE) and its staff, for hosting and allowing me to do my research at the institution and for assisting me with more information whenever I needed.

I would like to thank all the professors, graduate students, and staff from the Faculty of Agriculture and Department of Statistics and Actuarial Science, who contributed in one way or another to my experience at Jomo Kenyatta University of Agriculture and Technology.

Last but not least, I want to thank my family, who have supported me throughout my entire academic studies. Without the help of all the mentioned people, this project could not have been completed!

Thank you very much!

ABSTRACT

Dose-response studies in arthropod research usually involve observing and collecting successive information at different times on the same group of organisms (insects) exposed to different concentrations of stimulus such as botanical extracts. When successive observations are made on the same group of organisms at several concentrations over time the data becomes correlated. Correlated insect mortality data cannot be analyzed using Probit Analysis technique which is the usual way of analyzing data from bioassay experiments. In addition, when the speed of kill is of interest since mortality varies over time, estimating lethal time is the best. The objective of the study was to evaluate a complementary approach, Survival Analysis and repeated measures logistic regression using Generalized Estimating Equations, for lethal time determination in mosquito dose response. Mortality data from anopheles larva exposed to botanical extracts of different concentrations were used in the study. The Kaplan-Meier survival analysis technique and repeated measures logistic regression using GEE were used in estimating lethal time (LT_{50}) of the botanical extracts for control of mosquito larva and their performances were then compared. Results of this study suggest that different botanical extracts and the different concentration levels were significantly different from each other. Concentration 500 mg/ml was found to the most virulent chemical across all extracts, followed by concentration 250 mg/ml and concentration 50 mg/ml was the least active. The confidence intervals of the LT_{50} estimates from repeated measures logistic regression using GEE were consistently wider compared to those from Kaplan-Meier. Kaplan-Meier survival function and repeated measures logistic regression using GEE were effective tools for analyzing repeated measures data from mosquito dose response. Repeated measures logistic regression using GEE was a better method of estimating LT_{50} since it consistently gave precise LT_{50} estimates with a smaller confidence interval. It’s suggested that the repeated measures logistic regression using GEE method can be incorporated into arthropod dose response studies for repeated measures together with other existing methods for analyzing data from bioassay experiments.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
CLL	Complementary log-log
CPH	Cox Proportional Hazard
GEE	Generalized Estimating Equations
GLM	Generalized Linear Model
LT	Lethal Time
MSE	Mean Square Error
PL	Product-Limit
QIC	Quasi-likelihood Information Criterion

CHAPTER ONE

INTRODUCTION

1.1 Background information

Dose-response studies in arthropod research usually involve observing and taking measurements of insects' mortality on groups of insects subjected to different concentrations of stimulus over time, giving rise to repeated measures data. The same measure (number of dead insects) is collected several times in a single group of insects exposed to one or several concentrations of a stimulus e.g. pesticides or botanical extracts. The same insects subjected to each or several doses are followed over time.

Insects are exposed to a particular botanical extracts and the effect (mortality/death) from the same experimental unit is observed at different time intervals, say, t_1, t_2, t_3…. Data collected this way are usually correlated because successive observations are made on the same group of organisms at several concentrations over time (Robertson and Preisler, 1992; Thorne *et al.*, 1995; Nowierski *et al.*, 1996; Thomsen and Eilenberg, 2000).

Correlated mortality data cannot be an analyzed using standard probit analysis technique (Finney, 1964; Finney, 1971) which is the usual way of analyzing data from bioassay experiment (Thorne *et al.*, 1995; Robertson and Preisler, 1992). Probit Analysis is adequate if the responses are independent, true for data collected at once after a given time point.

In arthropod dose response studies, samples of insets are usually exposed to several concentrations of insecticide to determine the concentration that will kill 50% of the insects within a given time span (Finney, 1971; Eaton and Kells, 2009). Effects of time on the percentage of kill at one or several concentration (serial-time-mortality or time-dose-mortality data) may be of interest when the speed of kill is important as might occur with pests which lay all of its eggs within few days (Thorne *et al.*, 1995). This is also because mortality varies with time. Any tests comparing lethal time values should include confidence limits of the estimated statistics (Thorne *et al.*, 1995).

Given the correlated measurements in dose-response studies in addition to taking interest in the speed of kill, one has to move on to alternative methods that take care of the correlation in the data while estimating lethal time. Two of such methods are Survival Analysis and Generalized Estimating Equations (GEE).

Survival analysis encompasses a wide variety of methods for analyzing time to an event data as stated by Cox and Oakes (1984). The response is often referred to as a failure time, survival time, or event time. In Survival Analysis, the object of primary interest is estimating the survival function (Hosmer and Lameshow, 1999). The objectives of survival analysis are to estimate time to event, to compare time to event between two or more groups and to assess the relationship of co-variables to time to event. Survival time can be defined as time to the occurrence of a given event. Event of interest may be disease development, response to treatment, relapse or death. The survival data are typically not fully covered but rather censored. Censoring is present when there is no full information about a subject’s event time. For example, in mosquito dose response study some mosquitoes may still be alive at the end of the study period or the exact time when they died might not be known. If there is no censoring, standard regression procedures could be used. However, these may be inadequate because time to event is also restricted to be positive and has a skewed distribution (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 1980).

In mosquito dose-response study, while observing insects' response to plants extracts over time, the time to event (death) may be recorded and hence the survival analysis techniques may be applied to such data to estimate the median survival time. Survival models are therefore more useful than conventional dose-response methods for predicting the effects of a stimulus (stressor) on field of populations. For example, Borsuk *et al.* (2002) used survival model in their study on organisms' response towards a given concentration.

Generalized Estimating Equations (GEE) were introduced by Liang and Zeger (1986) as an extension of Generalized Linear Model (GLM) method (McCullagh and Nelder, 1983; McCullagh and Nelder, 1989) to handle correlated data. Generalized linear models (GLM) are a generalization of standard linear regression that allows the response variables to have a distribution other than the normal distribution.

In many arthropod dose-response studies the assumptions for Probit Analysis such as the independence of variables and their normal distribution are hardly met yet it is being used in the analysis and estimating LT_{50} (Robertson and Preisler, 1992). According to Ziegler *et* *al.* (1998) neglecting dependencies in these situations can lead to false conclusions. The independence of outcome variables, for example, is not guaranteed when different measurements are taken from the same experimental unit.

With GEE correlated data can be modeled with output that looks similar to generalized linear models (GLMs) with independent observations. The primary difference is their ability to account for the within-subject covariance structure for the various types of response data (Ziegler *et al.*, 1998; Zeger and Liang, 1986). The available covariance structures specify how observations within a subject or cluster are correlated with each other. Correlated data are modeled with the same link functions and linear predictor equation (systematic component) as found with independent data. The random component of GEEs is also described by the same variance functions, but now the covariance structure of the correlated measurements must also be modeled (Zuur *et al.*, 2009; Zeger and Liang, 1986). GEE's a quasi-likelihood method meaning the assumption of being a member of the exponential family is discarded and only the first and second moments need to be specified. There are two types of GEEs, the subject specific and population averaged or marginal models. GEE incorporates a correlation structure for the correlated errors to obtain consistent estimators which are not biased. GEE have also been used in arthropod studies to fit repeated measures logistic regression in estimating Lethal Time (LT) as shown by Bugeme *et al.* (2009).

Logistic regression is a GLM method for analyzing binary outcome but ignores the correlated nature of the data. The standard errors may be incorrectly estimated and thus certain covariates may be incorrectly identified as significant predictors in a model. Since the arthropod dose-response data has a binary repeated measures response, generalized estimating equations (GEE) in a logistic regression setting is a good way to model the data.

In this study the use of Survival Analysis and repeated measures logistic regression using GEE are considered as complementary approach to LT_{50} estimation to address the limitation of Probit Analysis in estimating LT_{50} for correlated mosquito dose response data. The performances of the two methods were then compared based on their confidence intervals. GEE for repeated measures logistic regression was used because the data was binary and correlation due to time was to be taken into account whereas Survival Analysis was used because the event of interest was time to death of insect.

1.2 Statement of the Problem

Repeated measures from dose-response studies are a challenge because the analysis methods are not straight forward. Indeed usual methods such as the use of standard probit techniques cannot be used for analyzing repeated dose-response data from arthropod studies because the data are correlated. However, researchers resort to analysis after a given time point which is not very efficient especially when the interest is also in the speed of kill (LT_{50}) since mortality varies with time. Survival analysis and repeated measures logistic regression using GEE are two possible methods for analyzing repeated arthropod dose-response data. This study provides insights to these two methods so as to guide researchers in the analysis of repeated dose-response data.

1.3 Overall Objective

The overall objective was to evaluate the potential of using GEE and Survival Analysis for Lethal Time determination in arthropod dose response studies.

1.3.1 Specific objectives

1. To estimate LT_{50} of three botanical extracts for control of mosquito larva using Survival Analysis techniques.
2. To estimate LT_{50} of three botanical extracts for control of mosquito larva using Generalized Estimating Equations in a logistic regression setting.
3. To compare the performance of survival analysis and generalized estimating equations approaches.

1.4 Justification

The probit analysis model has received criticism as to its validity in estimating lethal time for arthropod dose response studies where insects' mortality is observed over time. Alternative models (methods or approaches) need to be evaluated for use in such studies and hence the importance of this study in arthropod research. Such methods include survival analysis and repeated measure logistic regression using GEE since they are robust to estimating lethal time when the dose response mortality data is correlated.

1.5 Dissertation Layout

This chapter introduced dose response studies and the methods of analysis for the repeated dose response data. The statement of the problem, objectives and the justification of the study have been given in this chapter. The second chapter focuses on the literature review on the existing methods for analyzing dose response data in biological research. The third chapter deals with estimation procedures for LT_{50} using survival techniques and repeated measures logistic regression using GEE. Chapter four has the results with Chapter five focusing on the discussion and conclusions.

CHAPTER TWO

LITERATURE REVIEW

2.1 Analysis of Arthropod Dose-response Data

Dose-response can be defined as the change in effect caused by differing level of exposure to a stressor (Finney, 1971). Dose-response studies are important tools for investigating the existence, nature and extent of a dose effect on efficacy. Data from dose-response studies can either be independent or has some repeated measurement depending on the aims of the different studies being conducted, which in turn gives rise to different designs for data collection. Repeated measures arise when the response outcome is measured on the same experimental unit at different times or under different conditions.

Different kinds of data collected from arthropod dose-response studies will require different kinds of analysis depending on the nature of the data. Arthropod dose-response (mortality) data are usually analyzed to evaluate efficacy of insect control agents in terms of estimating lethal time (LT), lethal dose (LD) and lethal concentration (LC). These are usually estimated by different methods depending on the methodology and the assumptions made. The proportion of the estimates could be 50%, 90% or 95% and these are some of the standard measurements of efficacy.

Some of the commonly used methods in analyzing dose-response data (arthropod mortality data) include Probit analysis (Finney, 1971; Hubert, 1992), logistic regression analysis (Robertson and Preisler, 1992), serial-time-mortality model (Preisler and Robertson, 1989;

Thorne *et al.* 1995), Life-table analysis, Kaplan-Meier Product Limit estimator, Time-dose-mortality model (Robertson and Preisler, 1992), Aalen-Nelson estimator, Cox Proportional hazard model and GEE for repeated measures logistic regression.

The methods that have been used in analyzing arthropod dose-response data in biological research are described below.

2.2 Probit Analysis

Probit Analysis is used to analyze data from bioassay experiments. Probit Analysis is the commonly used method of estimating lethal doses or lethal time or lethal concentration (Finney, 1971). When subjects are exposed to several concentrations of an agent one can determine the time taken by a particular dose to kill 50% of the insects (LT_{50}). In probit analysis, different sets of insects are treated with varying amounts of insecticides and the insects are inspected for mortality at a single point in time or analyzed for a given dose or concentration separately. In this design, death of an individual is measured once and all observations on mortality are independent, an important assumption that must be met for probit or logit modeling as prescribed by Robertson and Preisler (1992). The independence assumption is violated if data are repeated measures. In many bioassay experiments (arthropod dose-response studies) investigators record at several points in time the number of subjects that have died giving rise to percentage insects dying. Probit analysis usually models the mortality data as a function of dose and hence it's ineffective when the data are repeatedly taken at several time points (Roberston and Preisler, 1992).

Sahaf and Moharrmipour (2008) used probit analysis proposed by Finney (1971) to estimate the time required for 50% and 95% kill (LT_{50} and LT_{95} respectively) of extracts on cowpea beetle.

Eaton and Kells (2009) used probit analysis to calculate LT_{50} and LT_{90} from the mortality curves (mortality versus time) which displayed sigmoidal curves for a given temperature on mortality of mold mites.

Osbrink *et al.* (2001) used probit analysis described by Finney (1971) to determine LT_{50} and LT_{90} of insecticides subjected to termites

Probit analysis has some limitations in that standard probit analysis techniques are not applicable on serial-time mortality data because observations made on the same group of organisms at different times are correlated and ignoring the correlation aspects will lead to giving false estimates and conclusions (Thorne *et al.*, 1995). Robertson and Preisler (1992) and Thorne *et al.* (1995) stated that alternatives to logit and probit analysis do exists, these are the ones that directly address the problem of correlation of serial-time mortality data. One of the approaches as described by Priesler and Robertson (1989) and Nowierski *et al.* (1996) is to use complementary log-log model (time-dose-mortality model). Other approaches involve using survival analysis (Holbrook *et al.*, 1999) and generalized estimating equations for repeated measures (Stokes *et al.*, 2000).

2.3 Time-dose-mortality model

Time-mortality regression model (Thorne *et al.,* 1995) is used in analyzing correlated serial time-mortality data using complementary log-log, logit, or probit transformations since observations made on the same group of organisms at different times are correlated. This method is also referred to as the Probit Analysis for correlated data at one concentration. The method involves regressing complementary log-log, logit, or probit transformations on untransformed or logarithmic transformations of time. The covariances of the probits are also estimated to account for correlation among observations. Lethal time value is obtained by calculating z (using appropriate formula for complementary log-log, logit or probit transformations) and slope i.e. $LT = (z - \mathrm{int}\,ercept)/slope$. The method is only used for estimating lethal time at one concentration and is applied to serial time mortality designs

For analyzing data with multiple concentrations of pesticides and multiple observations over time (Modeling time-dose-mortality relationships from bioassay tests), the model which describes the relationship between time, dose and the mortality probability as proposed by Preisler and Robertson (1989), Robertson and Preisler (1992), Feng *et al.* (1996), Feng *et al.* (1998), and Nowierski *et al.* (1996) is employed. The model can be used to estimate LT_{50} in serial-time-dose-mortality data where the effect of time on percentage of kill at one concentration or at several concentrations is estimated.

Complementary log-log model (Robertson and Preiesler, 1992) is a more general model for time-concentration-mortality designs. The dose-time-mortality is a method for analyzing correlated serial time-mortality data using complementary log-log or probit or logit transformation for the proportion of insects killed. The conditional mortality probability is estimated, maximum likelihood estimates of the conditional response parameters are obtained. LC_{50} or LD_{50} are then estimated from the given formula. LT_{50} values are obtained by linear interpolation (Nowerski *et al.,* 1996; Feng *et al.*, 1998)

The LT_{50} values are usually computed on the basis of the slope and the estimators of the slope and the parameter for the time effect of the dose. The goodness of fit is usually Pearson's chi-square.

Xu *et al.* (1999) used time-dose mortality modeling technique described by Feng *et al.* (1996), Nowierski *et al.* (1996), and Robertson and Preisler (1992) to analyze the resulting time-dose-mortality data of aphids. A weakness of time-dose-mortality model is that in some studies the estimation may be biased because of insufficient sample size. Sample size usually becomes smaller as the death of the target subject goes near the end of the observation.

Wang *et al.* (2004) conducted a study on time-dose-mortality and used complementary log-log (CLL) time-dose-mortality model to analyze the time-dose trends for the different five concentrations while studying virulence against sweet potato white fly. They noted that the classical ways (probit analysis, logit analysis and weibull function) revealed inefficiency in procedure for the complete data and appear inappropriate for estimating effectiveness of pathogen on the target (trends of time for each dose and trends of dose for each time). This was after analyzing the time-dose-mortality data by separate models of trends of time for each dose or of dose trends for each time. Thomsen and Eilenberg (2000) analyzed the mortality data using time-concentration-mortality regression based on complementary log-log (CLL) model while studying the effect of dustxtrins.

2.4 Survival Analysis

Survival analysis involves methods of analyzing time to event (survival) data. Of interest in survival analysis is to estimate the survival function (time) which is the probability that an individual will survive or die after a given time t. The estimated survival time ($S(t)$) is then used to estimate the median survival time (LT_{50}) i.e. the time when $S(t)=0.5$. A plot of $S(t)$ against time t can also be used to estimate the median survival time through interpolation

Survival analysis is used in arthropod dose-response studies to estimate the median lethal time (median time to death, LT_{50}) while taking into account censoring and time. This is because the interest is in the speed of kill.

Survival methods in arthropod dose response data use time to death data of all individuals in a study to characterize the probability of death as it relates to the level of a stressor and exposure time (Holbrook *et al.*, 1999). If the concentration is not constant, the effect on survivorship can be accounted for by explicitly including a time varying stressor in the hazard function (Borsuk *et al.*, 2002).

Hansen and Lambert, (2003) after using survival analysis in their study pointed out that Probit analysis can used to determine the dose response but they do not give significant results.

Holbrook *et al.* (1999) suggested that survival analysis may be more appropriate to measure resistance since logit and probit analysis were inadequate (Roush and Miller, 1986).

Several survival analysis techniques for estimating survival function (survival time), $S(t)$ include the life table analysis, Kaplan-Meier estimator, Aalen Nelson Estimator and Cox proportional hazard

2.4.1 Life-table Analysis

Life-table analysis uses the grouped time intervals to estimate the survival times (Culter and Edearer, 1958; Gehan, 1969). The survival times are the plotted and the median survival times can be estimated via interpolation

The life-table method requires a fairly large number of observations, so that survival times can be grouped into intervals. The distributions of survival time is divided into a certain number of interval, the number and the proportions of individual that enter the respective interval alive, the number and the proportion of subjects that died in the respective interval and the number of case that were censored in the respective interval. On the basis of proportion and the numbers one could compute proportion or cumulative proportion of surviving subjects, hazard rate and the median survival time (LT_{50}).

Life-table analysis has not been applied in arthropod dose-response studies. Life-table analysis has limitations; it requires very large number of observations so that they can be grouped into intervals. Life-table plots intervals (groupings) and hence median survival times cannot be adequately estimated. It also relies on censoring and has standard errors just like any other estimators.

2.4.2 Kaplan-Meier Estimator

It's a nonparametric way of estimating the survival function while taking into account the censoring time (Kaplan and Meier, 1958). The Kaplan-Meier estimate is a step function which is constant on the interval defined by the death times and changes at every distinct death times but does not change at the censoring times unless a death time happens to be simultaneous with censoring time. Kaplan-Meier estimator has been used in dose-response studies to estimate median lethal times when there is only one grouping variable. Hansen and Lambert (2003) used Kaplan-Meier survival analysis to determine median survival time in evaluating the dose response relationship. They used large censored data drawn from multiple sources and noted that traditional statistical approaches (probit analysis) could not deal adequately with censored data. Kaplan-Meier estimator has some limitations in that it's mainly descriptive, it does not control for covariates, it requires categorical predictors, and it can't accommodate time-dependent variables.

The Log-rank test is a non parametric way of comparing two or more survival curves.

Cabanillas and Jones (2009) used Kaplan-Meier product limit estimate to calculate median time to death (LT_{50}) of instar nymphs. They used Logrank test to test for equality of the estimated survival times

Zurek *et al.* (2002) used Kaplan-Meier to estimate survival probabilities and LT_{50} of the German cockroaches while implementing alternatives to organic insecticide control agents to cockroaches. The logrank test to compare the survival curves for different treatments

2.4.3 Aalen-Nelson Estimator

The Aalen-Nelson estimator is used as a non-parametric estimator of the cumulative hazard function (Aalen, 1978; Nelson, 1972). This method is closely related to Kaplan-Meier, the only difference is the way the survival function, $S(t)$, is calculated. It uses the hazard function to estimate $S(t)$. The Aalen-Nelson estimator has not been applied in dose-response studies.

2.4.4 Cox Proportional Hazard Model

Cox proportional hazards regression model is a method of survival analysis that is used when one wants to accounts for some covariates in the model while estimating the survival function (Cox, 1972). This model does not require knowledge of the underlying distribution. The Cox model assumes that the individual hazard function depends on a common baseline hazard and the values of the covariates. The hazard function or death rate is normally the instantaneous probability of deaths for individuals still alive. The hazard function in this model can take on any form, including that of a step function, but the hazard functions of different individuals are assumed to be proportional and independent of time. In arthropod dose response studies the vector of regression variables or covariate components may depend upon both the time and dose. Cox Proportional Hazard model is a semi-parametric model and makes no assumptions about the form of the $h(t)$.

Moncharmont *et al.* (2003) used Cox proportional hazard model and Kaplan-Meier to estimate the survival function to assess honey bees survival after exposure to insecticides.

Hansen and Lambert (2003) used Cox regression and a nonparametric Kaplan-Meier survival analysis procedure to asses longitudinal data by estimating the survival function while assessing patients recovery rate.

Scholte *et al.* (2003) used pairwise Kaplan-Meier comparison and Cox regression analysis to estimate the survival function for mosquito dose response.

2.4 Logistic Regression

Logistic regression is a class of generalized linear model that employs the logit-binomial link distribution canonical pairing to model the effects of one or more continuous or categorical predictor variables on a binary (dead/alive, presence/absence) response variable.

Generalized linear model allows modeling when response variables have a distribution other than Gaussian (McCullagh and Nelder, 1983). Data for mortality is a set of Bernoulli trials which is a form of binomial distribution. When data follows a binomial distribution, the appropriate GLM is logistic regression where the unknown probability is estimated using a set of regressors.

In dose response data, most of the outcomes are often binary. For example, in a mortality study, the outcome is usually died or survived and the independent variables are either dose or time. The LC_{50} at each time point or LT_{50} for each concentration or for each dose is usually estimated based on the parameter estimates from the logistic regression. $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta^{'}x$ is the general form of logistic regression, where α is the estimate of the intercept and x denotes one or more independent variables.

Logistic regression has been used to characterize the relationship between mortality, dose and time. For example, Preisler and Robertson (1992) used logistic regression to analyze relationship between dose and mortality and between time and mortality to estimate LD_{50} and LT_{50} respectively

Finney, 1971 used logistic regression to estimate LD_{50} for bioassay data

Repeated measures logistic regression can be used to analyze repeated measures binary outcome data so as to account for the correlation among a subject's outcomes. This method uses generalized estimating equations (GEE) as an implementing tool. Generalized Estimating Equations (GEE) was introduced by Liang and Zeger (1986) and Zeger and Liang (1986) to extend generalized linear model to account for repeated measures and correlated responses. The term generalized estimating equations indicates that an estimating equation is obtained by generalizing another estimating equation. In the GEE approach, one considers the mean and the variance of the vector of the response then specifies the 'working correlation matrix'

Stokes *et al.* (2000) estimated lethal time to 50% mortality (LT_{50}) and lethal time to 90% (LT_{90}) mortality for repeated measures using Generalized Estimating Equations.

Bugeme *et al.* (2009) used GEE in logistic regression analysis described by Stokes *et al.* (2000) to estimate lethal time of virulence of fungal isolates on spider mite.

Mburu *et al.* (2009) used repeated measure logistic regression via GEE to estimate lethal time to mortality (LT_{50}) of termites to virulent fungi isolates

Latifian and Rad (2012) used repeated measures logistic regression using generalized estimating equations to estimate the lethal time 50% mortality of fungal species

Limitations of GEE are as follows; Generalized Estimating Equation (GEE) approach is based on a working correlation matrix to obtain efficient estimators of regression parameters in the class of generalized linear models for repeated measures data. Because of uncertainty of the definition of the working correlation matrix, the GEE approach may lead to the loss of efficiency of the regression estimators due to misspecification of the correlation structures (Crowder, 1995). The independent and exchangeable working correlation structure produces the same parameter estimates as the one obtained if GLM is fitted to the data.

From the review Kaplan-Meier survival analysis and repeated measures logistic regression using GEE has not been used comparatively in dose response involving botanical extracts and mosquito larva.

In this study the application of repeated measures logistic regression using GEE and Kaplan-Meier survival analysis techniques on the mosquito larva mortality data from botanical extracts to estimate LT_{50} is explored.

CHAPTER THREE

METHODOLOGY

3.1 Study Overview

In this dissertation, the application of survival analysis and repeated measures logistic regression using GEE methods in analyzing repeated measures arthropod dose response data is described. The survival analysis technique used is the Kaplan-Meier estimator which is a non-parametric method that takes into account censoring while estimating the survival function. The GEE for repeated measure technique was employed in a logistic regression setting since mosquito mortality data was binary and the correlation due to time was to be accounted for. Data on mosquito larva mortality as a result of exposure to the botanical extract was used. The response variable was mosquito larva mortality observed at 12 hrs, 24 hrs, 36 hrs, 48 hrs, 60 hrs and 72 hrs. The interest was in the speed of kill since mortality varies with time. The LT_{50} for the different concentration levels for the different botanical extracts were estimated using the survival analysis and GEE techniques. R statistical software version R 2.14.1 was used in the data analysis.

3.2 Definitions and Parameters of Interest

The factors of interest under study are the different botanical extracts, their concentration, and time.

We are looking at time-dose-mortality relationship. The botanical extracts were the insects' control agents (chemicals). The botanical extracts had different concentration levels. Time was the different durations taken by the insect (mosquito) to dies after being exposed to the different concentration levels for the different botanical extracts. Time is considered as a random variable since mortality varies with time hence one will estimate the lethal time to mortality. Repeated measures rises when successive observations are made on the same groups of insects exposed to one or several concentrations of a stimulus e.g. pesticides or botanical extracts. The same insects subjected to each or several concentrations are followed over time.

Lethal Time (LT) is the period of time required for a proportion of a large group of organisms to respond after being exposed to a specific dose of an injurious agent, such as a drug or radiation or pathogen at a given concentration under a defined set of conditions.

3.3 Description of the Data

The data used in this study were from a laboratory experiment on the effect of botanical extracts on mortality of larvae of anopheles mosquito (*Anopheles gambiae*) as part of malaria control project. Several botanical products were studied but in this study we chose only three botanicals namely B,C,E and control D. The botanicals were studied at four concentration levels: 12.5mg/ml, 50 mg/ml, 250mg/ml and 500 mg/ml. Fifty larvae were dipped in glass beaker containing the specific botanical products at a specific concentration. Each concentration with specific botanical extracts was replicated three times. The response variable was larval mortality observed at 12 hrs, 24 hrs, 36 hrs, 48 hrs, 60 hrs and 72 hrs after exposure. There was no death in control which consisted of water only and hence does not appear in the analysis. The data collected had three factors; botanical extracts, concentrations and time.

3.4 Data Organization

The data set for survival analysis was organized in a column format for each extracts having all the concentration levels and the time. The column variables were extract, concentration, subjects, time and censor. The variable censor was coded zero if the event occurred and the exact time of the event occurrence was not known as shown in Appendix 1.

For the repeated measures logistic regression implemented using GEE a variable ‘ID’ which identifies the time clusters was created to give five clusters corresponding to the time points. The data set was created for each extract at each concentration level as shown in Appendix 2. Other variables in the column were replicate, total number of larva that died. The control was omitted in the data set because there was no mortality for larva exposed to water only (control treatment).

Descriptive analysis was done using box plots in terms of comparing the different botanical extracts and the different concentration levels. Cochrane Q test for three or more matched groups was also used to test for the differences between the three extracts (B,C,E) and differences between the four concentration levels (12.5 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml).

The analytical procedure of estimating LT_{50} using Survival Analysis and repeated measures logistic regression via GEE is shown below.

3.5 Survival Analysis

In survival analysis, the survival function ($S(t)$), and hazard function ($h(t)$), are used in estimating time to event. The statistic that is used in survival analysis is the Median Survival Time (LT$_{50}$) because survival data is skewed (not normally distributed). The survival model used in this study is the Kaplan-Meier estimator which is a non parametric approach for estimating survival function $S(t)$ when there are censored data (Hosmer and Lameshow 1999). Censoring is when there is no complete information for example, the exact time of insect death was not known. Survival data are non-normal (skewed) hence no distributional assumption. In arthropod dose response study when the interest is in the speed of kill, one will then use the $S(t)$ to estimate median survival time or the median time to death (LT$_{50}$) which is the time when $S(t) = 0.5$ (Cabanillas and Jones, 2009). This is the estimated time when 50% of the total exposed mosquito larva population will be dead

Survival function, according to Lee and Wang (2003), is the probability that an individual will survive longer than a specified time t.

$$S(t) = P(T > t) \tag{3.1}$$

Where T is the period of exposure

Survival function, $S(t)$, expresses the proportion of the total still alive or dead at time t. It can also be defined as the probability that an individual fails before time t

$$S(t) = 1 - F(t) \tag{3.2}$$

Where

$$F(t) = \frac{Number\ dead\ at\ time\ t}{Total\ number\ exposed} \tag{3.3}$$

The survival function in this study will be estimated using Kaplan-Meier estimator

3.5.1 Kaplan-Meier Estimator

Consider the dose-response study performed and a sample of size n mosquito larva is collected and for each mosquito larva the time to an event (time to death) is recorded after being exposed to the botanical extract (insecticide agent) of different dosages or concentration levels. The exact time of insects' death is not known or some of the insects may be still alive after the study period (72hrs) hence one doesn't know what happens to them. In these cases the mosquito larva are considered censored at the time where there is no complete information. Let $S(t)$ be the probability that the mosquito larva from a population of size n will survive after exceeding time t. Let the observed times until death of n sample members be $t(0) < t(12) < t(24) < ... < t(72)$

Corresponding to each t_j is n_j which is the number of mosquito larva alive (at risk) just before time t_j (n_j includes those who will die at time $t(j)$), and d_j is the number of failures (deaths) at time $t(j)$.

For any time t, where $t(k) \leq t < t(k+1)$, the Kaplan-Meier or the Product Limit estimate of the survivor function as described by Kalbfleisch and Prentice (1980) is given by :

$$S(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{3.4}$$

The variance of the survival function computed using the delta method is

$$V\left\{S(t)\right\} = S(t)^2 \sum_{t(j)\leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{3.5}$$

This is the Greenwood's formula for estimating variance of the survival function. Hence an approximate 95% confidence interval for the survival time (median) is given by

$$\left[S(t) - 1.96.\sqrt{V\left\{S(t)\right\}}, S(t) + 1.96.\sqrt{V\left\{S(t)\right\}}\right] \tag{3.6}$$

Estimating LT$_{50}$ using Kaplan-Meier Estimator

The median survival time (LT$_{50}$) is the time when $S(t) = 0.5$ and its respective confidence interval is obtained using the above equation (3.6) as used by Zurek *et al.* (2002) and Cabanillas and Jones (2009).

Survival curve (a plot of $S(t)$ against t) was also be used to estimate the median survival time by interpolation (Borsuk *et al.,* 2002; Hansen and Lambert, 2003) or one may use the statistical software to estimate the parameters. This is the time when $S(t) = 0.5$ from the survival plot.

The interest is to estimate the time of kill by the different concentrations hence with the Kaplan-Meier estimator one will estimate the time taken by the different concentration levels to kill 50% of the mosquito larva population (LT_{50}) which will be the median survival time (Zurel *et al.*, 2002)

The data set was sorted by extract and the analysis stated with creating survival object (appendix 1). The interest was the survival time after mosquito larvas have been exposed to different concentration levels and the status of the insects whether dead or alive. The survival times were calculated step wise together with their confidence intervals as shown in equation (3.4) and equation (3.6). The estimated survival times were then plotted against t .

3.5.2 Comparing Survival Curves for the Different Concentrations

The non-parametric test for comparison of two or more survival distributions is the log-rank test (Cabanillas and Jones, 2009). The log-rank test to test the hypothesis for two survival curves for the concentration levels is.

$$H_o : S_1(t) = S_2(t)$$

$$H_1 : S_1(t) \neq S_2(t)$$

If there are K ≥ 2 concentration levels of interest, similar tests can be constructed by generalizing notation to use matrix algebra.

The test hypothesis is then stated as

$$H_0 : S_1(t) = S_2(t) = ... = S_k(t)$$

H_1 : There is at least one pair of survival curve for concentrations k and j such that $S_j(t) \neq S_k(t)$

The Log-Rank test statistics according to Kleinbaum (1995), and Klein and Moeschberger (1997) is:

$$\log rank = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)} \tag{3.7}$$

$$Var(O_1 - E_1) = \frac{n_1 \times n_2 \times d \times (n - d)}{n^2 \times (d - 1)} \tag{3.8}$$

The Log-Rank statistic approaches to chi-square distribution with 1 degree of freedom. The log-rank test extends to G>2 concentration levels for the different extracts but requires the covariances of $(O_i - E_i)$ and the log-rank statistics will be a chi-square with G-1 degrees of freedom

3.6 Generalized Linear Models

Generalized linear models (GLM) are a generalization of standard linear regression so that the response variables may have a distribution other than the Gaussian. GLMs have two assumptions about the distribution of the responses. First, given the X_i, responses Y_i are assumed to be independent of one another. Secondly, the distribution of the Y_i must belong the exponential family. GLM can be characterized by the three parts; member of the exponential family being used, the link function and the design vector.

3.6.1 Logistic Regression

Logistic (logit) regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables (McCullagh and Nelder, 1983; McCullagh and Nelder, 1989). In arthropod dose-response mortality data is a set of Bernoulli trials which is a special case of Binomial distribution. The values of response Y_i (mortality status) are 1 if there is a success and 0 otherwise. In this case a 'success' would be if the mosquito larva has died within a given time period after being exposed to a particular concentration level of a given botanical extract. There exists some probably π that an observation would be a success. When the data follows Bernoulli or Binomial distribution the an appropriate GLM is the logistic regression

The scale parameter is set to one and the link function to be equal to $\log\left(\frac{\mu}{1-\mu}\right)$

In the logistic regression unknown probability π is estimated using a set of regressor variables in this case the doses or time

For outcome variable Y (mortality status), and a set of n predictor variables (dose or time), X_1. Consider a binary response variable with a logistic transformation or logit function, then logistic regression is

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \varepsilon \qquad (3.9)$$

Where β_0 is the intercept, β_1 is the regression coefficient for each corresponding predictor variable, X_1 (dose or time), and ε is the error of the prediction

The logit of a probability is simply the log of the odds of the response taking the value one. The above Equation can be rewritten as

$$\pi(x_1) = \frac{\exp(\beta_0 + \beta_1 X_1)}{1 + \exp(\beta_0 + \beta_1 X_1)} \tag{3.10}$$

The logistic regression model indirectly models the response variable based on probabilities associated with the values of Y. The logit function can take any real value, but the associated probability always lies in the required $[0,1]$ interval. The parameters of the logistic regression model (the vector of regression coefficients β) are estimated by maximum likelihood.

The logistic regression is used in arthropod dose response studies to model proportion of mortality as a function of time or concentration

Estimating LT$_{50}$ Using Logistic Regression

In this case time element need to be considered since mortality is a function of both time and dose. In addition mortality varies with time and the speed ok kill is of importance and therefore a more meaningful approach was to estimate the time it takes for 50% of the test animals to be killed as a function of dose or concentration (LT$_{50}$) (Robertson and Preisler, 1992).

Consider equation (3.9) having time as an explanatory variable and the response variable as the proportion of mortality

$$\log it(\pi(time)) = \beta_0 + \beta_1(time) \tag{3.11}$$

The LT$_{50}$ is, by definition, the time at which $\pi(time)$ equals 0.5. (Preisler and Robertson, 1989; Finney, 1971) by substituting $\pi(time)$ with 0.5 in the above equation one gets

$$LT_{50} = -\frac{\beta_0}{\beta_1} \tag{3.12}$$

Based on the asymptotic approximation, the variance of the LT$_{50}$ was computed using the delta method (Bugeme *et al.,* 2009).

$$Var(LT_{50}) = \frac{1}{\beta^2}Var(\beta_0) + \frac{\beta_0{}^2}{\beta^4}Var(\beta) + 2.\frac{1}{\beta^2}.\frac{\beta_0{}^2}{\beta^4}.Cov(\beta_0, \beta) \tag{3.13}$$

and hence an approximate 95% confidence interval for the LT$_{50}$ is given by

$$\left[LT_{50} - 1.96.\sqrt{Var(LT_{50})}, LT_{50} + 1.96.\sqrt{Var(LT_{50})}\right] \tag{3.14}$$

To account for correlation effect due to time (repeated measures) we use Generalized Estimating Equations to estimate the parameters β_0 and β_1

When there are repeated measures on the same individual, the observations on the same individual are highly correlated which must be taken into account in the statistical procedures. This permits the calculation of robust estimates for the standard error of the regression coefficients, accounting for the correlation of the outcomes.

Logistic regression which is GLM can be used to estimate the LT_{50} but since it cannot account for the correlation the GLM was extended using GEE to estimate the parameters (β_0 and β_1) by specifying the correlation structure (Zurr *et al,*. 2009; Zeger and Liang, 1986; Liang and Zeger, 1986). The parameters are estimated using different correlation structures for the fitted models and QIC was used to choose which correlation structure is giving the least QIC using GEE and then fit it to the logistic regression model. LT_{50} was estimated using repeated measures logistic regression which uses GEE

3.7 Generalized Estimating Equations

GLM logistic regression works under the assumptions that the data are independent. However often the data are cluster e.g. multiple responses from same individuals over time. The correlated data must be taken into account when estimating the regression coefficients and standard errors during the analysis. Extension of GLM to handle correlated data an example is GEE (Liang and Zeger, 1986)

Elements of the marginal models introduced in the study design section are mathematically defined using the following notations

Let Y_{ij}, $i=1,...,n$, $j=1,...n_i$ denote the mortality status of mosquito larva j after exposure to at a given concentration i for a given botanical extract. ($Y_{ij}=1$, dead and $Y_{ij}=0$, alive)

Let X_{ij} be the time taken by mosquito larva j to die after being exposed to concentration i

Y_{ij} is assumed to follow a Bernoulli distribution when the probability that mosquito larva is dead be denoted by μ_{ij}, that is $\mu_{ij}=P(Y_{ij})$ and this is also equal to the expected death $E(Y_{ij})=\mu_{ij}$

The marginal logistic regression model for the data is

$$\log it(\mu_{ij})=\beta_0+\beta X_{ij}$$

$$Var(Y_{ij})=\mu_{ij}(1-\mu_{ij})$$

$$Corr(Y_{ij},Y_{ik})=\alpha_{jk}$$

In this model the number of observations per cluster (time intervals) was small in a balanced and complete design and hence unstructured correlation matrix. The observations are correlated with no assumptions of the structure.

Let $E[Y_{ij}\,|\,X_{ij}]=\mu_{ij}$ be the conditional expectation of the outcome variable given the covariates

The binary response is the mortality status of 50 mosquito larva at time 12 hrs, 24 hrs, 36 hrs, 48 hrs, 60 hrs and 72 hours. The mean response is modeled as a logistic regression model by using the explanatory variables time for each concentration level for a given botanical extract. For any two pairs of binary responses for any individual mosquitoes there was no assumption about the correlation coefficients thus implying an unstructured correlation structure.

To use GEE in estimating, there are three-part specification; the conditional expectation of each response, the conditional variance of each Y_{ij} given the covariates and the covariance (correlation) matrix (Zuur *et al*,. 2009; Liang and Zeger, 1986). Let the marginal regression model to be:

$$g(E[Y_{ij} / X_{ij}]) = X'_{ij}\beta \qquad (3.15)$$

Where X_{ij} is a $p \times 1$ vector of covariates, β consists of the p regression parameters of interest (time) $g(.)$ is the link function, and Y_{ij} denotes the j^{th} outcome $(for\ j = 1,...,J)$ for the i^{th} mosquito larva/ subject$(for\, i = 1,...,N)$. The common choices for the link function are;

a. $g(a) = a$ [identity link] ,

b. $g(a) = \log(a)$ [for count data] ,

c. $g(a) = \log\left(\frac{a}{1-a}\right)$[logit link for binary data]

For this study the link function chosen was the logit link for binary data

In marginal models it is useful to specify the distribution of the outcome variable so that the variance can be calculated as a function of the mean. The marginal models may still lead to consistent variance estimates even when there is some misspecification of the variance.

The variance-covariance matrix, part of the model used in the estimating equation, is:

$$V_i = \phi A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \tag{3.16}$$

Where ϕ is a glm dispersion parameter - allows for over dispersion, A_i is a diagonal matrix of variance functions $(Var(Y_{ij}))$ i.e. $Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$, and $R_i(\alpha)$ is the correlation matrix of Y_i.

The over-dispersion parameter is estimated using the formula

$$\phi = \frac{1}{N-P}\sum_{i}^{k}\sum_{j}^{n_i}\frac{y_{ij}-\mu_{ij}}{\sqrt{Var(\mu_{ij})}} \tag{3.17}$$

Where $N = \sum_{i=1}^{k} n_i$ is the number of measurements and P is the number of regression parameters. The square root of the over dispersion parameter is called the scale parameter.

Probability of success is the expected value of Y_{ij} given explanatory data X

$$P(Y_j = 1 | X_j) = \pi_j(\beta_j) = E(Y_{ij} | X_{ij})$$

The GEE equation for vector β or the regression model (score) is given by

$$g(\beta)=\sum_{i=1}^{k} D_i^T V_i^{-1}(Y_i-\mu_i)=0 \tag{3.18}$$

Where D_i is the matrix of derivatives $\frac{\partial\mu_i}{\partial\beta_j}$, V_i is the "working" covariance matrix of Y_i.

Given a mean model, μ_{ij}, and variance structure, V_i , ("working" covariance matrix of Y_i), the parameter estimates will be given by solving $g(\beta)=0$ (Liang and Zeger, 1986; Zurr *et al.*, 2009) and are typically obtained via the Newton-Raphson algorithm, otherwise we need iterations to solve $g(\beta)=0$.

3.7.1 Newton-iteration

Fitting algorithm becomes an instance of the Newton algorithm to solving a system of equations

1. Compute an initial estimate of β from a GLM (i.e.by assuming independence)
2. Compute an estimate of $R(\alpha)$ of the working correlation on the basis of the current Pearson residuals (standardized residuals), the current estimate of β and the assumed structure of $R(\alpha)$
3. Compute an estimate of covariance as

$$V_i=\sum\nolimits_i=\phi A_i^{\frac{1}{2}}R(\alpha)A_i^{\frac{1}{2}} \tag{3.19}$$

4. Compute an updated estimate of β based on the Newton-step

$$\beta_{r+1}=\beta_r+\left[\sum_i^k\frac{\partial\mu_i}{\partial\beta}{}'V_i^{-1}\frac{\partial\mu_i}{\partial\beta'}\right]^{-1}\left[\sum_i^k\frac{\partial\mu_i}{\partial\beta}{}'V_i^{-1}(Y_i-\mu_i(\beta))\right] \tag{3.20}$$

5. Repeat/Iterate 2-4 until convergence

The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process by using the current value of the parameter vector β to compute appropriate functions of the Pearson's residuals

$$e_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{V(\mu_{ij})}} \tag{3.21}$$

3.7.2 Working Correlation Matrix

Choices for the correlation structure within GEE (Zuur *et al.*, 2009; Liang and Zeger, 1986) include the following:

1. Independent: The observations in an individual are uncorrelated with every other observation in that individual.
2. Exchangeable: It's when all measurements on the same units are equally correlated
3. Autoregressive, AR(1): The observations taken closer in time are more correlated than the observations taken far apart in the same individual. At times the correlation depends on time or distance between measurements.
4. Unstructured correlation: This is when there are no assumptions made about the correlation coefficients between any two pairs of observations.
5. M-dependent: Pairs of data elements separated by consecutive repeated measurements have a common correlation coefficient.

6. User fixed: Correlation coefficients are fixed by the user rather than being estimated from the data and the values are fixed prior to the analysis.

A useful feature of the GEE model is that the estimators are robust to departures from the true correlation patterns. A loss in estimator efficiency can occur but this loss decreases as the sample becomes larger

There are a number of issues guiding the choice of correlation structures. If the number of observations per cluster is small in a balanced and complete design, then an unstructured matrix is recommended. For data sets with miss-timed measurements, then one considers a model where the correlation is a function of the time between observations (i.e., M-dependent or Auto-regressive). For data sets with clustered observations, there may be no logical ordering for observations within a cluster and an exchangeable structure may be most appropriate. For both the independence working structure and the fixed working structure, no estimation of α is performed (for the fixed structure, the user must specify a $t \times t$ matrix mat). The use of the exchangeable (also referred to as compound symmetry) working correlation matrix with measured data and identity link function is equivalent to a random effects model with a random intercept per cluster.

GEE works best if the numbers of observations per subject is small and the number of the subjects is large, and also if in the longitudinal studies the measurements are taken at the same time for all subjects. The exchangeable correlation matrix was used since the assumption of that correlation is different for each pair was made.

3.7.3 Covariances of β

In GEE both model-based and empirical covariances are produced.

3.7.3.1 Model-based Estimate

The model-based estimator of the covariance matrix of β is given by

$$Cov(\beta)_m = \sum\nolimits_m(\beta) = I_0^{-1} \tag{3.22}$$

$$I_0 = \sum_i^k \frac{\partial\mu_i}{\partial\beta}' V_i^{-1} \frac{\partial\mu_i}{\partial\beta} \tag{3.23}$$

I_0 can also be written as

$$I_0 = D^T V^{-1} D \tag{3.24}$$

In this case $Cov(\beta)_m$ consistently estimates $Cov(\beta)$ if the mean model and the working correlation are correct.

3.7.3.1 Empirical-sandwich Estimate

The empirical or robust estimator of the covariance matrix of β is given by

$$Cov(\beta)e = \sum\nolimits_e(\beta) = I_0^{-1} I_1 I_0^{-1} \tag{3.25}$$

$$I_1 = \sum_{i=1}^{k} \frac{\partial \mu_i}{\partial \beta} V_i^{-1} Cov(y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \tag{3.26}$$

I_1 can also be written as

$$I_1 = D^T V^{-1} (y-\mu)(y-\mu)^T V^{-1} D \tag{3.27}$$

Here $Cov(\beta)_e$ is a consistent estimator of $Cov(\beta)$ even if the working correlation is misspecified, i.e. $C0v(y_i) \neq \sum_i$. In computing $\sum_e$, β and ϕ are replaced by estimates, and $Cov(y_i)$ is replaced by the estimate $(y_i - \mu(\beta))(y_i - \mu(\beta))'$

The robust or model-base standard errors are estimated in the GEE model.

The data set was sorted by the extract and for each dose level (appendix 2). The logistic regression model was fit with the proportion of insect mortality as the response variable and the different extracts and time being the explanatory variables for dose level. GEE models were fitted using the unstructured working correlation matrix was used and since the data was binary it belongs to the binomial family and the link function used was the logit. The fitted models estimates the intercept (β_0) and time (β). The variance covariance (empirical) matrix was extracted from the regression with the parameters labeled as variance of beta zero $Var(\beta_0)$, covariance of beta zero and beta $Cov(\beta_0, \beta)$, and variance of beta $Var(\beta)$. These estimated parameters are used to estimate LT$_{50}$, standard deviation of LT$_{50}$ and the lower and the upper limits for LT$_{50}$ using the given formulas in equations 3.12

and equation 3.14 to estimate the LT_{50} and the confidence interval for the estimated LT_{50} respectively (Thomsen and Eilenberg, 2000)

3.7.4 Choosing the Correlation Structure in GEE

Correlation structure was chosen to reflect the manner in which the data was collected. Other ways of choosing the best model in GEE include the likelihood ratio test which is a chi-square value, the deviance ratio and the information criteria test (Quasi-likelihood information criterion). Quasi-likelihood Information Criterion (QIC) which is an extension of the Akaike Information Criterion (AIC) to apply to models fit by GEE was used to find an acceptable working correlation structure (Hardin and Hilbe, 2003).

$$QIC = -2Q(\mu; I) + 2trace(A_I^{-1}V_R)$$

I is the independent covariance structure used to calculate the quasi-likelihood. $\mu = g^{-1}(X\beta)$ and $g^{-1}()$ is the inverse link function for the the model (logit). A_I^{-1} is the variance matrix under the assumption of independence model. V_R is the robust variance estimator obtained from a general working covariance structure R. The model with the smaller statistic is preferred

3.8 Comparing the Two Methods (GEE and Survival)

Kaplan-Meier estimator is a survival analysis technique for estimating survival time while taking into account censoring. Kaplan-Meier is a step function and it’s mainly descriptive

Repeated measures logistic regression is implemented using GEE which is an extension of GLM to allow parameters to be estimated while taking into account the correlation in data. Repeated measures logistic regression involves regressing parameters of interest

The two approaches estimated the LT_{50} together with their respective confidence intervals. The performances of the two methods are compared using the width of their confidence intervals for the estimated LT_{50} (van Zaane *et al.,* 2012). Confidence intervals ideally are usually estimated using the estimated standard errors. The method that had the consistent lesser or narrower confidence interval was the best method since narrow confidence intervals reflects smaller standard errors.

Narrow confidence interval implies that the estimates are precise and hence a better method (van Zaane *et al.,* 2012; Burton *et al.,* 2006)

CHAPTER FOUR

RESULTS

4.1 Descriptive Results

Box plots shown in figure 4.1 below suggest that the three botanical extracts are significantly different from each other in terms of insect mortality across all time points.

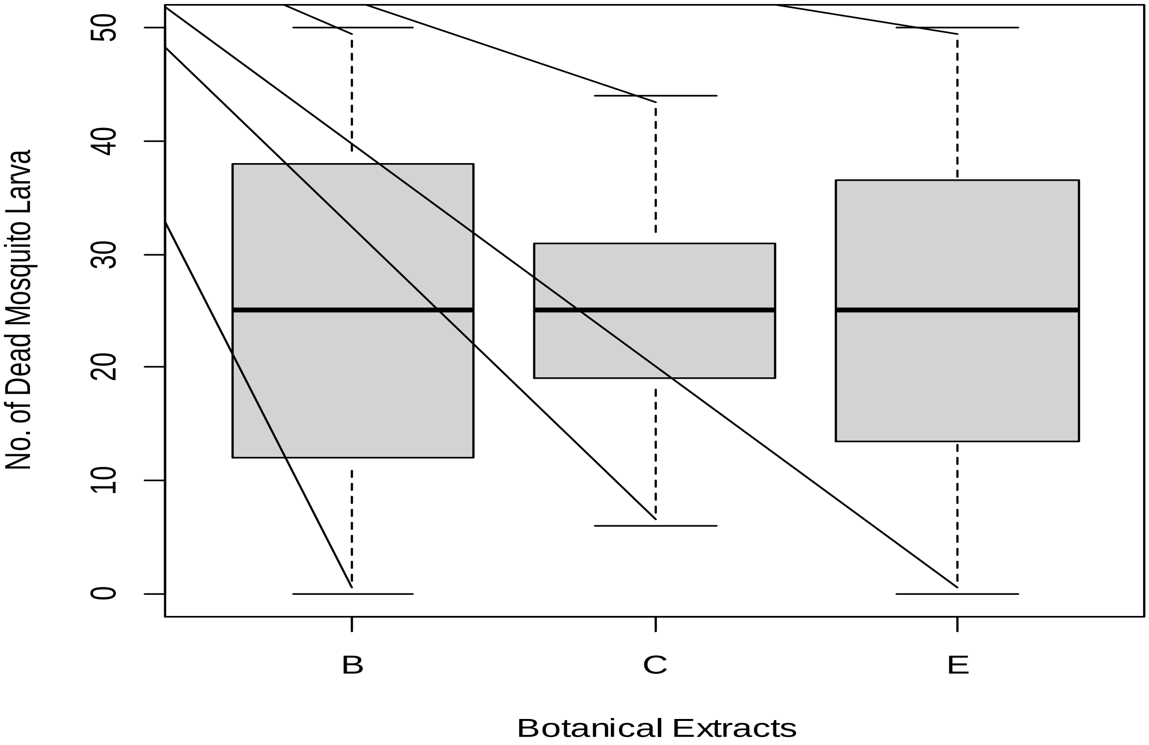

Figure 4.1 Box plot for botanical extracts B, C and E

Cochrane Q test for three or more matched suggests that the three extracts are significantly different from each other (Q test statistics was 8.9385 with a p-value = 0.03647)

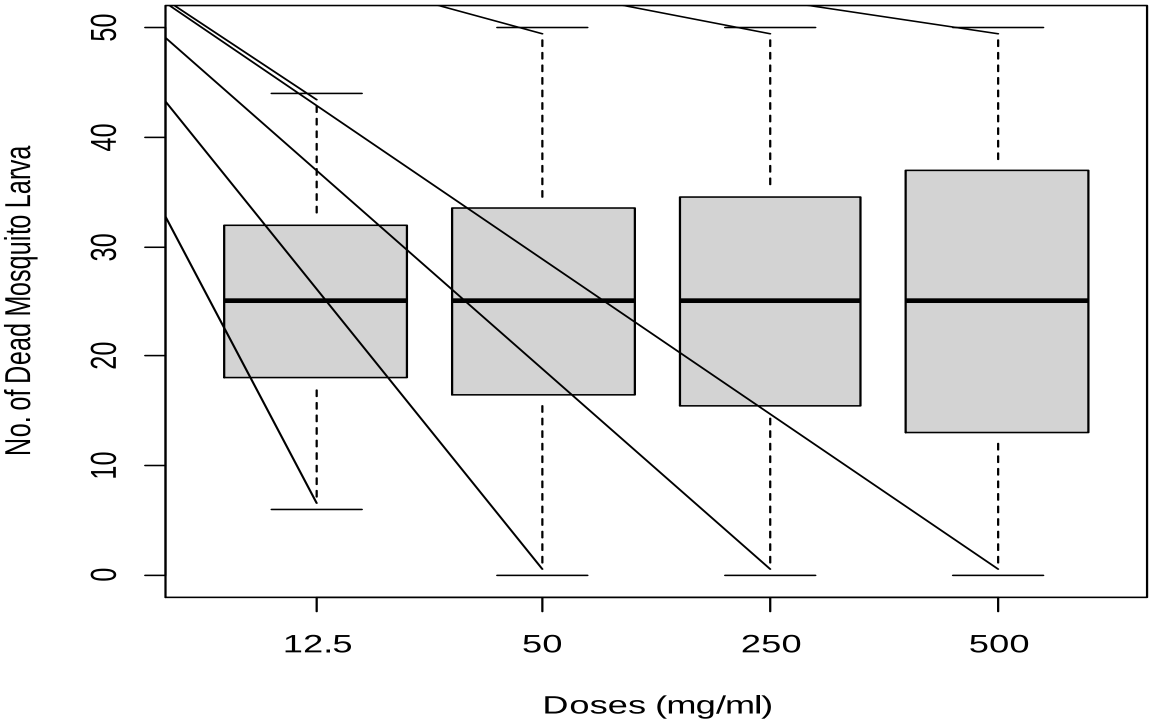

Figure 4.2 Box plot for concentration 12.5 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml

From the graphical representation in figure 4.2 above suggests that concentration 50 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml are different from each other in terms of insect mortality across all the time intervals. The Cochrane Q test for four different concentration levels was 5.4385 (p-value = 0.00292) hence the four concentration levels are significantly different from each other.

4.2 LT_{50} Estimates from Repeated Measures Logistic Regression using GEE

Insect mortality may vary with time (and other factors) and therefore a more meaningful approach was to estimate the time it takes for 50% of the test animals to be killed as a function of dose (LT_{50}). The unstructured correlation structure was used to reflect the manner in which the data was collected. The estimated LT_{50} together with their confidence intervals for the different concentration levels for extracts B, C and E are summarized in table 4.3 below

Table 4.1 LT_{50} Estimates from repeated measures logistic using GEE

Extract	Concentration (mg/ml)	LT_{50}(hrs)	95% CI for LT_{50}
B	12.5	52.1	50.5 – 53.7
B	50	23.0	16.8 – 29.3
B	250	12.3	7.4 – 17.2
B	500	10.3	1.51 – 19.1
C	12.5	70.7	69.3 – 72.0
C	50	43.4	42.0 – 44.7
C	250	21.5	19.1 – 23.9
C	500	7.2	4.3 – 10.1
E	12.5	55..0	52.6 – 57.3
E	50	16.6	11.57 – 21.7
E	250	12.2	6.16 – 18.2
E	500	10.3	9.5 – 11.3

The lethal time (LT_{50}) ranged between 10.3 hrs to 52.1 hrs for extract B; 7.2 hrs to 70.7 hrs for extract C and between 10.3 hrs to 55 hrs for extract E. The LT_{50} values for the different concentration levels ranged between 52.1 hrs to 70.7 hrs for concentration 12.5 mg/ml; 16.6 hrs to 43.4 hrs for concentration 50 mg/ml; 12.2 hrs to 21.5 hrs for concentration 250 mg/ml; and between 7.2 hrs to 10.3 hrs for concentration 500 mg/ml. From table 4.3 above concentration 500mg/ml was the most potent chemical since it had the lower LT_{50} value across all the extracts.

4.3 Median Survival Time (LT_{50}) from Kaplan-Meier Estimator

The Kaplan-Meier median survival time (time to death) estimates for each concentration in each of the botanical extract are presented in table 4.4 below. The same is presented graphically by figures 4.4, 4.5 and 4.6. Formal test for the differences in the survival curves using logrank has also been presented test was done. The median survival time was taken as the time taken to kill half of the population

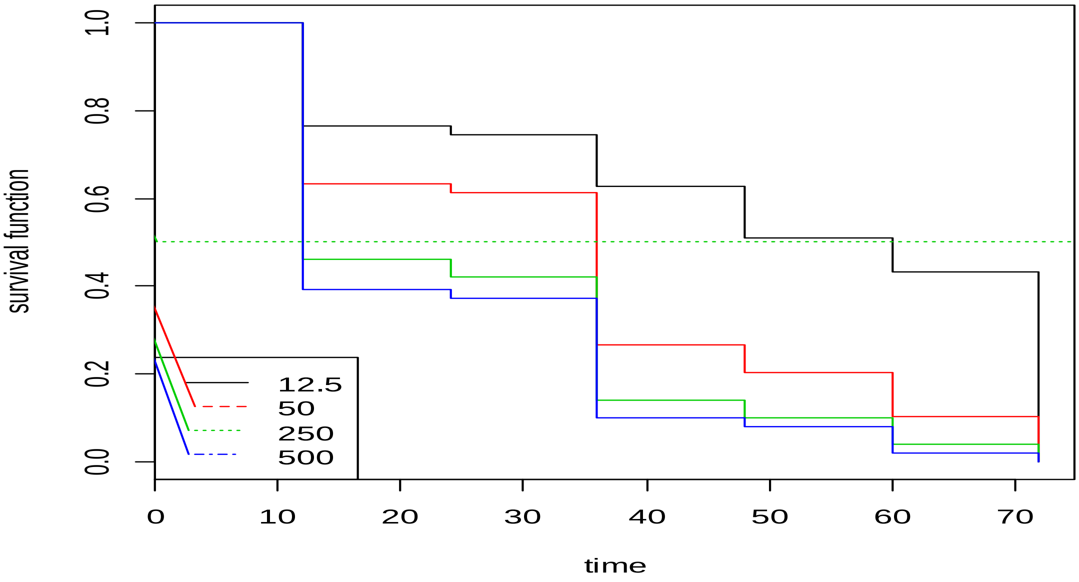

Figure 4.3 Kaplan-Meier survival plots for extracts B

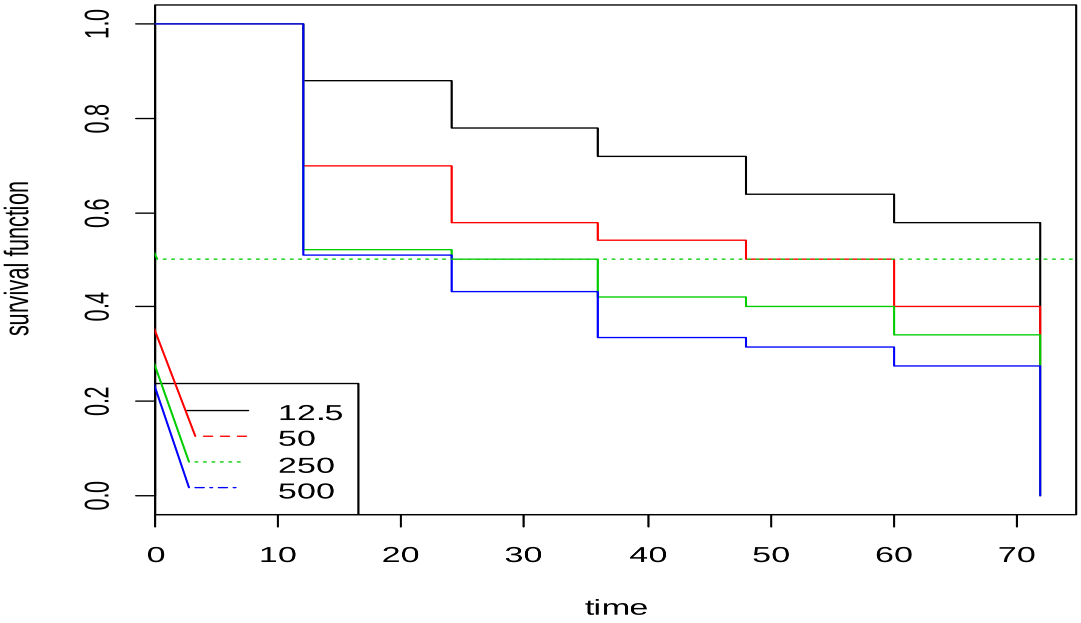

Figure 4.4 Kaplan-Meier survival plots for extracts C

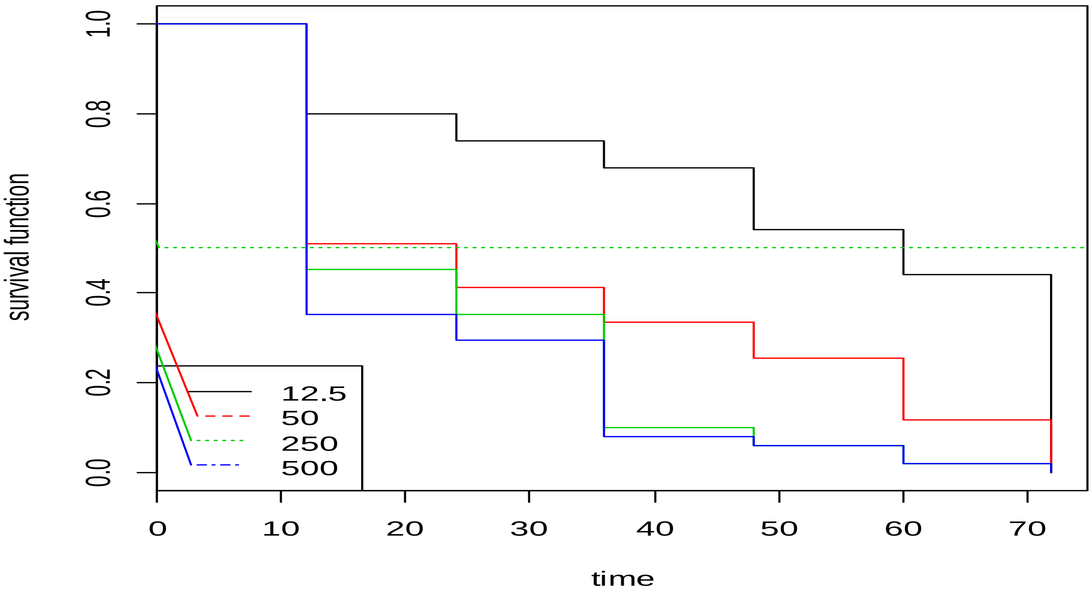

Figure 4.5 Kaplan-Meier survival plots for extract E

The above figures (4.3 to 4.5) show the Kaplan-Meier survival plots for the different concentrations in extracts B, C and E. Across all the extracts, as suggested by the figures concentration 500mg/ml was the most potent chemical since it falls faster in the graph meaning that it's taking shorter time to kill than other doses. Concentration 12.5 mg/ml was less potent since it's taking longer time to kill half of the mosquito population as shown in above in figures (4.3 to 4.5). The log-rank test was used to find out how the different survival curves were different from each other. From the three different survival plots for the four concentrations for three different extracts, the log-rank test for the survival curves showed that the survival curves were significantly different from each other since p-value <

0.0001 for each of the extracts. Table 4.4 below gives the summary table for the median survival time estimates together with their confidence intervals

Table 4.2 Median Survival Time (LT_{50}) from Kaplan-Meier Estimator

Extract	Concentration (mg/ml)	Median (LT_{50}) (hrs)	95% CI for median
B	12.5	60	(48-inf)
B	50	36	(12-36)
B	250	12	(12-36)
B	500	12	(12-36)
C	12.5	72	(60-inf)
C	50	54	(24-inf)
C	250	30	(12-inf)
C	500	24	(12-36)
E	12.5	60	(48-inf)
E	50	24	(12-36)
E	250	12	(12-36)
E	500	12	(12-24)

From table 4.4 above, concentration 500mg/ml was the most potent chemical since it takes the shortest time to kill 50% of the population in all the extracts (12hrs for extract B, 24 hrs for extract C and 12 hrs for extract E).

4.4 Comparing the Two Methods

The table 4.5 below shows the comparison of the confidence intervals of the LT_{50} estimates from the two techniques

Table 4.3 Comparing GEE and survival estimates

		Survival (Kaplan-Meier)		GEE	
Extract	**Concentration (mg/ml)**	**Estimate (hrs)**	**95% CI**	**Estimate (hrs)**	**95% CI**
B	12.5	60	(48-inf)	52.1	50.5 – 53.7
B	50	36	(12-36)	23.0	16.8 – 29.3
B	250	12	(12-36)	12.3	7.4 – 17.2
B	500	12	(12-36)	10.3	1.51 – 19.1
C	12.5	72	(60-inf)	70.7	69.3 – 72.0
C	50	54	(24-inf)	43.4	42.0 – 44.7
C	250	30	(12-inf)	21.5	19.1 – 23.9
C	500	24	(12-36)	7.2	4.3 – 10.1
E	12.5	60	(48-inf)	55..0	52.6 – 57.3
E	50	24	(12-36)	16.6	11.57 – 21.7
E	250	12	(12-36)	12.2	6.16 – 18.2
E	500	12	(12-24)	10.3	9.5 – 11.3

Table 4.5 above suggests that the confidence intervals of the LT_{50} estimates from Kaplan-Meier Survival analysis are unbounded and were consistently larger/ wider compared to the confidence intervals of the LT_{50} estimates from the repeated measures logistic regression

using GEE. This therefore means that the standard errors for GEE were better than those of the Survival as shown by the consistent narrow ranges of their confidence intervals hence the LT_{50} estimates from GEE were precise and therefore the method was better.

CHAPTER FIVE

DISCUSSION

5.1 Discussion

This study discusses Kaplan-Meier survival analysis and repeated measures logistic regression using GEE approaches for estimating LT_{50} in repeated measures dose response in arthropod studies.

Estimating LT_{50} is of importance when the interest is in the speed of kill (mortality vary with time) is of interest. It's also important because observations made on the same group of organisms at different times are correlated and hence standard probit analysis will not be applicable. Therefore this dissertation study contributes significantly and uniquely to methods of analyzing correlated dose response data from arthropod studies as it performs the comparisons of Kaplan-Meier survival analysis and repeated measure logistic regression using GEE when the speed of kill is of importance.

The analysis showed that different botanical extracts were significantly different from each other and different concentration levels were also significantly different from each other. The LT_{50} estimated corresponds to specific extracts and their different concentration levels. From both methods concentration 500mg/ml was the most potent chemical since it took shorter time to kill half of the insects' population. There is a strong body of knowledge regarding the lethal effects of concentrations on mortality in that the higher concentration levels are more effective in regards to mortality and this might have reflected in the estimated LT_{50} for the different concentrations. Further research should be done to ascertain the claim of the estimated LT_{50} to rule out if the estimates may have been affected by some other factor.

The results of the study were compared with results from the same methods but applied in a different setting to show that the methods were versatile for analyzing repeated measures dose response data from arthropod studies. The LT_{50} and the confidence intervals of the estimates in this study were similar to those given by Bugeme *et al.* (2009) and Mburu *et al.* (2009) who also used repeated measure logistic regression via GEE. The results from Kaplan-Meier estimator method in this study were also similar to those given Borsuk *et al.* (2002), Zurek *et al.* (2002) and Cabanillas and Jones (2009) who also used Kaplan-Meier estimator and the log-rank tests in their studies. The graphical representation of the estimates of the median survival time was also the same as those shown by Borsuk *et al.* (2002) and Hansen and Lambert (2003). In all the three botanical extracts the log-rank test was used to test for the differences in the different survival curves as used by Hansen and Lambert (2003). From the study results all the different survival curves were significantly different from each other as shown by Zurek *et al.* (2002). The same results also showed that higher concentrations are more effective than lower concentrations (Thomsen and Eilenberg, 2000; Nowierski *et al.*, 1995).

The two methods presented in the study are among the best methods for estimating time taken to kill half (50%) of the mosquito larva population in the test as a function of dose

(LT_{50}). This was because time was treated as a random variable since mortality varies with time in addition to taking into account the correlation aspect of the data.

The statistical procedure in this dissertation involves analyzing repeated dose response mortality data from arthropod studies for the three botanical extracts at four different concentration levels. The parameters of interest in our study were the same as those required in time-dose-mortality designs. These are time, concentration levels and proportion of dead insects as those used by Thorne *et al.* (2005) and Robertson and Preisler (1992), Bugeme *et al.* (2002) and Borsuk *et al*. (2002).

Kaplan-Meier survival analysis and repeated measures logistic regression using GEE are currently applied in arthropod dose response studies. This dissertation study analysis is unique in that there is a step-wise procedure on how to analyze repeated measures dose response data from arthropod studies using R free statistical software especially in estimating LT_{50} of three botanical extracts for control of mosquito larva.

Just as other approaches of estimating lethal doses or lethal time, this research approach started by first evaluating if the extracts were significantly different from each other and if the concentrations and the botanical extracts had significant effect on insect mortality (Finney, 1971; Preisler and Robertson, 1989). Since the interest is in the speed of kill and that mortality varies with time estimating LT_{50} was the best. In addition, since the mosquito mortality data is correlated, then there was need to estimate LT_{50} using GEE and survival analysis to ensure that the correlation is taken into account while estimating the effect of time on the percentage of kill at one or different concentration.

In this dissertation two approaches have been used to estimate LT_{50} in repeated measures dose response data from arthropod studies. The first approach (repeated measures logistic regression using GEE) is applied which assumed that the response variable is binary (mosquito larva is either dead or alive). There were several observations made from the mosquito population which were also being monitored over time giving rise to correlated measurements. This then led to using repeated measures logistic regression via GEE to account for correlation (Liang and Zeger, 1986). The interest was in the speed of kill since mortality varies with time and hence the need to estimate lethal time (LT). GEE was used to get the parameter estimates (beta note and beta) which are then used in estimating LT_{50}. Delta method was used to estimate variances of the estimated LT_{50}. The estimated variances were then used to construct the confidence intervals of the LT_{50}. The second approach (Kaplan-Meier survival analysis) assumed the use of time to death of mosquito larva. The mosquito larva responses were monitored within a given time period and their survival probability were estimated. Since survival data was skewed the non-parametric approach for estimating the median survival time while taking into account censoring was used (Hosmer and Lameshow, 1999). The interest is in the speed of kill hence the need to estimate the median time to death (LT_{50}). In both cases the datasets were prepared differently for the different analysis. The application of these two methods in analyzing dose-response data identified a common result (LT_{50}) which is one of the outcomes while analyzing dose response data.

The two methods estimated the LT_{50} and their confidence intervals thus showing similarities among the methods. The LT_{50} estimates of Kaplan-Meier and repeated measures GEE were compared in terms of their confidence intervals (van Zaane *et al.*, 2012; Burton *et al.*, 2006). Repeated measures logistic regression using GEE gave consistent smaller/ lesser confidence interval compared to those form the Kaplan-Meier survival estimate. The confidence intervals of the Kaplan-Meier LT_{50} estimates were unbounded confidence intervals. This therefore means that the standard errors for LT_{50} estimates from repeated measures logistic using GEE from this dissertation study were smaller and hence the better method as shown by van Zaane *et al.*, (2012) but further research should be done to ascertain the claim.

The implications of these findings are that there is need to improve on the way repeated measures data from arthropod dose-response studies is being analyzed by adopting or using the methods (GEE or Kaplan-Meier) in the analysis. In addition, the kind of the information given will be important in guiding how to analyze and interpret results from repeated measures data.

Since LT_{50} is the problem of estimating the duration of time of kill by different concentration or doses, then survival analysis and GEE techniques can be used effectively to estimate the effect of time on the percentage of kill at one or different concentrations in arthropod dose response studies. The standard method which usually estimates LD_{50} or LC_{50} for a given time point or LT_{50} for a given concentration or dose level ignores the correlation aspect of the data brought about by time. Since mortality varies with time it will also be of importance to estimate LT_{50}.

Approaches to correlated dose response data in arthropod studies while taking interest in the speed of kill represent the strong part of this dissertation.

One criticism of these approaches is that the exact times of kill is not known since time is used cumulatively to estimate if the mosquito larva has been killed at a particular time point. To address this concern effective data collection methods and use of existing methods of estimating LT_{50} should be used in a complementary fashion.

In this study one survival analysis method was used (Kaplan-Meier) to estimate LT_{50}. In addition, unstructured correlation matrix was the only one used in repeated measures logistic regression using GEE. Wider comparisons should be considered to make the research more representative.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATION

6.1 Conclusions

The results and the discussions of this dissertation leads to one concluding that the different botanical extracts used in the study were significantly different from each other. The Kaplan-Meier survival estimator and repeated measures logistic regression using GEE estimated the LT_{50} and confidence intervals for different concentration levels of the different botanical extracts. The estimated LT_{50} gave the estimated time taken by a particular concentration for the different botanical extract to kill 50% of the mosquito larva. Across all the botanical extracts concentration 500 mg/ml was the most potent chemical. The confidence intervals for the LT_{50} estimates from repeated measures logistic regression using GEE were consistently bounded and narrow compared to those from the Kaplan-Meier survival analysis hence precise estimates and therefore a better method.

The study results demonstrated that repeated measures logistic regression using GEE and Kaplan-Meier survival function provide a highly useful alternative for analysis of repeated measures dose response (correlated) data from arthropod studies when the interest is in the speed of kill.

Repeated measures logistic regression using GEE method combined with other existing methods will be helpful in estimating lethal time of repeated dose response data from arthropod studies

6.2 Recommendation

Kaplan-Meier survival function and repeated measures logistic regression using GEE are effective tools for analyzing repeated measures dose response data.

Since repeated measures logistic regression using GEE has given LT_{50} estimates with consistently bounded and smaller confidence level while taking into account the correlation when the interest is in the speed of kill; it should therefore be used in complementary fashion together with other existing methods in estimating LT_{50} so as to get meaningful results.

For further research one should try and widely explore other regression methods which incorporate GEE in estimating LT_{50}. Future studies should also consider other correlation matrices in GEE so as to have a wider range of comparison. Comparison of estimates from two survival analysis methods (Kaplan-Meier estimates and the Aalen-Nelson estimates) should also be considered in future studies.

6.3 Arthropod Dose-response Importance

This study is important for arthropod dose response studies as it has implications regarding the method for analyzing correlated dose-response data specifically when the speed of kill is of interest.

REFERENCES

Aalen, O.O. (1978), ‘Nonparametric inference for a family of counting processes’, *Annals of Statistics* **6**, 701–726.

Borsuk, M.E., Powers S.P. and Peterson C.H. (2002), ‘A survival model of the effects of bottom-water hypoxia on population density of an estuarine clam (macoma balthica)’, *Can. J. Fish. Aquat. Sci* **59**, 1266–1274.

Bugeme, D.M., Knap H.I.B., Wanjoya A.K. and Maniani N.K. (2009), ‘Influence of temperature on virulence of fungal isolates of metarhizium anissopliae and beaureria bassiana to the two-spotted spider mite tetranychus uriticae’, *Mycopathologia* **167**, 221–227.

Burton A, Altman DG, Royston P, Holder RL 2006: The design of simulation studies in medical statistics**.** *Stat Med*, **25**(24)**:**4279-4292

Cabanillas, H.E. and Jones W.A. (2009). ‘Pathogenicity of *Isaria sp* (Hypocreales: Claricipitaceae) against the sweet potato whitefly B biotype, *Bemisia tabaci* (Hemiptera: Aleyodidae)’, *Crop Protection* **28**, 333-377

Cox, D.R. (1972), ‘Regression models and life-tables’, *Biometrics* **38**, 67–77.

Cox, D.R. and Oakes D. (1984), *Analysis of Survival Data.*, Chapman and Hall.

Crowder, M. (1995), ‘On the use of a working correlation matrix in using generalized linear models for repeated measures.’, *Biometrika* **82**, 407–410.

Cutler, S.J. and Ederer F. (1958), ‘Maximum utilization of life-table method in analyzing survival.’ *J. Chronic Dis* **8**, 699–712.

Eaton, M. and Kells S.A. (2009), ‘Use of vapor pressure deficit to predict humidity and temperature effects on the mortality of mold mites, tyrophagus putrescentiae’, *Exp.Appl Acarol* **47**, 201–213.

Feng, M.G., Liu C.L., Xu J.H. and Xu Q. (1998), ‘Modeling and biological implication of time-dose-mortality data for entomophthoralen fungus, zoophthora anhuiensia, on green peach aphid myzus persicae.’, *J. Invertebr. Pathol* **72**, 246–255.

Feng, M.G., Tang Q.Y., Hu G.C. and Huang S.W. (1996), ‘Susceptibility of seven species of aphids to a beaureria bassiana isolate: analysis of time-dose-mortality model.’, *J. Basic Sc. Eng* **4**, 22–33.

Finney, D.J. (1964), *Statistical Methods in Biological Assay.*, second edn, Griffine, London.

Finney, D.J. (1971), *Probit analysis*, Cambridge University Press, Cambridge, UK.

Gehan, E.A. (1969), ‘Estimating survival functions from the life-table.’, *J. Chronic Dis* **21**, 629–644.

Hardin, J.W. and Hilbe, J.M. (2003), *Generalized Estimating Equations.*, Chapman & Hall: New York.

Hansen, N.B. and Lambert M.J. (2003), ‘An evaluation of the dose response relationship in naturalistic treatment settings using survival analysis.’, *Mental Health Services Research* **5**, 1–12.

Holbrook, G., Berger K., Steed K. and Leiewicz D. (1999), ‘Survival analysis for the detection of low-level insecticide resistance’, *J. Pest pp.* 159–166.

Hosmer, D.W. and Lameshow S. (1999), *Applied survival analysis*, Wiley and Sons, New York.

Kalbfleisch, J.D. and Prentice R.L. (1980), *The Statistical Analysis of Failure Time Data*, John Wiley, New York, NY.

Kaplan, E.L. and Meier P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**, 457–481.

Klein, J. and Moeschberger M. (1997), *Survival Analysis. Techniques for Censored and Truncated data*, New York, Springer-Verlag Inc.

Kleinbaum, G.D. (1995), *Survival Analysis. A Self-Learning Text*, Springer-Verlag, New York, Inc.

Latifian, M. and Rad B. (2012). 'Pathogenicity of entomopathogenic fungi *Beaureria bessiana* (Balsamo) Vuillmin, Beaveria brongniartii Saccardo and *Metarhizium anisopliae* Metsch to adult Oryctes elegans Prell and effects on feeding and fundicity', *Intl J Agri Crop sci* **4**, 1026-1032

Lee, T.W. and Wang J.W. (2003), *Statistical Methods for Survival Data Analysis*, third edn, A John Wiley and Sons, Inc. Publications.

Liang, K.Y. and Zeger S.L. (1986), 'Longitudinal data analysis using generalized linear models.', *Biometrika* **73**, 13–22.

Mburu, D.M., Ochola L., Maniania N.K., Njagi P.G.N., Gitonga L.M., Ndungu M.W., Wanjoya A.K. and Hassanali A. (2009). 'Relationship between virulence and repellency of entomopathogenic isolates of *Metarhizium anisopliae* and *Beaureria* *bassiana* to the termite *Macrotermes michaelseni.*', *Journal of Insect Physiology* **55**, 774-780

McCullagh, P. and Nelder J.A. (1983), *Generalized Linear Models*, first edn, Chapman and Hall.

McCullagh, P. and Nelder J.A. (1989), *Generalized Linear Models*, second edn, Chapman and Hall.

Moncharmont, F.D., Decourtyre A., Hennequet-Hantier C., Pons O. and Pham-Delegue M. (2003), 'Statistical analysis of honeybees survival after chronic exposure to insecticides', *Environmental Toxicology and Chemistry* 22, 3088–3094.

Nelson,W. (1972), 'Theory and applications of hazard plotting for censored failure data.', *Technometrics* **14**, 945–965.

Nowierski, R.M., Zeng Z., Jaronski S., Delgado F. and Swearingen W. (1996), 'Analysis and modeling of time-dose-mortality of melanoplus sanguinipes, locusta megretoria migratorioides, and schistocerca gregaria (orthoptera: Aicirdidae) from beaureria, metarhizium, and paecilomyces isolates from madagascar', *J. Invertebr. Pathol* **67**, 236–252.

Osbrink, W.L.A., Lax A.R. and Brenner J.R. (2001), 'Insecticides susceptibility in coptotermes formosanus and reticlitermes virginicus (isoptera: Rhinotermitidae)', *J. Econ. Entomol* **94**, 1217–1228.

Preisler, H.K. and Robertson J.L. (1989), 'Analysis of time-dose-mortality data.', *J. Econ. Entomol* **82**, 1534–1542.

Robertson, J.L. and Preisler H.K. (1992), *Pesticide bioassays with arthropods*, CRC. Boca Ratn, FL.

Roush, R.T. and Miller G.L. (1986), ‘Considerations for design of insecticide resistance monitoring programs’, *J. Econ. Entomol* **79**, 293–298.

Sahaf, B.Z. and Moharrmipour S. (2008), ‘Fumigant toxicity of carum copticum and vitex pseudo-negundo essential oils against eggs larvae and adults of callosobruchus macalatus’, *J Pest. Sci* **81**, 213–220.

Scholte, E.J., Njiru B.N., Smallegange R.C., TakkenW. and Knols B.G. (2003), ‘Infection of malaria (*anopheles gambiae s.s*.) and filariasis (*culex quinquefasciatus*) vectors with the entomopathogenic fungus *Metarhizium anisopliae*.’, *Malaria* **2**, 1–8.

Stokes, M.E., Davis C.S. and Koch G.G. (2000), *Categorical Data Analysis Using the SAS System,* Cary: SAS Institute, Inc.

Thomsen, L. and Eilenberg J. (2000), ‘Time-concentration mortality of pieris brassicae (lepidoptera: Pieridae) and agrotis segetum (lepidoptera: Noctuidae) larvae from different destruxins’, *Entomological Society of America* **29**, 1041–1047.

Thorne, J.E., Weaver D.K., Chew V. and Baker J.E. (1995), ‘Probit analysis for correlated data: Multiple observations over time at one concentration’, *J. Econ. Entomol* **88**, 1510–1512.

van Zaane1, B., Vergouwe, Y., Donders, A.R.T. and Moons, K.G.M. (2012). ‘Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study’. *BMC Medical Research Methodology* **12**, 1-9

Wang, L., Huang J., You M. and Liu B. (2004), 'Time-dose-mortality modeling and virulence indices for six strans of verticillium lecanie against sweet potato white fly, bemisia tabaci (gemadius)', *JEN* **127**, 494–500.

Xu, J.F., Feng M., and Yu Q. (1999), 'The virulence of entomophthoralean fungus pandora delphacis to the brown planthopper, nilaparrata lugens', *Entomologia sinics* **6**, 233–341.

Zeger, S.L. and Liang K.Y. (1986), 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics* **42**, 121–130.

Ziegler, A., Kastner C. and Blettner M. (1998), 'The generalised estimating equations: An annotated bibliography', *Biometrical Journal* **40**, 115–139

Zurek, L., Watson D.W., and Schal C. (2002). 'Synergism between *Metarhizium anisopliae* (Deuteromycota: Hyphomycetes) and Boric Acid against the German Cockroach (Dictyoptera: Blattellidae). *Biological Control* **23**, 296-302

Zuur, A.F., Ieno E.N., Walker N.J., Saveliev A.A. and Smith G.M. (2009), *Mixed Effects Models and Extensions in R.,* Springer Science and Business Media, LLC, New York.

APPENDICES

Appendix 1: Data format for survival analysis

```
> plantsurva

    Extract  Conc subjects time censo

1         B  12.5        1   12     0

2         B  12.5        1   12     0

3         B  12.5        1   12     0

4         B  12.5        1   12     0

5         B  12.5        1   12     0

6         B  12.5        1   12     0

7         B  12.5        1   12     0

8         B  12.5        1   12     0

9         B  12.5        1   12     0

10        B  12.5        1   12     0

11        B  12.5        1   12     0

12        B  12.5        1   12     0

13        B  12.5        0   24     0

14        B  12.5        1   36     0

15        B  12.5        1   36     0

16        B  12.5        1   36     0

17        B  12.5        1   36     0

18        B  12.5        1   36     0

19        B  12.5        1   36     0

20        B  12.5        1   48     0

21        B  12.5        1   48     0
```

Appendix 2: Data format for repeated measure logistic regression using GEE

```
>geedata

   extract time conc rep total success prop IDD

1        E   12 12.5   1    50       8 0.16   1

2        E   12 12.5   2    50      10 0.20   1

3        E   12 12.5   3    50      12 0.24   1

4        E   24 12.5   1    50      12 0.24   2

5        E   24 12.5   2    50      15 0.30   2

6        E   24 12.5   3    50      13 0.26   2

7        E   36 12.5   1    50      16 0.32   3

8        E   36 12.5   2    50      15 0.30   3

9        E   36 12.5   3    50      18 0.36   3

10       E   48 12.5   1    50      25 0.50   4

11       E   48 12.5   2    50      20 0.40   4

12       E   48 12.5   3    50      23 0.46   4

13       E   60 12.5   1    50      30 0.60   5

14       E   60 12.5   2    50      28 0.56   5

15       E   60 12.5   3    50      26 0.52   5

16       E   72 12.5   1    50      33 0.66   6

17       E   72 12.5   2    50      30 0.60   6

18       E   72 12.5   3    50      29 0.58   6
```

Appendix 3: R-program

The following R program was used to generate LT_{50} using repeated measures logistic regression using GEE and survival analysis

```
#logistic regression and estimating LC50

#importing the data set into R

plant=read.csv(file.choose(),header=T)

plant

attach(plant)

names(plant)


#boxplots for the extracts

boxplot(cbind(succ,total-succ)~extract, data = plant, col = "lightgray",

ylab="No of dead mosquito larva",xlab="Botanical Extracts")


#boxplots for the concentration

boxplot(cbind(succ,total-succ)~conc,  data  =  plant,  col  =  "lightgray",

ylab="No of dead mosquito larva",xlab="Doses(mg/ml)")


###Fitting GEE

geedata=read.csv(file.choose(),header=T)

geedata

attach(geedata)

names(geedata)

library(gee)

library(geepack)
```

```
#fitting gee models using the 4 correlation matrices
#choosing correlation structure
#independent
fitg.ind=gee(cbind(dead,Tot_subjects-
dead)~time_hrs,id=IDD,data=geedata,family=binomial,corstr="independence",
scale.fix=T)

#working correlation matrix
round(summary(fitg.ind)$working.correlation,2)

#coefficient table of parameters
round(summary(fitg.ind)$coefficients,2)

#exchangeable
fitg.ex=gee(cbind(dead,Tot_subjects-
dead)~time_hrs,id=IDD,data=geedata,family=binomial,corstr="exchangeable",
scale.fix=T)

#working correlation matrix
round(summary(fitg.ex)$working.correlation,2)

#coefficient table of parameters
round(summary(fitg.ex)$coefficients,2)

#AR-1
```

```
fitg.ar1=gee(cbind(dead,Tot_subjects-
dead)~time_hrs,id=IDD,data=geedata,family=binomial,corstr="AR-M",Mv=1,
scale.fix=T)

#working correlation matrix
round(summary(fitg.ar1)$working.correlation,2)

#coefficient table of parameters
round(summary(fitg.ar1)$coefficients,2)

#using unstructured correlation matrix for each concentration level
#########################################################################
##Repeated measures logistic regression using GEE
library(gee)
library(geepack)
plant=read.csv("G:\\plant.csv",header=T)
plant
names(plant)
output<-NULL

for(i in unique( plant$Plant_extract)){
  subset<-plant[plant$Plant_extract==i,]
  for(j in unique(subset$Dose)){
    subset2<-subset[subset$Dose==j,]
    fitg2=geeglm(cbind(dead,Tot_subjects-
dead)~time_hrs,id=IDD,data=subset2,family=binomial,corstr="unstructured")
    #initial parameter estimates
```

```
fitg2
#GEE parameter estimates
summary(fitg2)

#working correlation matrix
summary(fitg.un)$working.correlation

#GEE parameter estimates
summary(fitg2)

#coefficient table of parameters
summary(fitg.un)$coefficients

#getting the variance covariance matrix
summary(fitg2)$cov.unscaled
summary(fitg2)$cov.scaled

#equating parameter estimates
alpha=fitg2$coefficients[1]
beta=fitg2$coefficients[2]
var_alpha=summary(fitg2)$cov.scaled[1,1]
cov_al_bt=summary(fitg2)$cov.scaled[1,2]
var_beta=summary(fitg2)$cov.scaled[2,2]

lt50_est=(-alpha)/beta
sd_lt50_est=sqrt((1/beta)*(1/beta)*(var_alpha-
(2/beta)*(alpha)*cov_al_bt+(alpha)*(alpha)*(1/beta)*(1/beta)*var_beta))
```

```
    low_LT50=(lt50_est-1.96*sd_lt50_est)
    upper_LT50=(lt50_est+1.96*sd_lt50_est)
    output<-rbind(output,c(i,j,lt50_est,low_LT50,upper_LT50))

  }
}
output #gives the LT50 estimate, Lower 95% CI and Upper 95% CI

#fitting survival model for each extract
attach(dose)
attach(plantsurv)
names(plantsurv)
library(survival)
#creating survival object
msurv=with(plantsurv,Surv(time,censor == 0))
fit=survfit(Surv(time,censor==0)~conc)
#test for difference in dose survival curves - logrank test
survdiff(Surv(time,censor==0)~conc)
plot(fit, main="Kaplan-Meier estimate for Extract B", xlab="time",
ylab="survival function", col=c(1,2,3,4,5))
legend("bottomleft",legend = c("12.5","50","250","500"),lty = c(1,2,3,4),
col = c(1,2,3,4))
abline(h=0.5, col=3,lty=3) #drawing the horizontal curve at 0.5 point
```