A Multistage Model for Prediction of Sequence of Events

By

Rafiqul Islam Chowdhury
Ph.D. Candidate (Reg. No: 111 Session: 2014-2015)

A thesis submitted in total fulfillment of the requirements for the degree of ***DOCTOR OF PHILOSOPHY (Ph.D.).***

Institute of Statistical Research and Training (ISRT)
University of Dhaka
Dhaka, Bangladesh

December, 2017

Supervisor's declaration

The undersigned hereby certify that this thesis titled, **"A Multistage Model for Prediction of Sequence of Events"** submitted as a requirement for the degree od Ph.D. is the result of Rafiqul Islam Chowdhury's (Reg. No: 111 Session: 2014-2015) research work under our supervision and that this study in whole or in part has not been submitted for an award, including a higher degree, to any other University or Institution.

Dated: December, 2017

Supervisors:

M. Ataharul Islam, **Ph.D**
Q. M. Husain Professor
Institute of Statistical Research
and Training (ISRT)
University of Dhaka
Dhaka, Bangladesh.

A. H. M. Mahbub Latif, **Ph.D**
Professor
Institute of Statistical Research
and Training (ISRT)
University of Dhaka
Dhaka, Bangladesh.

Abstract

This dissertation investigates the existing methods for risk prediction of a sequence of events from longitudinal studies for the continuous time data, in addition to, proposing a simple alternative method. These outcomes (events) can change status at different follow-ups that may produce a large number of paths or trajectories. Also, regressive models for multinomial and ordinal outcomes for discrete time data to obtain a joint model for a sequence of events for risk prediction is proposed. A key challenge is the simplification and generalization of the existing method for continuous time data for risk prediction for a large sequence of events at different stages. Most of the models are proposed to solve the problem arising from the progression of specific diseases process.

The proposed alternative multistage procedure simplifies the transition models for risk prediction of a sequence of events for continuous time data. This framework provides the estimates for each stage in the process conditionally and the conditional estimates are linked based on marginal and conditional models to obtain the joint probabilities needed for predicting the status of disease based on the potential risk factors. The proposed method of prediction is a new development using a series of events in conditional setting arising from the beginning to the endpoint. Also, a general form of integral is developed for predicting the joint probability of a sequence of events from longitudinal studies for (i) different types of trajectories and (ii) any segment of a trajectory along with the generalization to any number of stages which is a new development.

In follow-up or panel studies, multinomial outcomes may occur within an interval where transition times are not exactly known, or the time of the event is itself discrete. Available models for risk prediction for multinomial outcomes with specified risk factors are only for a single response and are not extended for prediction of a sequence of events for discrete time data for different stages.

The regressive models for multinomial outcomes are proposed and then a modeling framework is developed to predict the joint probabilities for a sequence of events. The proposed models link the marginal and sequence of conditional models to provide the joint model needed for predicting the probability of a trajectory based on specified covariate patterns. The marginal model uses the outcome variable at the baseline and the models at the subsequent follow-ups provide the estimates of the parameters of the conditional models. The major improvement of the proposed framework is that one needs to fit a significantly smaller number of models compared to the conditional models such as Markov models.

The independence of the repeated outcomes will allow using simpler models, and the goodness-of-fit of the joint model is required for model performance. The proposed

goodness-of-fit test for joint model is obtained by linking marginal and conditional models. The test for independence uses marginal models for each repeated outcomes. The simulation study and application using real data prove the usefulness and illustrate the performance of these tests.

For ordinal outcomes from longitudinal studies regressive proportional odds model, and in the case of violation of proportional odds assumption regressive partial proportional odds model are proposed. Then a framework is developed to predict joint probabilities for a sequence of ordinal outcomes. The major improvement of the proposed model is that only one model is required for each repeated outcome compared to the sequence of conditional models such as Markov models. Results from these two models are compared to that from the proposed regressive multinomial logistic model. Also, test for goodness-of-fit and test for independence are shown. The proposed models provide the estimates for each stage in the process conditionally, and the joint model can be obtained for any order to predict the risk of a sequence of events. Proposed regressive partial proportional odds model and regressive multinomial models showed better performance compared to the regressive proportional odds model when proportional odds assumption is violated. Simulation studies showed satisfactory performance of the proposed regressive models for ordinal outcomes.

All the proposed model and the risk prediction framework for both continuous and discrete time data are a new development. The major improvement of the proposed model is that it reduces the over-parameterization. One can easily add interaction terms among previous outcomes, and predictors in the proposed framework which may provide a better understanding of the underlying process and the relationships between outcomes and risk factors. Using the developed framework, modeling and risk prediction for a sequence of events can be performed in many fields of studies such as epidemiology, public health, survival analysis, genetics, reliability, environmental studies, etc. This model would be very useful for analyzing big data. One can use the existing software for model fitting, and risk prediction of a sequence of events.

Acknowledgements

All praises and thanks are due to All mighty God who gave me the ability to make this dissertation possible, and for surrounding me with people who are true friends and well-wishers who continuously encouraged and helped me accomplishing this goal, therefore, they deserve a very special recognition.

My sincere gratitude goes to my supervisor Dr. M. Ataharul Islam, Q. M. Husain Professor, at the Institute of Statistical Research and Training (ISRT), University of Dhaka. I am honored and privileged to have the opportunity to work under his supervision. For me, he is a mentor, who taught me how to do research and take it as a passion.

I also would like to express my gratitude to my co-supervisor Professor Dr. A. H. M. Mahbub Latif for his support.

I want to express my gratitude and appreciation to the Professor Dr. Md. Israt Rayhan, Director, ISRT for his persistent help. I want to thank to Supernumerary Professor Dr. Pk. Md. Motiur Rahman, Supernumerary Professor Dr. M. Sekander Hayat Khan and Professor Dr. Ohidul Siddiqui for their continuous suggestions during my research work.

My sincere thanks to Professor Dr. Syed Shahadat Hossain for his continuous encouragement. I would like to offer my special thanks to all of the faculty members and staff at the Institute of Statistical Research and Training (ISRT) for their help and support throughout my Ph.D. research work. Thanks to Jahida Gulshan, my fellow researcher for her supportive role.

I would like to acknowledge the support that I received from the HEQEP subproject 3293 of the University Grants Commission of Bangladesh and the World Bank. I also acknowledge the University of Michigan for the permission for using the Health and Retirement Study (HRS) data and BIRPERHT for the data on maternal morbidity in Bangladesh.

I am indebted to Dr. Jagadish Chakraborty my mentor's, friend's and former colleague's at the Kuwait University who continuously encouraged me to pursue my Ph.D. and research work.

My special thanks go to my beloved wife, Farida Yeasmeen (Neela) for her unconditional love and support. She shared the pain of staying away for a long period during my research work. My debt and love for her are everlasting.

Before I finish, I will never forget my entire family who always been in my thoughts and their prayers and hope throughout my doctoral studies alleviated me to get this dissertation done.

Rafiqul I Chowdhury

Contents

List of Figures

List of Tables

List of Abbreviations

ADL	**A**ctivity of **D**aily **L**iving
BIRPERHT	**B**angladesh **I**nstitute of **R**esearch for **P**romotion of **E**ssential & **R**eproductive **H**ealth and **T**echnologies
CIC	**C**umulative **I**ncidence **C**urve
CIF	**C**umulative **I**ncidence **F**unction
EBMT	**E**uropean **G**roup for **B**lood and **M**arrow **T**ransplantation
GVHD	**Graft V**ersus **H**ost **D**isease
HRS	**H**ealth and **R**etirement **S**tudy
NIA	**National I**nstitute of **A**ging
SSA	**S**ocial **S**ecurity **A**dministration

Dedicated to

The departed Soul of My Father,
Siraj Uddin Chowdhury

Chapter 1

Introduction

1.1 Background

In longitudinal, panel or cohort studies, responses and covariates are repeatedly collected over time on each study participant or experimental unit. This repeated measures data are collected in various disciplines such as biomedical sciences, epidemiology, reliability, econometrics, environment, social science, etc. The responses may be qualitative (categorical) or quantitative (discrete or continuous) and time can be continuous or discrete. For example, a leukemia patient after bone marrow transplantation may experience multiple events, such as platelet recovery, acute graft versus host disease (GVHD), relapse or death. In this case, events are discrete and are observed over continuous time. When subjects move from one state to another, a transition occurs. In many instances, event occurrence of patients is only observed within an interval, e.g., transition times are not exactly known or the time of an event is itself discrete. In cohort studies, for example, depression status can repeatedly be measured over regular interval produces a sequence of discrete events at the discrete times. Multi-state models are the most common statistical technique to describe the occurrence of multiple events over time or disease progression longitudinally (Hougaard, 1999). When describing a sequence of similar or distinct types of events, multi-state models can be viewed as a series of nested models at different stages.

There is a growing interest on prediction of the probability of a sequence of events for a subject with specified covariate values and event history using multi-state model. Klein *et al.* (1994) first illustrated the prediction probability calculation for a future event from the multi-state model. Putter *et al.* (2006) demonstrated how to use the results of multi-state model to obtain predictions at a certain time after surgery for a patient. Computational aspects of prediction is illustrated in Putter *et al.* (2007). A key challenge is the simplification and generalization of the existing method for continuous time data for prediction for a large number of events or stages. For example, uncontrolled diabetes can lead to nephropathy, diabetic retinopathy, pulmonary tuberculosis, and coronary heart disease may occur in a large number of stages.

In social and behavioral science applications, discrete-time survival analysis is often more natural where time is likely to be measured discretely. Such data are better handled by the discrete time models (Commenges, 2002; Klein and Moeschberger, 2003; Sun, 2006). Prentice and Gloeckler (1978) and Pierce *et al.* (1979) discussed the use of the grouped proportional hazards model for the regression analysis of right censored data where life-times are partitioned into intervals. Lawless (2003) suggested a flexible and convenient method to analyze discrete time data using logistic regression. D'Agostino *et al.* (1990) proposed pooled logistic model to analyze discrete time data. By pooling the observations over multiple intervals into a single sample, logistic regression is employed to relate the risk factors to the occurrence of the event. The pooled logistic regression produced similar results as the Cox model. Barnett *et al.* (2009) provided practical accounts for the use of multinomial logit model for discrete time data. Beyersmann *et al.* (2012) suggested additional refinement for varying baseline risk for that model. However, this model is not extended for prediction for the competing risks for discrete time data under multi-state modeling framework. The motivation for this model stems from the need for generalization of competing risks models at different stages for discrete time data and prediction of future events. This can be achieved by extending the regressive model (Bonney, 1986, 1987) for prediction of disease status proposed by Islam and Chowdhury (2010) for multinomial and ordinal outcomes. This formulation can easily handle a large number of states emerging from different follow-ups from longitudinal data. Here, the difference is in the formulation of models based on multiple outcomes at various stages starting from an initial state.

Following Data sets are examples of longitudinal, cohort or panel studies and are based on biological and health sciences that represent specific data analysis challenges.

1.1.1 Data set I: Bone Marrow Transplantation (BMT) data

Complete data set is presented in the book by Klein and Moeschberger (2003), also, freely available in the 'mstate' package in R. The study consists of transplant patients from four hospitals in the USA, conducted from 1 March 1984 to 30 June 1989. A total of 137 patients went through transplantation. The maximum follow-up was seven years with 42 patients who relapsed and 41 who died in remission. Both pre and post transplant risk factors were recorded for patients, for example, recipient and donor sex and age, waiting time from diagnosis to transplantation and GVHD status. Researchers are interested in answering various questions from this data. Questions like following few are unanswered due to the complexity of the existing methods or unavailability of proper models.

(1) what is the probability of survival at some point after transplantation of a patient who is in remission;

(2) what is the probability of relapse whose platelet has not yet recovered;

(3) what is the probability of survival past two years for a patient who has first GVHD then platelet recovery after transplant.

1.1.2 Data set II: Maternal morbidity data

This data set comes from a prospective study on maternal morbidity in Bangladesh conducted by the Bangladesh Institute of Research for Promotion of Essential & Reproductive Health and Technologies (BIRPERHT), during November 1992 to December 1993 (Akhter *et al.*, 1996). A total of 1020 pregnant women was followed during the antenatal, delivery and postnatal stages.The occurrence of different types of complications, for example, excessive haemorrhage or fits/convulsion during pregnancy were recorded. Age, age at marriage, whether regular visits to a doctor for check-ups, planned index pregnancy are some of the risk factors. The health of a woman during pregnancy or childbirth has an impact on the health and development of the next generation and well-being of the family both economically and socially. Another consequence after delivery is that, there may be severe obstetric complication, including excess mortality and mental health problems (Filippi *et al.*, 2007). Therefore, clinician's and researcher's are interested in finding the answer to different questions. For some questions it is not be possible to find answer due to the unavailability of proper methods, for example,

(1) what is the impact of selected risk factors on the occurrence of complications during the three stages;

(2) Are the hazard rate for different stages differ;

(3) what is the probability of delivery complication for a woman with or without any antenatal.

1.1.3 Data set III: Health and Retirement Study (HRS) data

The Health and Retirement Study (HRS) is a panel study on retirement and health among the elderly born between 1931 and 1941 in the United States (HRS, 2014). This study was conducted by the University of Michigan and supported by the National Institute of Aging (NIA) and the Social Security Administration (SSA). In the HRS, the first wave of data is collected in 1992, which includes 12,652 individuals over age 50 and their spouses at two years apart. So far twelve waves of data are available. This survey collected data on demographics, income, assets, employment, and the outcome variables are health status (e.g., depression, self-reported health condition), chronic conditions (e.g., diabetes, stroke, heart, cancer, etc.), and health care service utilization. Details can be found at the HRS website (http://hrsonline.isr.umich.edu/). The special characteristics of HRS data are that the event time is discrete with a large number of follow-ups. There is a growing interest to predict the risk of a sequence of events from this type of data, and methods

are either unavailable or cannot handle large number of repeated outcomes. From this Data set following are some questions of interest that may not be answered in the existing methods:

(1) what is the risk of different diseases at various follow-ups or stages;

(2) what is the effect of selected variables effect on different transitions;

(3) what is the predicted risk of a sequence of events with covariate values and history.

1.2 Objectives

My doctoral research will focus on developing a simple framework for risk prediction of a sequence of events with specified covariate vector for both continuous and discrete time data. The proposed framework is expected to perform better than the existing methods and will be useful in answering questions of interest like the ones raises for the example Data sets in the previous section. Following are the specific objectives of the research:

(1) Devlop statistical model using multistage modeling framework to simplify the existing method for risk prediction for a sequence of events for continuous time data.

(2) Propose regressive models for multinomial and ordinal outcomes to predict the joint probability of a sequence of events at different stages for discrete time data.

(3) Use marginal and regressive models to obtain the joint probability of multinomial and ordinal outcomes by linking marginal and conditional probabilities.

(4) Propose test of independence for repeated outcomes and test for goodness-of-fit for the joint model.

(5) Conduct simulation studies to check the proposed model's performance.

(6) Illustrate the proposed models using real-life data described previously.

Rest of the chapters of the thesis are organized as follows. In chapter 2, a framework using the multistage model for risk prediction of a sequence of events for continuous time data is proposed. Chapter 3, describe the proposed regressive models for risk prediction of a sequence of multinomial outcomes from repeated measures for discrete time data. Test of independence and goodness-of-fit of a joint model for multinomial outcomes from discrete time data is proposed in Chapter 4. In chapter 5, regressive models for ordinal outcomes from repeated measures, for risk prediction of a sequence of events are proposed extending proportional odds, partial proportional odds and multinomial regression models. Chapter six, covers general conclusions and possible future research directions.

Chapter 2

Risks Prediction of Sequence of Events: Multistage Proportional Hazards Model

2.1 Introduction

In recent years, there is a growing interest to analyze the sequence of events that occurs over time. The occurrence of events and covariate information on each individual are collected at several time points. For example, the occurrence of maternal complications during the three stages of the childbearing process, namely, the antenatal, delivery and postnatal stages produce a sequence of events (Islam *et al.*, 2004; Islam and Chowdhury, 2017). If more than one complications are of concern, then it is the problem of competing risk. For example, after bone marrow transplantation, patients are subject to several competing risk, including the platelet recovery, relapse of leukemia, acute graft versus host disease (GVHD), or death, so that experience of such events may prevent the occurrence of the event of interest, or vice versa (Klein *et al.*, 1994). Analysis of these types of data involves the problems of censoring and repeated observations. Multi-state models are the most common statistical technique to describe the occurrence of these sequence of events or disease progression longitudinally (Hougaard, 1999). This is the model for a continuous time stochastic process that describes the movement of individuals among a finite number of states (e.g., healthy, disease, death, etc.). Once an individual moves from one state to another, then it is a transition, or an event occurs. In practice, it is often assumed that multi-state model follows Markov property though this assumption should be checked. To assess the covariates effect on each transition the Cox proportional hazards model is usually used (Cox, 1972).

Gail (1975) reviewed the actuarial model of competing risk and introduced notations to define competing risk models and independence assumption of competing risk. Holt (1978) illustrated the use of time dependent covariates along with the generalization of the Cox model to cause-specific hazard functions. Prentice *et al.* (1978) focused on the analysis of failure time in the presence of competing risk and interrelations among competing events and estimation of hazard rates after the elimination of causes. Farewell (1979)

discussed the use of the Cox proportional hazard model to study multiple infections following bone marrow transplantation for patients with aplastic anemia and leukemia. Kay (1982) proposed an extension of the proportional hazards model for some transient disease states between the initial state and death along with possible competing causes of death. Islam (1994) extended the Kay's model to several transitions, reverse transitions, and repeated transitions and proposed a method for testing the equality of parameters for transitions and repeated transitions. Hougaard (1999) presented general discussions about multi-state models including various state structure and related assumptions. The role of different time scales and assumptions used in the multi-state model discussed elsewhere (Commenges, 1999; Putter *et al.*, 2007; Meira-Machado *et al.*, 2009). Andersen and Perme (2008) provides a review of methods for analyzing data from patients with bone marrow transplantation. In the presence of competing risk, Andersen (2002) suggested using cause-specific hazards model as a starting augmented by other available methods, e.g., sub-distribution hazards proposed by Fine and Gray (1999).

In the estimation of cause-specific hazard model competing events are considered as censored, in addition to existing lost to follow-ups, withdrawal or censored cases, assuming independent censoring. This is a hypothetical situation where a competing event is impossible to occur for a subject, where the competing cause is eliminated from the population. For correct estimation, the cause-specific hazard approach requires the strong assumption of independence among competing events, which is unverifiable (Kalbfleisch and Prentice, 1980, p .250). Non-independence might produce biased parameter estimates in survival analysis hence misleading conclusions (Kleinbaum and Klein, 2012). The cumulative incidence function (CIF) also called 'crude cumulative incidence function' or 'sub-distribution function' is another summary measure which extends survival function for competing risk. No assumption of the independence of the competing risk is needed. Overall survival is computed by combining survival from all competing events. The cumulative incidence function estimates the probability that the event of interest occurs before time t and that it happens before any of the competing events. It is the estimate of the probability of the event of interest in the real world where any of the competing events can occur for a subject (Klein and Moeschberger, 2003). In the presence of the competing risk, the CIF provides a marginal probability of an event (Kalbfleisch and Prentice, 1980). Plots of hazard or survival from cumulative incidence function, cause- specific hazard function, and Kaplan-Meier estimate against time provide some insight about the dependency in the competing risk.

The CIF is extensively used in the calculation of prediction probabilities in the multi-state models. The effect of one or two binary risk factors can be assessed by estimating cumulative curve non-parametrically and testing whether the curves differ by risk factor category. Gray (1988) developed a log-rank type test to compare two or more Cumulative Incidence Curves (CICs) without covariates adjustment for one or two binary covariates vectors. Fine and Gray (1999) proposed a methodology to regress covariate effects directly on the

cumulative incidence function. In analogy with the relation between hazard and survival function, they defined a sub-distribution hazard. Fine and Gray (1999) imposed a proportional hazards assumption on the sub-distribution hazards. This method differs in the risk set compared to the cause-specific hazard. For the sub-distribution hazard, failure of a subject from the competing events remain in the risk set forever. The failure time of all subjects from the competing events is replaced by the which is larger than the highest observed event time in the data set. Putter *et al.* (2007) conducted a goodness-of-fit test by comparing the predicted cumulative incidence curve of the regression model with the non-parametric regression curve to the subset of a covariates.

There is a growing interest in using multi-state model for the clinical prognosis of a patient at a pre-specified time given the event history and the covariate values. This predicted risk or the transition probability is the probability of an individual of moving from one state to another state. The transition probability estimates the risk for a future event to a specified time for a subject who is event free at time 0. The probability of an event at time u is estimated. Hence, this transition probability is the probability for an interval. Klein *et al.* (1994) first presented the calculation of the transition probability in terms of hazards for the transition from the multi-state model by appropriately combining baseline hazards and regression coefficients based on the work of Arjas and Eerola (1993). Dabrowska *et al.* (1994) discussed the estimation and prediction of the realization of a process for a new subject in a Markov renewal model and used the Cox model to estimate transition hazards. More recently, Putter *et al.* (2006) reanalyzed the data arising from a study by the European Organization for Research and Treatment of Cancer, using the multi-state model to predict the risk of future events. In another study, Putter *et al.* (2007) presented a comprehensive illustration of prediction of probability of events based on the multi-state model. Aalen *et al.* (2008) noted that this predicted probability is a reasonable estimate of the transition probability. Meira-Machado *et al.* (2009) reviewed modeling approaches for multi-state models and available software.

Due to the complexity involved in using multi-state model along with related software, it's application is still limited. A key challenge is the simplification and generalization of the existing method for prediction for the large number of events that occurs at different stages. For example, uncontrolled diabetes can lead to nephropathy, diabetic retinopathy, pulmonary tuberculosis, and coronary heart disease among others that may occur in several stages. The existing framework for prediction involves multiple complex integrals and needs special computer skills to use multistate models for prediction. Generalization of these models for large number of stages becomes complicated and even difficult to handle as one has to derive separate prediction integrals for different trajectories. Another problem with some existing software those use same duration of time for events in a trajectory to ease the probability calculations that may not be appropriate for all types of problems. For example, in case of the maternal complications, the end point time for antenatal and delivery periods are more or less fixed, whereas for the postnatal period, it

could be much longer. Making the same duration for all these periods would be inappropriate as the total time for the antenatal and delivery periods would be unrealistic. Also, use of existing softwares need to define transition matrix as the first step among others. It would be problematic for maternal morbidity example as events (or-nonevent) in a stage are stratified depending on both events and non-event in previous stages.

The progression in a disease process may involve occurrence of a series of events over time. It is of great interest to predict the disease status at different stages and endpoints. In predicting the disease process, we need to link the likely transitions at different stages of the process through potential trajectories. In the previous attempts, the events were not considered in multistage framework and hence the underlying theory remained complex. As a result, the theories were demonstrated on the basis of specific problems. If the disease process involves several stages till the endpoint, those theories cannot be generalized as a simplified approach. This prediction procedure can be simplified and generalized for the large number of stages in a multi-state model if events in a stage are stratified. This will provide a better conditional model with application of Markov property. Also, probability calculations can be simplified from multiple integrals to single integrals for joint events risk prediction.

At this backdrop we proposed a simple framework for risk prediction of sequence of events using marginal and conditional models based on the work of Islam *et al.* (2004) and Islam and Chowdhury (2017). A general form of integral for risk prediction of a sequence of events is developed for (i) different types of trajectories and (ii) any segment of a trajectory along with the generalization to any number of stages. Section 2.2, focuses on the multistate models for prediction using existing framework. In section 2.3, we present the proposed method. Results using proposed method and comparison with the existing method are illustrated using 'European Group for Blood and Marrow Transplantation' (EBMT) data in Section 2.4. The second example from Maternal Morbidity data are presented in Section 2.5. Finally, conclusions are presented in Section 2.6.

2.2 Existing framework of multi-state models for risk prediction

In this section, the multi-state model is reviewed with the following state structure (Figure 2.1) for risk prediction proposed by Putter *et al.* (2006). Five states in Figure 1 are: i) event free and alive after surgery (State 0), ii) local recurrence only and alive (State 1), iii) distant metastasis only and alive (State 2), iv) both local recurrence and distant metastasis and alive (State 3) and v) the absorbing state death (State 4). This study used 'clock reset' approach (i.e., time start from zero once an individual moves to a new state). The Cox proportional hazards model is employed to evaluate the covariates effect on transition hazards $i \rightarrow j$. The hazard function for a subject with covariates vector $\mathbf{Z}$ for transition

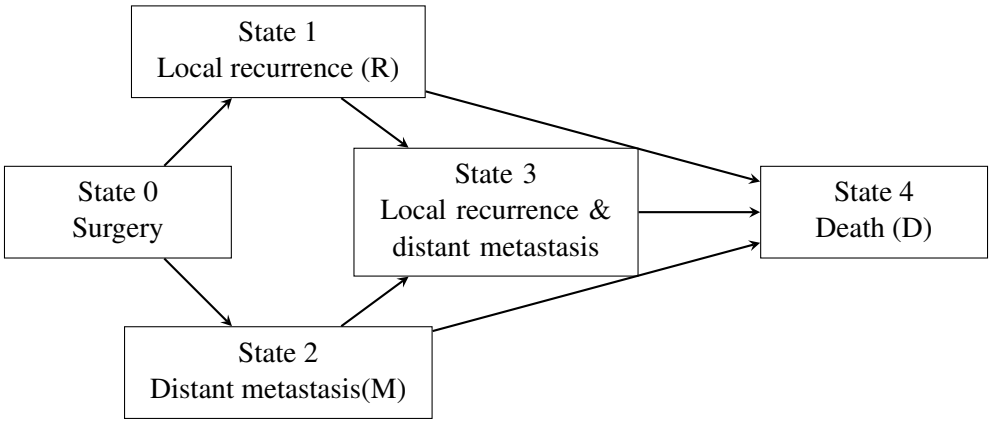


FIGURE 2.1: State structure of multi-state model used by Putter et al. (2006).

$i \to j$ is defined as

$$\lambda_{ij}(t) = \lambda_{ij,0}(t) exp(\boldsymbol{\beta}_{ij}^T \boldsymbol{Z}), \tag{2.1}$$

where $\lambda_{ij,0}(t)$ is the baseline hazard for transition $i \to j$, $\boldsymbol{\beta}_{ij}$ is the vector of regression coefficients and time t refers to time from entering state i, not from the begining of the study. Therefore, t should be understood as a different variable for each transition. Using the notation from Andersen *et al.* (1991) above hazard function can be re-written as:

$$\lambda_{ij}(t) = \lambda_{ij,0}(t) exp(\boldsymbol{\beta}^T \boldsymbol{Z}_{ij}), \tag{2.2}$$

where $\boldsymbol{Z}_{ij}$ is a vector of covariates specific to the transition $i \to j$, with $\boldsymbol{Z}_{ijk}$ denoting vector of covariates of subject k, estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Lambda}}_{ij,0}$ can be obtained by maximizing generalized Cox partial likelihood.

Klein *et al.* (1994) defined notations for patient history and is used by Putter *et al.* (2006), where $H_{i,rm}(t)$ denotes the event history of a subject who is at state i at time t after surgery, and, if appropriate, has experienced a local recurrence at time r and/or distant metastasis at time m. For example, $H_{3,rm}(t)$ represents the history that a patient had both a local recurrence at time r and a distant metastasis at time m and is still alive at time t. The local recurrence and/or distant metastasis can occur in any order. The probability that a subject who is in state 3 at time t after surgery, will make a transition to state 4 at time u $(u > t)$,

is given as

$$\begin{aligned}
\pi^4_{3,rm}(u,t) &= P(D \le u \mid H_{3,rm}(t)) \\
&= \int_{t-\max(r,m)}^{u-\max(r,m)} \lambda_{34,\mathrm{r,m}}(\mathrm{s}) exp\left[-\int_{t-\max(r,m)}^{s} \lambda_{34,rm}(v)dv\right] ds \\
&= \int_{t-\max(r,m)}^{u-\max(r,m)} \lambda_{34,\mathrm{r,m}}(\mathrm{s}) S_{34,rm}(s)ds / S_{34,rm}(t-\max(r,m)) \\
&= \frac{S_{34,rm}(t-\max(r,m)) - S_{34,rm}(u-\max(r,m))}{S_{34,rm}(t-\max(r,m))} \\
&= 1 - \frac{S_{34,rm}(u-\max(r,m))}{S_{34,rm}(t-\max(r,m))}. \qquad (2.3)
\end{aligned}$$

This is the conditional probability of death at time u for a subject who was at risk at time t after experiencing both local recurrence and distant metastasis. The probability that this subject is still alive is, $\pi_{3,rm}(u,t) = 1 - \pi^4_{3,rm}(u,t)$. Note that, in the left hand side of the equation (2.3), t is the lower limit and u is the upper limit of the integral. All time variables t, r, m and u in the integral are measured from the begining of the study (i.e., surgery). Here, $\{t - max(r,m)\} = l$ is the length of time for a subject since entered into state 3 who was at risk before making a transition 3→4. Similarly, $\{u - max(r,m)\} = l'$ gives the length of the time for a subject since entered into state 3 up to time u, to where the prediction for death is made with the length of interval for the prediction $\{u - max(r,m)\} - \{t - max(r,m)\} = l''$. It may be noted that time t defined in the hazards model in equation (2.1) refers time since entry into a state not from the beginning of the study.

The time scale for the above equation for selected transitions are displayed in the Figure 2.2. Line number 1 in the figure gives the time $\max(r,m) = m$ where a local recurrence occurred first at time r then a distant metastasis at time m. Therefore, time t as defined in equation (2.1) starts from distant metastasis time m (since entry into state 3, i.e., clock is reset) and $\{t - max(r,m)\} = l$ is the length of time at risk at state 3 that always starts from zero. Line 2 in the the figure explains the length $\{u - max(r,m)\} = l'$ up to prediction time u starting from m. Line line 3 shows the starting time of prediction interval and it's length l'' used in the equation (2.3). Line number 4 shows the scenario when probability of death is predicted since entry into a state up to u $(u > t)$ for the interval $[0 - u]$ with the length of the interval $l''' = u - t$. In this case the prediction time coincides with either 'r' or 'm'. In the figure, time t measured from the time as soon as distant metastasis occurs. In this case, $t = max(r,m) = m$ and then $t - max(r,m) = t - t = 0$. If distant metastasis occurs first at time m and next local recurrence at time r, then $t = max(r,m) = r$ and

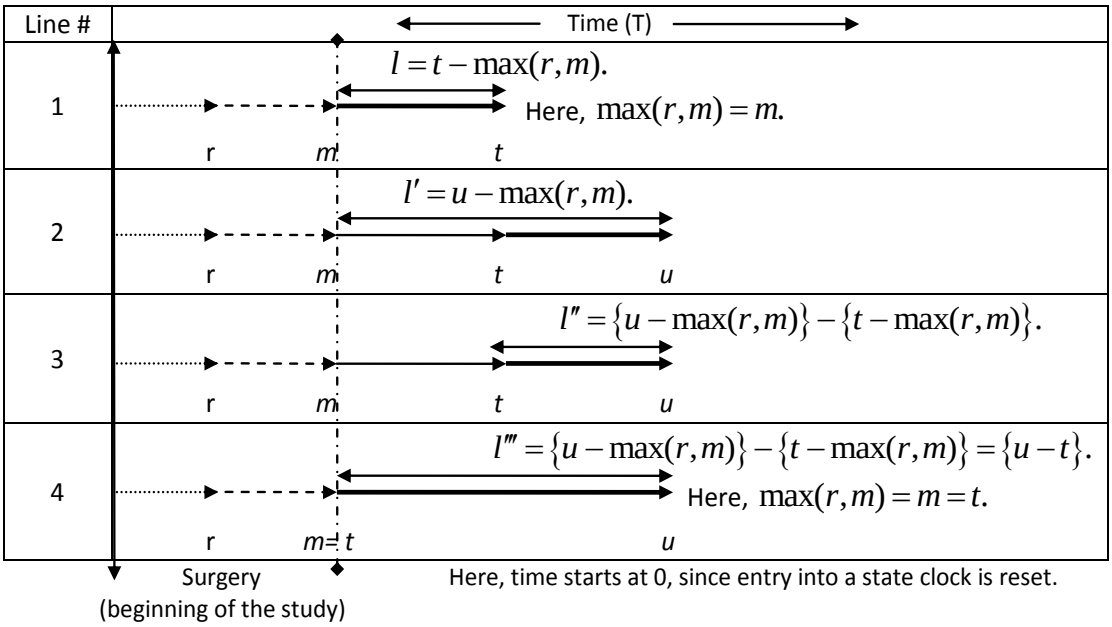


FIGURE 2.2: Time-scale for multi-state model.

$t - max(r,m) = 0$, then equation (2.3) can be written as

$$\begin{aligned}
\pi^4_{3,rm}(u,t) &= \int_0^{u-t} \lambda_{34,\mathrm{r,m}}(\mathrm{s}) exp\left[-\int_0^s \lambda_{34,rm}(v)dv\right] ds \\
&= \int_0^{l'''} \lambda_{34,\mathrm{r,m}}(\mathrm{s}) exp\left[-\int_0^s \lambda_{34,rm}(v)dv\right] ds \\
&= \int_0^{u-t} \lambda_{34,\mathrm{r,m}}(\mathrm{s}) S_{34,rm}(s)ds = S(0) - S_{34,rm}(u-t) \\
&= 1 - S_{34,rm}(u-t), \qquad (2.4)
\end{aligned}$$

where, $S_{34,rm}(t) = 1$, since $t = r = 0$ and $u - t$ is the length of the interval for which the prediction is made and $\pi_{3,rm}(u,t) = 1 - \pi^4_{3,rm}(u,t) = S_{34,rm}(u-t)$ is the probability of staying in state 3.

Consider a patient only with the local recurrence at time r and alive without distant metastasis at time t $[H_{1,r}(t) = \{R = r, M \geq t, D \geq t\}]$. With this history, four mutually exclusive transitions may occur through different paths. For example, the conditional probability, given that a subject is in state 1 at time t, and has arrived there at time r, with $t = r$ will visit state 3 before or at time u is

$$\begin{aligned}
\pi^3_{1,r}(u,t) &= P(M \leq u, D > u \mid H_{1,r}(t)) \\
&= \int_t^u \lambda_{13,r}(m-r) exp\left[-\int_t^m \{\lambda_{13,r}(s-r) + \lambda_{14,r}(s-r)\}ds\right] \\
&\times \pi_{3,rm}(u,m)\, dm \\
&= \int_t^u \lambda_{13,r}(m-r)\pi_{3,rm}(u,m) S_{1,r}(m-r, t-r)\, dm, \qquad (2.5)
\end{aligned}$$

where,
$S_{1,r}(m,t) = exp\left[-\int_t^m \{\lambda_{13,r}(v) + \lambda_{14,r}(v)\}dv\right] = \frac{S_{13,r}(m)S_{14,r}(m)}{S_{13,r}(t)S_{14,r}(t)}$ and
$\pi_{3,rm}(u,m)$ is the probability of staying in state 3.

The conditional probability, given that a subject is in state 1 at time t, and has arrived there at time r, with $t = r$, first will visit state 3 then state 4 before or at time u is

$$\begin{aligned}\pi_{1,r}^{34}(u,t) &= P(M < D \leq u \mid H_{1,r}(t)) \\ &= \int_t^u \breve{}_{13,r}(\mathrm{m}) exp\left[-\int_t^m \{\lambda_{13,r}(s) + \lambda_{14,r}(s)\}ds\right] \pi_{3,rm}^4(u,m)\, dm \\ &= \int_t^u \lambda_{13,r}(m-r)\pi_{3,rm}^4(u,m) S_{1,r}(m-r, t-r)\, dm. \end{aligned} \tag{2.6}$$

Following this approach, derivation for $\pi_0^{134}(u,t)$ will be complicated as three transitions are involved. This notational complexity arises as two transitions ($1 \to 3$ & $2 \to 3$) were made into the same state (i.e., non-progressive state structure). The generalized Cox partial likelihood function based on this approach for the example in Figure 1 is

$$L(\boldsymbol{\beta}) = \prod_{\substack{transition \\ i \to j}} \prod_{\substack{k=1 \\ d_{ij,k}=1}} \frac{exp(\boldsymbol{\beta}^T \mathbf{Z_{ij,k}})}{\sum_{l \in R_i(t_{ij,k})} exp(\boldsymbol{\beta}^T \mathbf{Z_{ij,l}})}, \tag{2.7}$$

where $t_{ij,k}$ is the failure or censoring time of individual k for transition $i \to j$ and $R_i(t)$ is the risk set of state i at time t (since entry at state i). Same time variable t is used for both integral and in equation (2.1). With this approach, it will be difficult to derive and handle necessary integrals for a large number of stages, for example, the case of diabetes complications given in the introduction section.

2.3 Proposed method

Islam *et al.* (2004) and Islam and Chowdhury (2017) proposed a multistage model for analyzing potential risk factors associated with complications at different stages of child-bearing process and interrelationships among the incidence of complications during these stages. At the beginning of the study, individuals are event free. Events in stage two onwards are conditional on the occurrence of events (or non events) at previous stages. The time of an event in a stage starts from 0, i.e., 'clock reset' approach. Model with this state structure is a progressive multi-state model. Proportional hazards model is used to assess the impact of covariates on each transition. The proposed method of risk prediction is an extension of Islam *et al.* (2004) and Islam and Chowdhury (2017) model by extending hazard, survival, density function and predicting risk of a sequence of events for the required time interval.

The multi-state models in Figure 2.1 is re-structured and shown in Figure 2.3, where events at stage two and onwards are stratified on the occurrence of the events in previous stages. The events are local recurrence (State 1), distant metastasis (State 2), both local recurrence and distant metastasis (State 3), death (State 4) and study begins after surgery (State 0). Here, all the events that occurred for the first time are in stage 1. After having first event, any second event following that is in stage 2. Similarly, events in stage 3 are conditional on all events from stage 2 and stage 1. Hence, this is an effective stratification on the outcomes at different stages. In Figure 2.3, there are six trajectories, for example, $0 \rightarrow 1 \rightarrow 2 \rightarrow 4$ is the first trajectory with state 0 as the beginning of the study and state 4 as the endpoint. Part of the above trajectory $0 \rightarrow 1 \rightarrow 2$ is termed as the segment of a trajectory, i.e., transition from state 0 to state 1 and to state 2 and being in the same state without any further transition to state 4. The Markov assumption, in particular the Markov process (continuous time) is assumed for this multi-state model which should be checked from the data. If in a stage, there are two or more events then it is a competing risk. This multi-stage model can be viewed as series of models which are nested in previous stages.

Consider a group of event free individuals can make transitions from the beginning of the study i $(i = 0)$ to state j $(j = 0, 1, \cdots, J)$ in stage one to state k $(k = 0, 1, \cdots, K)$ in stage two and to state l $(l = 0, 1, \cdots, L)$ in stage three, m $(m = 1, 2, 3)$ denotes the stages, with $i, j, k, l = 0$ indicates non-event and $j = 1, \cdots, J$; $K = 1, \cdots, K$; $l = 1, \cdots, L$ indicate the events. Let us denote $\boldsymbol{Z}_{kl|j}(t)$ for the regression vector for those who make transitions from state k at $(m-1)$th stage to state l at mth stage for given state j at stage $m-2$ and $\boldsymbol{\beta}_{kl|j}$ as the corresponding regression coefficients. Here t refers to a specific transition time (T_{ij}) in any stage. The hazard function for transition $k \rightarrow l$ can be defined as follows:

$$\lambda_{kl|j}(t;\mathbf{z}) = \lambda_{0,kl|j}(t) e^{\mathbf{Z}_{kl|j}(t)\boldsymbol{\beta}_{kl|j}} \tag{2.8}$$

where $\lambda_{kl|j}(t;\boldsymbol{z})$ is the hazard for a transition from a state k within a stage to state l in the next stage and $\lambda_{0,kl|j}(t;\mathbf{z})$ is the corresponding baseline hazard. The vector of regression coefficients corresponding to the covariate vector $\mathbf{Z}_{kl|j}(t)$ is $\boldsymbol{\beta}_{kl|j}$ and $Z(t) = [Z_1(t), Z_2(t), ..., Z_p(t)])$. Similarly, $\lambda_{ij}(t;\boldsymbol{z})$ refers particular transition during stage 1 and $\lambda_{jk|i}(t;\boldsymbol{z})$ refers particular transition during stage 2. For simplicity, we will use $\mathbf{Z}$ instead of $\mathbf{Z}(\mathbf{t})$ and $\lambda_{kl|j}(t)$ for $\lambda_{kl|j}(t;\mathbf{z})$ henceforth. Also, let $\hat{\boldsymbol{\beta}}_{kl|j}$ is the vector of the estimated regression parameters.

The subscript $kl \mid j$ of $\lambda_{kl|j}(t)$ identifies a specific transition. Also, it identifies the history of all states visited before making a transition to a state in next stage and the order of the state visited before. This history also identifies the corresponding conditional models. For example, $\lambda_{12|1}(t)$ is the hazards of transition to state 2 in stage 3 from state 1 in stage 2. Time T_{kl} for a transition in a stage starts from 0 as soon as an event takes place. The

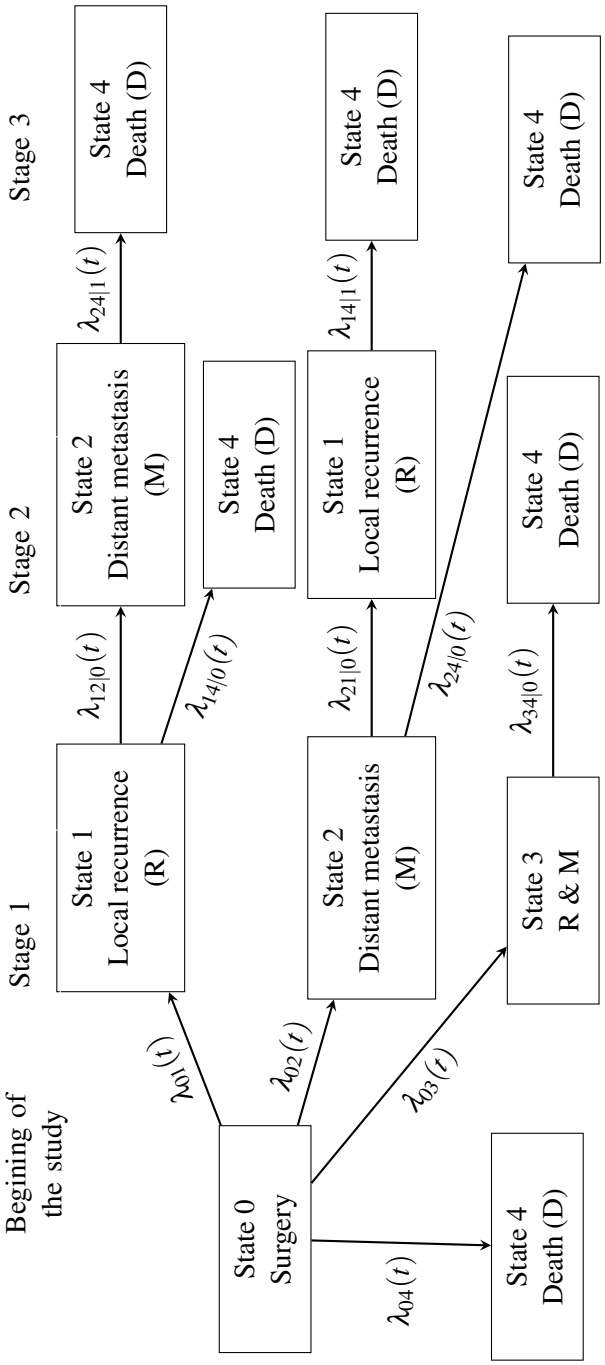


FIGURE 2.3: Re-structured multi-state model using proposed approach for Putter *et al.* (2006).

generalized Cox partial likelihood function for transition $k \to l$ can be shown as:

$$L(\boldsymbol{\beta}_{kl|j}) = \prod_{transitions} \prod_{r=1}^{n} \frac{exp(\boldsymbol{\beta}_{kl|j}\mathbf{Z_{kl|j,r}})}{\sum_{l \in R_i(t_{kl|j,r})} exp(\boldsymbol{\beta}_{kl|j}\mathbf{Z_l})}, \tag{2.9}$$

where $\boldsymbol{\beta}_{kl|j}$ is the vector of regression coefficients $\boldsymbol{Z}_{kl|j}$, the ordered transitions are $t_{kl|j,1} < t_{kl|j,2} < ... < t_{kl|j,n}$, n is the number of transition from k to l in a particular stage and R is the risk set. The censoring indicator $d_{kl,r} = 1$ if an individual r has an event for any transition, 0 otherwise.

Following notations are defined for cumulative hazard function, the survival function, the density function and cumulative distribution function for the random variable T_{kl} as

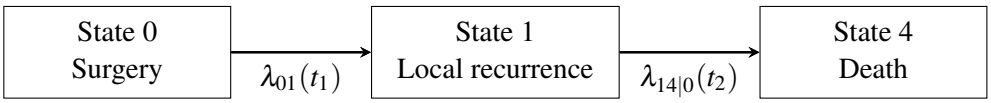
$$\Lambda_{kl|j}(t) = \int_0^t \lambda_{\mathrm{kl|j}}(\mathrm{t})\, dt, \quad S_{kl|j}(t) = e^{-\Lambda_{kl|j}(t)}, \quad f_{kl|j}(t) = \lambda_{kl|j}(t) S_{kl|j}(t)$$
$$\text{and } F_{kl|j}(t) = 1 - S_{kl|j}(t). \tag{2.10}$$

Similarly, for stage one and stage two we can define the same quantities as below:

$$\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(t)\, dt, \quad S_{ij}(t) = e^{-\Lambda_{ij}(t)}, \quad f_{ij}(t) = \lambda_{ij}(t) S_{ij}(t), \text{ and}$$
$$F_{ij}(t) = 1 - S_{ij}(t). \quad \Lambda_{jk|i}(t) = \int_0^t \lambda_{jk|i}(t)\, dt, \quad S_{jk|i}(t) = e^{-\Lambda_{jk|i}(t)},$$
$$f_{jk|i}(t) = \lambda_{jk|i}(t) S_{jk|i}(t) \text{ and } F_{jk|i}(t) = 1 - S_{jk|i}(t).$$

2.3.1 Predicted Risk (Transition probability)

The transition probability or the predicted risk is the probability of an event or sequence of events of interest within an interval. For homogeneous Markov process (i.e., transition to next state depends only on current state not on the arrival time to current state), the probability calculations depend on the length of the time interval (Beyersmann et al., 2012, pp. 30). To derive equations to calculate this probability, for simplicity, consider two transitions from the first trajectory of Figure 2.3 as shown below:



We consider the first transition as the local recurrence from surgery and the second transition as death from the local recurrence. Let T_1 $(0 \le T_1 \le t_1)$ be the time for transition to the local recurrence from surgery and T_2 $(0 \le T_2 \le t_2)$ be the time to death measured from the local recurrence. Thus $T = T_1 + T_2$ is the total time till death starting from surgery, where $T_1 = t_1$ is the time for the local recurrence and $T_2 = t_2$ is the time to death since the

time of the local recurrence. Since, $T_2 = T - T_1$ there is an inbuilt dependence between T_1 and T_2 through the fixed value of T.

Denote $f_{01}(t_1)$ as the density function of T_1 for the local recurrence and $f_{14|0}(t_2 \mid t_1)$ as the density function of T_2 for the second event (death) given the first event at t_1. Here, T_2 is the time for occurrence of the second event since the occurrence of the first event. Since T_1 is non-negative continuous random variable, the transition probability is obtained by integrating the corresponding density function on the required interval. We add up the "point probabilities" by integrating the density, $f_X(x)$, to obtain cumulative probability. The probability of the local recurrence by a given time $T_1 = t_1$ starting from surgery is obtained using the following integral:

$$\begin{aligned} Q_{01}(t_1) = \pi_0^1(0,t_1) &= \int_0^{t_1} \lambda_{01}(s_1) exp\left(-\int_0^{s_1} \lambda_{01}(v_1)\, dv_1 \right) ds_1 \\ &= \int_0^{t_1} f_{01}(s_1)\, ds_1. \end{aligned} \tag{2.11}$$

The notation $\pi_0^1(0,t_1)$ is used by Putter *et al.* (2006). This is the transition probability of the local recurrence for the interval $[0,t_1]$. It is also cumulative incidence of local recurrence as time started from 0. It may be of interest to predict probability for the interval $[u_1,t_1]$, $u_1 < t_1$, which is obtained using the following integral:

$$\begin{aligned} Q_{01}(u_1,t_1) = \pi_0^1(u_1,t_1) &= \int_{u_1}^{t_1} \lambda_{01}(s_1) exp\left(-\int_{u_1}^{s_1} \lambda_{01}(v_1)\, dv_1 \right) ds_1 \\ &= \int_{u_1}^{t_1} \lambda_{01}(s_1) S_{01}(s_1)\, ds_1 / S_{01}(u_1) \\ &= \frac{1}{S_{01}(u_1)} \left(\int_0^{t_1} \lambda_{01}(s_1') S_{01}(s_1')\, ds_1' - \int_0^{u_1} \lambda_{01}(t_1') S_{01}(t_1')\, dt_1' \right) \\ &= \frac{F_{01}(t_1) - F_{01}(u_1)}{S_{01}(u_1)} = \frac{S_{01}(u_1) - S_{01}(t_1)}{S_{01}(u_1)} \\ &= 1 - S_{01}(t_1)/S_{01}(u_1), \text{ since } S_{01}(t_1) = 1 - F_{01}(t_1). \end{aligned}$$

The conditional probability of the second event given the first event occurred at a given time for an interval can be obtained by integrating the conditional density function. In the following conditional pdf, t_1 is any given value for which the corresponding marginal probability can be obtained. In this way, the probability of different values of T_2 can be estimated given the knowledge that first event is observed at $T_1 = t_1$. The conditional probability of death by time t_2 (measured from the time of local recurrence) given that the

local recurrence has occurred at time t_1 is obtained using the following integral:

$$Q_{14|0}(t_2 \mid t_1) = \pi_1^4(0,t_2) = P(T_2 \leq t_2 \mid T_1 = t_1) = \int_0^{t_2} f_{14|0}(s_2 \mid t_1)\, ds_2. \tag{2.12}$$

Here, the time t_1 is the history of local recurrence given at time t_1. Note that the integral is taken from 0 to t_2 which is the time since local recurrence.

The probability of both events $[Q_{14}(t_1,t_2) = P(T_1 \leq t_1, T_2 \leq t_2)]$ i.e., the local recurrence at time $T_1 = t_1$ and death by time $T_2 = t_2$ where $T = T_1 + T_2$ can be obtained using the relation of the conditional marginal and joint probability (Moreira and Meira-Machado, 2012). The probability of both events for an interval (0 to t) can be obtained as follows:

$$\begin{aligned} Q_{14}(t_1,t_2) = P(T_1 \leq t_1, T_2 \leq t_2) = Q_{01}(t_1) Q_{14|0}(t_2 \mid t_1) = \pi_0^1(0,t_1)\pi_1^4(t_1,t_2), \\ \text{since, } P_{Y|X}(Y \leq y \mid X = x) = \frac{P_{XY}(x,y)}{P_X(x)}, \end{aligned} \tag{2.13}$$

where $Q_{14|0}(t_2 \mid t_1)$ is the conditional probability for the second event for the subset of $T_1 = t_1$ among those with the first event. $Q_1(t_1)$ is the probability of the first event for the interval $[0,t_1]$. This fixed value could be thought as the length of the interval for the first event. Time t_1 is an arbitrary value for which prediction can be made. Using the above equation, it is possible to predict the probability of the both events for any values of $T_1 = t_1$ and $T_2 = t_2$.

Alternatively, the probability of both events i.e., the local recurrence at time T_1 and death by time T where $T = T_1 + T_2$ can be expressed as the probability of occurrence of the first event at time $T_1 = t_1$ and the second event by time $T = t$ using the following integral:

$$\pi_0^{14}(t_1,t) = \int_0^{t_1} \int_{t_1}^{t} f_{01}(s_1) f_{14|0}(s \mid s_1)\, ds\, ds_1\,. \tag{2.14}$$

Now, let $T_2 = T - t_1$ for given value of $T_1 = t_1$. Then the equation (2.14) can be shown as follows:

$$\begin{aligned} \pi_0^{14}(t_1,t) = \int_0^{t_1} \int_0^{t_2} f_{01}(s_1) f_{14|0}(s_2 \mid t_1)\, ds_2\, ds_1 = \int_0^{t_1} f_{01}(s_1)\, ds_1 \times \\ \int_0^{t_2} f_{14|0}(s_2 \mid t_1)\, ds_2 = \pi_0^1(0,t_1)\pi_1^4(0,t_2) = Q_{14}(t_1,t_2). \end{aligned} \tag{2.15}$$

Here $\pi_0^{14}(t_1,t)$ is the probability for both events for the interval $[0,t]$ with the local recurrence at time $T_1 = t_1$ and death by time $T = t$. Equation (2.15) can easily be generalized to more than two stages. We can generalize (2.15) for trajectory with three stages using

the relationship shown in equation (2.13) as follows:

$$Q_{123}(t_1,t_2,t_3) = Q_{01}(t_1)Q_{12|0}(t_2 \mid t_1)Q_{23|01}(t_3 \mid t_2,t_1).$$

Further generalization for m stages is quite straightforward. For example, the predicted risk for a trajectory, i.e., for a sequence of events from beginning to the endpoint can be shown as follows:

$$Q_{123\ldots L}(t_1,t_2,t_3,\ldots,t_m) = Q_{01}(t_1)Q_{12|0}(t_2 \mid t_1)Q_{23|01}(t_3 \mid t_2,t_1), \tag{2.16}$$

$$\ldots, Q_{KL|0,1,\cdots,J}(t_m \mid t_{m-1},\cdots,t_1). \tag{2.17}$$

We may need to predict the risk of a segment of a trajectory, i.e., transition from one state to another state and being in the same state without any further transition. For instance, predicted probability from surgery to local recurrence and being alive, could be estimated as:

$$Q_{01}(t_1) \times [1 - Q_{14|0}(t_2 \mid t_1)] = Q_{01}(t_1) \times S_{14|0}(t_2 \mid t_1). \tag{2.18}$$

First part of the above equation is the predicted probability to move to local recurrence after surgery. Then the subject should be alive, so multiplication by the probability of surviving from death. The predicted risk for a segment of a trajectory as shown in equation (2.18) readily generalizes for several stages.

To compute these prediction probabilities using multistage model following steps are used. First, estimate of the regression coefficient $\hat{\boldsymbol{\beta}}$ for each transition is obtained by fitting the Cox regression model. Proportional hazard assumption should be checked for each variable. Next, cumulative baseline hazard $\hat{\Lambda}_0$ is obtained for each transition. Using this estimate the cumulative hazard for a patient with specified covariate vector ($\mathbf{Z}^*$) can be obtained. For example, for the first transition the cumulative hazard for a subject with covariate vector ($\mathbf{Z}^*$) is $\hat{\Lambda}(t) = \hat{\Lambda}_0(t)exp(\hat{\boldsymbol{\beta}}^\top \mathbf{Z}^*)$. Let τ_1 denote the event time-points for t_1 for which $d\hat{\Lambda}_1(t_1) = \hat{\Lambda}_1(t_1) - \hat{\Lambda}_1(t_1-) > 0$. This is the estimator of the hazard function $\hat{\lambda}_{01}(t_1)$. Then using these estimates the density function is estimated for each transition. The estimator of the prediction probabilities are obtained by replacing each integral by a sum and by replacing hazard and survival function by their estimated counterparts (Klein *et al.*, 1994; Klein and Moeschberger, 2003; Putter *et al.*, 2006, 2007). For instance, $Q_{01}(t_1)$ in equation (2.11) is estimated by $\hat{Q}_{01}(t_1) = \sum_{\substack{0<s_1\le t_1 \\ s_1\in\tau_1}} \hat{\lambda}_{01}(s_1)\hat{S}(s_1-)$. Finally, using equations (2.11), (2.12) and (2.13) the required prediction probabilities are obtained. Overall predicted probability of the event of interest through different trajectories can be obtained by adding the probability of the event through each trajectory.

2.4 Application: EBMT data

To demonstrate the proposed method, data on 2204 patients from the 'European Group for Blood and Marrow Transplantation' (EBMT) collected between 1995 and 1998 are used. This data set is available from the 'mstate' package. In the paper by Putter *et al.* (2007) used the same data for the prediction using multi-state model. The computational details for clock reset method for this data set can be found in Putter (2011). The state structure used for this multi-state model is presented in Figure 2.4. The transitions are: Transplant (TX)→Platelet Recovery (PR), PR→Relapse or death and TX→ Relapse or death. Platelet recovery and relapse or death are competing events. Covariates used are: disease subclassification (AML, ALL, CML); patient age at transplant (20, 20-40, >40); donor-recipient gender match (No gender mismatch, Gender mismatch) and GvHD prevention: T-celll depletion (No TCD, TCD), reader's are referred to Putter (2011) for details. Interest is on the risk prediction of relapse or death through two paths.

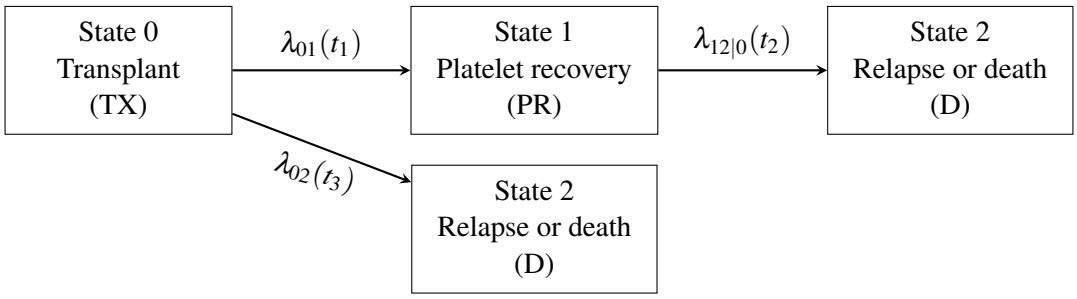


FIGURE 2.4: Progressive multi-state model for EBMT data.

For comparison, prediction is made using five methods: (i) First, we estimated cumulative baseline hazards using the model in the paper by Putter *et al.* (2007) and predicted probabilities using the 'mstate' package as shown in Putter (2011). Both SAS and R package 'mstate' are used for estimation and predictions.

(ii) SAS is used to fit the proposed model and cumulative hazards for all transitions are estimated by setting the covariates value to their reference value to demonstrate the method. Estimated coefficients of cause-specific hazards for all three transitions are presented in the upper panel of Table 2.1. Since Platelet recovery (PR) and Relapse or Death (D) in

stage 1 are competing risk, cumulative incidence function (CIF) is estimated as follows:

$$\begin{aligned}
\hat{Q}_{01}(t_1) &= \int_0^{t_1} \hat{\lambda}_{01}(s_1) exp\left(-\int_0^{s_1} \left[\hat{\lambda}_{01}(v_1) + \hat{\lambda}_{02}(v_1) \right] dv_1 \right) ds_1 \\
&= \int_0^{t_1} \hat{\lambda}_{01}(s_1) \hat{S}(s_1^-) ds_1, \text{ where} \\
\hat{S}(s_1) &= exp\left(-\int_0^{s_1} \left[\hat{\lambda}_{01}(v_1) + \hat{\lambda}_{02}(v_1) \right] dv_1 \right). \\
\hat{Q}_{02}(t_3) &= \int_0^{t_3} \hat{\lambda}_{02}(s_3) \hat{S}(s_3^-) ds_3 \text{ and} \\
\hat{Q}_{12|0}(t_2) &= \int_0^{t_2} \hat{\lambda}_{12|0}(s_2) exp\left(-\int_0^{s_2} \hat{\lambda}_{12|0}(v_2) dv_2 \right) ds_2 \\
&= \int_0^{t_2} \hat{\lambda}_{12|0}(s_2) \hat{S}_{12|0}(s_2^-) ds_2.
\end{aligned}$$

(iii) Kaplan-Meier (KM) method (Kaplan and Meier, 1958) is used to estimate the survival and cumulative hazard function assuming independence between two competing risk. The transition probability is estimated as follows:

$$\begin{aligned}
\hat{Q}'_{01}(t_1) &= \sum_{s_1 : s_1 \le t_1} \hat{\lambda}_{01}(s_1) \hat{S}_{01}(s_1^-), \; \hat{Q}'_{02}(t_3) = \sum_{s_3 : s_3 \le t_3} \hat{\lambda}_{03}(s_3) \hat{S}_{03}(s_3^-) \text{ and} \\
\hat{Q}'_{12|0}(t_2) &= \sum_{s_2 : s_2 \le t_2} \hat{\lambda}_{12|0}(s_2) \hat{S}_{12|0}(s_2^-).
\end{aligned}$$

(iv) Again, Kaplan-Meier method is used to estimate the survival and cumulative hazard function. However, cumulative incidence function (CIF) is estimated as follows

$$\begin{aligned}
\hat{Q}_{01}(t_1) &= \sum_{s_1 : s_1 \le t_1} \hat{\lambda}_{1|0}(s_1) \hat{S}(s_1^-), \text{ and } \hat{Q}_{02}(t_3) = \sum_{s_3 : s_3 \le t_3} \hat{\lambda}_{02}(s_3) \hat{S}(s_3^-) \text{ where} \\
\hat{S}(t) &= \hat{S}_{01}(t) + \hat{S}_{02}(t).
\end{aligned}$$

(v) Finally, sub-distribution hazards were estimated by changing the time of competing events higher than the event time in the study (Fine and Gray, 1999) and the estimates are presented in the lower panel of Table 2.1.

Disease classification, age at transplantation and GvHD prevention is found to be significant ($p < 0.05$) for transition $TX \rightarrow PR$ as shown in Table 2.1. However, for transition $TX \rightarrow Rel/Death$ age at transplantation and GvHD prevention is found to be significant

TABLE 2.1: Parameter Estimates for different transitions using EBMT data.

Variables	TX->PR		TX->Rel/Death		PR->Rel/Death	
	Coef. (SE)	P	Coef. (SE)	P	Coef. (SE)	P
Cox proportional hazard models						
Disease classification						
AML						
ALL	-0.044 (0.078)	0.576	0.256 (0.135)	0.058	0.120 (0.148)	0.416
CML	-0.297 (0.068)	0.0001	0.017 (0.108)	0.877	0.252 (0.117)	0.031
Age at transplantation						
$\leqslant$ 20						
20-40	-0.165 (0.079)	0.037	0.255 (0.151)	0.091	0.066 (0.153)	0.668
$>$ 40	-0.090 (0.086)	0.299	0.526 (0.158)	0.001	0.582 (0.160)	0.0003
Donor-recipient						
No gender mismatch						
Gender mism.	0.046 (0.067)	0.492	-0.075 (0.110)	0.495	0.170 (0.115)	0.138
GvHD prevention						
No TCD						
+ TCD	0.429 (0.080)	0.0001	0.297 (0.150)	0.048	0.197 (0.126)	0.119
Fine and Gray models: sub-distribution hazard						
Disease classification						
AML						
ALL	-0.062 (0.078)	0.423	0.268 (0.135)	0.048		
CML	-0.274 (0.068)	0.000	0.273 (0.108)	0.012		
Age at transplantation						
$\leqslant$ 20						
20-40	-0.189 (0.079)	0.017	0.403 (0.150)	0.007		
$>$ 40	-0.130 (0.086)	0.133	0.587 (0.158)	0.000		
Donor-recipient						
No gender mismatch						
Gender mism.	0.058 (0.067)	0.386	-0.098 (0.110)	0.374		
GvHD prevention						
No TCD						
+ TCD	0.428 (0.080)	0.000	-0.226 (0.150)	0.131		

($p < 0.05$). For transition $PR \rightarrow Rel/Death$ disease classification, age at transplantation is significant ($p < 0.05$). It should be noted that same multistate hazard model is used for both proposed and Putter's method. Estimates from Fine and Gray's model for transition $TX \rightarrow PR$ were similar with higher significant level.

Figure 2.5 presents the cumulative baseline hazards for all three transition using cause-specific hazards model. Differing baseline hazard for transitions $TX \rightarrow PR$ and $TX \rightarrow Rel/Death$ justifies stratification. Predicted transition probabilities of trajectories and segment of trajectories for different scenarios using estimates from all five methods are plotted in the graphs. Predicted transition probabilities for $TX \rightarrow Rel/Death$ for different time points is displayed in Figure 2.6. Predicted probabilities for proposed and Puter's

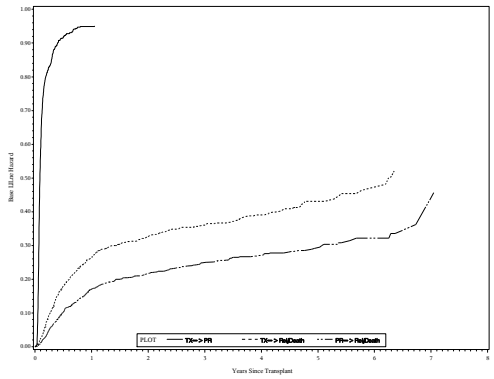


FIGURE 2.5: Cumulative baseline hazards for three transitions.

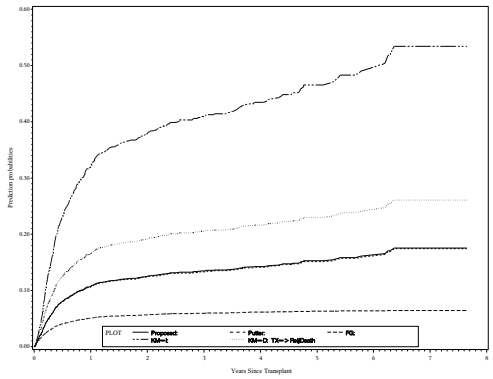


FIGURE 2.6: Predicted probability of TX→Rel/Death.

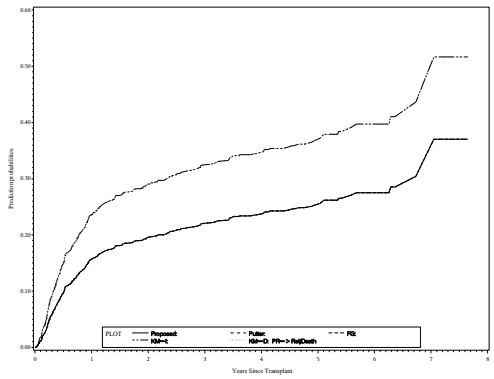


FIGURE 2.7: Predicted conditional probability of PR→Rel/Death.

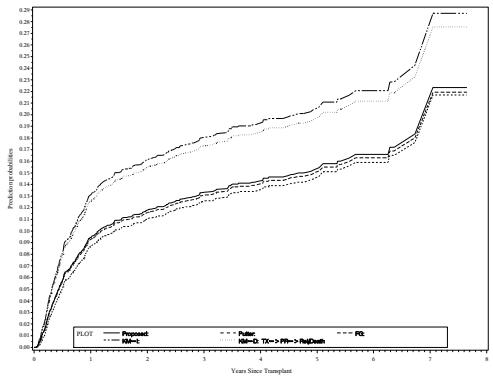


FIGURE 2.8: Predicted probability of trajectory TX→PR→Rel/Death.

approach coincides, the second line from bottom. The top line is based on the KM estimates assuming independence and the second line from the top is based on cumulative incidence function considering survival from all competing events. This difference is because of dependencies among the competing events. However, the predicted risk is much lower based on estimates from Fine and Gray approach. This is because 1159 cases from competing event PR remains in the risk set forever, which decreases this predicted risk.

Figure 2.7 shows the conditional predicted risk of relapse or death given the platelet recovery. The bottom line of figure using proposed, Putter and Fine & Gray approach coincides as expected, because of no competing events. The top line is based on KM method assuming both independence and using CIF. However, this prediction is overestimated based on KM hazards. Predicted risk of both platelet recovery and relapse or death (TX→PR→Rel/Death) are presented in Figure 2.8. The third line from the bottom based on proposed approach that very closely follows that of Putter (second line from bottom) and the first line from the bottom based on Fine and Gray estimates. The first line from top based on KM assuming independence and the second line using CIF those overestimates the predicted risk for both events.

Figure 2.9 displays the overall death through both paths. The first line from the bottom

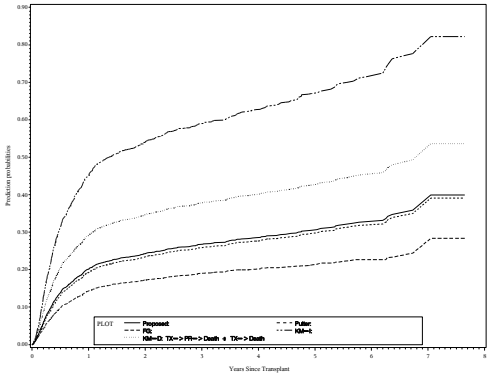


FIGURE 2.9: Total predicted probability of Rel/Death through two paths.

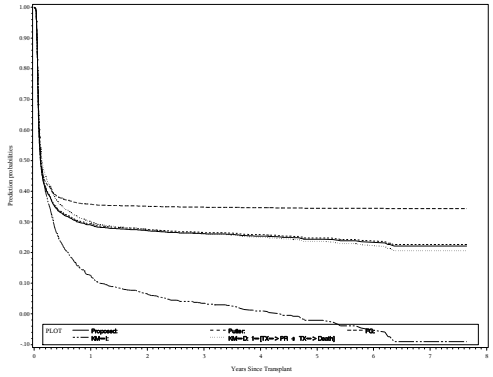


FIGURE 2.10: Predicted probability of being in state TX.

based on Fine and Gray approach; the second line is the proposed one, and the third one is using Putter's approach. Prediction using KM estimates assuming independence (first line from the top) are much higher as seen in the graph. The second line using CIF gives a closer prediction, though higher than the proposed one. Fine and Gray shows lowest predicted risks.

The predicted probability of being in state TX (0) is presented in Figure 2.10. Proposed and Putters coincides third and second line from the top closely followed by an estimate from KM assuming dependence, fourth line from the top. But using Fine and Gray's model predicted risks (first line from the top) are over estimated. Predicted risk using KM assuming independence goes to negative teritory at the tail due to dependent competing events. The predicted risk of having platelet recovery and being alive is shown in Figure 2.11. The proposed (top line) and Putters (second top) coincide and closely followed by Fine and Gray (third from top). The reason is only 458 patients (TX-> Relapse or Death) from competing events are considered as censored those remains in the risk set forever. The last two line using KM method underestimate this probability.

Predicted transition probability of trajectory or segment of a trajectory using proposed method either coincides to that using Putter's method or follows very closely. The same predicted risk using Fine and Gray's method is also similar. However, estimate using KM assuming independence overestimate the risk much higher rate compares to that from CIF.

2.4.1 Computational procedure

Steps involved in the computation of risk prediction are explained in this section. Tables 2.2, 2.3 and 2.4 provide examples of calculations based on the proposed method. All computations are performed using SAS. Table 2.2 shows the first few event times for three transitions and corresponding cumulative baseline hazards. Events for three transitions occurred at different time points (Table 2.2). To make computations simple event

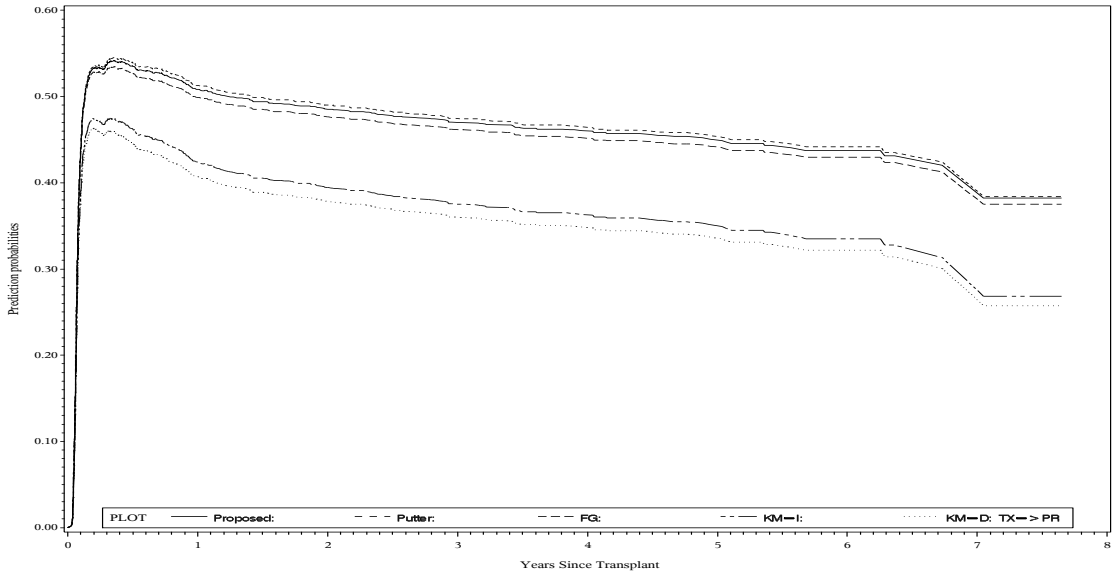


FIGURE 2.11: Predicted probability of being in state PR and being alive.

time points from all three transitions are combined considering unique time points (Table 2.3). This table provides the corresponding cumulative baseline hazards and all other quantities needed in the risk prediction, which is continued in Table 2.4. The time point for which there is no event for a transition cumulative baseline hazard is repeated. Difference between two time points $d\hat{\Lambda}_1(\tau_1) = \hat{\Lambda}_1(\tau_1) - \hat{\Lambda}_1(\tau_1-) > 0$ provides the estimates of $\hat{\lambda}(t)$ for a transition. Using this cumulative baseline hazard and hazard, the estimate of the survival probability for each transition and overall survival $S(t)$ in the presence of competing events and probability density function $f(t)$ are obtained using the equations shown earlier. Based on these quantities the prediction for a trajectory or segment is easily computed.

TABLE 2.2: Cumulative baseline hazard using SAS for three transitions.

TX->PR		TX->Rel/Death		PR->Rel/Death	
Time (τ)	$\hat{\Delta}_0$	Time (τ)	$\hat{\Delta}_0$	Time (τ)	$\hat{\Delta}_0$
1	0.00052777	1	0.00030470	5	0.00053859
3	0.00105609	4	0.00060974	10	0.00107757
7	0.00158576	6	0.00122040	12	0.00215679
9	0.00211674	7	0.00183162	14	0.00269708
10	0.00318014	8	0.00244385	17	0.00323805
11	0.00371233	9	0.00305655	20	0.00432073
12	0.00637587	11	0.00336354	21	0.00486320
13	0.01172665	12	0.00489959		
		13	0.00520808		

2.5 Maternal morbidity example

As a second example, the data on maternal morbidity is used to demonstrate the proposed method of prediction. Two stages (stage 1 and stage 2) are considered. The antenatal complications are: hemorrhage, edema, excessive vomiting, fits/convulsion; and delivery complications: excessive hemorrhage before or after delivery, retained placenta, obstructed labor, prolonged labor, other complications. Study sample comprises 993 pregnant women. Among them, 485 women were free from antenatal complications while 508 women suffered antenatal complications during stage 1. During stage 2 among 485 women, 364 women were free from delivery complications while 115 women suffered delivery complications. Out of 508 women with antenatal complications, 342 were complications free during delivery while 160 women suffered delivery complications. It should be noted that there is no competing risk for this example.

To keep the illustration simple only four explanatory variables are considered for different transitions, those are Z_1: whether the index pregnancy was desired or not, Z_2: age at marriage (15 years or lower, more than 15 years), Z_3: number of pregnancies prior to the current pregnancy (0, 1+) and Z_4: educational attainment of respondent (no education, primary or higher). The estimate of regression parameters for three transitions are presented in Table 2.5. Age at marriage with sixteen years or above decreases antenatal complications ($0 \rightarrow 1$), it is significant at ten percent level. Delivery complications given that antenatal complication has occurred ($0 \rightarrow 1 \rightarrow 1$) is reduced significantly if the pregnancy is desired, i.e., planned. This transition rate is also significantly lower for women with primary or higher education. Without antenatal complications, delivery complications ($0 \rightarrow 0 \rightarrow 1$) is lowered significantly for women with the experience of previous pregnancies. Figure 2.12 displays the cumulative baseline hazards for transitions $0 \rightarrow 1$ and $0 \rightarrow 0$ (i.e., censoring distribution) in stage-1. Conditional on stage 1, the cumulative baseline hazards for delivery complication for two transitions ($0 \rightarrow 1 \rightarrow 1$ and $0 \rightarrow 0 \rightarrow 1$) in stage-2 are presented in Figure 2.13. It is evident from the figure that the baseline hazards are different.

Predicted risk of antenatal, delivery and both complications together are estimated for a woman without education with the value of $Z_1 = Z_2 = Z_3 = Z_4 = 0$ and for a woman who has primary or higher education, i.e., $Z_1 = Z_2 = Z_3 = 0$ and $Z_4 = 1$. For example, the predicted probability of antenatal complications by 2.61 months for a woman with no education and a woman with primary or higher education are 0.003215 and 0.003376, respectively. This is expected as education was not significant for the transition $0 \rightarrow 1$. Similarly, the predicted conditional probability of delivery complications ($0 \rightarrow 1 \rightarrow 1$) given that antenatal complications has occurred by time 2.85 months for women without and with education are 0.959179 and 0.8435461, respectively. Which is expected as education affects this transition significantly. This predicted probability for different time points are shown in figure 2.14.

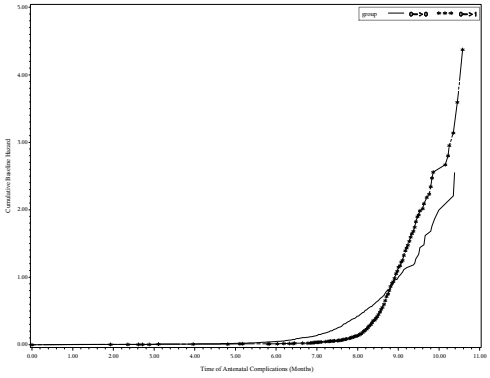


FIGURE 2.12: Baseline hazards of $0 \rightarrow 1$ transitions and $0 \rightarrow 0$transitions.

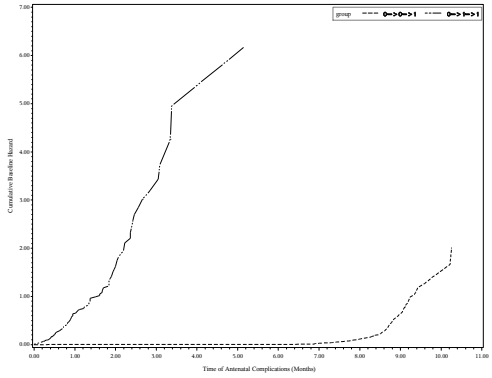


FIGURE 2.13: Baseline hazards of $0 \rightarrow 0 \rightarrow 1$ and $0 \rightarrow 1 \rightarrow 1$ transitions.

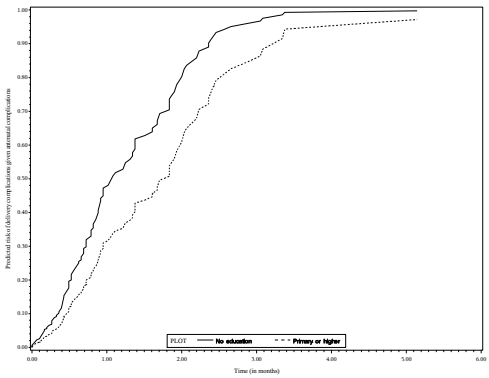


FIGURE 2.14: Predicted probability of delivery complications given antenatal complications.

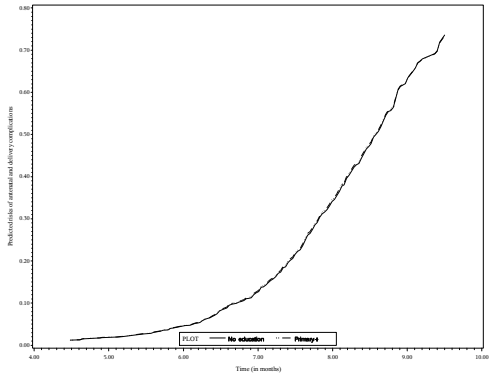


FIGURE 2.15: Predicted risk of both complications ($0 \rightarrow 1 \rightarrow 1$) by education.

The predicted probability of both antenatal and delivery complications i.e., the probability of both events are estimated next. The delivery complication by $t = 9.5$ month is fixed. Then the time of antenatal complications is varied ($4.16 \leq t_1 \leq 9.5$) and using equation (2.15) the joint probabilities are calculated for different values of t_1. The predicted risks of both antenatal and delivery complications for two categories of education are shown in figure 2.15 and both the lines coincides. The influence of education in stage 2 ($p<0.01$) disappear while predicting both events together. In the figure 2.16 the predicted risk of both events by desired or undesired pregnancy are shown. The bottom line of this figure displays antenatal complications for women with desired pregnancy, and this complication is lower than that of undesired one (the top line).

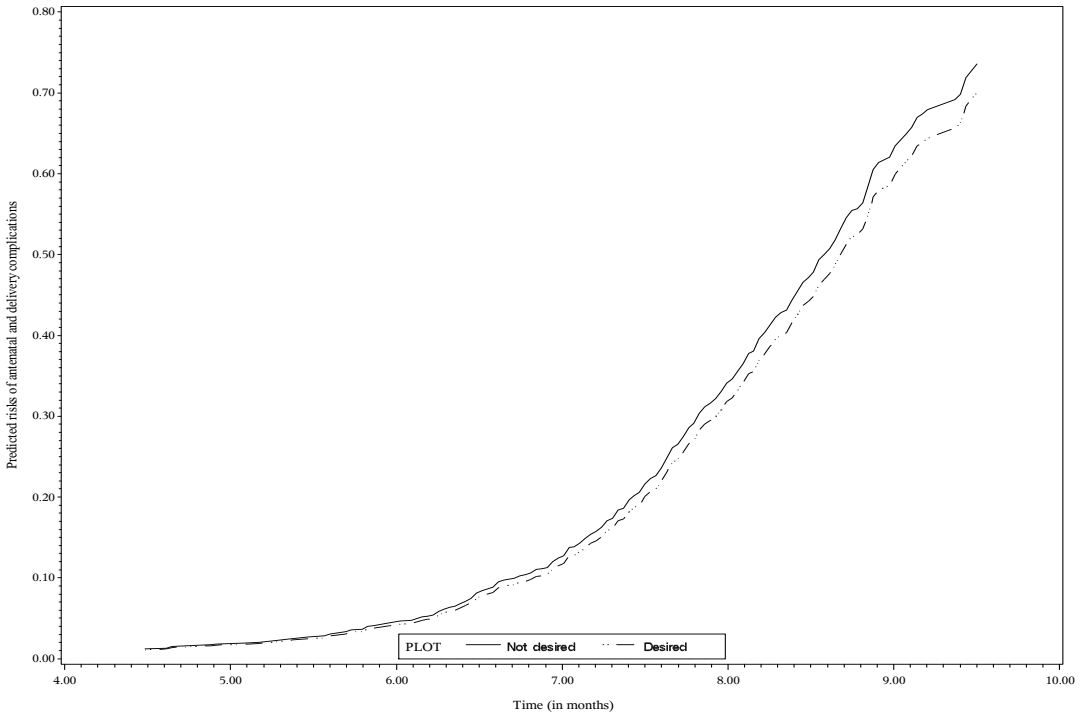


FIGURE 2.16: Predicted risk for antenatal and delivery complications for $0 \rightarrow 1 \rightarrow 1$ transition by desired pregnancy.

2.6 Conclusions

In this study, an alternative multistage procedure is developed in order to simplify the transition models for the underlying trajectories for risk prediction. The proposed alternative provides the estimates for each stage in the process conditionally and the conditional estimates is linked based on marginal-conditional models in order to provide the joint probabilities needed for predicting the status of disease based on the potential risk factors. This simplification will allow any number of intermediate stages without making the theory complicated. As compared to the existing method (Putter:2006) the proposed method provides a generalization that can be employed for any prediction model for any sequence of occurrence of events longitudinally. It may be noted here that previous methods require problem-specific modeling and a generalization cannot be shown due to lack of exposition of the probabilities of survival and failure over segments and linking the probabilities in a general form to arrive at the risk prediction of trajectories. On the other hand, the probabilities in segments as well as transitions to various states are expressed coherently for trajectories in a general form for a sequence of events in the proposed method using the multistage approach. Hence, the models for different transitions can be handled conveniently without making the exposition difficult for estimation and test of hypothesis.

The proposed method of prediction is a new development using a series of events in conditional setting arising from the beginning to the endpoint. This method used a marginal-conditional approach to link the events occurring in the trajectory. The main improvement of the proposed method is that it is simple, as a general form of integral is developed for predicting the joint probability of a sequence of events from longitudinal studies for (i) different types of trajectories and (ii) any segment of a trajectory along with the generalization to any number of stages. The timescale is easier to understand as it starts from zero for each transition in a stage (clock-reset method). A simple method of risk prediction from multi-state models for continuous time data is demanding. The proposed methods can be applied in many field of studies such as epidemiology, public health, survival analysis, genetics, reliability, environmental studies, etc.

TABLE 2.3: Computational steps for prediction of probability.

Time (τ)	$\hat{\Delta}_{0,12}(t)$	$\hat{\lambda}_{12}(t) = d\hat{\Lambda}_{0,12}(t)$	$\hat{S}_{12}(t)$	$\hat{f}_{12}(t) = \hat{\lambda}_{12}(t) \times \hat{S}(t-)$	$\hat{F}_{12}(t)$	$\hat{\Delta}_{0,14}(t)$	$\hat{\lambda}_{14}(t) = d\hat{\Lambda}_{0,14}(t)$	$\hat{S}_{14}(t)$	$\hat{f}_{14}(t) = \hat{\lambda}_{14}(t) \times \hat{S}(t-)$	$\hat{F}_{14}(t)$
1	2	3	4	5	6	7	8	9	10	11
0	0.00000	0.00000	1.00000	0.0000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1	0.00053	0.00053	0.99947	0.0005	0.00053	0.00030	0.00030	0.99970	0.00030	0.00030
3	0.00106	0.00053	0.99894	0.0005	0.00106	0.00030	0.00000	0.99970	0.00000	0.00030
4	0.00106	0.00000	0.99894	0.0000	0.00106	0.00061	0.00031	0.99939	0.00030	0.00061
5	0.00106	0.00000	0.99894	0.0000	0.00106	0.00061	0.00000	0.99939	0.00000	0.00061
6	0.00106	0.00000	0.99894	0.0000	0.00106	0.00122	0.00061	0.99878	0.00061	0.00122
7	0.00159	0.00053	0.99842	0.0005	0.00158	0.00183	0.00061	0.99817	0.00061	0.00183
8	0.00159	0.00000	0.99842	0.0000	0.00158	0.00244	0.00061	0.99756	0.00061	0.00244
9	0.00212	0.00053	0.99789	0.0005	0.00211	0.00306	0.00061	0.99695	0.00061	0.00305
10	0.00318	0.00106	0.99682	0.0011	0.00317	0.00306	0.00000	0.99695	0.00000	0.00305
11	0.00371	0.00053	0.99629	0.0005	0.00370	0.00336	0.00031	0.99664	0.00031	0.00335
12	0.00638	0.00266	0.99364	0.0026	0.00634	0.00490	0.00154	0.99511	0.00153	0.00488
13	0.01173	0.00535	0.98834	0.0053	0.01164	0.00521	0.00031	0.99481	0.00031	0.00518

Where, $S(t) = \hat{S}_{12}(t) \times \hat{S}_{14}(t)$, is the overall survival from both transitions TX$-$ >PR and TX$-$ > Relapse/Death. $\hat{S}_{12} = \exp[-\hat{\Delta}_{0,12}(t)]$; $\hat{S}_{14}(t) = exp[-\hat{\Delta}_{0,14}(t)]$.

TABLE 2.4: Continued...

Time (τ)	$\hat{\Delta}_{0,23}(t)$	$\hat{\lambda}_{23}(t) = d\hat{\Lambda}_{0,23}(t)$	$\hat{S}_{23}(t) = exp[-\hat{\Delta}_{0,23}(t)]$	$\hat{f}_{23}(t) = \hat{\lambda}_{23}(t) \times \hat{S}_{23}(t-)$	$\hat{F}_{23}(t)$	$TX \to PR$	$TX \to PR \to Rel.Death$		
	12	13	14	15	16	17	18	19	20
0	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000
1	0.00000	0.00000	1.00000	0.00000	0.00000	0.00053	0.00000	0.00030	0.99917
3	0.00000	0.00000	1.00000	0.00000	0.00000	0.00106	0.00000	0.00030	0.99864
4	0.00000	0.00000	1.00000	0.00000	0.00000	0.00106	0.00000	0.00061	0.99834
5	0.00054	0.00054	0.99946	0.00054	0.00054	0.00106	0.00000	0.00061	0.99834
6	0.00054	0.00000	0.99946	0.00000	0.00054	0.00106	0.00000	0.00122	0.99773
7	0.00054	0.00000	0.99946	0.00000	0.00054	0.00158	0.00000	0.00183	0.99659
8	0.00054	0.00000	0.99946	0.00000	0.00054	0.00158	0.00000	0.00244	0.99598
9	0.00054	0.00000	0.99946	0.00000	0.00054	0.00211	0.00000	0.00305	0.99484
10	0.00108	0.00054	0.99892	0.00054	0.00108	0.00317	0.00000	0.00305	0.99378
11	0.00108	0.00000	0.99892	0.00000	0.00108	0.00370	0.00000	0.00336	0.99295
12	0.00216	0.00108	0.99785	0.00108	0.00216	0.00633	0.00001	0.00489	0.98878
13	0.00216	0.00000	0.99785	0.00000	0.00216	0.01161	0.00003	0.00521	0.98318

Col. (17)=$TX \to PR = \hat{F}_{12}(t) \times \hat{S}_{12}(t)$; Col. (18)=$TX \to PR \to Rel./Death = \hat{F}_{12}(t) \times \hat{F}_{23}(t)$
Col. (19) $TX \to PR \to Rel./Death + TX \to Rel./Death$; Col. (20)= $1 - [\hat{F}_{12}(t) + \text{TX} \to$ Rel./Death]

TABLE 2.5: Estimates for three transitions from multi-state model.

Variables	β	S.E	p-value	HR
Antenatal complication: $0 \rightarrow 1$				
Desired pregnancy	-0.06277	0.09669	0.5162	0.939
Age at marriage	-0.15935	0.09670	0.0994	0.853
Number of previous pregnancy	0.01528	0.10782	0.8873	1.015
Education	0.04893	0.09098	0.5907	1.050
Delivery complication after antenatal complication: $0 \rightarrow 1 \rightarrow 1$				
Desired pregnancy	-0.46601	0.18861	0.0135	0.628
Age at marriage	0.10864	0.17543	0.5357	1.115
Number of previous pregnancy	0.13674	0.18933	0.4702	1.147
Education	-0.54482	0.17047	0.0014	0.580
Delivery complication given no antenatal complication: $0 \rightarrow 0 \rightarrow 1$				
Desired pregnancy	-0.13080	0.21838	0.5492	0.877
Age at marriage	-0.26930	0.20024	0.1787	0.764
Number of previous pregnancy	-0.61007	0.19948	0.0022	0.543
Education	0.32378	0.19371	0.0946	1.382
Model χ^2 (p-value, d.f.)	37.1 (0.0002,12)			

Chapter 3

Regressive Models for Risk Prediction for a Sequence of Multinomial Outcomes

3.1 Introduction

In longitudinal, cohort or panel studies repeated outcomes are observed along with various risk factors on the same subjects. The outcome variables may be discrete or continuous. If the outcome variables are categorical, either nominal or ordinal, then the sequence of outcomes for each subject may produce outcomes that may follow multivariate binary or multivariate multinomial distributions. For example, uncontrolled diabetes can lead to nephropathy, diabetic retinopathy, pulmonary tuberculosis, and coronary heart disease that may be observed longitudinally in a large number of follow-ups producing nominal outcomes. Activities of daily living (ADL) indices measure functional limitations as ordinal outcomes. Repeated outcomes over time may produce a large number of trajectories as transitions between states may occur at different follow-ups. A multinomial outcome with three states from three follow-ups produces a total of twenty-seven trajectories (paths) as shown in Figure 3.1. A growing area of interest is to understand the disease progression over time, predict risks of a sequence of events with risk factors and status at previous outcomes (Wen *et al.*, 2016; Islam and Chowdhury, 2010; Putter *et al.*, 2006, 2007; Rothman, 2002; Klein *et al.*, 1994). With the risk quantification of a sequence of events, health care providers could screen individuals that would help them to suggest necessary therapy and prevention (Tripepi *et al.*, 2013). Risk prediction would also allow a patient to be aware of the future course of disease (Tripepi *et al.*, 2013).

Prediction of risk of a sequence of events for multinomial outcomes with specified predictors is a challenge to the researchers. To understand this process we need to examine the sequence of events during subsequent follow-ups. One need to deal with transitions to a number of states over time generating a large number of trajectories from beginning to the end of the study (Figure 3.1). With the increased number of follow-ups, this problem

becomes even more difficult to model. For survival data, Klein *et al.* (1994) illustrated risk prediction for a sequence of events based on the work of Arjas and Eerola (1993). Putter *et al.* (2006) and Putter *et al.* (2007) provided a comprehensive illustration of risk prediction for a sequence of events using multistate modeling framework, among others. For discrete survival time data, two commonly used regression models are the grouped proportional hazard models (Pierce *et al.*, 1979; Prentice and Gloeckler, 1978) and the logistic model (Lawless, 2003). D'Agostino *et al.* (1990) illustrated pooled logistic regression model for discrete time data.

The multistate higher order Markov model is a natural choice to study the underlying property of dependence in consecutive follow-ups (Islam *et al.*, 2009). Using this model one can investigate the relationship between outcomes and predictors and risk could be calculated for a sequence of events (Islam *et al.*, 2012). For better prediction, it is important to understand how the transitions between states occur and how the covariates influence these transitions. However, Markov chain models appear to be restricted due to over-parameterization (Islam *et al.*, 2013). For example, outcomes from three follow-ups with three categories (Figure 3.1) one need to fit thirteen models, one marginal model for the outcome at follow-up one or baseline, three first order and nine second order Markov models. Outcomes from large number of follow-ups would be intractable and computationally infeasible (Wen *et al.*, 2016). Generalized estimating equations (GEE), a marginal model, is used for model parameter estimation for correlated outcomes. Yu (2003) used first-order Markov transition model to evaluate the impact of risk factors on longitudinal back-pain data for ordinal outcomes.

Another class of models is the regressive model. Muenz and Rubinstein (1985), Bonney (1986, 1987), Azzalini (1994), Islam *et al.* (2004), Islam and Chowdhury (2006), Islam and Chowdhury (2010) and Islam *et al.* (2014) proposed regressive logistic models under the Markovian assumptions to include both binary outcomes in previous times in addition to covariates in the conditional models. The regressive model for binary outcomes proposed by Bonney (1986, 1987) is extended by Islam and Chowdhury (2010). A framework to predict joint probability of a sequence of events for binary outcomes from repeated measures based on specified risk factors is proposed by Islam and Chowdhury (2010). In this paper, we extended the regressive models for multinomial outcomes from repeated measures and proposed a framework for risk prediction for a sequence of events for specified covariate values. The proposed framework links the conditional process and obtains predictive outcome based on the whole process through all possible trajectories. This model allows to include interaction between previous outcomes and covariates in the model as long as sample size permits.

3.2 Models

The method for risk prediction for a trajectory is based on the proposed regressive models for multinomial outcomes. Figure 3.1 displays the transitions between three outcome levels from three follow-ups. Outcome levels (0,1,2) are denoted inside the rectangles. Here, first column shows marginal probabilities and second onward are conditional probabilities.

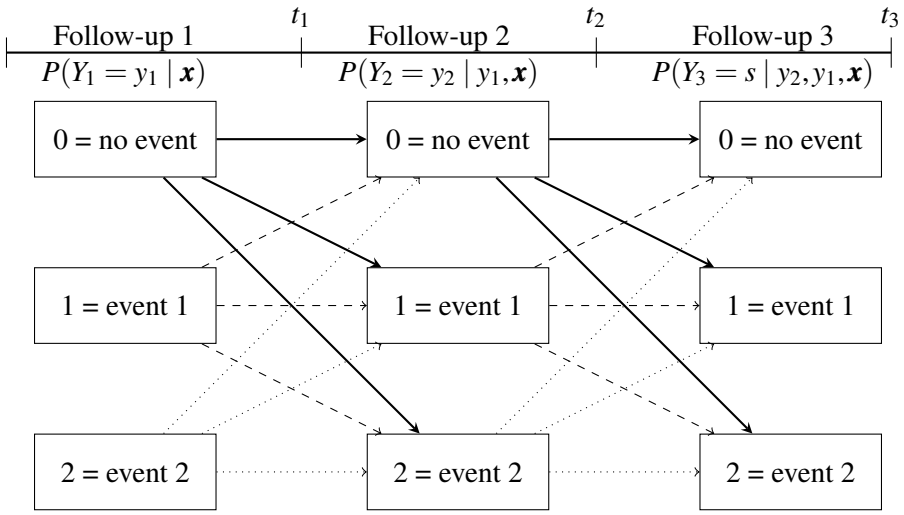


FIGURE 3.1: Transitions between states for regressive models.

3.2.1 Notations

Let $Y_{i1}, Y_{i2}, ..., Y_{iJ_i}$ represent the past and present responses for *i-th* subject at *j-th* follow-up where $(i = 1, 2, ..., n)$ and $(j = 1, 2, ..., J_i)$, J_i is the number of follow-ups for subject i. For simplicity, subscript i is omitted what follows next unless explicitly specified. Assume, $Y_j = s$ follows multinomial distribution where $(s = 0, 1, 2, ..., S)$ with $S+1$ outcome categories. The category 0 denotes non-event.

The joint probability mass function of $Y_1, Y_2, ..., Y_J$ with covariate vector $\boldsymbol{X} = \boldsymbol{x}$ can be expressed as:

$$\begin{aligned} & P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{X} = \boldsymbol{x}) \\ & \quad = P(Y_1 = y_1 \mid \boldsymbol{X} = \boldsymbol{x}) \times P(Y_2 = y_2 \mid Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) \\ & \quad \times ... \times P(Y_J = s \mid Y_{j-1} = y_{j-1}, ..., Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) \\ & \quad = P_{y_1}(\boldsymbol{x}) \times P_{y_2 \cdot y_1}(\boldsymbol{x}) \times ... \times P_{s \cdot y_{j-1}, ..., y_1}(\boldsymbol{x}), \end{aligned} \tag{3.1}$$

where $\boldsymbol{X}' = [1, x_1, ..., x_p]$ is vector of covariates for a subject at first follow-up. It should be noted that $\boldsymbol{X} = \boldsymbol{x}$ can be time dependent.

Explanations of the functions of the right hand side in equation (3.1) are as follows:

$P(Y_1 = s \mid \boldsymbol{X} = \boldsymbol{x}) = P_s(\boldsymbol{x})$ is the marginal probability function of Y_1 conditional on $\boldsymbol{x}$;

$P(Y_J = s \mid Y_{j-1} = y_{j-1}; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1}}(\boldsymbol{x})$ is the probability function of Y_j conditional on y_{j-1} and $\boldsymbol{x}$ of order one;

$P(Y_J = s \mid Y_{j-1} = y_{j-1}, Y_{j-2} = y_{j-2}; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1},y_{j-2}}(\boldsymbol{X} = \boldsymbol{x})$ is the probability function for Y_j conditional on y_{j-1},y_{j-2} and $\boldsymbol{x}$ of order two;

$P(Y_J = s \mid Y_{j-1} = y_{j-1}, Y_{j-2} = y_{j-2}, ..., Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1},y_{j-2},...,y_1}(\boldsymbol{x})$ is the probability function of Y_j conditional on $y_{j-1}, ..., y_1$ and $\boldsymbol{x}$ of order $k = j-1$.

The unconditional probabilitiy of the left hand side of equation (3.1) is defined as:

$P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{X} = \boldsymbol{x}) = P_{y_1,y_2,...,y_j}(\boldsymbol{x})$.

3.2.2 Multinomial logistic regression (Marginal model)

For simplicity, let us consider outcomes with three categories ($s = 0, 1, 2$) as shown in Figure 3.1, a natural choice is multinomial logistic regression to model outcome (Y_1) as a function of covariates vector $\boldsymbol{X} = \boldsymbol{x}$. Two sets of parameters would be estimated ($Y_1 = 1$ vs. $Y_1 = 0$ and $Y_1 = 2$ vs. $Y_1 = 0$). Then the marginal model $P(Y_1 = y_1 \mid \boldsymbol{Z})$ can be shown as:

$$P_s(\boldsymbol{Z}) = P(Y_1 = s \mid \boldsymbol{Z}) = \frac{e^{(\boldsymbol{Z}'\boldsymbol{\beta}_s)}}{\sum_{s=0}^{2} e^{(\boldsymbol{Z}'\boldsymbol{\beta}_s)}} = \frac{e^{g_s(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_s(\boldsymbol{Z}')}}, \quad s = 0, 1, 2, \tag{3.2}$$

$$\text{where } g_s(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_1=s|\boldsymbol{Z})}{P(Y_1=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_s(\boldsymbol{Z})$ is the first logit of s-*th* component of y_1 conditional on $\boldsymbol{Z}$ and

$$g_s(\boldsymbol{Z}) = \beta_{s0} + \beta_{s1}Z_1 + ... + \beta_{sp}Z_p, \quad s = 1, 2,$$

where $\boldsymbol{Z'} = [1, Z_1, ..., Z_p] = \boldsymbol{X}' = [1, X_1, ..., X_p]$ and $\boldsymbol{\beta'_s} = [\beta_{s0}, \beta_{s1}, ..., \beta_{sp}]$ are the parameter vectors of the *s-th* component for outcome Y_1 where $\boldsymbol{\beta'_1} = [\beta_{10}, \beta_{11}, ..., \beta_{1p}]$ and $\boldsymbol{\beta'_2} = [\beta_{20}, \beta_{21}, ..., \beta_{2p}]$, $\boldsymbol{\beta'_1}$ and $\boldsymbol{\beta'_2}$ are $1 \times (p+1)$ vectors totalling a $[(p+1)2]$ regression coefficients.

For ordinal outcome, the proportional odds model could be used. However, proportional odds assumption should be checked from the data which in many instances are not attainable. Hence to model ordinal outcome, a multinomial logistic regression model is one of the choices among other alternatives. Multinomial logistic regression disregards the ordering of the outcome levels.

3.2.3 Proposed first order multinomial regressive model

Consider a subject moves between states from $Y_1 = y_1$ at time t_1 to $Y_2 = y_2$ at time t_2 as shown in Figure 3.1. Then the first order regressive model $P(Y_2 = y_2 \mid Y_1 = y_1, \boldsymbol{Z})$ can be shown as:

$$P_{s.y_1}(\boldsymbol{Z}) = P(Y_2 = s \mid Y_1 = y_1, \boldsymbol{Z}) = \frac{e^{g_{s.y_1}(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_{s.y_1}(\boldsymbol{Z}')}}, \quad s, y_1 = 0, 1, 2, \tag{3.3}$$

$$\text{where } g_{s.y_1}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_2=s|\boldsymbol{Z})}{P(Y_2=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_{s.y_1}(\boldsymbol{Z})$ is the second logit of s-*th* component of y_2 conditional on previous outcome y_1, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_1}(\boldsymbol{Z}) =& \beta_{s.y_1 0} + \beta_{s.y_1 1} Z_1 + \ldots + \beta_{s.y_1 p} Z_p + \beta_{s.y_1 (p+1)} Z_{p+1} \\ &+ \beta_{s.y_1 (p+2)} Z_{p+2} \quad s = 1, 2, \end{aligned}$$

where $\boldsymbol{Z'} = [1, Z_1, \ldots, Z_p, Z_{p+1}, Z_{p+2}] = [\boldsymbol{X'}, \boldsymbol{D'}] = [1, X_1, \ldots, X_p, D_{11}, D_{12}]$. Here D_{11} and D_{12} are the dummy variables for categories 1 and 2 of outcome Y_1 with 0 as the reference category. Here $\boldsymbol{X'}$ is a $1 \times (p+1)$ and $\boldsymbol{D'}$ is a 1×2 vector producing a total of $[(p+1)+2]2$ regression coefficients.

3.2.4 Proposed second order multinomial regressive model

The second order regressive model $P(Y_3 = y_3 \mid Y_1 = y_1, Y_2 = y_2, \boldsymbol{Z})$ can be shown as:

$$P_{s.y_2,y_1}(\boldsymbol{Z}) = P(Y_3 = s \mid Y_1 = y_1, Y_2 = y_2, \boldsymbol{Z}) = \frac{e^{g_{s.y_2}(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_{s.y_2}(\boldsymbol{Z}')}}, \quad s = 0, 1, 2, \tag{3.4}$$

$$\text{where } g_{s.y_2}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_3=s|\boldsymbol{Z})}{P(Y_3=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_{s.y_2}(\boldsymbol{Z})$ is the third logit of s-*th* component of y_3 conditional on previous two outcomes y_1, y_2, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_2}(\boldsymbol{Z}) =& \beta_{s.y_2 0} + \beta_{s.y_2 1} Z_1 + \ldots + \beta_{s.y_2 p} Z_p + \beta_{s.y_2 (p+1)} Z_{p+1} \\ &+ \beta_{s.y_2 (p+2)} Z_{p+2} + \beta_{s.y_2 (p+3)} Z_{p+3}, + \beta_{s.y_2 (p+4)} Z_{p+4} \quad s = 1, 2, \end{aligned}$$

$$\begin{aligned} \text{where } \boldsymbol{Z'} &= [1, Z_1, \ldots, Z_p, Z_{p+1}, Z_{p+2}, Z_{p+3}, Z_{p+4}] \\ &= [\boldsymbol{X'}, \boldsymbol{D'}] = [1, X_1, \ldots, X_p, D_{11}, D_{12}, D_{21}, D_{22}]. \end{aligned}$$

Here D_{11} and D_{12} are the dummy variables for categories 1 and 2 of outcome variable Y_1 and D_{21} and D_{22} are the dummy variables for categories 1 and 2 of outcome variable Y_2 with 0 as the reference category. $\boldsymbol{X'}$ is a $1\times(p+1)$ and $\boldsymbol{D'}$ is a 1×4 vector producing $[(p+1)+4]$ regression coefficients with a total of $[(p+1)+4]2$ regression coefficients.

3.2.5 Proposed higher order multinomial regressive model

Above regressive model could be generalized for *k-th* order ($k=j-1$) for $s=0,1,2,...,S$ outcome levels as follows:

$$P_{s.y_{j-1},...,y_1}(\boldsymbol{Z}) = P(Y_j=s \mid Y_1=y_1,...,Y_{j-1}=y_{j-1},\boldsymbol{Z}) \qquad = \frac{e^{g_{s.y_{j-1}}(\boldsymbol{Z})}}{\sum_{s=0}^{S} e^{g_{s.y_{j-1}}(\boldsymbol{Z'})}},$$

$$s=0,1,2,...,S, \tag{3.5}$$

$$\text{where } g_{s.y_{j-1}}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s=0 \\ \ln\left[\frac{P(Y_j=s|\boldsymbol{Z})}{P(Y_j=0|\boldsymbol{Z})}\right] & \text{if } s=1,2,...,S, \end{cases}$$

is the *j-th* logit of the *s-th* component of y_j conditional on previous previous $j-1$ outcomes $y_1,y_2,...,y_{j-1}$, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_{j-1}}(\boldsymbol{Z}) &= \beta_{s.y_{j-1}0} + \beta_{s.y_{j-1}1}Z_1 + ... + \beta_{s.y_{j-1}p}Z_p + \beta_{s.y_{j-1}(p+1)}Z_{p+1} \\ &\quad + ... + \beta_{s.y_{j-1}(p+S)}Z_{p+S} + \beta_{s.y_{j-1}(p+S+1)}Z_{p+S+1} \\ &\quad + ... + \beta_{s.y_{j-1}(p+2S)}Z_{p+2S} + ... + \beta_{s.y_{j-1}[p+(j-1)S+1]}Z_{[p+(j-1)S+1]} \\ &\quad + ... + \beta_{s.y_{j-1}[p+(j-1)S+S]}Z_{[p+(j-1)S+S]}, \quad s=1,2,...,S, j>1 \end{aligned}$$

$$\begin{aligned} \text{where } \boldsymbol{Z'} &= [1,Z_1,...,Z_p,Z_{p+1},...,Z_{p+S}, \\ &\quad Z_{p+S+1},...,Z_{p+2S},...,Z_{[p+(j-1)S+1]},...,Z_{[p+(j-1)S+S]}\Big] = \left[\boldsymbol{X'},\boldsymbol{D'}\right] \\ &= \Big[1,X_1,...,X_p,D_{11},...,D_{1S},D_{21},...,D_{2S},...,D_{(j-1)1},...,D_{(j-1)S}\Big]. \end{aligned}$$

Here, $D_{11},...,D_{1S},D_{21},...,D_{2S},...,D_{(j-1)1},...,D_{(j-1)S}$ are the dummy variables for categories $1,2,...,S$ of outcomes $y_1,y_2,...,y_{j-1}$ with 0 as the reference category, respectively. $\boldsymbol{X'}$ is a $1\times(p+1)$ vector of covariates and $\boldsymbol{D'}$ is a $1\times[(j-1)S]$ vector of dummy variables for previous $y_{j-1},...,y_1$ outcomes with $S+1$ categories considering 0 as the reference category. There are $[(p+1)+(j-1)S]$ regression coefficients for *s-th* component of the model and with a total of $[(p+1)+(j-1)S]S$ regression coefficients.

3.2.6 Estimation

Let $\delta_s = 1$ if $Y_j = s$ otherwise $\delta_s = 0$, $s = 0,1,...,S$, is an indicator variable to identify observed levels of Y_j.

Then the likelihood function for a single subject of jth order model can be expressed as:

$$L = \prod_{s=0}^{S} \left[P_{s.y_{j-1}}(\boldsymbol{Z}) \right]^{\delta_s},$$

and the log likelihood for l-th component for a subject is given by

$$\ln L_l = \left[\sum_{s=1}^{S} \delta_s g_{s.y_{j-1}}(\boldsymbol{Z}) - \ln \left(\sum_{s=0}^{S} e^{g_{s.y_{j-1}}(\boldsymbol{Z})} \right) \right], \quad l = 1,2,...,S.$$

Differentiate with respect to the parameters and solving the following equations we obtain the likelihood estimates for S sets of parameters:

$$\frac{\partial \ln L_l}{\partial \beta_{sjq}} = \left[\delta_s - P_{s.y_{j-1}}(\boldsymbol{Z}) \right] \boldsymbol{Z}_q \quad q = 0,1,...,p.$$

Observed information matrix can be obtained using following second derivatives:

$$\frac{\partial^2 \ln L_l}{\partial \beta_{sjq} \partial \beta_{sjq'}} = -\left[P_{s.y_{j-1}}(\boldsymbol{Z}) \{1 - P_{s.y_{j-1}}(\boldsymbol{Z})\} \right] \boldsymbol{Z}'_q \boldsymbol{Z}_q,$$

and

$$\frac{\partial^2 \ln L_l}{\partial \beta_{sjq} \partial \beta_{s'jq'}} = \left[P_{s.y_{j-1}}(\boldsymbol{Z}) P_{s'y_{j-1}}(\boldsymbol{Z}) \right] \boldsymbol{Z}'_q \boldsymbol{Z}_q,$$

where $q, q' = 0,1,...,p; s,s' = 1,2,...,S; l = 1,2,...S$.

For n subjects, there will be summation over $i = 1,2,...,n$. Here, subscript i is omitted for notational convenience. The information matrix $I(\boldsymbol{\beta})$ is the $[(p+1)+(j-1)S]S \times [(p+1)+(j-1)S]S$ matrix where elements are the negative of the second derivatives and the asymptotic covariance matrix is $[I(\boldsymbol{\beta})]^{-1}$.

It may be noted that first and all higher order regressive models are equivalent to that of the marginal multinomial logistic regression models shown in equation (3.2). Regressive models for higher order shown in equation (3.5) can be estimated using appropriate data structure and usual SAS, STATA or R-package or other software capable of fitting multinomial logistic regression. It should be noted that the regressive model proposed by Bonney (1986, 1987) and Islam and Chowdhury (2010) are special cases of the model shown in equation (3.5) for s=0,1.

3.2.7 Predictive models and joint probabilities

Our objective is to predict the risks of occurring a sequence of events from repeated measures for a subject with specified covariate vector $\boldsymbol{X}^* = \boldsymbol{x}^*$. Which is the predicted risk that a subject with covariate vector $(\boldsymbol{X}^* = \boldsymbol{x}^*)$ would follow a particular trajectory as shown in the Figure 3.1.

The predicted joint probabilities of $\hat{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_j = y_j \mid \boldsymbol{x}^*)$ can be obtained as:

$$\begin{aligned}
&\hat{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{x}) \\
&\quad = \hat{P}(Y_1 = y_1 \mid \boldsymbol{x}) \times \hat{P}(Y_2 = y_2 \mid Y_1 = y_1; \boldsymbol{x}) \\
&\quad \times, ..., \times \hat{P}(Y_J = s \mid Y_{j-1} = y_{j-1}, ..., Y_1 = y_1; \boldsymbol{x}) \\
&\quad = \hat{P}_{y_1}(\boldsymbol{x}) \times \hat{P}_{y_2.y_1}(\boldsymbol{x}) \times ... \times \hat{P}_{s.y_{j-1},...,y_1}(\boldsymbol{x}).
\end{aligned} \tag{3.6}$$

Based on the equation (3.6) the predicted joint probabilities for Y_1 and Y_2 is

$$\hat{P}_{y_1,y_2}(\boldsymbol{x}) = \hat{P}(Y_1 = y_1, Y_2 = y_2 \mid \boldsymbol{x}),$$

the conditional probability for $Y_2 = s$ given Y_1 and $\boldsymbol{x}$ is

$$\hat{P}_{s.y_1}(\boldsymbol{x}) = \hat{P}(Y_2 = s, \mid Y_1 = y_1; \boldsymbol{x}),$$

and the marginal probability for Y_1 given $\boldsymbol{x}$ is

$$\hat{P}_{y_1}(\boldsymbol{x}) = P(Y_1 = y_1 \mid \boldsymbol{x}).$$

The joint probabilities can be predicted using marginal and conditional probabilities as:

$$\begin{aligned}
\hat{P}(Y_1 = y_1, Y_2 = y_2 \mid \boldsymbol{x}) = \hat{P}(Y_1 = y_1 \mid \boldsymbol{x}) \times \hat{P}(Y_2 = s, \mid Y_1 = y_1; \boldsymbol{x}) \\
\implies \hat{P}_{y_1,y_2}(\boldsymbol{x}) = \hat{P}_{y_1}(\boldsymbol{x}) \times \hat{P}_{s.y_1}(\boldsymbol{x}).
\end{aligned} \tag{3.7}$$

Then for outcomes y_1, y_2 with categories 0,1 and 2 and using equation (3.7) we can predict joint probabilities from conditional and marginal probabilities as follows:

$$\begin{aligned}
\hat{P}(Y_1 = y_1, Y_2 = s \mid \boldsymbol{x}) = \hat{P}(Y_1 = y_1; \boldsymbol{x}) \times \hat{P}(Y_2 = s \mid Y_1 = y_1; \boldsymbol{x}), \\
s = 0, 1, 2;\ y_1 = 0, 1, 2;\ j = 1, 2.
\end{aligned}$$

Similarly, for *j-th* outcomes $y_1, y_2, ..., y_j$ we can predict joint probabilities using marginal and conditional probabilities as:

$$\begin{aligned}
&\hat{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_J = s \mid \boldsymbol{x}) = \hat{P}(Y_1 = y_1; \boldsymbol{x}) \times \hat{P}(Y_2 = y_2 \mid Y_1 = y_1; \boldsymbol{x}), \\
&\times ... \times \hat{P}(Y_J = s \mid Y_{j-1} = y_{j-1}, ..., Y_1 = y_1; \boldsymbol{x}), s = 0, 1, 2;\ y_1, ...y_{j-1}, = 0, 1, 2; \\
&j = 1, 2, ..., J.
\end{aligned}$$

3.3 Steps involved in prediction

Using following steps, the marginal and conditional probabilities can be estimated from fitted marginal and regressive models for a subject with a specified covariates vector $\boldsymbol{X}^* = \left[1, X_1^*, ..., X_p^*\right]'$. Then using relation shown in equation (3.6), we can predict the risk of joint events for any number of follow-ups as follows:

(i) The predicted marginal probabilities $\hat{P}_2(\boldsymbol{x}^*); \hat{P}_1(\boldsymbol{x}^*); \hat{P}_0(\boldsymbol{x}^*)$ can be estimated from the fitted marginal model shown in equation (3.2). The first order conditional probabilities $\hat{P}_{s.y_1}(\boldsymbol{x}^*)$, $s, y_1 = 0, 1, 2$, can be estimated from the fitted first order regressive model using covariates vector $\boldsymbol{Z'} = [\boldsymbol{x}^*, D_{11}, D_{12}]'$ where $D_{11}, D_{12} = 0, 1$. For example, $\hat{P}_{2.0}(\boldsymbol{x}^*)$ is estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0]'$ in equation (3.3) for $s = 2$, $\hat{P}_{1.0}(\boldsymbol{x}^*)$ using covariates vector $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0]'$ for $s = 1$, and $\hat{P}_{0.0}(\boldsymbol{x}^*) = 1 - \hat{P}_{1.0}(\boldsymbol{x}^*) - \hat{P}_{2.0}(\boldsymbol{x}^*)$ and so on. Similarly, using appropriate covariates vector $\boldsymbol{Z}$ second order conditional probabilities can be estimated. For example, $\hat{P}_{2.00}(\boldsymbol{x}^*)$ is estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0, 0, 0]'$ in equation (3.4) for $s = 2$, and $\hat{P}_{1.00}(\boldsymbol{x}^*)$ using covariate vector $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0, 0, 0]'$ for $s = 1$.

(ii) Now using predicted marginal and conditional probabilities showed in equation (3.6) joint probabilities for events are obtained. For example, $\hat{P}_{01}(\boldsymbol{x}^*) = \hat{P}_0(\boldsymbol{x}^*) \times \hat{P}_{1.0}(\boldsymbol{x}^*)$, $\hat{P}_{001}(\boldsymbol{x}^*) = \hat{P}_0(\boldsymbol{x}^*) \times \hat{P}_{0.0}(\boldsymbol{x}^*) \times \hat{P}_{1.00}(\boldsymbol{x}^*)$ and so on.

3.4 Tests

3.4.1 Significance of the joint model

The significance of the joint model can be tested using likelihood ratio test between joint constant only model (Reduced) and joint full model (Full) as follows:

$$-2\left[\ln L_{\text{Reduced}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}) - \ln L_{\text{Full}}(\hat{\boldsymbol{\beta}})\right] \tag{3.8}$$

which is distributed asymptotically as χ^2 with $[\{(p+1)S\} + \{(p+1+S)S\} + \{(p+1+2S)S\} + ... + \{p+1+(j-1)S\}S] - jS$ degrees of freedom. Here $\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}'$ includes all the regression parameters from the constant only joint model and $\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}'$ includes all the parameters from the full joint model. Calculation of the degrees of freedom is shown in the following table. Each of the constant only marginal, first and higher order regressive models have S sets of constants as we are obtaining the number of categories minus one (S) sets of parameter estimates.

Number of parameters for different models.

Models	*Constant only*	*s-th component*	*Full*
Marginal	S	$[p+1]$	$[p+1]S$
First order regressive	S	$[p+1+S]$	$[p+1+S]S$
Second order regressive	S	$[p+1+2S]$	$[p+1+2S]S$
...	...	...	...
$j-1$ *th* order regressive	S	$[(p+1+(j-1)S)]$	$[(p+1+(j-1)S)]S$

3.4.2 Test for proportional odds assumption

If the outcome is ordinal it is common to use proportional odds model (McCullagh, 1980). Let, $P(Y_j \leq s) = \pi_0 + ... + \pi_s, s = 0,1,...,S$ where $P(Y_j \leq 0) \leq P(Y_j \leq 1) \leq ... \leq P(Y_j \leq S) = 1$. The ordinal logistic regression model can be shown as:

$$logit[P(Y_j \leq S)] = \ln\left[\frac{\pi_0+...+\pi_s}{\pi_{s+1}+...+\pi_S}\right] = \alpha_s + \beta_1 x_1 + ... + \beta_p x_p = \alpha_s + \boldsymbol{x'\beta},$$
$$s = 0,...,S-1 \text{ and } P(Y_j = s \mid \boldsymbol{x}) = P(Y_j \leq s+1 \mid \boldsymbol{x}) - P(Y_j \leq s \mid \boldsymbol{x}).$$

One of the important assumptions of this model is proportional odds assumption. In this model, the coefficients that describe the relationship between lower level versus all higher levels of the response variable are the same as those that describe the relationship between the next lowest level and all higher levels. Likelihood ratio test (Peterson and Harrell, 1990) and Brant test (Brant, 1990) is used to test the proportional odds assumption. However, these tests have been criticized for having a tendency to reject the null hypothesis (Harrell, 2001). If this assumption is violated the multinomial logistic regression is one option among others (Hosmer and Lemeshow 2013, McCullagh and Nelder 1989).

3.4.3 Brant test

Brant (1990) proposed a test by by creating $S-1$ binary logits on the outcomes defined by $Y^* = 1$ if $Y > s$ and $Y^* = 0$ if $Y \leq s$. An outcome Y with levels $s = 0,1,2$, one can define two binary outcomes $Y_1^* = 1$ if $Y > 0$, $Y_1^* = 0$ if $Y \leq 0$ and $Y_2^* = 1$ if $Y > 1$, $Y_2^* = 0$ if $Y \leq 1$.

Then one can estimate $\hat{\boldsymbol{\beta}}_{\mathbf{1}}' = [\hat{\beta}_{10}, \hat{\beta}_{11}, ..., \hat{\beta}_{1p}]$, $\hat{\boldsymbol{\beta}}_{\mathbf{2}}' = [\hat{\beta}_{20}, \hat{\beta}_{21}, ..., \hat{\beta}_{2p}]$, $\widehat{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{1}})$, $\widehat{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{2}})$, $\hat{\pi}_{i1}(\boldsymbol{x_i}) = P(Y_1^* = 1 \mid \boldsymbol{x_i})$ and $\hat{\pi}_{i2}(\boldsymbol{x_i}) = P(Y_2^* = 1 \mid \boldsymbol{x_i})$ from two binary logistic regression model. Define $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{*\prime} = [\hat{\beta}_{11}, \hat{\beta}_{12}, ..., \hat{\beta}_{1p}]$, $\hat{\boldsymbol{\beta}}_{\mathbf{2}}^{*\prime} = [\hat{\beta}_{21}, \hat{\beta}_{22}, ..., \hat{\beta}_{2p}]$ and $w_{iuv} = \hat{\pi}_{iv}(\boldsymbol{x_i}) - \hat{\pi}_{iu}(\boldsymbol{x_i})\hat{\pi}_{iv}(\boldsymbol{x_i})$, $u, v = 1,2$.

The null hypothesis for proportional odds assumption is $H_0 : \boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^*$. This hypothesis is equivalent to $H_0 : \boldsymbol{D}\boldsymbol{\beta}^* = \mathbf{0}$, where

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0} \\ \ldots & \ldots \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}.$$

$\boldsymbol{I}$ is a $(p+1)\times(p+1)$ identity matrix and $\mathbf{0}$ is a $(p+1)\times(p+1)$ matrix of 0's. The test statistic can be shown as:

$$X^2 = (\boldsymbol{D}\hat{\boldsymbol{\beta}}^*)'[\boldsymbol{D}\widehat{Var}(\hat{\boldsymbol{\beta}}^*)\boldsymbol{D}']^{-1}(\boldsymbol{D}\hat{\boldsymbol{\beta}}^*) \tag{3.9}$$

distributed as χ^2 with $[(S+1)-2]p = (3-2)p$ degrees of freedom. Where

$$\widehat{Var}(\hat{\boldsymbol{\beta}}^*) = \begin{bmatrix} \widehat{Var}(\hat{\boldsymbol{\beta}}_1^*) & \widehat{Var}(\hat{\boldsymbol{\beta}}_1^*, \hat{\boldsymbol{\beta}}_2^*) \\ \widehat{Var}(\hat{\boldsymbol{\beta}}_2^*, \hat{\boldsymbol{\beta}}_1^*) & \widehat{Var}(\hat{\boldsymbol{\beta}}_2^*) \end{bmatrix}.$$

The diagonal elements are the variance-covariance matrix from each binary logistic regression and the off-diagonal elements is estimated by deleting first row and column of $(\boldsymbol{X}'\boldsymbol{W}_{uu}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}_{uv}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}_{vv}\boldsymbol{X})^{-1}$, where $\boldsymbol{W}_{uv}$ is a $N\times N$ diagonal matrix whose diagonal element is w_{iuv} and $\boldsymbol{X}$ is a $N\times(p+1)$ covariate vector including constant. This test easily generalize for $s = 0,1,...,S$ and details can be found in Long (1997).

3.4.4 Goodness-of-fit

Islam and Chowdhury (2010) proposed modified deviance for repeated measures for binary outcome to test the goodness-of-fit for the joint model. We generalized this for multinomial outcome from repeated measures.

3.4.5 Modified deviance for repeated measures

Let the outcome y has $s = 0,1,...,S$ with $S+1$ categories. In the case of multinomial logistic regression, suppose the observed data is of the form $(z_1,y_1),(z_2,y_2),...,(z_i,y_i),...,(z_n,y_n)$ where y_i is a (S+1) indicator vector identifies which class an observation belongs to. From the fitted model the estimate of a vector of probabilities is $\hat{p}_i(z) = [\hat{p}_{i0}(z),\hat{p}_{i1}(z),...,\hat{p}_{iS}(z)]$.

Then the model-based likelihood for ith subject can be shown as:

$$\prod_{s=0}^{S}[\hat{P}_{is}(z_i)]^{y_{is}} \tag{3.10}$$

where $y_{is} = 1$ if the outcome is in category s, otherwise $y_{is} = 0$. $E(Y) = E(Y_1) = P_s(x)$ as defined in (3.1). The saturated model assigns probability one to each observed events. Here the vectors of probabilities $\hat{p}_i(z)$ is equal to y_i for each observation and the ratio of these likelihood can be written as:

$$\prod_{s=0}^{S} \left[\frac{\hat{P}_{is}(z_i)}{y_{is}} \right]^{y_{is}} \tag{3.11}$$

where y_{is} is observed outcome for subject i with sth category. By taking minus two times the log of this quantity, we find the deviance for all n subjects as

$$D_1 = -2 \sum_{i=0}^{n} \sum_{s=0}^{S} y_{is} log \left[\frac{\hat{P}_{is}(z_i)}{y_{is}} \right]. \tag{3.12}$$

The contribution of marginal model shown in equation (3.1) to the overall deviance is D_1. All first and higher order regressive models shown previously are equivalent to the marginal model. Similarly, the contribution of jth order regressive model to overall deviance is:

$$D_j = -2 \sum_{i=0}^{n} \sum_{s=0}^{S} y_{ijs} log \left[\frac{\hat{P}_{ijs}(z_i)}{y_{ijs}} \right], \quad j = 2,3,...J, \tag{3.13}$$

where y_{ijs} is ith subject from jth outcome of sth category. Following, Islam and Chowdhury (2010) the summary deviance statistic for $y_1, y_2, ..., y_J$ is the sum of deviance from marginal and all regressive models and can be written as:

$$\begin{aligned} D = D_1 + D_2 + ... + D_j = -2 \sum_{i=0}^{n} \sum_{s=0}^{S} \delta_{is} log \left[\frac{\hat{P}_{is}(z_i)}{y_{is}} \right] \\ + \sum_{j=2}^{J} \left\{ -2 \sum_{i=0}^{n} \sum_{s=0}^{S} \delta_{ijs} log \left[\frac{\hat{P}_{ijs}(z_i)}{y_{ijs}} \right] \right\}. \end{aligned} \tag{3.14}$$

3.4.6 Tests for order

Islam *et al.* (2009) proposed a simple and flexible test to check the order of the Markov model for binary outcomes. For *j-th* order regressive model, dummy variables for each category except for reference level from previous j-1 outcomes are incorporated as the covariates for investigating the adequacy of the order of the model as shown in equation (3.5). Then the null hypotheses

$$\begin{aligned} H_0 : &\beta_{s.y_{j-1}(p+1)} = \cdots = \beta_{s.y_{j-1}(p+S)} = \beta_{s.y_{j-1}(p+S+1)} = \cdots = \beta_{s.y_{j-1}(p+2S)} \\ &= \cdots = \beta_{s.y_{j-1}[p+(j-1)S+1]} = \cdots = \beta_{s.y_{j-1}[p+(j-1)S+S]} = 0, \\ &s = 1,2,\cdots,S; j = 2,\cdots,J, \text{ can be tested using following statistic:} \end{aligned}$$

$$-2 \left[\ln L(\hat{\boldsymbol{\beta}}_1) - \ln L(\hat{\boldsymbol{\beta}}) \right], \tag{3.15}$$

which is distributed asymptotically as χ^2 with $[\{p+1+(j-1)S\}S]-\{(j-1)S\}]$ degrees of freedom, $[\{p+1+(j-1)S\}S]$ is the total number of parameters of $(j-1)th$ order regressive model and $j-1$ in $(j-1)S$ are the number of previous outcomes $y_1,...,y_{j-1}$ multiplied by the number of dummy variables (S) for each outcomes. Let,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathbf{1}}' =&[\hat{\beta}_{s.y_{j-1}(p+1)},\cdots,\hat{\beta}_{s.y_{j-1}(p+S)},\beta_{s.y_{j-1}(p+S+1)},\cdots,\hat{\beta}_{s.y_{j-1}(p+2S)},\cdots,\\ &\hat{\beta}_{s.y_{j-1}[p+(j-1)S+1]},\cdots,\hat{\beta}_{s.y_{j-1}[p+(j-1)S+S]}], \quad s=1,2,\cdots,S\end{aligned}$$

be a $1\times(j-1)S$ vector of the regression coefficients of dummy variables $D_{11},\cdots,D_{1S}$, $D_{21},\cdots,D_{2S},\cdots,D_{(j-1)1},\cdots,D_{(j-1)S}$ corresponding to previous $y_1,\cdots,y_{j-1}$ outcomes and

$$\begin{aligned}\hat{\boldsymbol{\beta}}' =&[\hat{\beta}_{s.y_{j-1}0},\hat{\beta}_{s.y_{j-1}1},\cdots,\hat{\beta}_{s.y_{j-1}p},\\ &\hat{\beta}_{s.y_{j-1}(p+1)},\cdots,\hat{\beta}_{s.y_{j-1}(p+S)},\beta_{s.y_{j-1}(p+S+1)},\cdots,\hat{\beta}_{s.y_{j-1}(p+2S)},\cdots,\\ &\hat{\beta}_{s.y_{j-1}[p+(j-1)S+1]},\cdots,\hat{\beta}_{s.y_{j-1}[p+(j-1)S+S]}], \quad s=1,2,\cdots,S\end{aligned}$$

be a $1\times[p+1+(j-1)S]S$ vector of all regression coefficients of $(j-1)th$ order regressive model.

Alternatively, we can test the above hypothesis that some subset of parameters equal to zero and construct a Wald test. Let,

$$\boldsymbol{\beta}=\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\mathbf{0}}\\ \hat{\boldsymbol{\beta}}_{\mathbf{1}}\end{bmatrix}$$

$$\text{where } \hat{\boldsymbol{\beta}}_{\mathbf{0}}=[\hat{\beta}_{s.y_{j-1}0},\hat{\beta}_{s.y_{j-1}1},...,\hat{\beta}_{s.y_{j-1}p}]', \quad s=1,2,...,S$$

is a $[p+1]S\times 1$ vector of parameters corresponding to $[1,X_1,...,X_p]$ and $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ defined above is a $[\{p+1+(j-1)S\}S-(p+1)S]\times 1$ vector of parameters corresponding to the dummy variables $D_{11},...,D_{1S}$, $D_{21},...,D_{2S}$
$,...,D_{(j-1)1},...,D_{(j-1)S}$. Let,

$$\widehat{V(\boldsymbol{\beta})}=I(\hat{\boldsymbol{\beta}})^{-1}=\begin{bmatrix}\hat{V}_{00}(\hat{\boldsymbol{\beta}}) & \hat{V}_{01}(\hat{\boldsymbol{\beta}})\\ \hat{V}_{10}(\hat{\boldsymbol{\beta}}) & \hat{V}_{11}(\hat{\boldsymbol{\beta}})\end{bmatrix}$$

where $I(\hat{\beta})$ is the observed information matrix and $\hat{V}_{11}(\hat{\boldsymbol{\beta}})$ is the lower sub-matrix of $\widehat{V(\boldsymbol{\beta})}$. The Wald statistic is then

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}}'[\hat{V}_{11}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}}_{1} \tag{3.16}$$

which is asymptotically χ^2 with $[(j-1)S]S$ degrees of freedom.

Then one can perform the test as follows:

(i) The likelihood ratio test can be used to test the significance of the overall model at the first stage.

(ii) The Wald test can be used to test the significance of the parameter(s) corresponding to the previous outcomes as shown below:

$$W = \hat{\beta}_{s.y_{j-1}[p+(j-1)s+s]} / se(\hat{\beta}_{s.y_{j-1}[p+(j-1)s+s]}), \quad s = 1,2,...S.$$

Good fit models with the better discriminative ability and predictive power are expected to provide higher prediction accuracy. Predictive accuracy of models can be estimated based on confusion matrix and over(under)fitting can be evaluated using training and test data sets approach (James *et al.*, 2013).

3.5 Application

The panel data from the Health and Retirement Study (HRS), sponsored by the National Institute of Aging (grant number NIA U01AG09740), conducted by the University of Michigan (HRS, 2014) is used for the application. In wave one, a total of 12652 subjects were interviewed in the HRS cohort. Out of all these subjects, 9762 were age eligible (those with birth years 1931-1941). A total of six waves (follow-ups) of the RAND version of the data from wave six (2002) to wave 11 (2012) is considered for this application. At wave six minimum age of the subjects was 60. The outcome variables considered are Activity of daily living index (ADL) from wave six to wave eleven $(Y_1,...,Y_6)$. This index is the sum of five tasks (yes/no) ranging from 0 to 5: whether respondents faced difficulties in walking, dressing, bathing, eating and getting in/out of bed. Due to small frequencies 3 and higher values were coded as 2. The explanatory variables considered are: age (in years), marital status (married/partnered=1, single/separated=0), whether drink (yes=1, no=0), sex (male=1, female=0), number of conditions ever had (N.cond) ranges from 0 to 8, White (yes=1, no=0), Black (yes=1, no=0) with others as reference category, education (in years) and veteran status (1=yes, 0= no). The variable drink indicates whether the respondent drinks alcoholic beverages. After removal of cases with missing values for outcome variable at wave six, the number of subjects is 7130. Table 3.1 displays the frequency distribution of the outcomes for different waves.

The outcomes used here are ordinal in nature and it is common to use proportional odds model (McCullagh, 1980). One of the important assumptions of this model is proportional odds assumption. The proposed regressive models for the ordinal outcome for first and higher order are equivalent to the ordinal regression for a single outcome. Likelihood ratio test and Brant test for proportional odds assumption are shown in Table 3.2. The likelihood ratio test for marginal and first order models and Brant test for marginal model satisfies the assumption.

TABLE 3.1: Distribution of Activity of Daily Living Index, Waves 6 to 11.

Outcome	*Wave*					
	6		*7*		*8*	
Value	*N*	*%*	*N*	*%*	*N*	*%*
0	6210	87.1	5906	86.7	5459	84.9
1	477	6.7	462	6.8	503	7.8
2	443	6.2	445	6.5	467	7.3
Total	7130	100.0	6813	100.0	6429	100.0
	Wave					
	9		*10*		*11*	
Value	*N*	*%*	*N*	*%*	*N*	*%*
0	5459	84.9	4600	81.7	4262	81.5
1	503	7.8	460	8.2	479	9.2
2	467	7.3	569	10.1	491	9.4
Total	6429	100.0	5629	100.0	5232	100.0

TABLE 3.2: Test results for proportionality odds assumption.

	Approximate LRT		Brant test		
Wave	χ^2	p.v.	χ^2	p.v.	d.f.
6	13.4	0.145	13.7	0.134	9
7	18.2	0.076	23.9	0.013	11
8	22.8	0.044	25.5	0.020	13
9	41.8	0.000	45.9	0.000	15
10	31.3	0.016	34.4	0.008	17
11	42.4	0.002	43.8	0.001	19

Parameter estimates along with standard error and significance level for marginal and regressive models are shown in Table 3.3 and Table 3.4. Various predictors are found to be significantly associated with outcome variables for different models. All dummy indicators for previous outcomes are significantly and positively associated with the current outcomes except for fifth-order model. Model statistics are shown in Table 3.5. Likelihood ratio test for the joint model is statistically significant ($p < 0.001$) as shown in Table 3.5.The prediction accuracy based on confusion matrix for full data and test and training data varies between 0.86 to 0.89 which is reasonably high. Also, accuracy from full, training and test data are very close, which shows the absence of over(under)fitting for all models.

3.6 Predicted joint probabilities

Specified covariates vector were used to predict marginal and conditional probabilities and to predict the joint probability of outcomes for three selected trajectories. Three paths are: (i) $\hat{P}(Y_1 = 0, Y_2 = 0, Y_3 = 0, Y_4 = 0, Y_5 = 0, Y_6 = 0 \mid \boldsymbol{X}^* = \boldsymbol{x}^*)$ remains functional limitations free from wave six to eleven. (ii) $\hat{P}(Y_1 = 1, Y_2 = 1, Y_3 = 1, Y_4 = 1, Y_5 = 1, Y_6 = 1 \mid \boldsymbol{X}^* = \boldsymbol{x}^*)$ one functional limitations among all six waves. (iii) $\hat{P}(Y_1 = 2, Y_2 = 2, Y_3 =$

TABLE 3.3: Estimates of marginal and regressive models using multinomial logistic regression for different order.

	Waves											
	6				7				8			
	Category 1		Category 2		Category 1		Category 2		Category 1		Category 2	
Variables	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.
Constant	−1.480	1.068	−1.923	1.144	−4.664**	1.189	−4.513**	1.426	−6.160**	1.239	−5.104**	1.539
Age	−0.011	0.016	−0.004	0.017	0.017	0.017	0.012	0.020	0.038*	0.017	0.024	0.021
Mstat	−0.332**	0.107	−0.536**	0.113	−0.031	0.117	−0.580**	0.136	0.006	0.116	−0.234	0.142
N.cond	0.497**	0.034	0.670**	0.036	0.425**	0.037	0.466**	0.043	0.330**	0.037	0.415**	0.045
Drink	−0.321**	0.107	−0.790**	0.128	−0.163	0.115	−0.613**	0.150	−0.126	0.111	−0.471**	0.146
Gender	0.016	0.128	−0.008	0.141	0.116	0.135	0.476**	0.164	0.214	0.135	0.028	0.174
White	−0.070	0.253	−0.548*	0.235	−0.154	0.274	−0.428	0.290	0.237	0.315	−0.114	0.335
Black	0.025	0.268	−0.121	0.247	0.029	0.291	−0.257	0.309	0.492	0.330	0.411	0.350
Educ.	−0.091**	0.015	−0.087**	0.016	−0.018	0.017	−0.028	0.019	−0.038*	0.017	−0.083**	0.020
Veteran	−0.037	0.153	−0.157	0.176	−0.354*	0.166	−0.325	0.202	−0.133	0.158	−0.153	0.218
D61					1.802**	0.138	2.258**	0.163	1.658**	0.146	1.888**	0.175
D62					2.483**	0.183	4.245**	0.169	2.197**	0.212	3.558**	0.201
D71									1.204**	0.153	1.138**	0.183
D72									1.274**	0.215	2.096**	0.206

TABLE 3.4: Continued... Table 3.

	Waves											
	6				7				8			
	Category 1		Category 2		Category 1		Category 2		Category 1		Category 2	
Variables	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.
Constant	−6.828**	1.363	−9.071**	1.620	−6.917**	1.388	−9.174**	1.571	−5.912**	1.447	−11.06 **	1.862
Age	0.052**	0.018	0.074**	0.022	0.057**	0.018	0.088**	0.020	0.034	0.018	0.092**	0.024
Mstat	−0.025	0.124	−0.518**	0.143	−0.314**	0.120	−0.522**	0.134	−0.107	0.124	−0.391*	0.156
N.cond	0.276**	0.040	0.362**	0.046	0.280**	0.040	0.336**	0.044	0.278**	0.040	0.401**	0.051
Drink	−0.055	0.122	−0.449**	0.154	−0.076	0.118	−0.291*	0.138	−0.266*	0.122	−0.335*	0.161
Gender	0.037	0.147	0.102	0.179	0.241	0.146	0.230	0.165	0.181	0.150	0.179	0.190
White	0.104	0.307	0.149	0.355	−0.230	0.297	−0.282	0.318	−0.088	0.302	0.067	0.371
Black	−0.250	0.334	0.211	0.374	−0.115	0.319	−0.061	0.338	0.063	0.325	0.102	0.397
Educ.	−0.054**	0.018	−0.053**	0.021	−0.046*	0.018	−0.090**	0.020	−0.013	0.019	−0.051*	0.022
Veteran	−0.239	0.176	−0.032	0.216	0.078	0.165	0.024	0.194	0.083	0.167	−0.257	0.227
D61	1.762**	0.147	1.423**	0.183	1.459**	0.164	1.681**	0.176	1.481**	0.154	1.894**	0.192
D62	2.067**	0.220	3.116**	0.202	1.545**	0.240	2.878**	0.201	1.566**	0.211	3.226**	0.200
D71	0.913**	0.170	0.835**	0.194	1.157**	0.167	1.089**	0.183	1.037**	0.177	0.918**	0.213
D72	1.000**	0.266	1.980**	0.241	0.485	0.285	1.405**	0.241	0.872**	0.286	1.927**	0.260
D81	0.793**	0.176	0.949**	0.194	0.824**	0.186	0.724**	0.199	0.446*	0.189	0.432*	0.219
D82	0.697**	0.249	1.346**	0.236	0.181	0.317	0.758**	0.275	0.380	0.295	0.390	0.300
D91					0.370	0.198	0.613**	0.196	0.714**	0.199	0.041	0.245
D92					0.680*	0.278	0.701**	0.270	0.384	0.343	0.505	0.337
D101									0.393	0.210	0.692**	0.234
D102									0.323	0.315	0.969**	0.313

TABLE 3.5: Model statistics for marginal and regressive models.

	Constant only			Full model						
							L.R.T		Accuracy		
Wave	Log L.	Dev.	AIC	Log L.	Dev.	AIC	p.v.	d.f.	All	Train	Test
6	-3378.9	6757.7	6761.7	-2921.5	5842.9	5882.9	0.000	18	0.873	0.871	0.878
7	-3195.9	6391.7	6395.7	-2309.0	4618.0	4666.0	0.000	22	0.887	0.887	0.881
8	-3283.0	6566.0	6570.0	-2244.9	4489.8	4545.8	0.000	26	0.878	0.878	0.875
9	-3162.3	6324.6	6328.6	-2028.2	4056.4	4120.4	0.000	30	0.883	0.882	0.880
10	-3232.7	6465.4	6469.4	-2151.8	4303.6	4375.6	0.000	34	0.861	0.863	0.859
11	-3048.0	6096.0	6100.0	-1862.8	3725.6	3805.6	0.000	38	0.860	0.861	0.855
Joint model	-19300.8	38601.4	38625.4	-13518.2	27036.3	27396.3	0.000	72			

$2, Y_4 = 2, Y_5 = 2, Y_6 = 2 \mid \boldsymbol{X}^* = \boldsymbol{x}^*)$ two or more functional limitations from wave six to eleven. Figure 3.2-3.4 displays three joint predicted risks for selected trajectories along with the average from 10,000 bootstrap samples. It should be noted that predicted risk at wave six in the graphs are marginal probabilities and from wave seven onward are joint probabilities.

Figure 3.2 displays the predicted joint risks for three trajectories by number of previous conditions (0,2,4,6, and 8) and gender. The value of other predictors were set to: mstat=0, Age=65 years, whether drink=1, white=1, Educ. =12 years, and veteran status=1. The risk of functional limitations free for zero previous conditions was close to one at wave six this risk decreases in later waves and the risk was little higher for male compared to female. The risk to follow path (iii) remains flat for all six waves with zero previous conditions. However, this risk increases with increased number of previous conditions. Figure 3.3 displays the predicted joint risks for three trajectories by gender. The predicted risk for trajectories (i, ii, and iii) are shown in the graph. The value of other predictors was set to: age=65 years, mstat=0, N.cond =2, whether drink=1, white=1, Educ. =12 years, veteran status=1. Male subject has more risk compared to female subject. Predicted joint risks for three trajectories by veteran status is presented in Figure 3.4. The value of other predictors was set to: age=65 years, mstat=0, N.cond =2 N.cond, whether drink=1, white=1, Educ. =12 years, gender=1. Non-veteran subject has higher risk compared to veteran subject. A sample calculation of marginal, conditional and joint risk for a trajectory is shown in Table 3.6.

TABLE 3.6: Computation of predicted risk for a trajectory.

N.cond	Gender	P_0	$P_{0.0}$	$P_{0.00}$	$P_{0.000}$	$P_{0.0000}$	$P_{0.00000}$	P_{000000}
0	Female	0.97	0.98	0.98	0.97	0.96	0.97	0.85
2	Female	0.91	0.96	0.96	0.95	0.93	0.95	0.70
4	Female	0.76	0.91	0.91	0.92	0.88	0.91	0.47
6	Female	0.51	0.81	0.84	0.85	0.80	0.85	0.20
8	Female	0.24	0.64	0.72	0.76	0.68	0.75	0.04
0	Male	0.97	0.98	0.97	0.97	0.95	0.97	0.82
2	Male	0.91	0.95	0.95	0.95	0.91	0.94	0.67
4	Male	0.76	0.89	0.90	0.91	0.85	0.90	0.43
6	Male	0.50	0.77	0.82	0.85	0.76	0.83	0.17
8	Male	0.24	0.58	0.69	0.74	0.63	0.72	0.03

3.7 Bootstrapping

To measure the accuracy of sample estimates, bootstrapping is used. We performed 10,000 bootstraps and computed bias, standard error, and mean squared error for estimates. Estimates from Table 3.3 and Table 3.4 are considered as population parameters

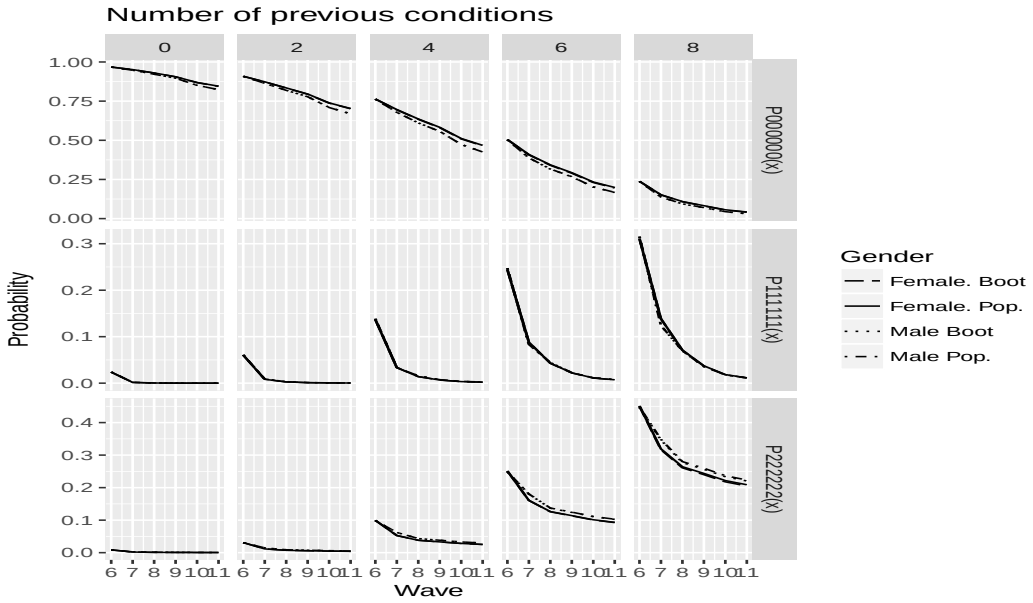


FIGURE 3.2: Predicted risk by gender and no. of conditions from original and bootstrap sample.

while bias, standard error and mean squared error are computed. Bias is very small generally (less than 1 percent) for the estimators of parameters of all models. Standard error and mean squared errors are also found to be very small (Table 3.7-3.9). Convergence is achieved for all 10,000 bootstrap samples. We also predicted joint risks as shown in Figures 3.2-3.4 from all 10,000 samples. Average of the predicted joint risks from all bootstrap samples and predicted risk are also shown in the Figure 3.2-3.4. Lines of predicted risk from original sample and bootstrap samples are overlapped in all three graphs. Therefore, very minimal or no bias in the case of predicted joint risk.

3.8 Conclusions

In this paper, a modeling framework is proposed to predict joint probabilities for a sequence of multinomial events from longitudinal studies that may change through different trajectories. The proposed models provide the estimates for each stage in the process conditionally, and the conditional estimates are linked using marginal and sequence of conditional models to provide the joint model needed for predicting the probability of a trajectory based on specified covariates pattern. The estimates of the parameters of the marginal models are obtained from the outcome variable at the baseline and the models at the subsequent follow-ups provide the estimates of the parameters of the conditional models. Proposed approach also allows interaction among previous outcomes and predictors. The interaction terms may provide a better understanding of the underlying disease process and the relationships between outcomes and related risk factors. The likelihood ratio test for the goodness of fit, deviance and AIC for the proposed model are shown in this paper. Also, 10,000 bootstrap simulation is undertaken to study the performance of

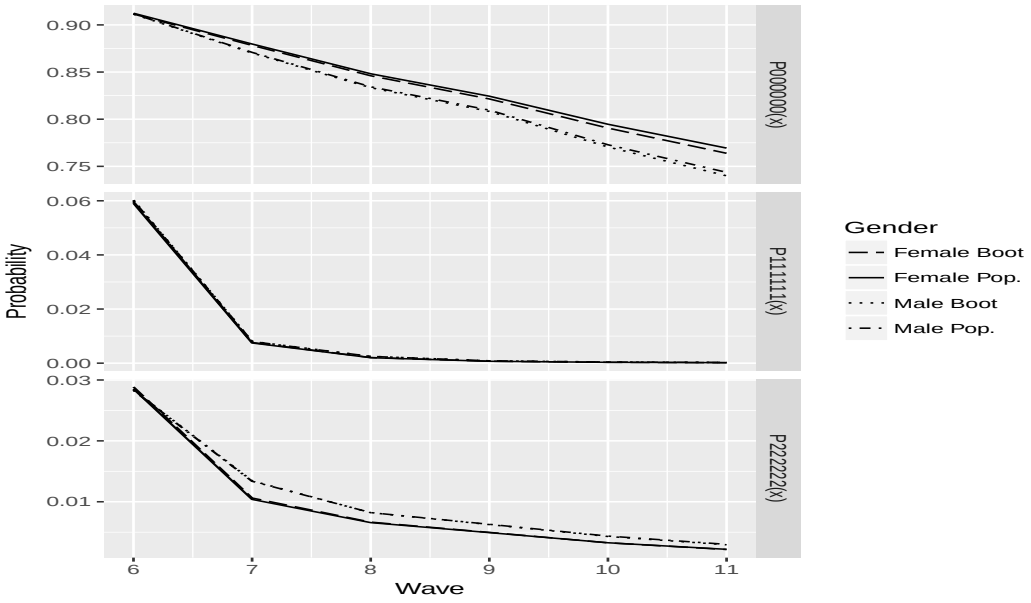


FIGURE 3.3: Predicted Risk by gender from original and bootstrap sample.

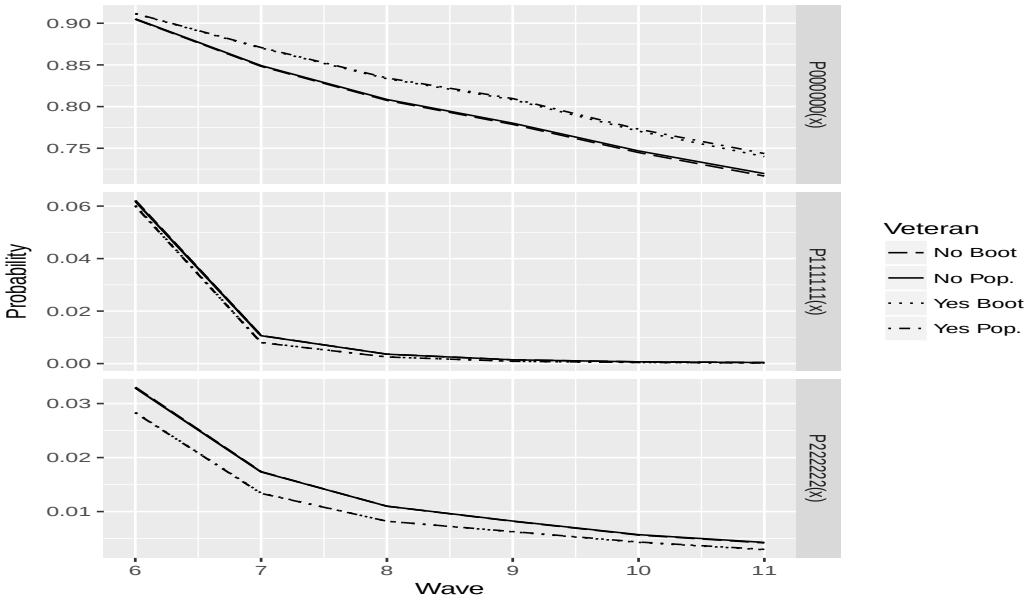


FIGURE 3.4: Predicted risk by veteran status from original and bootstrap sample.

the proposed model and predicted joint probabilities for a sequence of events. One can easily fit proposed models and predict the risk of a sequence of events using the available statistical softwares when there are multiple outcomes at each follow-up time.

The major improvement of the proposed framework is that one needs to fit a significantly smaller number of models compared to the conditional models such as Markov models. The bias of parameter estimates for all models from all bootstrap simulation is less than one percent in most of the cases except for intercepts and the explanatory variable, race. The estimated mean squared error is also very low. Predicted joint risks for trajectories from bootstrap simulation overlap with that of the assumed population as shown in Figures 3.2-3.4. The proposed methods can be applied in many fields of studies such as epidemiology, public health, survival analysis, genetics, reliability, environmental studies, etc. Also, we believe that the proposed framework would be very useful for analyzing big data.

TABLE 3.7: Bootstrap Results parameter estimates for models for wave 6 and 7.

Variables	Wave 6						Wave 7					
	Category 1			Category 2			Category 1			Category 2		
	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$
Constant	-0.0234	1.0461	1.0949	-0.0142	1.1666	1.3611	-0.0468	1.2011	1.4449	-0.0463	1.4315	2.0514
Age	-0.0001	0.0153	0.0002	-0.0002	0.0172	0.0003	0.0001	0.0172	0.0003	0.0001	0.0206	0.0004
Mstat	-0.0014	0.1090	0.0119	-0.0003	0.1148	0.0132	0.0025	0.1196	0.0143	-0.0030	0.1371	0.0188
N.cond	0.0009	0.0332	0.0011	0.0025	0.0355	0.0013	0.0007	0.0370	0.0014	0.0019	0.0464	0.0022
Drink	-0.0018	0.1087	0.0118	-0.0063	0.1324	0.0176	-0.0003	0.1167	0.0136	-0.0028	0.1520	0.0231
Gender	0.0002	0.1293	0.0167	-0.0020	0.1431	0.0205	-0.0045	0.1369	0.0188	0.0011	0.1649	0.0272
White	0.0258	0.2719	0.0746	0.0121	0.2323	0.0541	0.0302	0.2803	0.0795	0.0169	0.2862	0.0822
Black	0.0206	0.2860	0.0822	0.0107	0.2438	0.0595	0.0273	0.2985	0.0899	0.0159	0.3104	0.0966
Education	-0.0001	0.0153	0.0002	0.0003	0.0160	0.0003	0.0000	0.0165	0.0003	0.0002	0.0197	0.0004
Veteran	-0.0023	0.1545	0.0239	-0.0037	0.1741	0.0303	0.0010	0.1633	0.0267	-0.0022	0.1936	0.0375
D61							0.0042	0.1468	0.0216	0.0108	0.1744	0.0305
D62							0.0087	0.1919	0.0369	0.0303	0.1809	0.0337

TABLE 3.8: Bootstrap results parameter estimates for models for wave 8 and 9.

Variables	Wave 8						Wave 9					
	Category 1			Category 2			Category 1			Category 2		
	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$
Constant	-0.0711	1.2696	1.6168	-0.0671	1.6153	2.6138	-0.0620	1.3960	1.9527	-0.0887	1.6527	2.7393
Age	0.0002	0.0171	0.0003	0.0003	0.0217	0.0005	0.0001	0.0185	0.0003	0.0006	0.0221	0.0005
Mstat	0.0027	0.1179	0.0139	-0.0016	0.1519	0.0231	0.0034	0.1286	0.0165	-0.0020	0.1526	0.0233
N.cond	0.0015	0.0375	0.0014	0.0031	0.0477	0.0023	0.0023	0.0418	0.0018	0.0028	0.0475	0.0023
Drink	-0.0017	0.1118	0.0125	-0.0064	0.1477	0.0219	-0.0010	0.1234	0.0152	-0.0062	0.1565	0.0245
Gender	-0.0036	0.1408	0.0198	-0.0044	0.1843	0.0340	-0.0048	0.1548	0.0240	-0.0088	0.1877	0.0353
White	0.0390	0.3381	0.1159	0.0255	0.3367	0.1140	0.0368	0.3065	0.0953	0.0242	0.4214	0.1782
Black	0.0361	0.3519	0.1251	0.0261	0.3504	0.1235	0.0282	0.3406	0.1168	0.0263	0.4347	0.1896
Education	0.0001	0.0178	0.0003	-0.0007	0.0221	0.0005	-0.0004	0.0198	0.0004	-0.0005	0.0220	0.0005
Veteran	0.0002	0.1609	0.0259	-0.0014	0.2247	0.0505	-0.0030	0.1824	0.0333	0.0038	0.2261	0.0511
D61	0.0037	0.1617	0.0262	0.0010	0.1927	0.0371	-0.0005	0.1880	0.0353	0.0037	0.2273	0.0517
D62	-0.0001	0.2300	0.0529	0.0129	0.2185	0.0479	0.0015	0.2771	0.0768	0.0153	0.2590	0.0673
D71	0.0058	0.1545	0.0239	0.0078	0.1899	0.0361	0.0006	0.1878	0.0353	-0.0019	0.2115	0.0447
D72	0.0125	0.2211	0.0490	0.0321	0.2101	0.0452	0.0005	0.2996	0.0898	0.0279	0.2732	0.0754
D81							0.0094	0.1549	0.0241	0.0038	0.1952	0.0381
D82							0.0083	0.2372	0.0563	0.0278	0.2197	0.0491

TABLE 3.9: Bootstrap Results parameter estimates for models for wave 10 and 11.

Variables	Wave 10						Wave 11					
	Category 1			Category 2			Category 1			Category 2		
	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$
Constant	-0.0698	1.3749	1.8951	-0.1082	1.5977	2.5643	-0.0182	1.5155	2.2971	-0.1787	1.9738	3.9278
Age	0.0002	0.0179	0.0003	0.0010	0.0207	0.0004	-0.0004	0.0192	0.0004	0.0016	0.0249	0.0006
Mstat	0.0016	0.1227	0.0151	-0.0048	0.1342	0.0180	-0.0017	0.1227	0.0150	-0.0082	0.1594	0.0255
N.cond	0.0021	0.0405	0.0016	0.0033	0.0471	0.0022	0.0014	0.0396	0.0016	0.0039	0.0563	0.0032
Drink	-0.0026	0.1183	0.0140	-0.0068	0.1378	0.0190	-0.0018	0.1232	0.0152	-0.0048	0.1641	0.0270
Gender	-0.0053	0.1490	0.0222	-0.0004	0.1719	0.0295	-0.0008	0.1526	0.0233	0.0008	0.1936	0.0375
White	0.0365	0.3156	0.1009	0.0155	0.3132	0.0983	0.0262	0.3362	0.1137	0.0236	0.3978	0.1588
Black	0.0323	0.3382	0.1154	0.0138	0.3327	0.1109	0.0204	0.3570	0.1278	0.0106	0.4279	0.1832
Education	-0.0001	0.0183	0.0003	-0.0010	0.0211	0.0004	0.0001	0.0197	0.0004	-0.0005	0.0231	0.0005
Veteran	0.0062	0.1718	0.0295	-0.0011	0.2025	0.0410	0.0049	0.1721	0.0296	-0.0019	0.2376	0.0565
D61	-0.0067	0.2054	0.0422	-0.0035	0.2198	0.0483	-0.0034	0.2335	0.0546	0.0065	0.2629	0.0692
D62	0.0005	0.3086	0.0952	0.0067	0.2999	0.0900	-0.0005	0.3493	0.1220	0.0185	0.3417	0.1171
D71	0.0039	0.1942	0.0377	0.0012	0.2195	0.0482	0.0059	0.2262	0.0512	-0.0059	0.2665	0.0710
D72	-0.0126	0.3372	0.1139	0.0040	0.3098	0.0960	-0.0152	0.4337	0.1883	-0.0042	0.3923	0.1539
D81	0.0070	0.1740	0.0303	0.0082	0.1885	0.0356	0.0027	0.2184	0.0477	0.0033	0.2326	0.0541
D82	-0.0049	0.3441	0.1185	0.0218	0.2797	0.0787	-0.0010	0.3335	0.1112	0.0046	0.3474	0.1207
D91	0.0046	0.1731	0.0300	0.0088	0.1904	0.0363	0.0045	0.1949	0.0380	0.0001	0.2298	0.0528
D92	0.0065	0.2559	0.0655	0.0320	0.2057	0.0433	0.0088	0.3154	0.0996	0.0432	0.2867	0.0840
D101							0.0133	0.1658	0.0277	0.0110	0.1977	0.0392
D102							0.0133	0.2360	0.0559	0.0511	0.2165	0.0495

Chapter 4

Goodness-of-fit Test of Joint Model for a Sequence of Multinomial Outcomes from Repeated Measures

4.1 Introduction

There is a growing interest in the risk prediction model to predict the risk of a sequence of responses for a multinomial outcome from repeated measures based on patient-specific characteristics. Islam and Chowdhury (2010) proposed a regressive model to predict the risk of a sequence of binary outcomes from repeated measures. The predicted risk could be used to present the evidence to decision makers (e.g., clinicians and patients) and are highly relevant to clinical decision support, personalized health care, and shared decision making (Moons *et al.*, 2009). The development of robust and accurate risk prediction models are a resource-demanding task and their performance needs to be rigorously validated (Calster *et al.*, 2017; Wehberg and Schumacher, 2004). Core elements of performance include (i) discrimination and (ii) classification (calibration). Discrimination considers the ability how well the model discriminates between the different categories of outcome and classification which is not error free measures the reliability of the predicted risks (Steyerberg, 2009; Johnson and Wichern, 2008; Harrell, 2001). A good classification method should result in few misclassification. Classification techniques are often evaluated in terms of their misclassification rate ignoring misclassification cost. For example, misclassifying a diseases subject as a non-diseased may have serious implications. For dichotomous outcome, many discrimination performance measures exist such as the receiver operating characteristic curve (ROC), the area under the curve (AUC), sensitivity, specificity and accuracy among others. A comprehensive discussion regarding ROC and AUC can be found in the book of Krzanowski and Hand (2009). Toledano and Gatsonis (1996) generalized ROC curve for multiple category outcomes. Accuracy and overfitting are popularly estimated by splitting the data into training and test sets. This approach is fine for the very large data set (Johnson and Wichern, 2008). There are several variants

of this technique and can be used for the polytomous outcome (James *et al.*, 2013). For a multinomial risk prediction model, either a set of dichotomous measures or one overall measure can be used to assess the discriminative ability (Calster *et al.*, 2012). The bootstrap simulation is a very useful technique to derive bias (Efron and Tibshirani, 1997, 1993). AIC and BIC can be used for model comparison.

Calibration curve and goodness of fit test for the prediction model are commonly used as calibration measure (Austin and Steyerberg, 2014; Steyerberg, 2009; Goeman and le Cessie, 2006; Hand and Till, 2001). For example, Hosmer-Lemeshow test for goodness of fit (Hosmer and Lemeshow, 1980) for binary outcome. Fagerland *et al.* (2008) generalized the Hosmer-Lemeshow test for multinomial logistic regression. Various authors reported several drawbacks of Hosmer-Lemeshow statistic (Fagerland *et al.*, 2008; Peek *et al.*, 2007; Vergouwe *et al.*, 2005; Harrell, 2001; Hosmer *et al.*, 1997; Tsiatis, 1980). A score test is suggested by Tsiatis (1980), a generalized logistic model framework to test the adequacy of the fitted model is proposed by Stukel (1988), a class of test based on smoothed residuals by le Cessie and van Houwelingen (1995) and using partial sum of residuals by Royston (1992). A detailed overview can be found in Hosmer *et al.* (1997). The Brier score measures the accuracy (prediction error) of probabilistic predictions (Brier, 1950). It can be thought of as either a measure of the "calibration" of a set of probabilistic predictions. Two other well-known statistics are the deviance and Pearson chi-square for comparing the observed number with the expected number. Using a fitted model and saturated model deviance uses a likelihood ratio test. All the methods discussed above are for a single outcome and are not readily applicable to the joint model.

Muenz and Rubinstein (1985), Bonney (1986, 1987), Azzalini (1994), Islam and Chowdhury (2006), Islam *et al.* (2009), and Islam and Chowdhury (2010) proposed regressive logistic models under the Markovian assumptions to include both binary outcomes in previous times in addition to covariates in the conditional models (Islam *et al.*, 2014, 2012, 2009, 2004). This approach reduces the over-parameterization as occurs for conditional models such as Markov models (Islam *et al.*, 2013). The framework proposed by Islam and Chowdhury (2010) for binary responses from repeated measures data links the conditional process and obtains predictive outcome based on the whole process through all possible trajectories to obtain the joint model (Islam *et al.*, 2013, 2012; Islam and Chowdhury, 2010). They also proposed modified deviance, extended Hosmer-Lemeshow test and the ROC curve for repeated measures for binary outcomes to test the goodness-of-fit and discriminative power.

Most of the available measures to check the model performance are for a single binary outcome and are not directly applicable to test the goodness of fit of the joint model for multinomial outcomes. At this drop back, we proposed a test to check the goodness-of-fit for a joint model for multinomial outcomes from repeated measures. The proposed model takes account interdependence in the outcomes variables which applies to each

subject. Also, we showed a test of independence to check the association among repeated outcomes along with bootstrap simulation.

4.2 Regressive multinomial logistic models

Figure 4.1 displays the transitions between categories of three outcomes Y_1, Y_2 and Y_3 from three follow-ups. Outcome levels (0,1,2) are denoted inside the rectangles. Here, first column shows marginal probabilities and second onward are conditional probabilities. Marginal and conditional probabilities are estimated using marginal and regressive models.

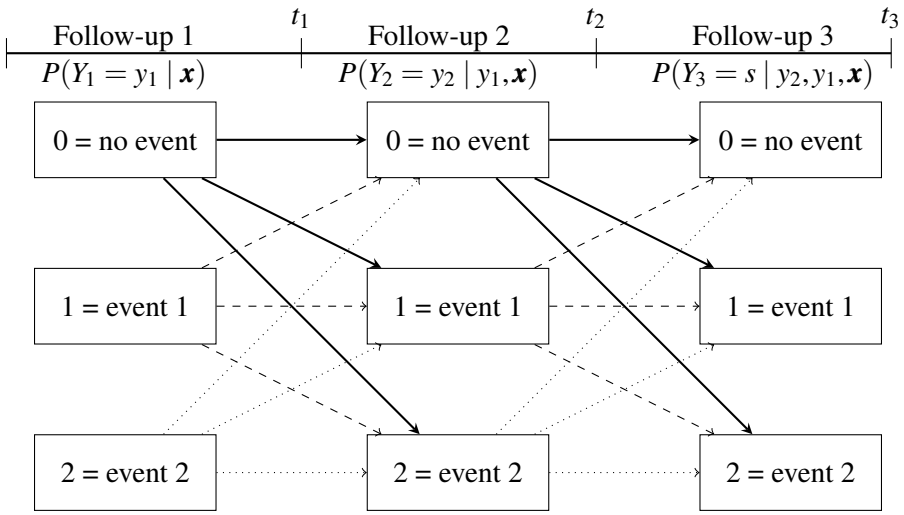


FIGURE 4.1: Transitions between states for regressive models.

4.2.1 Notations

Let $Y_{i1}, Y_{i2}, ..., Y_{iJ_i}$ are the responses from *i-th* subject at *j-th* follow-up where $(i = 1, 2, ..., n)$ and $(j = 1, 2, ..., J_i)$, J_i is the number of follow-ups for subject i. For simplicity, subscript i is omitted henceforth unless explicitly specified. Assume, $Y_j = s$ follows multinomial distribution where $(s = 0, 1, 2, ..., S)$ with $S + 1$ outcome categories and denoting non-event by category 0. For simplicity, consider outcomes with three categories $(s = 0, 1, 2)$. The risk of a sequence of events is estimated from the joint probability mass function of $Y_1, Y_2, ..., Y_J$ with covariates vector $\boldsymbol{X} = \boldsymbol{x}$ as follows:

$$\begin{aligned} &P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{X} = \boldsymbol{x}) \\ &\quad = P(Y_1 = y_1 \mid \boldsymbol{X} = \boldsymbol{x}) \times P(Y_2 = y_2 \mid Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) \\ &\quad \times ... \times P(Y_J = s \mid Y_{j-1} = y_{j-1}, ..., Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) \\ &\quad = P_{y_1}(\boldsymbol{x}) \times P_{y_2 \cdot y_1}(\boldsymbol{x}) \times ... \times P_{s \cdot y_{j-1}, ..., y_1}(\boldsymbol{x}), \end{aligned} \tag{4.1}$$

where $\boldsymbol{X}' = [1, x_1, ..., x_p]$ is vector of covariates for a subject at first follow-up and $\boldsymbol{X} = \boldsymbol{x}$ can be time dependent. We have used $Y_J = s$ as the observed outcome for the last follow-up to specify the category of a specified outcome at the endpoint.
$P(Y_1 = s \mid \boldsymbol{X} = \boldsymbol{x}) = P_s(\boldsymbol{x})$ is the marginal probability function of Y_1 conditional on $\boldsymbol{x}$;
$P(Y_J = s \mid Y_{j-1} = y_{j-1}; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1}}(\boldsymbol{x})$ is the probability function of Y_j conditional on y_{j-1} and $\boldsymbol{x}$ of order one;
$P(Y_J = s \mid Y_{j-1} = y_{j-1}, Y_{j-2} = y_{j-2}; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1}, y_{j-2}}(\boldsymbol{X} = \boldsymbol{x})$ is the probability function for Y_j conditional on y_{j-1},y_{j-2} and $\boldsymbol{x}$ of order two;
$P(Y_J = s \mid Y_{j-1} = y_{j-1}, Y_{j-2} = y_{j-2}, ..., Y_1 = y_1; \boldsymbol{X} = \boldsymbol{x}) = P_{s.y_{j-1}, y_{j-2}, ..., y_1}(\boldsymbol{x})$ is the probability function of Y_j conditional on $y_{j-1}, ..., y_1$ and $\boldsymbol{x}$ of order $k = j-1$.
The joint probabilitiy is defined as:
$P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{X} = \boldsymbol{x}) = P_{y_1, y_2, ..., y_j}(\boldsymbol{x})$.
The log-likelihood function of the joint mass function can be obtamed as:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ln P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iJ} = y_{iJ} \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{i=1}^{n} \Big[\ln P(Y_{i1} = y_{i1} \mid \boldsymbol{X} = \boldsymbol{x}) + \ln P(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}; \boldsymbol{X} = \boldsymbol{x}) \\ &\quad + ... + \ln P(Y_{iJ} = s \mid Y_{i(j-1)} = y_{i(j-1)}, ..., Y_{i1} = y_{i1}; \boldsymbol{X} = \boldsymbol{x}) \Big]. \end{aligned}$$

Here, $Y_{iJ} = s$ is used as the observed outcome for the last follow-up to specify the category of a specified outcome at the endpoint.

4.2.2 Marginal model

Multinomial logistic regression is a natural choice to model a nominal outcome Y_1 as a function of covariates vector $\boldsymbol{X} = \boldsymbol{x}$. This model, for the outcome Y_1 with three categories $(0, 1, 2)$ will produce two sets of parameter vector ($Y_1 = 1$ vs. $Y_1 = 0$ and $Y_1 = 2$ vs. $Y_1 = 0$). The marginal model $P(Y_1 = y_1 \mid \boldsymbol{Z})$ can be shown as:

$$P_s(\boldsymbol{Z}) = P(Y_1 = s \mid \boldsymbol{Z}) = \frac{e^{(\boldsymbol{Z}'\boldsymbol{\beta}_s)}}{\sum_{s=0}^{2} e^{(\boldsymbol{Z}'\boldsymbol{\beta}_s)}} = \frac{e^{g_s(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_s(\boldsymbol{Z}')}}, \quad s = 0, 1, 2, \tag{4.2}$$

$$\text{where } g_s(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_1 = s \mid \boldsymbol{Z})}{P(Y_1 = 0 \mid \boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_s(\boldsymbol{Z})$ is the first logit of s-*th* component of y_1 conditional on $\boldsymbol{Z}$ and

$$g_s(\boldsymbol{Z}) = \beta_{s0} + \beta_{s1} Z_1 + ... + \beta_{sp} Z_p, \quad s = 1, 2,$$

where $\boldsymbol{Z'} = [1, Z_1, ..., Z_p] = \boldsymbol{X'} = [1, X_1, ..., X_p]$ and $\boldsymbol{\beta'_s} = [\beta_{s0}, \beta_{s1}, ..., \beta_{sp}]$ are the parameter vectors of the *s-th* component for outcome Y_1 where $\boldsymbol{\beta'_1} = [\beta_{10}, \beta_{11}, ..., \beta_{1p}]$ and $\boldsymbol{\beta'_2} = [\beta_{20}, \beta_{21}, ..., \beta_{2p}]$, $\boldsymbol{\beta'_1}$ and $\boldsymbol{\beta'_2}$ are $1 \times (p+1)$ vectors totalling a $[(p+1)2]$ regression coefficients.

4.2.3 First order regressive model

The first order regressive model $P(Y_2 = y_2 \mid Y_1 = y_1, \boldsymbol{Z})$ can be shown as:

$$P_{s.y_1}(\boldsymbol{Z}) = P(Y_2 = s \mid Y_1 = y_1, \boldsymbol{Z}) = \frac{e^{g_{s.y_1}(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_{s.y_1}(\boldsymbol{Z'})}}, \quad s, y_1 = 0, 1, 2, \tag{4.3}$$

$$\text{where } g_{s.y_1}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_2=s|\boldsymbol{Z})}{P(Y_2=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_{s.y_1}(\boldsymbol{Z})$ is the second logit of s-*th* component of y_2 conditional on previous outcome y_1, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_1}(\boldsymbol{Z}) =& \beta_{s.y_1 0} + \beta_{s.y_1 1} Z_1 + ... + \beta_{s.y_1 p} Z_p + \beta_{s.y_1 (p+1)} Z_{p+1} \\ &+ \beta_{s.y_1 (p+2)} Z_{p+2}, \quad s = 1, 2, \end{aligned}$$

where $\boldsymbol{Z'} = [1, Z_1, ..., Z_p, Z_{p+1}, Z_{p+2}] = [\boldsymbol{X'}, \boldsymbol{D'}] = [1, X_1, ..., X_p, D_{11}, D_{12}]$. Here D_{11} and D_{12} are the dummy variables for categories 1 and 2 of outcome Y_1 with 0 as the reference category. Here $\boldsymbol{X'}$ is a $1 \times (p+1)$ and $\boldsymbol{D'}$ is a 1×2 vector producing a total of $[(p+1)+2]2$ regression coefficients.

4.2.4 Second order regressive model

The second order regressive model $P(Y_3 = y_3 \mid Y_1 = y_1, Y_2 = y_2, \boldsymbol{Z})$ is

$$P_{s.y_2,y_1}(\boldsymbol{Z}) = P(Y_3 = s \mid Y_1 = y_1, Y_2 = y_2, \boldsymbol{Z}) = \frac{e^{g_{s.y_2}(\boldsymbol{Z})}}{\sum_{s=0}^{2} e^{g_{s.y_2}(\boldsymbol{Z'})}}, \quad s = 0, 1, 2, \tag{4.4}$$

$$\text{where } g_{s.y_2}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_3=s|\boldsymbol{Z})}{P(Y_3=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, \end{cases}$$

here $g_{s.y_2}(\boldsymbol{Z})$ is the third logit of s-*th* component of y_3 conditional on previous two outcomes y_1, y_2, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_2}(\boldsymbol{Z}) =& \beta_{s.y_2 0} + \beta_{s.y_2 1} Z_1 + ... + \beta_{s.y_2 p} Z_p + \beta_{s.y_2 (p+1)} Z_{p+1} \\ &+ \beta_{s.y_2 (p+2)} Z_{p+2} + \beta_{s.y_2 (p+3)} Z_{p+3}, + \beta_{s.y_2 (p+4)} Z_{p+4} \quad s = 1, 2, \end{aligned}$$

$$\begin{aligned}\text{where } \boldsymbol{Z'} &= [1, Z_1, ..., Z_p, Z_{p+1}, Z_{p+2}, Z_{p+3}, Z_{p+4}] \\ &= \left[\boldsymbol{X'}, \boldsymbol{D'}\right] = [1, X_1, ..., X_p, D_{11}, D_{12}, D_{21}, D_{22}].\end{aligned}$$

Here D_{11} and D_{12} are the dummy variables for categories 1 and 2 of outcome variable Y_1 and D_{21} and D_{22} are the dummy variables for categories 1 and 2 of outcome variable Y_2 with 0 as the reference category. $\boldsymbol{X'}$ is a $1 \times (p+1)$ and $\boldsymbol{D'}$ is a 1×4 vector producing $[(p+1)+4]$ regression coefficients with a total of $[(p+1)+4]2$ regression coefficients.

4.2.5 Higher order multistate regressive model

Above regressive model could be generalized for *k-th* order ($k = j-1$) for S outcome levels is shown as:

$$P_{s.y_{j-1},...,y_1}(\boldsymbol{Z}) = P(Y_j = s \mid Y_1 = y_1, ..., Y_{j-1} = y_{j-1}, \boldsymbol{Z}) = \frac{e^{g_{s.y_{j-1}}(\boldsymbol{Z})}}{\sum_{s=0}^{S} e^{g_{s.y_{j-1}}(\boldsymbol{Z'})}},$$

$$s = 0, 1, 2, ..., S, \tag{4.5}$$

$$\text{where } g_{s.y_{j-1}}(\boldsymbol{Z}) = \begin{cases} 0 & \text{if } s = 0 \\ \ln\left[\frac{P(Y_j=s|\boldsymbol{Z})}{P(Y_j=0|\boldsymbol{Z})}\right] & \text{if } s = 1, 2, ..., S, \end{cases}$$

is the *j-th* logit of the *s-th* component of y_j conditional on previous $j-1$ outcomes $y_1, y_2, ..., y_{j-1}$, $\boldsymbol{Z}$ and

$$\begin{aligned} g_{s.y_{j-1}}(\boldsymbol{Z}) =& \beta_{s.y_{j-1}0} + \beta_{s.y_{j-1}1} Z_1 + ... + \beta_{s.y_{j-1}p} Z_p + \beta_{s.y_{j-1}(p+1)} Z_{p+1} \\ &+ ... + \beta_{s.y_{j-1}(p+S)} Z_{p+S} + \beta_{s.y_{j-1}(p+S+1)} Z_{p+S+1} + ... \\ &+ \beta_{s.y_{j-1}(p+2S)} Z_{p+2S} + ... + \beta_{s.y_{j-1}[p+(j-1)S+1]} Z_{[p+(j-1)S+1]} \\ &+ ... + \beta_{s.y_{j-1}[p+(j-1)S+S]} Z_{[p+(j-1)S+S]}, \quad s = 1, 2, ..., S, j > 1 \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{Z'} =& [1, Z_1, ..., Z_p, Z_{p+1}, ..., Z_{p+S}, \\ & Z_{p+S+1}, ..., Z_{p+2S}, ..., Z_{[p+(j-1)S+1]}, ..., Z_{[p+(j-1)S+S]}\Big] = \left[\boldsymbol{X'}, \boldsymbol{D'}\right] \\ =& \Big[1, X_1, ..., X_p, D_{11}, ..., D_{1S}, D_{21}, ..., D_{2S}, ..., D_{(j-1)1}, ..., D_{(j-1)S}\Big]. \end{aligned}$$

Here, $D_{11}, ..., D_{1S}, D_{21}, ..., D_{2S}, ..., D_{(j-1)1}, ..., D_{(j-1)S}$ are the dummy variables for categories $1, 2, ..., S$ of outcomes $y_1, y_2, ..., y_{j-1}$ with 0 as the reference category, respectively. $\boldsymbol{X'}$ is a $1 \times (p+1)$ vector of covariates and $\boldsymbol{D'}$ is a $1 \times [(j-1)S]$ vector of dummy variables for previous $y_{j-1}, ..., y_1$ outcomes with $S+1$ categories considering 0 as the reference category. There are $[(p+1)+(j-1)S]$ regression coefficients for *s-th* component

of the model and with a total of $[(p+1)+(j-1)S]S$ regression coefficients. Number of parameters in various models are shown in Table 4.1.

TABLE 4.1: Number of parameters for different models.

Models	*Constant only*	*s-th component*	*Full*
Marginal	S	$[p+1]$	$[p+1]S$
First order regressive	S	$[p+1+S]$	$[p+1+S]S$
Second order regressive	S	$[p+1+2S]$	$[p+1+2S]S$
...	...	...	...
$j-1$ *th* order regressive	S	$[(p+1+(j-1)S)]$	$[(p+1+(j-1)S)]S$

It may be noted that first and all higher order regressive models are equivalent to that of the marginal multinomial logistic regression models shown in equation (4.2). Regressive models for any order shown in equation (4.5) can be estimated using appropriate data structure and usual SAS, STATA or R-package or other software capable of fitting multinomial logistic regression. We used R software to do all the computations and 'multinom' function of R package "nnet" is used to fit all the marginal and regressive multinomial models.

4.2.6 Predictive models and joint probabilities

The predicted joint probabilities of $\hat{P}(Y_1=y_1,Y_2=y_2,...,Y_j=y_j \mid \boldsymbol{x})$ can be obtained as:

$$\begin{aligned}\hat{P}(Y_1=y_1,Y_2=y_2,...,Y_J=y_J \mid \boldsymbol{x}) &= \hat{P}(Y_1=y_1 \mid \boldsymbol{x})\times,...,\times \\ \hat{P}(Y_J=s \mid Y_{j-1}=y_{j-1},...,Y_1=y_1;\boldsymbol{x}) &\times \hat{P}(Y_2=y_2 \mid Y_1=y_1;\boldsymbol{x}) \\ = \hat{P}_{y_1}(\boldsymbol{x}) &\times \hat{P}_{y_2.y_1}(\boldsymbol{x}) \times ... \times \hat{P}_{s.y_{j-1},...,y_1}(\boldsymbol{x}). \end{aligned} \tag{4.6}$$

Based on the equation (4.6) the predicted joint probabilities for Y_1 and Y_2 is

$$\hat{P}_{y_1,y_2}(\boldsymbol{x}) = \hat{P}(Y_1=y_1,Y_2=y_2 \mid \boldsymbol{x}),$$

the conditional probability for $Y_2=s$ given Y_1 and $\boldsymbol{x}$ is

$$\hat{P}_{s.y_1}(\boldsymbol{x}) = \hat{P}(Y_2=s,\mid Y_1=y_1;\boldsymbol{x}),$$

and the marginal probability for Y_1 given $\boldsymbol{x}$ is

$$\hat{P}_{y_1}(\boldsymbol{x}) = P(Y_1=y_1 \mid \boldsymbol{x}).$$

The joint probabilities can be predicted using marginal and conditional probabilities as:

$$\begin{aligned}\hat{P}(Y_1=y_1,Y_2=y_2 \mid \boldsymbol{x}) &= \hat{P}(Y_1=y_1 \mid \boldsymbol{x}) \times \hat{P}(Y_2=s,\mid Y_1=y_1;\boldsymbol{x}) \\ \Longrightarrow \hat{P}_{y_1,y_2}(\boldsymbol{x}) &= \hat{P}_{y_1}(\boldsymbol{x}) \times \hat{P}_{s.y_1}(\boldsymbol{x}).\end{aligned} \tag{4.7}$$

Then for outcomes y_1, y_2 with categories 0,1 and 2 and using equation (4.7) we can predict joint probabilities from conditional and marginal probabilities as follows:

$$\hat{P}(Y_1 = y_1, Y_2 = s \mid \boldsymbol{x}) = \hat{P}(Y_1 = y_1; \boldsymbol{x}) \times \hat{P}(Y_2 = s \mid Y_1 = y_1; \boldsymbol{x}),$$
$$s = 0,1,2;\ y_1 = 0,1,2;\ j = 1,2.$$

Similarly, for *j-th* outcomes $y_1, y_2, ..., y_j$ we can predict joint probabilities using marginal and conditional probabilities as:

$$\hat{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_J = s \mid \boldsymbol{x}) = \hat{P}(Y_1 = y_1; \boldsymbol{x}) \times \hat{P}(Y_2 = y_2 \mid Y_1 = y_1; \boldsymbol{x})$$
$$\times ... \times \hat{P}(Y_J = s \mid Y_{j-1} = y_{j-1}, ..., Y_1 = y_1; \boldsymbol{x}), \quad s = 0,1,2;$$
$$y_1, ... y_{j-1} = 0,1,2;\ j = 1,2,...,J.$$

4.3 Tests

The prediction of the joint probability of events is based on the joint model, hence, we need to check the goodness-of-fit of the joint model. It is also of interests to check the association (independence) of repeated outcomes.

4.3.1 Independence of outcomes Y_1 and Y_2

The observed counts of Y_1 and Y_2 from two follow-ups each with 3 categories $(s = 0,1,2)$ as defined in Section (4.2.1) will produce $9 (= 3.3)$ possible outcomes which can be shown as a 3×3 cross-classification table using Y_1 as row and Y_2 as column variables. Let, n_{ab} and e_{ab} are observed and expected cell frequencies. The subscripts a $(a = 0,1,2)$ and b $(b = 0,1,2)$ denotes the categories of Y_1 and Y_2 and n_{a+}, n_{+b}, e_{a+} and e_{+b} are the marginal totals of observed and expected frequencies corresponding to Y_1 and Y_2. Assuming cell counts follow a multinomial sampling and joint probability p_{ab} of (Y_1,Y_2) with the restriction $\sum_a \sum_b e_{ab} = n$, the null hypothesis (H_0 :) is the statistical independence (Agresti, 2013) of Y_1 and Y_2. Under the null hypothesis H_0 :

$$p_{ab} = p_{a+} p_{+b}, \text{ for all } a \text{ and } b, \tag{4.8}$$

where $p_{a+} = n_{a+}/n$ and $p_{+b} = n_{+b}/n$ are the marginal probabilities corresponding to Y_1 and Y_2.

Under the null hypothesis (H_0 :); $e_{ab} = E(n_{ab}) = n p_{a+} p_{+b}$. The estimates of the unknown marginal probability can be obtained from the marginal model shown in equation (4.2). Those are $\hat{p}_{a+} = \hat{P}_{y_1}(\boldsymbol{z})$ and $\hat{p}_{+b} = \hat{P}_{y_2}(\boldsymbol{z})$. Then $\hat{e}_{ab} = n \hat{p}_{a+} \hat{p}_{+b}$.

In the presence of independent variables in the model there will be covariate patterns. Long (1997) proposed to use the mean predicted probability for each count after controlling for independent variables empirically. Islam and Chowdhury (2017) showed model based marginal probabilities can be estimated directly from multinomial distribution. Assuming cell counts follow the Poisson distribution they used the connection between Poisson and multinomial distribution to obtain the predicted probabilities for categories of Y_1 or Y_2 (Islam and Chowdhury, 2017).

4.3.2 Pearson X^2 statistic

Using the observed and expected frequencies the Pearson X^2 statistic to test the null hypothesis of independence (H_0) is:

$$X^2 = \sum_a \sum_b \frac{(n_{ab} - \hat{e}_{ab})^2}{\hat{e}_{ab}}, \tag{4.9}$$

4.3.3 Likelihood-ratio chi-squared statistic

The deviance G^2 can be used to test the above hypothesis. The deviance can be shown as:

$$G^2 = 2\sum_a \sum_b n_{ab} \log\left(\frac{n_{ab}}{\hat{e}_{ab}}\right), \tag{4.10}$$

both the statistics X^2 and G^2 using model based estimates are asymptotically distributed as χ^2 with $(3-1)(3-1) = 4$ degrees of freedom assuming total n fixed for each marginal.

Both the statistics X^2 and G^2 readily generalizes for more than two outcomes. For outcomes Y_1, Y_2 and Y_3 the estimates of joint probability p_{abc} can be obtained from marginal probabilities of Y_1, Y_2 and Y_3 as $\hat{p}_{abc} = \hat{P}_{y_1}(\boldsymbol{z})\hat{P}_{y_2}(\boldsymbol{z})\hat{P}_{y_3}(\boldsymbol{z})$. Then $\hat{e}_{abc} = n\hat{p}_{abc}$. Using observed frequency n_{abc} and expected frequency $\hat{e}_{abc}$ both the test statistics can be calculated as usual manner.

4.3.4 Tests for goodness-of-fit of the joint model of Y_1 and Y_2

In equation (4.1) the joint model shown is based on marginal and conditional models as a function of covariates to predict the sequence of events. A goodness-of-fit test is needed to assess the suitability of such models. The null hypothesis is, H_0 : the fitted model is correct. The observed (n_{ab}) and expected (e_{ab}) frequencies of joint outcome of Y_1 and Y_2 is same as defined previously. The joint probability p_{ab} can be obtain using the marginal and conditional probabilities as

$$p_{ab} = p_{a+}p_{b|a}, \text{ for all } a \text{ and } b \tag{4.11}$$

where, $p_{b|a}$ is the conditional probability, i.e., the probability of classification in column b of Y_2 given that a subject is classified in row a of Y_1 and p_{a+} is marginal probabilities of Y_1. The quantities $e_{ab} = np_{ab}$ are expected frequencies, where

$$\sum_a \sum_b p_{ab} = 1, \qquad \sum_a \sum_b e_{ab} = n.$$

Islam and Chowdhury (2017) proposed a goodness-of-fit test for repeated measures by estimating the joint probabilities using model based marginal and conditional probabilities. They showed that both the marginal probabilities of Y_1 and conditional probabilities of Y_2 for any given value of Y_1 follow multinomial distribution.

The estimated joint probability is $\hat{p}_{ab} = \hat{P}_{y_1}(\boldsymbol{z})\hat{P}_{y_2.y_1}(\boldsymbol{z})$, where $\hat{P}_{y_1}(\boldsymbol{x})$ is estimated from marginal model and $y_1 = s$ in equation (4.2). $\hat{P}_{y_2.y_1}(\boldsymbol{x})$ is estimated from first order regressive model in equation (4.3). It may be noted that $y_2 = s$ in equation (4.3). Then the estimated expected frequency is, $\hat{e}_{ab} = n\hat{p}_{ab}$. The Pearson X^2 and likelihood ratio G^2 can be calculated similarly shown in equations (4.9) and (4.10). Both the statistics X^2 and G^2 are asymptotically distributed as χ^2 with $(3-1)(3-1) = 4$ degrees of freedom as we are using the restriction for both the marginal and conditional models as $\sum_a \hat{p}_{a+} = n$ and $\sum_b \hat{p}_{b|a} = n$ for all a.

In the presence of covariates pattern one can use the predicted empirical means proposed by Long (1997). Alternatively, we can used the predicted probabilities shown by Islam and Chowdhury (2017) using the connection between the Poisson and multinomial.

4.3.5 Goodness-of-fit test of joint model for Y_1, Y_2 and Y_3

The proposed method of goodness-of-fit in previous section readily generalizes for more than two outcomes. For example, three repeated outcomes Y_1, Y_2 and Y_3 each with categories $s = 0, 1, 2$, the expected frequencies p_{abc} can be estimated as:

$$\hat{p}_{abc} = \hat{P}_{y_1,y_2,y_3}(\boldsymbol{x}) = \hat{P}_{y_1}(\boldsymbol{x})\hat{P}_{y_2.y_1}(\boldsymbol{x})\hat{P}_{y_3.y_2,y_1}(\boldsymbol{x}).$$

The estimated marginal probabilities for Y_1 can be obtained from the fitted marginal models shown in equation (4.2) and the estimated conditional probabilities from the fitted first and second order regressive models shown in equations (4.3) and (4.4), respectively. The estimated expected frequencies is then $\hat{e}_{abc} = np_{abc}$ and the Pearson X^2 statistic is:

$$X^2 = \sum_a \sum_b \sum_c \frac{(n_{abc} - \hat{e}_{abc})^2}{\hat{e}_{abc}}, \tag{4.12}$$

and the deviance G^2 can be shown as:

$$G^2 = 2\sum_a\sum_b\sum_c n_{abc}\log\left(\frac{n_{abc}}{\hat{e}_{abc}}\right), \tag{4.13}$$

both the statistics X^2 and G^2 using model based estimates are asymptotically distributed as χ^2 with $(3-1)(3-1)(3-1) = 8$ degrees of freedom due to single restriction on marginal and conditional models as shown previously.

4.3.6 Significance of the joint model

The significance of the joint model can be tested using likelihood ratio test between joint constant only model (Reduced) and joint full model (Full) as follows:

$$-2\left[\ln L_{\text{Reduced}}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - \ln L_{\text{Full}}(\hat{\boldsymbol{\beta}})\right] \tag{4.14}$$

which is distributed asymptotically as χ^2 with $[\{(p+1)S\} + \{(p+1+S)S\} + \{(p+1+2S)S\} + ... + \{p+1+(j-1)S\}S] - jS$ degrees of freedom. Here $\hat{\boldsymbol{\beta}}_{\mathbf{0}}'$ includes all the regression parameters from the constant only joint model and $\hat{\boldsymbol{\beta}}_{\mathbf{1}}'$ includes all the parameters from the full joint model. The above test can be extended to test the significance of a set of covariates which is important especially for the case where there is a group of covariates to choose from.

4.4 Application

For the application we used data from wave (follow-ups) six to eight from the Health and Retirement Study (HRS, 2014). At wave six minimum age of the respondents was 60. In wave one, a total of 12652 subjects were interviewed in the HRS cohort out of which 9762 were age eligible (those with birth years 1931-1941). After removal of cases with missing values for outcome variable at wave six, the number of subjects is 7130. The outcome variables are Activity of Daily Living Index (ADL) from wave six to wave eight (Y_1, Y_2, Y_3). This index is the sum of five tasks (yes/no) ranging from 0 to 5: whether respondents faced difficulties in walking, dressing, bathing, eating and getting in/out of bed. Due to small frequencies 3 and higher values were coded as 2. The explanatory variables considered are: age (in years), marital status (married/partnered=1, single/separated=0), whether drink (yes=1, no=0), sex (male=1, female=0), number of conditions ever had (N.cond) ranges from 0 to 8, White (yes=1, no=0), Black (yes=1, no=0) with others as reference category, education (in years) and veteran status (1=yes, 0= no). The variable drink indicates whether the respondent drinks alcoholic beverages. Table 4.2 displays the frequency distribution of the outcomes for different waves.

TABLE 4.2: Distribution of Activity of Daily Living Index.

	Outcomes					
Outcome	Y_1		Y_2		Y_3	
Value	N	%	N	%	N	%
0	6210	87.1	5906	86.7	5459	84.9
1	477	6.7	462	6.8	503	7.8
2	443	6.2	445	6.5	467	7.3
Total	7130	100.0	6813	100.0	6429	100.0

We assumed outcomes as nominal variables for application purpose. Parameter estimates along with standard error and significance level for marginal and regressive models are shown in Table 4.3. Various predictors are found to be significantly associated with outcome variables for different models. All dummy indicators for previous outcomes are significantly and positively associated with the current outcomes. Likelihood ratio test for the joint model is statistically significant ($p < 0.001$) as shown in Table 4.4. The prediction accuracy based on confusion matrix for full data and training (70% sample) data and test (30% sample) data varies between 0.86 to 0.89 which is reasonably high (Table 4.4). Parameters estimated from training data set were applied to test data set to predict the outcome. Also, accuracy from full, training and test data are very close, which shows the absence of over(under)fitting for all the models.

4.4.1 Tests for independence of outcomes

Both the statistics X^2 and G^2 showed highly significant ($p < 0.001$) association between Y_1 and Y_2 (Table 4.5). The association between Y_1, Y_2 and Y_3 are also found to by highly significant ($p < 0.001$) as shown by both the statistics X^2 and G^2 (Table 4.6) implying dependence in outcome variables.

4.4.2 Tests for goodness-of-fit for joint model

For joint model $P(Y_1, Y_2 \mid \boldsymbol{X})$ both the X^2 and G^2 statistics showed a good-fit ($p = 0.357$) as shown in Table 4.7. The accuracy of joint model is also found to be high (0.866). However, for the joint model $P(Y_1, Y_2, Y_3 \mid \boldsymbol{X})$ the null hypothesis of the good-fit were found to be rejected ($p < 0.001$) by both the statistics X^2 and G^2 (Table 4.8). The joint model accuracy for prediction is computed up to the last follow-up which is shown at the end of Table 4.8. The overall accuracy of the joint model for outcomes Y_1, Y_2 and Y_3 is 0.79 but appears to be higher (0.89) for Y_2 and Y_3 if $Y_1 = 0$. The accuracies between Y_2 and Y_3 are relatively lower for $Y_1 = 1$ and $Y_1 = 2$. The high accuracy for $Y_1 = 0$ may be attributed to the subjects starting without any ADL difficulties. This better prediction is a meaningful finding, because if someone starts without any ADL difficulties, then it is expected that the prediction would be affected less as compared to those who start initially

with minor or severe ADL difficulties at the beginning that might be subject to carryover effect.

4.5 Bootstrapping

To measure the accuracy of sample estimates and proposed test statistics, bootstrapping is used. We performed 10,000 bootstraps and computed bias, standard error, and mean squared error for estimates. Bias is estimated as $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ where $\hat{\theta}$ is an estimator of parameter θ and mean squared error is estimated as $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$. Estimates from Table 4.3 are considered as population parameters while bias, standard error and mean squared error are computed. Bias is very small generally (less than 1 percent) for the estimators of parameters of all models. Standard error and mean squared errors are also found to be very small (Table 4.9-4.10). For both marginal and regressive models, convergence are achieved for all 10,000 bootstrap samples. Density plot for all bootstrap estimates of parameters for all models are shown in Figure 4.2 to Figure 4.6.

For the test for independence of Y_1 and Y_2 and goodness-of-fit of joint model $P(Y_1, Y_2 \mid \boldsymbol{X})$, 9976 bootstrap samples showed valid computation of tests statistic. For remaining 24 bootstrap samples some cells frequencies were empty hence test statistics could not be computed. Test of independence were significant ($p < 0.05$) for all 9976 bootstrap samples. The bias for X^2 was 11.77 with estimated standard error 140.6 and the bias for G^2 was 6.65 with estimated standard error 84.2. For goodness-of-fit statistic (X^2) only 38 bootstrap samples (0.38%) rejected the hypothesis of goodness-of-fit with 0.216 estimated bias and 1.582 estimated standard error. For (G^2) only 70 bootstrap samples (0.70%) rejected the hypothesis of goodness-of-fit with 0.230 estimated bias and 1.664 estimated standard error.

For three outcomes (Y_1, Y_2, Y_3), 9976 and 9973 bootstrap samples produced valid test statistics for test of independence, X^2 and G^2, respectively. All 9976 bootstrap samples for X^2 and 9973 bootstrap samples for G^2 showed significant ($p < 0.05$) association between outcomes Y_1, Y_2, Y_3. The estimated bias and standard error of X^2 are 136.9 and 1241.1. For G^2 these are 23.46 and 143.42. For goodness-of-fit-test, out of 10000 bootstrap samples 9949 valid X^2 and 9947 valid G^2 were produced. For X^2 among 9949 bootstrap samples 9944 (99.95%) rejected the hypothesis of goodness-of-fit and for G^2 among 9947 bootstrap samples 9944 (99.97%) rejected the hypothesis of goodness-of-fit. The estimated bias and standard error of X^2 are 8.241 and 8.680. For G^2 these are 9.085 and 9.385.

TABLE 4.3: Estimates of marginal and regressive models using multinomial logistic regression.

	$\hat{P}(Y_1 \mid \boldsymbol{X})$				$\hat{P}(Y_2 \mid Y_1;\boldsymbol{X})$				$\hat{P}(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$			
	Category 1		Category 2		Category 1		Category 2		Category 1		Category 2	
Variables	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_2$	S.E.
Constant	−1.480	1.068	−1.923	1.144	−4.664**	1.189	−4.513**	1.426	−6.160**	1.239	−5.104**	1.539
Age	−0.011	0.016	−0.004	0.017	0.017	0.017	0.012	0.020	0.038*	0.017	0.024	0.021
Mstat	−0.332**	0.107	−0.536**	0.113	−0.031	0.117	−0.580**	0.136	0.006	0.116	−0.234	0.142
N.cond	0.497**	0.034	0.670**	0.036	0.425**	0.037	0.466**	0.043	0.330**	0.037	0.415**	0.045
Drink	−0.321**	0.107	−0.790**	0.128	−0.163	0.115	−0.613**	0.150	−0.126	0.111	−0.471**	0.146
Gender	0.016	0.128	−0.008	0.141	0.116	0.135	0.476**	0.164	0.214	0.135	0.028	0.174
White	−0.070	0.253	−0.548*	0.235	−0.154	0.274	−0.428	0.290	0.237	0.315	−0.114	0.335
Black	0.025	0.268	−0.121	0.247	0.029	0.291	−0.257	0.309	0.492	0.330	0.411	0.350
Educ.	−0.091**	0.015	−0.087**	0.016	−0.018	0.017	−0.028	0.019	−0.038*	0.017	−0.083**	0.020
Veteran	−0.037	0.153	−0.157	0.176	−0.354*	0.166	−0.325	0.202	−0.133	0.158	−0.153	0.218
D61					1.802**	0.138	2.258**	0.163	1.658**	0.146	1.888**	0.175
D62					2.483**	0.183	4.245**	0.169	2.197**	0.212	3.558**	0.201
D71									1.204**	0.153	1.138**	0.183
D72									1.274**	0.215	2.096**	0.206

Note: ** significant at 1 % level; * significant at 5 % level

TABLE 4.4: Model statistics for marginal and regressive models.

Out-	Constant only model			Full model			L.R.T
comes	Log L.	Dev.	AIC	Log L.	Dev.	AIC	p.v. (d.f.)
Y_1	-3378.9	6757.7	6761.7	-2921.5	5842.9	5882.9	0.000 (18)
Y_2	-3195.9	6391.7	6395.7	-2309.0	4618.0	4666.0	0.000 (22)
Y_3	-3283.0	6566.0	6570.0	-2244.9	4489.8	4545.8	0.000 (26)
Joint model	-19300.8	38601.4	38625.4	-13518.2	27036.3	27396.3	0.000 (72)

	Accuracy		
Model	All	Train	Test
$P(Y_1 \mid \boldsymbol{X})$	0.873	0.871	0.878
$P(Y_2 \mid Y_1; \boldsymbol{X})$	0.887	0.887	0.881
$P(Y_3 \mid Y_2, Y_1; \boldsymbol{X})$	0.878	0.878	0.875

TABLE 4.5: Observed and expected frequencies for independence test of Y_1 and Y_2.

	Y_2							
	0		1		2		Total	
Y_1	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$
0	5412	4997	268	390	121	373	5801	5761
1	247	384	108	30	88	29	443	443
2	79	357	72	28	220	27	371	411
Total	5738	5739	448	448	429	429	6615	6615
X^2	2305.92 (d.f=4, p<0.001)		G^2	1471.81 (d.f=4, p<0.001)				

Note: Expected frequencies are rounded to zero decimal place

TABLE 4.6: Observed and expected frequencies for independence test of Y_1, Y_2 and Y_3.

		Y_3							
		0		1		2		Total	
Y_1	Y_2	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$
0	0	4728	4624	231	234	100	98	5059	4956
	1	137	156	59	55	38	34	234	245
	2	35	37	21	25	36	50	92	111
1	0	161	168	43	40	21	20	225	228
	1	38	33	38	41	24	26	100	100
	2	10	11	22	24	44	46	76	81
2	0	42	50	16	14	14	17	72	81
	1	21	16	17	23	27	35	65	74
	2	9	17	27	35	140	173	176	224
Total								6099	6099
X^2	14357.2 (d.f=8, p<0.001)			G^2	3046.5 (d.f=8, p<0.001)				

Note: Expected frequencies are rounded to zero decimal place

TABLE 4.7: Goodness-of-fit test for joint model $P(Y_1, Y_2 \mid \boldsymbol{X})$.

	Y_2							
	0		1		2		Total	
Y_1	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$	n_{ab}	$\hat{e}_{ab}$
0	5412	5375	268	266	121	120	5801	5761
1	247	247	108	108	88	88	443	443
2	79	88	72	80	220	243	371	411
Total	5738	5710	448	454	429	452	6615	6615
X^2	4.24 (d.f=4, p<0.357)		G^2	(4.38 d.f=4, p<0.357)				
Accuracy	0.8657							

Note: Expected frequencies are rounded to zero decimal place

TABLE 4.8: Goodness-of-fit test for joint model $P(Y_1, Y_2, Y_3 \mid \boldsymbol{X})$.

		Y_3							
		0		1		2		Total	
Y_1	Y_2	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$	n_{abc}	$\hat{e}_{abc}$
0	0	4728	4624	231	234	100	98	5059	4956
	1	137	156	59	55	38	34	234	245
	2	35	37	21	25	36	50	92	111
1	0	161	168	43	40	21	20	225	228
	1	38	33	38	41	24	26	100	100
	2	10	11	22	24	44	46	76	81
2	0	42	50	16	14	14	17	72	81
	1	21	16	17	23	27	35	65	74
	2	9	17	27	35	140	173	176	224
Total								6099	6099
X^2		30.6 (d.f=8, p<0.001)		G^2	(32.6 d.f=8, p<0.001)				

Overall Accuracy	0.7932
Accuracy between Y_2 and Y_3 for $Y_1 = 0$	0.8901
Accuracy between Y_2 and Y_3 for $Y_1 = 1$	0.6235
Accuracy between Y_2 and Y_3 for $Y_1 = 2$	0.6674

Note: Expected frequencies are rounded to zero decimal place

TABLE 4.9: Bootstrap parameter estimates for marginal and first order regressive models.

	$\hat{P}(Y_1 \mid \boldsymbol{X})$						$\hat{P}(Y_2 \mid Y_1; \boldsymbol{X})$					
	Category 1			Category 2			Category 1			Category 2		
Variables	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$
Constant	-0.023	1.046	1.095	-0.014	1.167	1.361	-0.047	1.201	1.445	-0.046	1.432	2.051
Age	0.000	0.015	0.000	0.000	0.017	0.000	0.000	0.017	0.000	0.000	0.021	0.000
Mstat	-0.001	0.109	0.012	0.000	0.115	0.013	0.003	0.120	0.014	-0.003	0.137	0.019
N.cond	0.001	0.033	0.001	0.003	0.034	0.001	0.001	0.037	0.001	0.002	0.046	0.002
Drink	-0.002	0.109	0.012	-0.006	0.132	0.018	0.000	0.117	0.014	-0.003	0.152	0.023
Gender	0.000	0.129	0.017	-0.002	0.143	0.021	-0.005	0.137	0.019	0.001	0.165	0.027
White	0.026	0.272	0.075	0.012	0.232	0.054	0.030	0.280	0.080	0.017	0.286	0.082
Black	0.021	0.286	0.082	0.011	0.244	0.060	0.027	0.299	0.090	0.016	0.310	0.097
Education	0.000	0.015	0.000	0.000	0.016	0.000	0.000	0.017	0.000	0.000	0.020	0.000
Veteran	-0.002	0.155	0.024	-0.004	0.174	0.030	0.001	0.163	0.027	-0.002	0.194	0.038
D61							0.004	0.147	0.022	0.011	0.174	0.031
D62							0.009	0.192	0.037	0.030	0.181	0.034

4.6 Conclusions

Use of multinomial outcomes from repeated measures data to predict the risk of a sequence of events are growing in recent years. Markov chain is used to link the marginal and conditional probabilities for estimating joint probability of a sequence of events. Conditional probabilities can be obtained using conditional models or regressive models. The goodness-of-fit of the joint model need to be checked for model performance. In this paper, we proposed a goodness-of-fit test for joint model obtained by linking marginal and conditional models. Tests for the independence of repeated outcomes are also shown. Application of the proposed tests are shown using the HRS data from the USA. Activity of Daily Living Index (ADL) from follow-up six to eight (Y_1, Y_2, Y_3) are used as multinomial outcome variables. Test of independence of outcomes showed significant departure from null hypothesis for both bi-variate and tri-variate outcomes. The hypothesis of goodness-of-fit is not rejected for the joint model $\hat{P}(Y_1, Y_2 \mid \boldsymbol{X})$. However, for the joint model $\hat{P}(Y_1, Y_2, Y_3 \mid \boldsymbol{X})$ for the selected covariates the hypothesis for goodness-of-fit is rejected. The acceptance of the hypothesis of goodness-of-fit for the joint model $\hat{P}(Y_1, Y_2 \mid \boldsymbol{X})$ and rejection for the joint model $\hat{P}(Y_1, Y_2, Y_3 \mid \boldsymbol{X})$ are also confirmed by the bootstrap simulation results.

To measure the performance of the test statistics and regression parameters 10000 bootstrap simulation is performed. Bootstrap estimates of the most of the regression parameters showed less than 1 percent bias along with the low estimated standard errors. Both the test statistics (X^2 and G^2) from all bootstrap sample for independence showed significance association for two and three outcomes. This is in line with the significant result found from full sample (Table 4.5 and Table 4.6). Bootstrap estimates of both the goodness-of-fit statistics (X^2 and G^2) were in agreement to those from Table 4.7 and Table 4.8). The proposed tests readily generalize for more than three outcomes and can easily be performed using existing software.

TABLE 4.10: Bootstrap parameter estimates for second order regressive model.

Variables	$\hat{P}(Y_3 \mid Y_1, Y_2; \boldsymbol{X})$					
	Category 1			Category 2		
	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$	$\widehat{Bias}$	$\widehat{S.E.}$	$\widehat{MSE}$
Constant	-0.071	1.270	1.617	-0.067	1.615	2.614
Age	0.000	0.017	0.000	0.000	0.022	0.001
Mstat	0.003	0.118	0.014	-0.002	0.152	0.023
N.cond	0.002	0.038	0.001	0.003	0.048	0.002
Drink	-0.002	0.112	0.013	-0.006	0.148	0.022
Gender	-0.004	0.141	0.020	-0.004	0.184	0.034
White	0.039	0.338	0.116	0.026	0.337	0.114
Black	0.036	0.352	0.125	0.026	0.350	0.124
Education	0.000	0.018	0.000	-0.001	0.022	0.001
Veteran	0.000	0.161	0.026	-0.001	0.225	0.051
D61	0.004	0.162	0.026	0.001	0.193	0.037
D62	-0.000	0.230	0.053	0.013	0.219	0.048
D71	0.006	0.155	0.024	0.008	0.190	0.036
D72	0.013	0.221	0.049	0.032	0.210	0.045

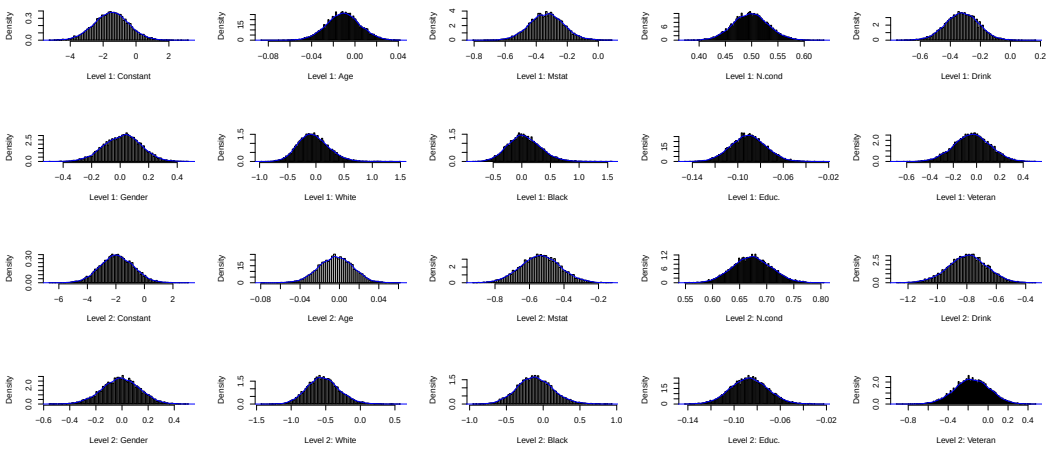


FIGURE 4.2: Density plot of bootstrap estimates for marginal model $P(Y_1 \mid \boldsymbol{X})$.

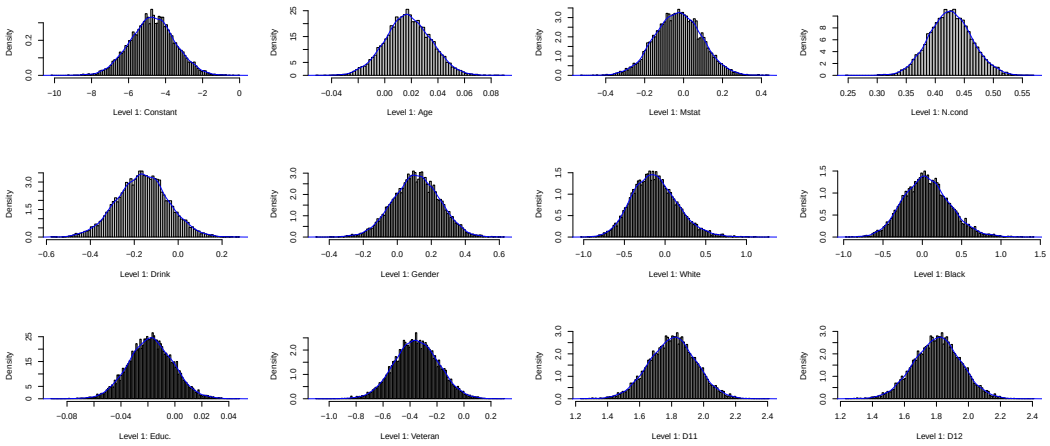


FIGURE 4.3: Density plot of bootstrap estimates for regressive model $P(Y_2 = 1 \mid Y_1; \boldsymbol{X})$.

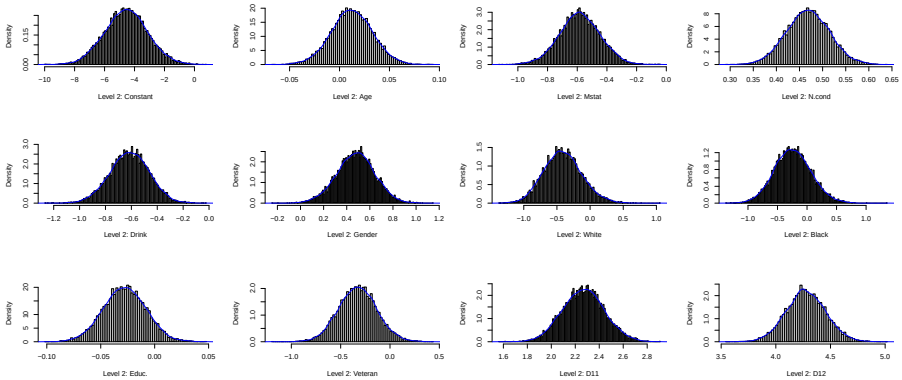


FIGURE 4.4: Density plot of bootstrap estimates for regressive model $P(Y_2 = 2 \mid Y_1; \boldsymbol{X})$.

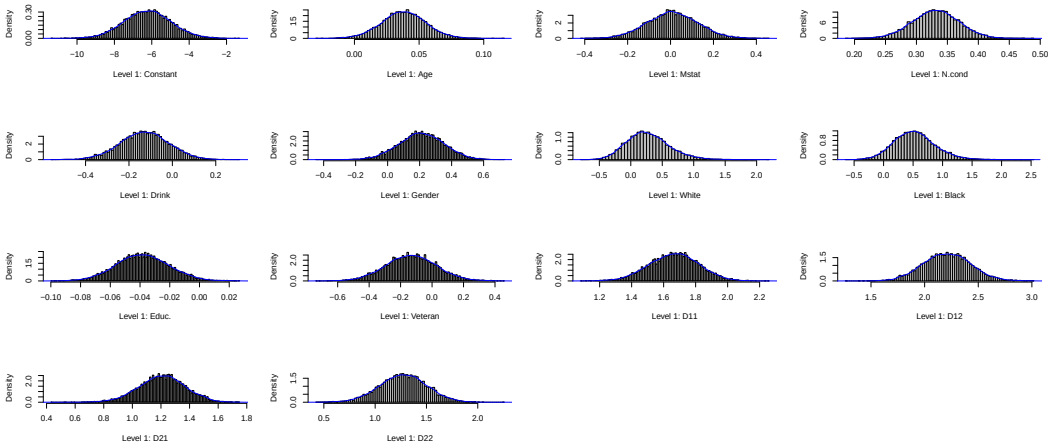


FIGURE 4.5: Density plot of bootstrap estimates for regressive model $P(Y_3 = 1 \mid Y_1, Y_2; \boldsymbol{X})$.

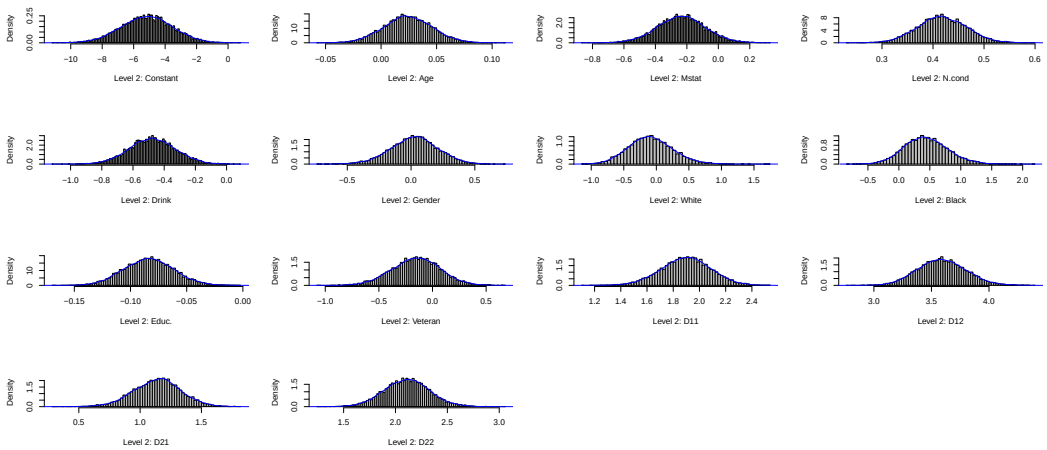


FIGURE 4.6: Density plot of bootstrap estimates for regressive model $P(Y_3 = 2 \mid Y_1, Y_2; \boldsymbol{X})$.

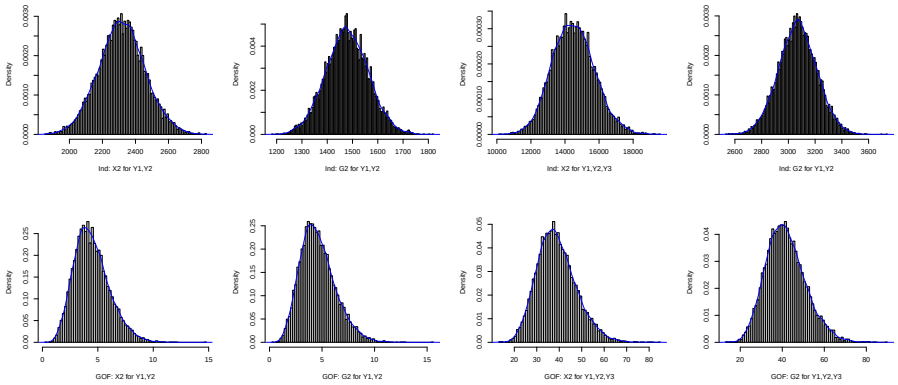


FIGURE 4.7: Density plot of bootstrap estimates for the test statistics.

Chapter 5

Regressive Models for Risk Prediction of a Sequence of Ordinal Outcomes from Repeated Measures

5.1 Introduction

The ordinal outcomes from longitudinal studies are repeatedly observed over time and increasingly uses in many fields of studies such as epidemiology, public health, genetics, reliability, environmental studies, ecology. The outcomes may represent disease status at different stages which can be viewed as a long sequences of discrete events over time. The interest is to model an outcome at specific follow-up with risk factors and status at previous outcomes recorded before that follow-up to understand the disease progression over time and risk of outcome prediction (Bodilsen *et al.*, 2016; Barnes *et al.*, 2013; Wallace *et al.*, 2014; Gundersen *et al.*, 2009; Fox *et al.*, 2016; Bovelstad *et al.*, 2009). Another growing area of interest is to predict the joint probability of a sequence of events based on specified covariates vector (Wen *et al.*, 2016; Islam and Chowdhury, 2010; Lee and Daniels, 2007; Miller *et al.*, 2001; Liski and Nummi, 1996; Yu, 2003). For example, physical activity may prospectively relate to the progression of functional limitations and disability among elderly (Beddoes-Ley *et al.*, 2016) and may increase the utilization of health care services. Modeling these sequences, allow us to predict likely future outcomes. The estimation and prediction resulting from a sequence of ordinal outcomes based on specified covariates from repeated measures data is a challenge to the researchers. To predict the joint probability of a sequence of outcomes we need to examine the sequence of events during subsequent follow-ups using a joint model (multivariate) for ordinal outcomes. From an application point of view, a multivariate approach is often complicated and would be difficult to develop for a large number of follow-ups (Gottschau, 1994).

The multistate higher order Markov model (conditional model) can be used to study the underlying dependence in consecutive follow-ups (Islam *et al.*, 2009). Using this model

one can investigate the relationship between recent outcomes and predictors including previous outcomes status and risk could be calculated for a sequence of events (Islam and Chowdhury, 2010; Islam *et al.*, 2012). However, for a large number of repeated outcomes, this approach involves fitting many conditional models which appear to be restricted due to over-parameterization (Gottschau, 1994; Islam *et al.*, 2013). Figure 5.1 displays three repeated outcomes each with three categories and twenty-seven possible trajectories (paths). To obtain the joint model, one needs to fit thirteen models, one marginal model for the outcome at follow-up one or baseline, three first order and nine-second order Markov models which could be computationally cumbersome and explodes for a large number of repeated outcomes (Gottschau, 1994; Islam *et al.*, 2013). Another choice is the regressive logistic models under the Markovian assumption which include both binary outcomes in previous times in addition to covariates in the conditional models proposed by various authors (Muenz and Rubinstein, 1985; Bonney, 1986, 1987; Azzalini, 1994; Islam *et al.*, 2004; Islam and Chowdhury, 2006, 2010; Islam *et al.*, 2013; Tripepi *et al.*, 2013; Islam *et al.*, 2014). Islam and Chowdhury (2010) proposed a regressive logistic model to predict the joint probability of a sequence of binary outcomes based on specified covariates which reduce the fitting of conditional models significantly.

Several types of regression models were proposed considering the ordinal nature of the outcome, for example, mixed models or probit models (Walters *et al.*, 2001; Lall *et al.*, 2002). The ordinal logistic regression models with different variants is a popular approach to model ordinal response (McCullagh, 1980; McCullagh and Nelder, 1983; Anderson, 1984; Brant, 1990; Ananth and Kleinbaum, 1997; Hosmer and Lemeshow, 2000). For example, proportional odds, partial proportional odds, continuation ratio, stereotype, adjacent category, baseline category and multinomial regression models. However, these are univariate models only for the single ordinal outcome.

At this backdrop, we proposed three regressive models for repeated ordinal outcomes and joint model (multivariate) model is shown which are new developments. The proposed model includes covariates, as well as the ordinal responses from previous follow-ups, and a re-parameterization is suggested that reduces the number of parameter sets need to be estimated. First, we propose proportional odds regressive model for repeated ordinal outcomes by extending the POM model for a single outcome. For POM the proportional odds assumption needs to be tested (Brant, 1990). Second, in the case of violations of proportional odds assumption for some covariates, we proposed partial proportional odds regressive model for repeated ordinal outcomes. Finally, the multinomial logistic regressive model is shown for repeated ordinal outcomes by ignoring the ordinal nature of the response variables. The risk for a sequence of events for specified covariates value is estimated by linking marginal and conditional probabilities. Marginal probability is obtained using proportional odds, partial proportional odds and multinomial models for the outcome from the first follow-up or baseline. The conditional probabilities are estimated from the proposed regressive models and the prediction of a sequence of outcomes is

shown. A Goodness-of-fit test for the joint model is also proposed. Using data partitioning (training and tests data) prediction accuracy is shown to check over(under)fitting. Finally, an application is shown using follow-up data from the Health and Retirement Study (HRS), in the USA.

5.2 Repeated Outcomes and Trajectories

Consider three repeated ordinal outcomes $(Y_1, Y_2$ and $Y_3)$ from a longitudinal study with three categories (0,1,2). Figure 5.1 displays the possible transitions between three outcome categories from three follow-ups. A total of twenty-seven distinct trajectories (paths) are possible. Outcome categories are shown inside the rectangles. Here, first column shows marginal probabilities and second onward are conditional probabilities.To model such outcomes natural choice is proportional odds model (ordinal logistic regression) assuming proportional odds assumption holds. When this assumption violates partial proportional odds, and multinomial logistic regression models are alternative choices among others.

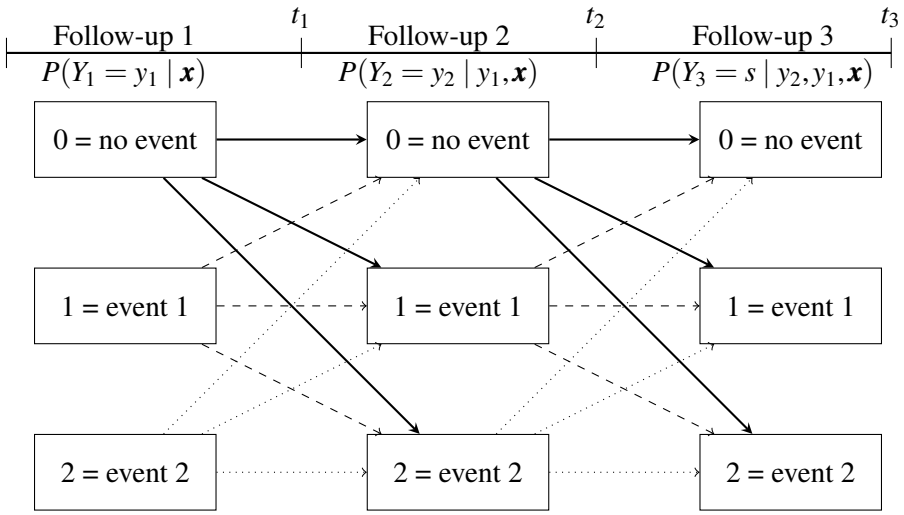


FIGURE 5.1: Transitions between states for regressive models.

5.2.1 Notations

Let $Y_{i1}, Y_{i2}, ..., Y_{iJ_i}$ represent the past and present responses for *i-th* subject at *j-th* follow-up where $(i = 1, 2, ..., n)$ and $(j = 1, 2, ..., J_i)$, J_i is the number of follow-ups for subject i. For simplicity, subscript i is omitted what follows next unless explicitly specified. Define, $Y_j = s$ where $(s = 0, 1, 2, ..., S)$ with $S+1$ outcome categories. The category 0 may denote non-event.

The joint probability mass function of $Y_1, Y_2, ..., Y_J$ with covariate vector $\boldsymbol{X} = \boldsymbol{x}$ can be expressed as:

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_J = y_J \mid \boldsymbol{x}) &= P(Y_1 = y_1 \mid \boldsymbol{x}) \times P(Y_2 = y_2 \mid y_1; \boldsymbol{x}) \\ &\times \cdots \times P(Y_J = s \mid y_1, \cdots, y_{j-1}; \boldsymbol{x}) = P_{y_1}(\boldsymbol{x}) \times P_{y_2.y_1}(\boldsymbol{x}) \\ &\times \cdots \times P_{s.y_1}, \cdots, y_{j-1}(\boldsymbol{x}), \end{aligned} \tag{5.1}$$

where $\boldsymbol{X}' = [1, x_1, ..., x_p]$ is vector of covariates for a subject at first follow-up. It should be noted that $\boldsymbol{X} = \boldsymbol{x}$ can be time dependent. Where

$P(Y_1 = s \mid \boldsymbol{x}) = P_s(\boldsymbol{x})$ is the marginal probability function of Y_1 conditional on $\boldsymbol{x}$;

$P(Y_J = s \mid y_{j-1}; \boldsymbol{x}) = P_{s.y_{j-1}}(\boldsymbol{x})$ is the probability function of Y_j conditional on y_{j-1} and $\boldsymbol{x}$ of order one;

$P(Y_J = s \mid y_{j-2}, y_{j-1}; \boldsymbol{x}) = P_{s.y_{j-2}, y_{j-1}}(\boldsymbol{x})$ is the probability function for Y_j conditional on y_{j-2},y_{j-1} and $\boldsymbol{x}$ of order two;

$P(Y_J = s \mid y_1, ..., y_{j-2}, y_{j-1}; \boldsymbol{x}) = P_{s.y_1, \cdots, y_{j-2}, y_{j-1}}(\boldsymbol{x})$ is the probability function of Y_j conditional on $y_1, ..., y_{j-1}$ and $\boldsymbol{x}$ of order $k = j - 1$.

The unconditional probability of the left hand side of equation (5.1) is defined as:

$P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J \mid \boldsymbol{x}) = P_{y_1, y_2, ..., y_J}(\boldsymbol{x})$.

The log-likelihood function of the joint mass function in (5.1) can be obtained as:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \sum_{j=1}^{J} \ln P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \cdots, Y_{iJ} = y_{iJ} \mid \boldsymbol{x}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \Big[\ln P(Y_{i1} = y_{i1} \mid \boldsymbol{x}) + \ln P(Y_{i2} = y_{i2} \mid y_{i1}; \boldsymbol{x}) \\ &+ \cdots + \ln P(Y_{iJ} = s \mid y_{i1}, \cdots, y_{i(j-1)}; \boldsymbol{x}) \Big]. \end{aligned} \tag{5.2}$$

5.2.2 Models

To obtain the joint model in Equation (5.1), we need to fit marginal and a series of conditional models depending on the order of the joint model. Then we can estimate the marginal and conditional probability from the marginal and conditional models and predict the joint probability with a specified covariates vector. In this Section, the alternative marginal and conditional models as displayed in Equation (5.1) are proposed and the predictive and joint models are proposed later. With increasing number of follow-ups, a large number of conditional models are required to fit which may be impractical or intractable. A better choice is to use the regressive model from which conditional probability can be estimated (Islam and Chowdhury, 2010). This approach requires to fit only one model for

each repeated outcomes by incorporating previous outcomes as covariates along with the risk factors. Following subsection details the proposed proportional odds regressive, partial proportional odds regressive and regressive multinomial logistic models for repeated ordinal outcomes.

5.2.3 Proportional odds model (POM)

Proportional odds model was proposed by McCullagh (1980) to analyze ordinal outcomes as a function of covariates. The proportional odds model (POM), also known as the cumulative odds or cumulative logit model is the most commonly used ordinal logistic model which is based on cumulative probabilities. POM assumes that the coefficients that describe the relationship between the lowest versus all higher categories of the outcomes are the same as those that describe the relationship between the next lowest category and all higher categories (proportional odds assumption or the parallel regression assumption). The proportional odds assumption needs to be tested (Brant, 1990). Fitting of POM using baseline outcome as a function of covariates will provide a marginal model and hence marginal probability. Let, the outcome Y_1 having s categories $(s = 0,1,\cdots,S)$ with associated probabilities $\pi_0+\pi_1+\cdots+\pi_S$ and $P(Y_1 \leq s) = \pi_0+\cdots+\pi_s$ where $P(Y_1 \leq 0) \leq P(Y_1 \leq 1) \leq \cdots \leq P(Y_1 \leq S)$. Then the proportional odds model can be shown as:

$$P(Y_1 \leq s \mid \boldsymbol{x}) = \frac{exp(\alpha_j - \boldsymbol{\beta_1'X})}{1+exp(\alpha_j - \boldsymbol{\beta_1'X})}, \quad s = 1,2,\cdots,S \tag{5.3}$$

or equivalently can be expressed in logit form as

$$\begin{aligned} logit[P(Y_1 \leq s \mid \boldsymbol{x})] = \ln\left[\frac{\pi_0+\cdots+\pi_s}{\pi_{s+1}+\cdots+\pi_S}\right] &= \alpha_s - (\beta_1 X_1+\cdots+\beta_p X_p) \\ &= \alpha_s - \boldsymbol{\beta_1'X} \end{aligned} \tag{5.4}$$

where α_s's are the threshold parameters (cut points) and $\boldsymbol{\beta_1} = [\beta_1,\beta_2,\cdots,\beta_p]'$ is the vector of regression coefficients corresponding to the covariate vectors $\boldsymbol{X} = [X_1,X_2,\cdots,X_p]'$. This model assumes that the effects of the covariates are same for all categories (proportional odds). Then the marginal probability of s-th category is

$$P_s(\boldsymbol{x}) = P(Y_1 = s \mid \boldsymbol{x}) = P(Y_1 \leq s+1 \mid \boldsymbol{x}) - P(Y_1 \leq s \mid \boldsymbol{x}), \quad s = 0,1,\cdots,S. \tag{5.5}$$

5.2.4 Proposed first order proportional odds regressive model

For first order conditional model, we need to fit three proportional odds models for Y_2 as a function of $\boldsymbol{x}$ by stratifying on Y_1. However, as in regressive model (Islam and Chowdhury, 2010), we can fit single proportional odds model for Y_2 as a function of $\boldsymbol{x}$ and Y_1. Then

from the fitted model, we can estimate the conditional probabilities for different categories of Y_2 given Y_1 and $\boldsymbol{x}$. Consider two repeated outcomes Y_1 and Y_2 each having s categories $(s = 0, 1, \cdots, S)$. Then following Islam and Chowdhury (2010) the first order proportional odds regressive model can be shown as:

$$\begin{aligned} &logit[P(Y_2 \leq s \mid \boldsymbol{z})] = \alpha_{s.y_1} - (\beta_{2.y_1 1} Z_1 + \cdots + \beta_{2.y_1 p} Z_p + \beta_{2.y_1 (p+1)} Z_{11} \\ &+ \cdots + \beta_{2.y_1 (p+S)} Z_{1S}) = \alpha_{s.y_1} - \boldsymbol{\beta}'_{2.y_1} \boldsymbol{Z}, \quad s = 1, 2, \cdots, S \end{aligned} \tag{5.6}$$

above logit is conditional on the previous outcome Y_1 and $\boldsymbol{x}$ where $\alpha_{s.y_1}$'s are the threshold parameters and

$$\boldsymbol{\beta}_{2.y_1} = [\beta_{2.y_1 1}, \cdots, \beta_{2.y_1 p}, \beta_{2.y_1 (p+1)}, \cdots, \beta_{2.y_1 (p+S)}]' \tag{5.7}$$

is the vector of regression coefficients corresponding to the covariate vectors

$$\boldsymbol{Z} = [Z_1, \cdots, Z_p, Z_{p+1}, \cdots, Z_{p+S}]' = [\boldsymbol{X}', \boldsymbol{D}'] = [X_1, X_2, \cdots, X_p, D_{11}, \cdots, D_{1S}]'. \tag{5.8}$$

Here, $D_{11}, D_{12}, \cdots, D_{1S}$ are the dummy variables for categories $1, 2, \cdots, S$ for Y_1 with 0 as the reference category. The conditional probability of s-th category is

$$\begin{aligned} P_{s.y_1}(\boldsymbol{z}) = P(Y_2 = s \mid y_1; \boldsymbol{x}) = P(Y_2 \leq s+1 \mid y_1; \boldsymbol{x}) - P(Y_2 \leq s \mid y_1; \boldsymbol{x}), \\ s, y_1 = 0, 1, \cdots, S. \end{aligned} \tag{5.9}$$

5.2.5 Proposed second order proportional odds regressive model

Similarly the second order proportional odds regressive model for outcomes Y_1, Y_2 and Y_3 can be shown as:

$$\begin{aligned} &logit[P(Y_3 \leq s \mid \boldsymbol{z})] = \alpha_{s.y_2} - (\beta_{3.y_2 1} Z_1 + \cdots + \beta_{3.y_2 p} Z_p + \beta_{3.y_2 (p+1)} Z_{p+1} \\ &+ \cdots + \beta_{3.y_2 (p+S)} Z_{p+S} + \beta_{3.y_2 (p+S+1)} Z_{p+S+1} + \cdots + \\ &\beta_{3.y_2 (p+2S)} Z_{p+2S}) = \alpha_{s.y_2} - \boldsymbol{\beta}'_{3.y_2} \boldsymbol{Z}, \quad s = 1, 2, \cdots, S \end{aligned} \tag{5.10}$$

logit in equation (5.10) is conditional on previous two outcomes Y_1 and Y_2 and $\boldsymbol{x}$ where $\alpha_{s.y_2}$'s are the threshold parameters and

$$\begin{aligned} \boldsymbol{\beta}_{3.y_2} = [&\beta_{3.y_2 1}, \cdots, \beta_{3.y_2 p}, \beta_{3.y_2 (p+1)}, \cdots, \beta_{3.y_2 (p+S)}, \beta_{3.y_2 (p+S+1)} \\ &, \cdots, \beta_{3.y_2 (p+2S)}]' \end{aligned} \tag{5.11}$$

is the vector of regression coefficients corresponding to the covariate vectors

$$\begin{aligned} \boldsymbol{Z} &= [Z_1, \cdots, Z_p, Z_{p+1}, \cdots, Z_{p+S}, Z_{p+S+1}, \cdots, Z_{p+2S}]' \\ &= [\boldsymbol{X}', \boldsymbol{D}'] = [X_1, X_2, \cdots, X_p, D_{11}, \cdots, D_{1S}, D_{21}, \cdots, D_{2S}]'. \end{aligned} \tag{5.12}$$

Here, $D_{11}, D_{12}, \cdots, D_{1S}, D_{21}, \cdots, D_{2S}$ are the dummy variables for categories $1, 2, \cdots, S$ for Y_1 and Y_2 with 0 as the reference category. Then the conditional probability of s-th category is

$$\begin{aligned} P_{s.y_1y_2}(\boldsymbol{z}) &= P(Y_3 = s \mid y_1, y_2; \boldsymbol{x}) = P(Y_3 \leq s+1 \mid y_1, y_2; \boldsymbol{x}) \\ &\quad - P(Y_3 \leq s \mid y_1, y_2; \boldsymbol{x}), \quad s, y_1, y_2 = 0, 1, \cdots, S. \end{aligned} \tag{5.13}$$

5.2.6 Proposed higher order proportional odds regressive model

Above regressive model readily generalizes for outcomes $Y_1, Y_2, \cdots, Y_j$ as:

$$\begin{aligned} &logit[P(Y_j \leq s \mid \boldsymbol{z})] = \alpha_{s.y_{j-1}} - (\beta_{j.y_{j-1}1} Z_1 + \cdots + \beta_{j.y_{j-1}p} Z_p + \beta_{j.y_{j-1}(p+1)} Z_{p+1} \\ &+ \cdots + \beta_{j.y_{j-1}(p+S)} Z_{p+S} + \beta_{j.y_{j-1}(p+S+1)} Z_{p+S+1} + \cdots + \beta_{j.y_{j-1}(p+2S)} Z_{p+2S} \\ &+ \cdots + \beta_{j.y_{j-1}[p+(j-2)S+1]} Z_{p+(j-2)S+1} + \cdots + \beta_{j.y_{j-1}[p+(j-1)S]} Z_{p+(j-1)S}) \\ &= \alpha_{s.y_{j-1}} - \boldsymbol{\beta}'_{j.y_{j-1}} \boldsymbol{Z}, \quad s = 1, 2, \cdots, S \end{aligned} \tag{5.14}$$

where α_s's are the threshold parameters and

$$\begin{aligned} \boldsymbol{\beta}_{j.y_{j-1}} &= [\beta_{j.y_{j-1}}, \cdots, \beta_{j.y_{j-1}p}, \beta_{j.y_{j-1}(p+1)} \cdots, \beta_{p+S}, \beta_{j.y_{j-1}(p+S+1)}, \cdots, \beta_{p+2S}, \\ &\cdots, \beta_{j.y_{j-1}(p+2S)}, \cdots, \beta_{j.y_{j-1}[p+(j-2)S+1]}, \cdots, \beta_{j.y_{j-1}[p+(j-1)S]}]' \end{aligned} \tag{5.15}$$

is the vector of regression coefficients corresponding to the covariate vectors

$$\begin{aligned} \boldsymbol{Z} &= [Z_1, \cdots, Z_p, Z_{p+1}, \cdots, Z_{p+S}, Z_{p+S+1}, \cdots, Z_{p+2S}, \cdots, Z_{p+(j-2)S+1}, \\ &\cdots, Z_{p+(j-1)S}]' = [\boldsymbol{X}', \boldsymbol{D}'] = [X_1, X_2, \cdots, X_p, D_{11}, \cdots, D_{1S}, D_{21}, \cdots, \\ &D_{2S}, D_{(j-1)1}, \cdots, D_{(j-1)S}]'. \end{aligned} \tag{5.16}$$

Here, $D_{11}, \cdots, D_{1S}, D_{21}, \cdots, D_{2S}, \cdots, D_{(j-1)1}, \cdots, D_{(j-1)S}$ are the dummy variables for categories $1, 2, \cdots, S$ for $Y_1, \cdots Y_{j-1}$ with 0 as the reference category. Then the conditional probability of s-th category is

$$\begin{aligned} &P_{s.y_1y_2,\ldots,y_{j-1}}(\boldsymbol{z}) = P(Y_3 = s \mid y_1, y_2, \ldots, y_{j-1}; \boldsymbol{x}) \\ &= P(Y_3 \leq s+1 \mid y_1, y_2, \ldots, y_{j-1}; \boldsymbol{x}) - P(Y_3 \leq s \mid y_1, y_2, \ldots, y_{j-1}; \boldsymbol{x}), \\ &\quad s, y_1, \cdots, y_{j-1} = 0, 1, \cdots, S. \end{aligned} \tag{5.17}$$

5.2.7 Partial proportional odds model (PPOM)

If the proportional odds assumption violates for some predictors then alternative models are unconstrained or constrained partial proportional odds (Peterson and Harrell, 1990) or multinomial logistic regression models among others (Agresti, 2013; Hosmer and

Lemeshow, 2013). The unconstrained partial proportional odds model (Peterson and Harrell, 1990) allows non-proportional odds for a subset of q predictors (q<p, p is the total number of predictors in the model) for those proportional odds assumption violates. Then the marginal model using baseline outcome can be shown as:

$$P(Y_1 \leq s \mid \boldsymbol{x}) = \frac{exp(\alpha_j - \boldsymbol{\beta_1'X} - \boldsymbol{\gamma_s'T})}{1 + exp(\alpha_j - \boldsymbol{\beta_1'X} - \boldsymbol{\gamma_s'T})}, \quad s = 1, 2, \cdots, S. \tag{5.18}$$

or equivalently can be expressed in logit form as

$$logit[P(Y_1 \leq s \mid \boldsymbol{x})] = \alpha_s - \boldsymbol{\beta_1'X} - \boldsymbol{\gamma_s'T} \tag{5.19}$$

where α_s are the cut points, $\boldsymbol{T}$ is the subset of covariates vector for which the proportional odds assumption is violated and γ_s is a vector of regression coefficients corresponding to the q covariates in $\boldsymbol{T}$, $\boldsymbol{\beta_1'}$ is the vector of the regression coefficients of covariates those are not in q. Then the marginal probability of s-th category can be obtained using equation (5.5).

5.2.8 Proposed first order regressive PPOM

First order partial proportional odds regressive models for two repeated outcomes Y_1 and Y_2 can be shown as

$$logit[P(Y_2 \leq s \mid \boldsymbol{z})] = \alpha_{2.s} - \boldsymbol{\beta'_{2.y_1}Z} - \boldsymbol{\gamma'_{2.s}T} \tag{5.20}$$

where $\alpha_{2.s}$ are the cut points, $\boldsymbol{T}$, $\gamma_{2.s}$, $\boldsymbol{\beta'_{2.y_1}}$ are equivalent as explained in equation (5.19) and $\boldsymbol{Z}$ is a covariates vector as defined in equation (5.8). The conditional probability of s-th category of Y_2 for given Y_1 and $\boldsymbol{x}$ can be estimated using equation (5.9).

5.2.9 Proposed second order regressive PPOM

Similarly, for outcomes Y_1, Y_2 and Y_3 the second order regressive PPOM can be shown as

$$logit[P(Y_3 \leq s \mid \boldsymbol{z})] = \alpha_{3.s} - \boldsymbol{\beta'_{3.y_2}Z} - \boldsymbol{\gamma'_{3.s}T} \tag{5.21}$$

where $\alpha_{3.s}$ are the cut points, $\boldsymbol{T}$, $\gamma_{3.s}$, $\boldsymbol{\beta'_{3.y_2}}$ are equivalent as explained in equation (5.19) and $\boldsymbol{Z}$ is a covariates vector as defined in equation (5.11). The conditional probability of s-th category can be estimated using equation (5.13).

5.2.10 Proposed higher order regressive PPOM

Higher order regressive PPOM for $Y_1, \cdots, Y_j$ can be shown as

$$logit[P(Y_j \leq s \mid \boldsymbol{z})] = \alpha_{j.s} - \boldsymbol{\beta}'_{j.y_{j-1}}\boldsymbol{Z} - \boldsymbol{\gamma}'_{j.s}\boldsymbol{T} \tag{5.22}$$

where $\alpha_{j.s}$ are the cut points, $\boldsymbol{T}$, $\gamma_{j.s}$, $\boldsymbol{\beta}'_{j.y_{j-1}}$ are equivalent as explained in equation (5.19) and $\boldsymbol{Z}$ is a covariates vector as defined in equation (5.15). The conditional probability of s-th category can be estimated using equation (5.17).

5.2.11 Multinomial logistic regression model (MNOM)

Multinomial logistic regression disregards the ordering of the outcome categories (Agresti, 2013; Hosmer and Lemeshow, 2013). For baseline outcome Y_1 with categories $(s = 0, \cdots, S)$ the marginal multinomial logistic regression model $P(Y_1 = y_1 \mid \boldsymbol{x})$ as a function of covariates $\boldsymbol{x}$ can be shown as

$$P_s(\boldsymbol{x}) = P(Y_1 = s \mid \boldsymbol{x}) = \frac{e^{(\boldsymbol{\beta}'_s \boldsymbol{X})}}{\sum_{s=0}^{S} e^{(\boldsymbol{\beta}'_s \boldsymbol{X})}} = \frac{e^{g_s(\boldsymbol{X})}}{\sum_{s=0}^{S} e^{g_s(\boldsymbol{X})}}, \quad s = 0, 1, \cdots, S, \tag{5.23}$$

$$\text{where } g_s(\boldsymbol{X}) = \beta_{s0} + \beta_{s1}X_1 + ... + \beta_{sp}X_p, \quad s = 1, \cdots, S,$$

and $\boldsymbol{X}' = [1, X_1, ..., X_p]$ is a covariates vector and $\boldsymbol{\beta}'_{\boldsymbol{s}} = [\beta_{s0}, \beta_{s1}, ..., \beta_{sp}]$ are the parameter vectors of the *s-th* component for outcome Y_1 totaling a $[(p+1)S]$ regression coefficients.

5.2.12 Proposed first order regressive multinomial logistic model

For outcomes Y_1 and Y_2 the first order regressive multinomial logistic model $P(Y_2 \mid y_1; \boldsymbol{z})$ can be shown as:

$$P_{s.y_1}(\boldsymbol{z}) = P(Y_2 = s \mid y_1; \boldsymbol{z}) = \frac{e^{g_{s.y_1}(\boldsymbol{Z})}}{\sum_{s=0}^{S} e^{g_{s.y_1}(\boldsymbol{Z})}}, \quad s, y_1 = 0, 1, \cdots, S, \tag{5.24}$$

where

$$\begin{aligned} g_{s.y_1}(\boldsymbol{Z}) =& \beta_{s.y_10} + \beta_{s.y_11}Z_1 + ... + \beta_{s.y_1p}Z_p + \beta_{s.y_1(p+1)}Z_{p+1} + \cdots \\ &+ \beta_{s.y_1(p+S)}Z_{p+S}, \quad s = 1, 2, \cdots, S \text{ and} \end{aligned}$$

$\boldsymbol{Z}' = [1, Z_1, ..., Z_p, Z_{p+1}, \cdots, Z_{p+S}] = [\boldsymbol{X}', \boldsymbol{D}'] = [1, X_1, ..., X_p, D_{11}, \cdots, D_{1S}]$. Here $D_{11}, \cdots, D_{12}$ are the dummy variables for categories $1, \cdots, 2$ of outcome Y_1 with 0 as the reference category and producing a total of $[(p+1)+S]S$ regression coefficients.

5.2.13 Proposed second order regressive multinomial logistic model

The second order regressive multinomial logistic model $P(Y_3 \mid y_1, y_2; \boldsymbol{z})$ can be shown as:

$$P_{s.y_1,y_2}(\boldsymbol{z}) = P(Y_3 = s \mid y_1, y_2; \boldsymbol{z}) = \frac{e^{g_{s.y_2}(\boldsymbol{Z})}}{\sum_{s=0}^{S} e^{g_{s.y_2}(\boldsymbol{Z})}}, \quad s = 0, 1, \cdots, S, \tag{5.25}$$

where

$$\begin{aligned} g_{s.y_2}(\boldsymbol{Z}) &= \beta_{s.y_20} + \beta_{s.y_21} Z_1 + ... + \beta_{s.y_2p} Z_p + \beta_{s.y_2(p+1)} Z_{p+1} + \cdots \\ &+ \beta_{s.y_2(p+S)} Z_{p+S} + \beta_{s.y_2(p+S+1)} Z_{p+S+1} + \cdots + \beta_{s.y_2(p+2S)} Z_{p+2S}, \\ &\quad s = 1, \cdots, S, \text{ and} \\ \boldsymbol{Z'} &= [1, Z_1, ..., Z_p, Z_{p+1}, \cdots, Z_{p+S}, Z_{p+S+1}, \cdots, Z_{p+2S}] \\ &= \left[\boldsymbol{X'}, \boldsymbol{D'}\right] = [1, X_1, ..., X_p, D_{11}, \cdots, D_{1S}, D_{21}, \cdots, D_{2S}]. \end{aligned} \tag{5.26}$$

Here $D_{11}, \cdots, D_{1S}$ are the dummy variables for categories $1, \cdots, S$ of outcome Y_1 and $D_{21}, \cdots, D_{2S}$ are the dummy variables for categories $1, \cdots, S$ of outcome Y_2 with 0 as the reference category and producing a total of $[(p+1) + 2S]S$ regression coefficients.

5.2.14 Proposed higher order regressive multinomial logistic model

For outcomes $Y_1, \cdots, Y_j$ higher order regressive multinomial logistic model can be shown as

$$P_{s.y_1,...,y_{j-1}}(\boldsymbol{z}) = P(Y_j = s \mid y_1, \cdots, y_{j-1}; \boldsymbol{z}) = \frac{e^{g_{s.y_{j-1}}(\boldsymbol{Z})}}{\sum_{s=0}^{S} e^{g_{s.y_{j-1}}(\boldsymbol{Z})}},$$

$$s = 0, 1, 2, ..., S, \tag{5.27}$$

where

$$\begin{aligned} g_{s.y_{j-1}}(\boldsymbol{Z}) &= \beta_{s.y_{j-1}0} + \beta_{s.y_{j-1}1} Z_1 + ... + \beta_{s.y_{j-1}p} Z_p + \beta_{s.y_{j-1}(p+1)} Z_{p+1} \\ &+ ... + \beta_{s.y_{j-1}(p+S)} Z_{p+S} + \beta_{s.y_{j-1}(p+S+1)} Z_{p+S+1} + ... + \\ &\beta_{s.y_{j-1}(p+2S)} Z_{p+2S} + ... + \beta_{s.y_{j-1}[p+(j-1)S+1]} Z_{[p+(j-1)S+1]} + ... + \\ &\beta_{s.y_{j-1}[p+(j-1)S+S]} Z_{[p+(j-1)S+S]}, \quad s = 1, 2, ..., S, j > 1 \end{aligned} \tag{5.28}$$

and

$$\boldsymbol{Z'} = \Big[1, Z_1, ..., Z_p, Z_{p+1}, ..., Z_{p+S}, Z_{p+S+1}, ..., Z_{p+2S}, ..., Z_{[p+(j-1)S+1]}, ..., Z_{[p+(j-1)S+S]}\Big] = [\boldsymbol{X'}, \boldsymbol{D'}] = \Big[1, X_1, ..., X_p, D_{11}, ..., D_{1S}, D_{21}, ..., D_{2S}, ..., D_{(j-1)1}, ..., D_{(j-1)S}\Big].$$

Here, $D_{11}, ..., D_{1S}, D_{21}, ..., D_{2S}, ..., D_{(j-1)1}, ..., D_{(j-1)S}$ are the dummy variables for categories $1, 2, ..., S$ of outcomes $Y_1, Y_2, ..., Y_{j-1}$ with 0 as the reference category, respectively. There are $[(p+1)+(j-1)S]$ regression coefficients for *s-th* component of the model and with a total of $[(p+1)+(j-1)S]S$ regression coefficients.

It may be noted that first and all higher order regressive models for POM, PPOM and MNOM are equivalent to the corresponding marginal models shown in equations (5.3, 5.17 and 5.22). Regressive models for marginal or higher order can be estimated using appropriate data structure and usual SAS, STATA or R-package or other software capable of fitting all these model. It should be noted that the regressive model for binary outcomes proposed earlier (Islam and Chowdhury, 2010; Bonney, 1986, 1987) are special case for s=0,1.

5.2.15 Predictive models and joint probabilities

We can predict the risks of a sequence events from repeated measures for a subject with specified covariates vector $\boldsymbol{X}^* = \boldsymbol{x}^*$ for a particular trajectory as shown in the Figure 5.1. The predicted joint probabilities of $\hat{P}(Y_1 = y_1, Y_2 = y_2, ..., Y_j = y_j \mid \boldsymbol{x}^*)$ can be obtained using predictive models as:

$$\begin{aligned}
&\hat{P}(Y_1 = y_1, Y_2 = y_2, \cdots, Y_J = y_J \mid \boldsymbol{x}^*) = \hat{P}(Y_1 = y_1 \mid \boldsymbol{x}^*) \times \hat{P}(Y_2 = y_2 \mid y_1; \boldsymbol{x}^*) \\
&\times, \cdots, \times \hat{P}(Y_J = s \mid y_1, \cdots, y_{j-1}; \boldsymbol{x}^*) = \hat{P}_{y_1}(\boldsymbol{x}^*) \times \hat{P}_{y_2.y_1}(\boldsymbol{x}^*) \times \cdots \times \\
&\hat{P}_{s.y_1, \cdots, y_{j-1}}(\boldsymbol{x}^*). \qquad (5.29)
\end{aligned}$$

For simplicity, let the repeated outcomes have categories $s = 0, 1, 2$. Then using equation (5.29) the predicted joint probabilities $P(Y_1 = y_1, Y_2 = y_2 \mid \boldsymbol{x}^*)$ is

$$\begin{aligned}
\hat{P}_{y_1, y_2}(\boldsymbol{x}^*) &= \hat{P}(Y_1 = y_1, Y_2 = y_2 \mid \boldsymbol{x}^*) = \hat{P}(Y_1 = y_1 \mid \boldsymbol{x}^*) \times \hat{P}(Y_2 = s, \mid y_1; \boldsymbol{x}^*) \\
&= \hat{P}_{y_1}(\boldsymbol{x}^*) \times \hat{P}_{s.y_1}(\boldsymbol{x}^*), \quad y_1, y_2 = 0, 1, 2. \qquad (5.30)
\end{aligned}$$

For POM model the predicted marginal probabilities $\hat{P}_0(\boldsymbol{x}^*); \hat{P}_1(\boldsymbol{x}^*); \hat{P}_2(\boldsymbol{x}^*)$ can be estimated from the fitted model shown in equation (5.4) and in equation (5.5). The first order conditional probabilities $\hat{P}_{s.y_1}(\boldsymbol{x}^*)$ can be estimated from the fitted first order regressive POM shown in equation (5.6) and equation (5.8) using covariates vector $\boldsymbol{Z} = [\boldsymbol{x}^*, D_{11}, D_{12}]'$ where $D_{11}, D_{12} = 0, 1$. For example, $\hat{P}_{1.0}(\boldsymbol{x}^*)$ and $\hat{P}_{2.0}(\boldsymbol{x}^*)$ are estimated

using $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0]'$; $\hat{P}_{1.1}(\boldsymbol{x}^*)$ and $\hat{P}_{2.1}(\boldsymbol{x}^*)$ are estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 1, 0]'$; $\hat{P}_{1.2}(\boldsymbol{x}^*)$ and $\hat{P}_{2.2}(\boldsymbol{x}^*)$ are estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 1]'$ and so on. Then the joint probabilities for two outcomes $\hat{P}_{00} = \hat{P}_0 \times \hat{P}_{0.0}$; $\hat{P}_{01} = \hat{P}_0 \times \hat{P}_{1.0}$ and $\hat{P}_{02} = \hat{P}_0 \times \hat{P}_{2.0}$ and so on.

To estimate the joint probabilities $\hat{P}_{y_1 y_2 y_3}(\boldsymbol{x}^*)$, the required second order conditional probabilities can be estimated using fitted POM model in (5.9) and equation in (5.13). For example, $\hat{P}_{1.00}(\boldsymbol{x}^*)$ and $\hat{P}_{2.00}(\boldsymbol{x}^*)$ is estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 0, 0, 0, 0]'$; $\hat{P}_{1.10}(\boldsymbol{x}^*)$ and $\hat{P}_{2.10}(\boldsymbol{x}^*)$ is estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 1, 0, 0, 0]'$ and $\hat{P}_{1.11}(\boldsymbol{x}^*)$ and $\hat{P}_{2.11}(\boldsymbol{x}^*)$ is estimated using $\boldsymbol{Z} = [\boldsymbol{x}^*, 1, 0, 1, 0]'$ and so on. Then the joint probabilities for three outcomes $\hat{P}_{000} = \hat{P}_0 \times \hat{P}_{0.0} \times \hat{P}_{0.00}$; $\hat{P}_{001}(\boldsymbol{x}^*) = \hat{P}_0(\boldsymbol{x}^*) \times \hat{P}_{0.0}(\boldsymbol{x}^*) \times \hat{P}_{1.00}(\boldsymbol{x}^*)$ and $\hat{P}_{002} = \hat{P}_0 \times \hat{P}_{0.0} \times \hat{P}_{2.00}$ and so on.

Similarly, we can estimate the joint probabilities of a sequence of events by estimating the marginal and conditional probabilities from partial proportional odds and multinomial models, respectively.

5.3 Tests

5.3.1 Significance of the joint model

The significance of the joint model can be tested using likelihood ratio test between joint constant only model (Reduced) and joint full model (Full) as follows:

$$= -2\left[\ln L_{\text{Reduced}}(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - \ln L_{\text{Full}}(\hat{\boldsymbol{\beta}})\right] \tag{5.31}$$

which is distributed asymptotically as $\chi^2_{(d)}$.

The degrees of freedom (d) for three models are as follows:

$$\begin{aligned}
d_{POM} =& [\{(p+S)\} + \{(p+S+S)\} + \{(p+2S+S)\} + ... + \\
& \{p+(j-1)S\} + S] - jS. \\
d_{PPOM} =& [\{(p'+S)\} + \{(p'+S+S)\} + \{(p'+2S+S)\} + ... + \\
& \{p'+(j-1)S\} + S] - jS. \\
d_{MNOM} =& [\{(p+1)S\} + \{(p+1+S)S\} + \{(p+1+2S)S\} + ... + \\
& \{p+1+(j-1)S\}S] - jS.
\end{aligned}$$

Here $\hat{\boldsymbol{\beta}}'_{\mathbf{0}}$ includes all the regression parameters from the constant only joint model and $\hat{\boldsymbol{\beta}}'$ includes all the parameters from the full joint model. The degrees of freedom for different models are shown in the following table.

Number of parameters for different models.

Models	*Constant only*	*s-th component*	*Full*
Proportional odds models			
Marginal	S		$[p+S]$
First order	S		$[p+S+S]$
Second order	S		$[p+2S+S]$
...	...	...	...
$j-1$ *th* order	S		$[(p+(j-1)S)+S]$
Partial proportional odds models			
Marginal	S		$[p'+S]$
First order	S		$[p'+S+S]$
Second order	S		$[p'+2S+S]$
...	...	...	...
$j-1$ *th* order	S		$[(p'+(j-1)S)+S]$
Multinomial logistic regression models			
Marginal	S	$[p+1]$	$[p+1]S$
First order	S	$[p+1+S]$	$[p+1+S]S$
Second order	S	$[p+1+2S]$	$[p+1+2S]S$
...	...	...	...
$j-1$ *th* order	S	$[(p+1+(j-1)S)]$	$[(p+1+(j-1)S)]S$

Note: p' will depends on the scale and number of covariates in $\boldsymbol{T}$

5.3.2 Test for proportional odds assumption

One of the important assumptions of POM is proportional odds assumption which should be tested. In this model, the regression coefficients for models from different cut points are same, only threshold parameters varies. Likelihood ratio test (Peterson and Harrell, 1990) and Brant test (Brant, 1990) can be used to test the proportional odds assumption. However, these tests have been criticized for having a tendency to reject the null hypothesis (Harrell, 2001).

5.3.3 Goodness of fit

It is important to check the goodness of fit for all models to have a more precise estimate and refined predictions. As all marginal and regressive models boil down to the univariate case, we used available tests for goodness-of-fit. Lipsitz *et al.* (1996) proposed a goodness-of fit test for ordinal response regression model. Fagerland *et al.* (2008) proposed a goodness-of-fit test for multinomial logistic regression. Fagerland and Hosmer (2013) proposed another goodness-of-fit test for proportional odds regression model. Also, the goodness-of-fit test for ordinal regression is applied to test the fit of the partial proportional odds model.

5.3.4 Proposed tests for goodness-of-fit of the joint model

All the above tests for goodness-of-fit are for a single outcome. Islam and Chowdhury (2010) proposed a test for goodness of fit of the joint model for repeated binary outcomes. They estimated the transition probabilities from the outcome status from previous time point to the current time point and the current time point is considered as the end point. The goodness of fit is tested at the endpoint of the joint model. Here we proposed a test for repeated ordinal outcomes. The null hypothesis is, H_0 : the fitted joint model is correct. Both the Pearson and Likelihood-ratio χ^2 statistics using observed and expected frequencies from the joint model can be shown as

$$X^2 = \sum_{s=0}^{S}\sum_{s'=0}^{S}\frac{(n_{ss'}-\hat{e}_{ss'})^2}{\hat{e}_{ss'}}, \quad s,s' = 0,1,\cdots,S \text{ and} \tag{5.32}$$

$$G^2 = 2\sum_{s=0}^{S}\sum_{s'=0}^{S} n_{ss'}\log\left(\frac{n_{ss'}}{\hat{e}_{ss'}}\right), \quad s,s' = 0,1,\cdots,S, \tag{5.33}$$

where $n_{ss'}$ and $\hat{e}_{ss'}$ are observed and expected transition counts at the end points, $\sum n_{ss'} = n$, $\hat{e}_{ss'} = n\hat{p}_{ss'}$. The joint probability $\hat{p}_{ss'} = \hat{P}_{y_1}(\boldsymbol{x})\hat{P}_{y_2.y_1}(\boldsymbol{x})$, where $\hat{P}_{y_1}(\boldsymbol{x})$ and $\hat{P}_{y_2.y_1}(\boldsymbol{x})$ are estimated from fitted marginal and first order regressive models for POM or PPOM or MNOM. We imposed a single restriction on total sample size n fixed summing total joint probability to 1. Both the statistics X^2 and G^2 are asymptotically distributed as χ^2 with $s^j - 1$ degrees of freedom where s^j is the total number of end points $(s = 0,1,\cdots,S;\ j = 2,\cdots,J)$. Some instances there might not be any observed counts for a trajectory. Then we can merge those end points with another trajectory. This test readily generalizes for any number of repeated outcomes.

In the presence of covariates pattern one can use the predicted empirical means proposed by Long (1997). Alternatively, we can used the predicted probabilities shown by Islam and Chowdhury (2017) using the connection between the Poisson and multinomial. They showed that both the marginal probabilities of Y_1 and conditional probabilities of Y_2 for any given value of Y_1 follow a multinomial distribution.

5.3.5 Proposed tests for order

We extended a test for binary outcomes proposed by Islam *et al.* (2009) to test the order of the Markov model for ordinal outcomes. For *k-th* $(k = j-1)$ order regressive model, dummy variables for each category except for reference level from previous j-1 outcomes are incorporated as the covariates to test the order of the model. For higher

order regressive POM the null hypothesis is

$$H_0 : \beta_{j.y_{j-1}(p+1)} = \cdots = \beta_{p+S} = \beta_{j.y_{j-1}(p+S+1)} = \cdots = \beta_{p+2S} = \cdots = \beta_{j.y_{j-1}(p+2S)} = \cdots = \beta_{j.y_{j-1}[p+(j-2)S+1]} = \cdots = \beta'_{j.y_{j-1}[p+(j-1)S]} = 0 \tag{5.34}$$

which can be tested using following statistic:

$$-2\left[\ln L(\hat{\boldsymbol{\beta}}_{\mathbf{1}}) - \ln L(\hat{\boldsymbol{\beta}})\right], \tag{5.35}$$

is distributed asymptotically as χ^2 with $[p+(j-1)S+S]-\{(j-1)S\}]$ degrees of freedom, $[p+(j-1)S+S]$ is the total number of parameters of $(j-1)th$ order regressive model and $(j-1)S$ are the number of previous outcomes $y_1,...,y_{j-1}$ multiplied by the number of dummy variables (S) for each outcome. Similarly, we can test the order for PPOM and MNOM. Then the test can be performed as follows:

(i) The likelihood ratio test can be used to test the significance of the overall model at the first stage.

(ii) The Wald test can be used to test the significance of the parameter(s) corresponding to the previous outcomes as shown below:

5.3.6 Overfitting, underfitting and predictive accuracy

Good fit models with the better discriminative ability and predictive power are expected to provide higher prediction accuracy. Predictive accuracy of models is estimated from confusion matrix and over(under)fitting is evaluated using training and test data sets approach (James *et al.*, 2013, p. 21, 29).

5.4 Application

The panel data from the Health and Retirement Study (HRS), sponsored by the National Institute of Aging (grant number NIA U01AG09740), conducted by the University of Michigan (HRS, 2014) is used for the application. In wave one (first follow-up), a total of 12652 subjects were interviewed in the HRS cohort. A total of six waves (follow-ups) of the RAND version of the data from wave six (2002) to wave 11 (2012) is used for this application. At the wave six minimum age of subjects were 60 years. The outcome variables considered are Activity of Daily Living Index (ADL) based on Wallace and Herzog (1995) from wave six to wave eleven $(Y_1,...,Y_6)$. This index is the sum of three tasks (yes/no) ranging from 0 to 3: whether respondents faced difficulties in bathing, dressing and eating. Due to small frequencies 3 was coded as 2. The explanatory variables

considered are: age (in years), marital status (married/partnered=1, single/separated=0), whether drink (yes=1, no=0), gender (male=1, female=0), number of conditions ever had (N.cond) ranges from 0 to 8, White (yes=1, no=0), Black (yes=1, no=0) with others as reference category, education (in years) and veteran status (1=yes, 0= no). The variable drink indicates whether the respondent drinks alcoholic beverages. After removal of cases with missing values for outcome variable at wave six, the number of subjects is 7130. Table 5.1 displays the frequency distribution of the outcomes for different waves.

TABLE 5.1: Distribution of Activity of Daily Living Index, Waves 6-11.

	Outcomes											
Y_j	Y_1		Y_2		Y_3		Y_4		Y_5		Y_6	
levels	N	%	N	%	N	%	N	%	N	%	N	%
0	6424	90.1	5934	89.7	5437	87.8	5125	87.2	4567	84.7	4252	84.7
1	439	6.2	416	6.3	482	7.8	436	7.4	455	8.4	449	8.9
2	267	3.7	265	4.0	273	4.4	317	5.4	370	6.9	317	6.4
Total	7130	100	6615	100	6192	100	5878	100	5392	100	5018	100

Parameter estimates, significance level, standard errors and Brant p-value to test proportional odds assumption from POM for marginal and regressive models are shown in Table 5.2. Various predictors are found to be significantly associated with outcome variables for different models. All dummy indicators for previous outcomes are significantly and positively associated with the current outcomes except for fifth-order model. For fifth order model D_{12}, D_{21} and D_{32} were not statistically significant. The overall test for proportional odds assumption was violated for all models. Specifically, for marginal model proportional odds assumption were violated for marital status, drink habit and black subjects. For other models, different variables violated this assumption. Parameter estimates, standard errors and significance level from PPOM for marginal and regressive models are shown in Table 5.3 and Table 5.4. The PPOM models were fitted to tackle the variables those violated the proportional odds assumption in POM. Finally, we considered ordinal outcomes as nominal and used multinomial logistic regression. Parameter estimates, significance level, standard errors from MNOM for marginal and regressive models are shown in Table 5.5 and Table 5.6. Various predictors are found to be significantly associated with outcome variables for different models. For all three fifth order models (POM, PPOM, MNOM) dummy variables from the previous outcome (Y_5) were statistically significant justifies fifth order model. AIC for marginal and all higher order models were lowest for PPOM followed by MNOM and POM.

Model statistics are shown in Table 5.7. Log-likelihood value for the constant only model and full model for marginal and all higher order are shown for POM, PPOM, MNOM. Likelihood ratio test between joint constant only and full models are statistically significant ($p < 0.001$) for POM, PPOM, MNOM. The prediction accuracy based on confusion matrix for full data and test and training data varies between 0.87 to 0.90 which is reasonably high. Prediction accuracy for POM, PPOM and MNOM are overly similar. Also,

TABLE 5.2: Parameter estimates of proportional odds models (POM) for different order.

	Models								
	$P(Y_1 \mid \boldsymbol{X})$			$P(Y_2 \mid Y_1;\boldsymbol{X})$			$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$		
Variables	$\hat{\beta}_1$	$\widehat{S.E.}$	Brant	$\hat{\beta}_{2.1}$	$\widehat{S.E.}$	Brant	$\hat{\beta}_{3.12}$	$\widehat{S.E.}$	Brant
$(\boldsymbol{X})$			p.v.			p.v.			p.v.
Age	0.008	0.013	0.22	0.007	0.015	0.32	0.031*	0.015	0.12
Mstat	−0.380**	0.091	0.02	−0.336**	0.102	0.20	−0.203*	0.102	0.04
N.Cond	0.568**	0.029	0.19	0.398**	0.032	0.51	0.342**	0.032	0.19
Drink	−0.578**	0.096	0.00	−0.354**	0.106	0.50	−0.263**	0.101	0.01
Gender	0.125	0.109	0.78	0.278*	0.120	0.25	0.163	0.122	0.76
White	−0.401*	0.194	0.34	−0.155	0.224	0.17	−0.138	0.241	0.11
Black	−0.085	0.205	0.04	0.028	0.238	0.35	0.197	0.254	0.55
Educ.	−0.073**	0.013	0.33	−0.016	0.014	0.89	−0.070**	0.015	0.65
Veteran	−0.207	0.134	0.22	−0.274	0.147	0.82	−0.115	0.146	0.56
D_{11}				2.132**	0.118	0.01	1.093**	0.135	0.01
D_{12}				3.725**	0.164	0.04	1.715**	0.195	0.87
D_{21}							1.984**	0.127	0.17
D_{22}							3.324**	0.186	0.86
Intercepts									
0|1	2.426**	0.912		3.591**	1.054		4.591**	1.093	
1|2	3.580**	0.913		5.017**	1.056		6.326**	1.097	
Brant Overall			0.00			0.03			0.00

	Models								
	$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$			$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$			$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$		
Variables	$\hat{\beta}_{4.123}$	$\widehat{S.E.}$	Brant	$\hat{\beta}_{5.1234}$	$\widehat{S.E.}$	Brant	$\hat{\beta}_{6.12345}$	$\widehat{S.E.}$	Brant
$(\boldsymbol{X})$			p.v.			p.v.			p.v.
Age	0.054**	0.016	0.71	0.065**	0.015	0.48	0.067**	0.016	0.41
Mstat	−0.254*	0.106	0.00	−0.422**	0.101	0.31	−0.237*	0.109	0.98
N.Cond	0.310**	0.034	0.05	0.288**	0.033	0.77	0.332**	0.035	0.84
Drink	−0.236*	0.108	0.00	−0.220*	0.101	0.06	−0.190	0.109	0.47
Gender	0.138	0.127	0.77	0.261*	0.124	0.27	0.303*	0.131	0.57
White	0.300	0.271	0.10	−0.188	0.239	0.29	−0.021	0.262	0.14
Black	0.429	0.286	0.14	−0.089	0.256	0.07	0.156	0.283	0.71
Educ.	−0.059**	0.015	0.76	−0.058**	0.015	0.15	−0.047**	0.016	0.80
Veteran	−0.079	0.151	0.08	0.105	0.140	0.02	−0.163	0.151	0.04
D_{11}	0.747**	0.150	0.00	0.674**	0.157	0.03	0.568**	0.176	0.70
D_{12}	1.076**	0.229	0.82	0.565*	0.269	0.21	0.373	0.302	0.60
D_{21}	0.701**	0.153	0.12	0.475**	0.163	0.04	0.288	0.185	0.41
D_{22}	1.691**	0.227	0.43	0.620*	0.291	0.31	0.740*	0.319	0.64
D_{31}	1.745**	0.127	0.01	1.196**	0.139	0.00	0.382*	0.163	0.40
D_{32}	3.147**	0.203	0.84	1.094**	0.253	0.31	0.230	0.285	0.30
D_{41}				1.526**	0.132	0.07	0.861**	0.152	0.02
D_{42}				3.051**	0.200	0.92	1.764**	0.233	0.20
D_{51}							1.839**	0.132	0.08
D_{52}							2.743**	0.180	0.01
Intercepts									
0|1	6.825**	1.170		7.033**	1.161		8.027**	1.286	
1|2	8.397**	1.174		8.455**	1.165		9.711**	1.291	0.00
Brant overall			0.00			0.00			

* Significant at 5% level; ** Significant at 1% level.

TABLE 5.3: Parameter estimates of partial proportional odds models (PPOM) for different order.

	Models								
	$P(Y_1 \mid \boldsymbol{X})$			$P(Y_2 \mid Y_1;\boldsymbol{X})$			$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$		
Variables $(\boldsymbol{X})$	$\hat{\beta}_1$	$\widehat{S.E.}$	p.value	$\hat{\beta}_{2.1}$	$\widehat{S.E.}$	p.value	$\hat{\beta}_{3.12}$	$\widehat{S.E.}$	p.value
Threshold coefficients									
0|1.Intercept	2.432	0.911	0.008	3.530	1.056	0.001	4.678	1.099	0.000
1|2.Intercept	3.438	0.914	0.000	4.920	1.060	0.000	5.919	1.103	0.000
0|1.Mstat	0.355	0.092	0.000				0.146	0.106	0.167
1|2.Mstat	0.619	0.133	0.000				0.524	0.161	0.001
0|1.Drink	0.560	0.097	0.000				0.194	0.103	0.061
1|2.Drink	1.002	0.173	0.000				0.710	0.179	0.000
0|1.Black	0.125	0.205	0.544						
1|2.Black	-0.203	0.226	0.369						
0|1.D_{11}				-2.216	0.122	0.000	-1.271	0.143	0.000
1|2.D_{11}				-1.799	0.189	0.000	-0.603	0.205	0.003
0|1.D_{12}				-3.392	0.181	0.000	-1.611	0.229	0.000
1|2.D_{12}				-3.823	0.184	0.000	-1.634	0.229	0.000
Coefficients									
Age	0.008	0.013	0.553	0.006	0.015	0.669	0.031	0.015	0.039
Mstat				-0.339	0.102	0.001			
N.Cond	0.563	0.029	0.000	0.401	0.032	0.000	0.341	0.032	0.000
Drink				-0.351	0.106	0.001			
Gender	0.126	0.109	0.246	0.287	0.120	0.017	0.162	0.122	0.187
White	-0.399	0.194	0.039	-0.163	0.225	0.467	-0.162	0.243	0.503
Black				0.016	0.239	0.946	0.177	0.255	0.489
Educ.	-0.073	0.013	0.000	-0.018	0.014	0.225	-0.071	0.015	0.000
Vateran	-0.208	0.134	0.121	-0.280	0.147	0.057	-0.114	0.146	0.438
D_{11}									
D_{12}									
D_{21}							1.986	0.127	0.000
D_{22}							3.339	0.188	0.000

TABLE 5.4: Parameter estimates of partial proportional odds models.

	$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$			$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$			$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$		
Variables $(\boldsymbol{X})$	$\hat{\beta}_{4.123}$	$\widehat{S.E.}$	p.value	$\hat{\beta}_{5.1234}$	$\widehat{S.E.}$	p.value	$\hat{\beta}_{6.12345}$	$\widehat{S.E.}$	p.value
Threshold coefficients									
0|1.Intercept	6.993	1.179	0.000	7.186	1.170	0.000	8.167	1.293	0.000
1|2.Intercept	7.599	1.190	0.000	8.334	1.173	0.000	9.561	1.299	0.000
0|1.Mstat	0.182	0.109	0.097						
1|2.Mstat	0.633	0.159	0.000						
0|1.N.Cond	-0.328	0.035	0.000						
1|2.N.Cond	-0.226	0.050	0.000						
0|1.Drink	0.184	0.110	0.095						
1|2.Drink	0.620	0.185	0.001						
0|1.Vateran				-0.156	0.143	0.275	0.084	0.154	0.583
1|2.Vateran				0.068	0.193	0.725	0.647	0.236	0.006
$0\|1.D_{11}$	-0.974	0.161	0.000						
$1\|2.D_{11}$	-0.219	0.220	0.318						
$0\|1.D_{12}$	-1.052	0.269	0.000						
$1\|2.D_{12}$	-0.928	0.268	0.001						
$0\|1.D_{21}$				-0.743	0.178	0.000			
$1\|2.D_{21}$				-0.089	0.221	0.689			
$0\|1.D_{22}$				-0.498	0.348	0.153			
$1\|2.D_{22}$				-0.606	0.319	0.058			
$0\|1.D_{31}$	-1.850	0.131	0.000	-1.460	0.151	0.000			
$1\|2.D_{31}$	-1.356	0.197	0.000	-0.727	0.192	0.000			
$0\|1.D_{32}$	-2.906	0.251	0.000	-0.790	0.300	0.008			
$1\|2.D_{32}$	-3.066	0.226	0.000	-1.050	0.279	0.000			
$0\|1.D_{41}$							-1.028	0.164	0.000
$1\|2.D_{41}$							-0.450	0.211	0.033
$0\|1.D_{42}$							-2.024	0.306	0.000
$1\|2.D_{42}$							-1.468	0.262	0.000
$0\|1.D_{51}$							-1.948	0.138	0.000
$1\|2.D_{51}$							-1.549	0.207	0.000
$0\|1.D_{52}$							-2.366	0.197	0.000
$1\|2.D_{52}$							-2.946	0.214	0.000
Coefficients									
Age	0.055	0.016	0.001	0.067	0.015	0.000	0.069	0.016	0.000
Mstat				-0.431	0.102	0.000	-0.243	0.110	0.027
N.Cond				0.285	0.033	0.000	0.329	0.036	0.000
Drink				-0.217	0.102	0.033	-0.190	0.110	0.085
Gender	0.136	0.128	0.289	0.259	0.125	0.038	0.293	0.132	0.027
White	0.271	0.273	0.321	-0.223	0.238	0.349	-0.061	0.263	0.816
Black	0.408	0.288	0.157	-0.124	0.256	0.628	0.114	0.284	0.687
Educ.	-0.059	0.015	0.000	-0.058	0.015	0.000	-0.046	0.016	0.004
Vateran	-0.081	0.152	0.594						
D_{11}				0.709	0.157	0.000	0.577	0.179	0.001
D_{12}				0.578	0.272	0.033	0.375	0.304	0.217
D_{21}	0.711	0.155	0.000				0.305	0.187	0.103
D_{22}	1.753	0.231	0.000				0.757	0.322	0.019
D_{31}							0.414	0.164	0.012
D_{32}							0.204	0.287	0.477
D_{41}				1.550	0.133	0.000			
D_{42}				3.075	0.198	0.000			

TABLE 5.5: Parameter estimates of multinomial logistic regression models for different order.

	Models											
	$P(Y_1 \mid \boldsymbol{X})$				$P(Y_2 \mid Y_1;\boldsymbol{X})$				$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$			
	Category 1		Category 2		Category 1		Category 2		Category 1		Category 2	
Variables $(\boldsymbol{X})$	$\hat{\beta}_1$	S.E.	$\hat{\beta}_1$	S.E.	$\hat{\beta}_{2.1}$	S.E.	$\hat{\beta}_{2.1}$	S.E.	$\hat{\beta}_{3.12}$	S.E.	$\hat{\beta}_{3.12}$	S.E.
Intercept	−1.740	1.100	−5.065**	1.453	−2.677*	1.230	−5.994**	1.730	−4.044**	1.262	−8.210**	1.910
Age	−0.005	0.016	0.027	0.021	−0.007	0.018	0.019	0.024	0.011	0.017	0.066*	0.026
Mstat	−0.183	0.112	−0.671**	0.142	−0.232	0.120	−0.569**	0.164	−0.064	0.119	−0.526**	0.176
N.Cond	0.485**	0.035	0.676**	0.044	0.374**	0.038	0.483**	0.051	0.349**	0.037	0.420**	0.054
Drink	−0.324**	0.112	−1.093**	0.177	−0.305*	0.123	−0.525**	0.185	−0.069	0.114	−0.679**	0.190
Gender	0.098	0.129	0.181	0.178	0.150	0.142	0.485*	0.197	0.179	0.139	0.192	0.218
White	−0.493*	0.223	−0.231	0.330	−0.433	0.250	0.016	0.384	0.250	0.312	−0.402	0.375
Black	−0.470*	0.243	0.369	0.339	−0.195	0.269	0.156	0.403	0.397	0.328	0.215	0.392
Educ.	−0.080**	0.015	−0.073**	0.020	−0.014	0.017	−0.017	0.023	−0.069**	0.017	−0.090**	0.024
Veteran	−0.321*	0.162	−0.044	0.218	−0.226	0.173	−0.385	0.247	−0.143	0.164	−0.042	0.273
D_{11}					2.218**	0.135	2.206**	0.197	1.336**	0.154	1.240**	0.221
D_{12}					2.501**	0.230	4.391**	0.213	1.262**	0.270	2.195**	0.278
D_{21}									1.950**	0.146	2.261**	0.209
D_{22}									2.693**	0.264	4.274**	0.271

* Significant at 5% level; ** Significant at 1% level.

TABLE 5.6: Parameter estimates of multinomial logistic regression models for different order.

	Models											
	$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$				$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$				$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$			
	Category 1		Category 2		Category 1		Category 2		Category 1		Category 2	
Variables $(\boldsymbol{X})$	$\hat{\beta}_{4.123}$	S.E.	$\hat{\beta}_{4.123}$	S.E.	$\hat{\beta}_{5.1234}$	S.E.	$\hat{\beta}_{5.1234}$	S.E.	$\hat{\beta}_{6.12345}$	S.E.	$\hat{\beta}_{6.12345}$	S.E.
Intercept	−7.306**	1.395	−9.519**	1.904	−6.129**	1.411	−11.224**	1.824	−7.707**	1.508	−12.998**	2.143
Age	0.047*	0.019	0.080**	0.026	0.054**	0.018	0.105**	0.024	0.054**	0.019	0.114**	0.027
Mstat	0.003	0.127	−0.688**	0.172	−0.462**	0.123	−0.444**	0.156	−0.226	0.128	−0.325	0.181
N.Cond	0.345**	0.040	0.355**	0.054	0.248**	0.040	0.354**	0.050	0.292**	0.041	0.434**	0.058
Drink	0.021	0.126	−0.594**	0.192	−0.059	0.122	−0.457**	0.163	−0.284*	0.127	−0.216	0.188
Gender	0.121	0.149	0.129	0.217	0.132	0.153	0.457*	0.187	0.325*	0.155	0.329	0.214
White	0.525	0.350	0.007	0.387	−0.479	0.276	−0.264	0.367	0.370	0.346	−0.168	0.402
Black	0.665	0.367	0.235	0.408	−0.589	0.303	0.011	0.388	0.281	0.371	0.207	0.428
Educ.	−0.053**	0.018	−0.068**	0.024	−0.073**	0.018	−0.065**	0.022	−0.038*	0.019	−0.065**	0.025
Vateran	−0.236	0.179	0.200	0.259	0.331	0.171	−0.145	0.223	0.040	0.172	−0.630*	0.270
D_{11}	1.129**	0.174	0.769**	0.237	0.946**	0.191	0.842**	0.228	0.570**	0.216	0.815**	0.266
D_{12}	0.787*	0.316	1.291**	0.324	0.764*	0.364	0.671	0.380	0.372	0.418	0.470	0.451
D_{21}	0.819**	0.181	0.741**	0.240	0.787	0.198	0.616*	0.238	0.430	0.225	0.396	0.283
D_{22}	1.500**	0.334	2.456**	0.331	0.085	0.429	0.901*	0.393	0.839	0.478	1.284**	0.497
D_{31}	1.881**	0.146	1.772**	0.206	1.499**	0.168	1.400**	0.206	0.404*	0.198	0.391	0.247
D_{32}	2.277**	0.298	3.788**	0.281	0.276	0.374	1.045**	0.341	0.475	0.409	0.350	0.437
D_{41}					1.609**	0.161	1.802**	0.193	1.091**	0.181	0.927**	0.236
D_{42}					2.081**	0.304	3.736**	0.271	1.621**	0.352	2.311**	0.356
D_{51}									1.905**	0.153	2.103**	0.220
D_{52}									1.627**	0.246	3.356**	0.246

* Significant at 5% level; ** Significant at 1% level.

accuracy from full, training and test data are very close for POM, PPOM, MNOM., which shows the absence of over(under)fitting for all models.

Table 5.8 displays the goodness-of-fit test results of joint models based on proposed Pearson χ^2 and Likelihood ratio χ^2. None of the joint models using POM showed good fit. However, for PPOM fifth order joint model and for MNOM fourth order joint models showed a good fit based on Pearson χ^2. Many variables were not included in the model to keep the application simple which may be the reason for lack of good fit. More than thirty 33 percent of the expected frequencies are less than 5 or zeros for Joint models of sixth order. Goodness-of-fit results are in line with the high prediction accuracy as shown in the Table 5.7. As our objective is to develop a modeling framework for risk prediction of a sequence of events we did not do further modeling exercise to obtain good fitted models.

TABLE 5.7: Proportional odds, partial proportional odds and multinomial models statistics.

Model	Constant only model	Full Model		Likelihood ratio test			Accuracy		
	Log L.	AIC	Log L.	χ^2	d.f.	p.value	All	Train	Test
POM									
$P(Y_1 \mid \boldsymbol{X})$	-2770.6	4859.8	-2418.9	703.5	9	0.000	0.902	0.902	0.902
$P(Y_2 \mid Y_1;\boldsymbol{X})$	-2648.1	3969.9	-1972.0	1352.3	11	0.000	0.906	0.907	0.906
$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$	-2789.7	3866.4	-1918.2	1743.1	13	0.000	0.895	0.893	0.898
$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$	-2762.4	3636.7	-1801.4	1922.1	15	0.000	0.895	0.896	0.897
$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$	-2874.6	3887.6	-1924.8	1899.6	17	0.000	0.872	0.874	0.871
$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$	-2663.6	3346.1	-1652.1	2023.1	19	0.000	0.875	0.877	0.867
Joint model	-16509.1		-11687.2	9643.7	84	0.000			
PPOM									
$P(Y_1 \mid \boldsymbol{X})$	-2770.6	4836.9	-2404.5	732.3	12	0.000	0.902	0.902	0.901
$P(Y_2 \mid Y_1;\boldsymbol{X})$	-2648.1	3956.8	-1963.4	1369.4	13	0.000	0.906	0.905	0.908
$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$	-2789.7	3843.7	-1902.9	1773.8	17	0.000	0.896	0.895	0.898
$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$	-2762.4	3598.3	-1775.1	1974.6	22	0.000	0.900	0.899	0.898
$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$	-2874.6	3854.6	-1903.3	1942.7	22	0.000	0.877	0.876	0.875
$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$	-2663.6	3320.0	-1634.0	2059.2	24	0.000	0.875	0.879	0.866
Joint model	-16509.1		-11583.1	9851.9	110	0.000			
MNOM									
$P(Y_1 \mid \boldsymbol{X})$	-2770.6	4841.3	-2400.7	739.9	18	0.000	0.902	0.901	0.902
$P(Y_2 \mid Y_1;\boldsymbol{X})$	-2648.1	3967.2	-1959.6	1377.0	22	0.000	0.906	0.910	0.907
$P(Y_3 \mid Y_1,Y_2;\boldsymbol{X})$	-2789.7	3844.6	-1894.3	1790.9	26	0.000	0.896	0.899	0.893
$P(Y_4 \mid Y_1,Y_2,Y_3;\boldsymbol{X})$	-2762.4	3585.2	-1760.6	2003.6	30	0.000	0.902	0.898	0.903
$P(Y_5 \mid Y_1,Y_2,Y_3,Y_4;\boldsymbol{X})$	-2874.6	3841.6	-1884.8	1979.7	34	0.000	0.879	0.876	0.880
$P(Y_6 \mid Y_1,Y_2,Y_3,Y_4,Y_5;\boldsymbol{X})$	-2663.6	3332.0	-1626.0	2075.3	38	0.000	0.878	0.864	0.883
Joint model	-16509.1		-11526.0	9966.3	168	0.000			

TABLE 5.8: Goodness-of-fit test results of joint models for POPM, PPOM and MNOM.

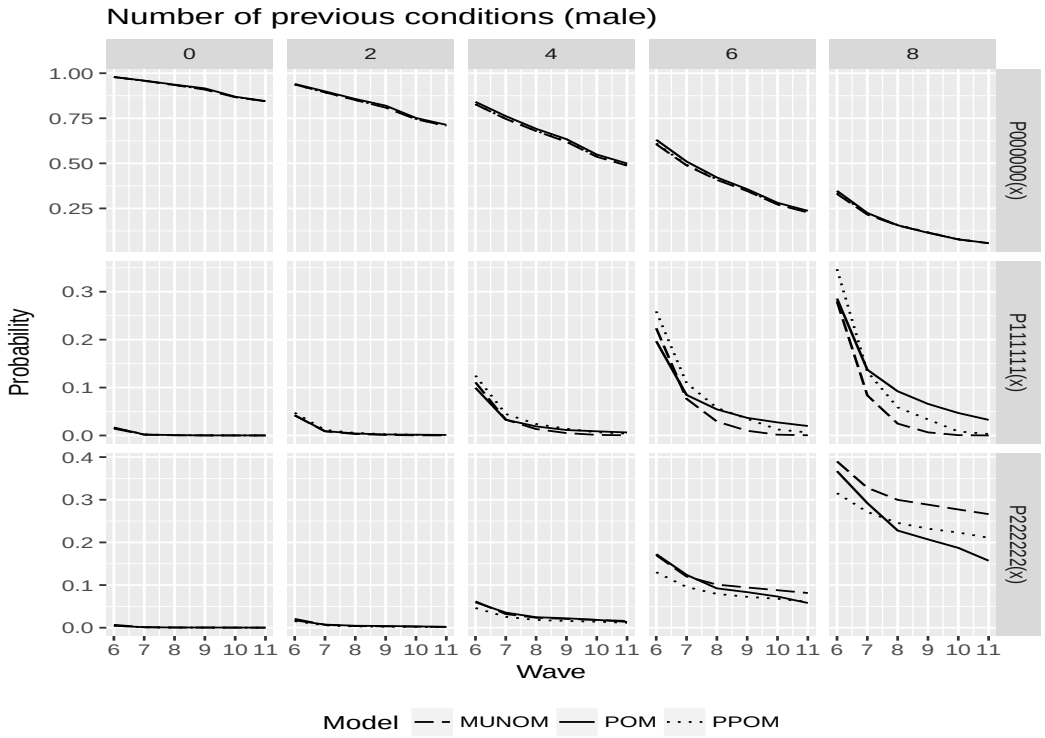
Models	POM			PPOM			MNOM		
	χ^2	d.f.	p.value	χ^2	d.f.	p.value	χ^2	d.f.	p.value
Goodness-of-fit of joint models (Pearson χ^2)									
$P(Y_1,Y_2 \mid \boldsymbol{X})$	20.07	8	0.010	4.96	8	0.762	4.36	8	0.823
$P(Y_1,Y_2,Y_3 \mid \boldsymbol{X})$	63.82	26	0.000	35.08	126	0.110	26.97	26	0.411
$P(Y_1,Y_2,Y_3,Y_4 \mid \boldsymbol{X})$	182.66	77	0.000	122.25	77	0.001	98.34	77	0.051
$P(Y_1,Y_2,Y_3,Y_4,Y_5 \mid \boldsymbol{X})$	439.80	160	0.000	172.22	160	0.241	299.55	160	0.000
$P(Y_1,Y_2,Y_3,Y_4,Y_5,Y_6 \mid \boldsymbol{X})$	958.56	254	0.000	301.23	254	0.022	768.16	254	0.000
Goodness-of-fit of joint models (Likelihood ratio χ^2)									
$P(Y_1,Y_2 \mid \boldsymbol{X})$	21.41	8	0.006	5.20	8	0.736	4.55	8	0.804
$P(Y_1,Y_2,Y_3 \mid \boldsymbol{X})$	69.06	26	0.000	37.27	26	0.071	29.23	26	0.301
$P(Y_1,Y_2,Y_3,Y_4 \mid \boldsymbol{X})$	200.70	77	0.000	133.76	77	0.000	111.55	77	0.006
$P(Y_1,Y_2,Y_3,Y_4,Y_5 \mid \boldsymbol{X})$	492.54	160	0.000	431.00	160	0.000	364.39	160	0.000
$P(Y_1,Y_2,Y_3,Y_4,Y_5,Y_6 \mid \boldsymbol{X})$	850.42	254	0.000	633.49	254	0.000	717.16	254	0.000

5.4.1 Predicted joint probabilities

First, marginal and conditional probabilities were predicted using various specified covariate vector and then the joint probability of outcomes are predicted for three selected trajectories. Those trajectories are: (i) $\hat{P}(Y_1=0,Y_2=0,Y_3=0,Y_4=0,Y_5=0,Y_6=0 \mid \boldsymbol{x}^*)$ remains functional limitations free from wave six to eleven. (ii) $\hat{P}(Y_1=1,Y_2=1,Y_3=1,Y_4=1,Y_5=1,Y_6=1 \mid = \boldsymbol{x}^*)$ one functional limitations among all six waves. (iii) $\hat{P}(Y_1=2,Y_2=2,Y_3=2,Y_4=2,Y_5=2,Y_6=2 \mid \boldsymbol{x}^*)$ two or more functional limitations from wave six to eleven. Figure 5.2-5.7 displays three joint predicted risks. The predicted risk at wave six in the graphs are marginal probability while from wave seven onward are joint probability.

Figure 5.2 displays the predicted joint risks of events free using POM, PPOM and MNOM by the number of previous conditions (0,2,4,6, and 8) and for a male subject. The value of other predictors were, mstat=0, Age=65 years, whether drink=1, white=1, Educ. =12 years, and veteran status=1. The predicted joint risk of functional limitations free $P_{000000}(\boldsymbol{x}^*)$ from all three models for varying number of previous conditions was overly similar. This probability was very high during early waves and gradually decreased at later waves. For $P_{111111}(\boldsymbol{x}^*)$ and $P_{222222}(\boldsymbol{x}^*)$ predicted risks of events differs noticeably for a subject with 8 previous conditions at later waves. Later waves predicted risks of joint events are much lower compared to early waves. The highest predicted risks were based on MNOM followed by PPOM and POM, respectively. A similar pattern was found for the female sample (Figure 5.3).

The predicted joint risks from all three models by varying age (60, 65, 70, 75 and 80)



H

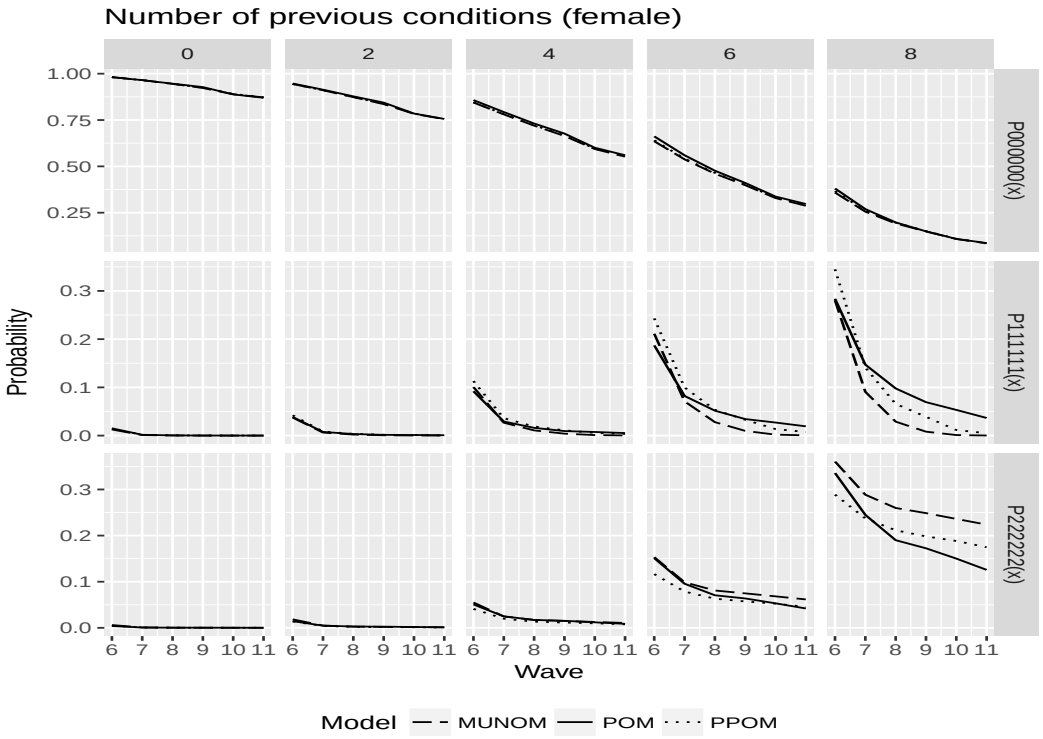
FIGURE 5.2: Predicted joint probability for male from three models.

for a male subject by setting mstat=0, mean(N.cond), whether drink=1, white=1, Educ. =12 years, and veteran status=1 is shown in Figure 5.4. The predicted risks were very close for $P_{000000}(\boldsymbol{x}^*)$ and $P_{111111}(\boldsymbol{x}^*)$ for all ages. For $P_{222222}(\boldsymbol{x}^*)$ were very similar up to age 70 years while a noticeable difference was observed for age 75 and 80 years. In this case, the highest probabilities were estimated using POM followed by MNOM and PPOM, respectively. Similar trends are observed for female sample (Figure 5.5).

Predicted joint risks for three trajectories using three models for a veteran subject is presented in Figure 5.6 by setting age=65 years, mstat=0, N.cond =2, whether drink=1, white=1, Educ. =12 years, gender=1. For $P_{000000}(\boldsymbol{x}^*)$ trajectory there is no differences between the predicted risks of event free from all three models. Slight differences were observed at early waves between the predicted risks of joint events from MNOM and POM or PPOM for $P_{111111}(\boldsymbol{x}^*)$ and $P_{222222}(\boldsymbol{x}^*)$ paths. The difference disappears at later waves. A similar trend is observed for the non-veteran subject (Figure 5.7).

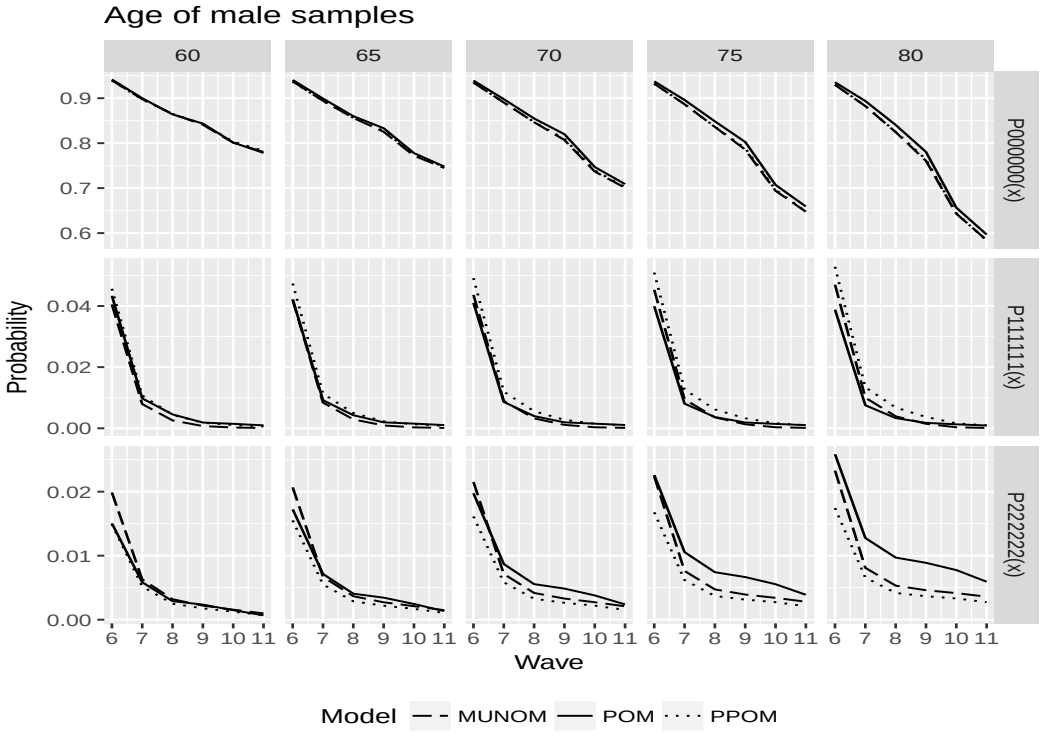
5.5 Conclusions

Ordinal repeated outcomes are collected from longitudinal studies in many disciplines. There is a great demand for the prediction of the joint probability of a sequence of ordinal events. Usually, marginal and sequence of conditional models such as the Markov models are employed and marginal and conditional probabilities are estimated from those



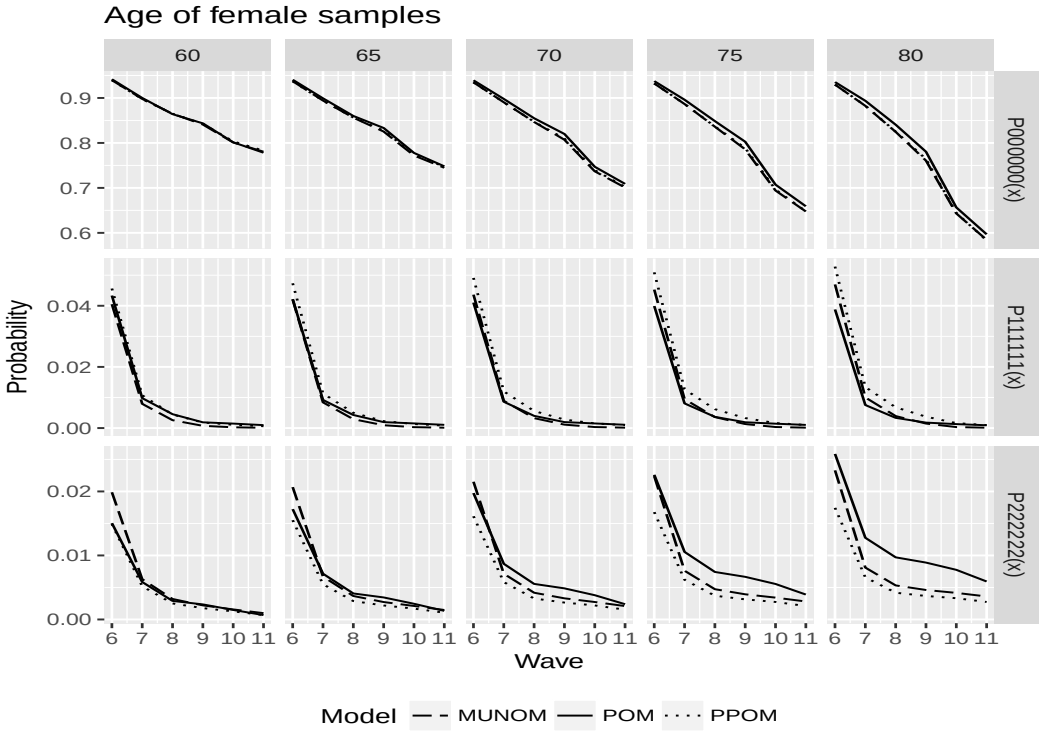
H

FIGURE 5.3: Predicted joint probability for female from three models.



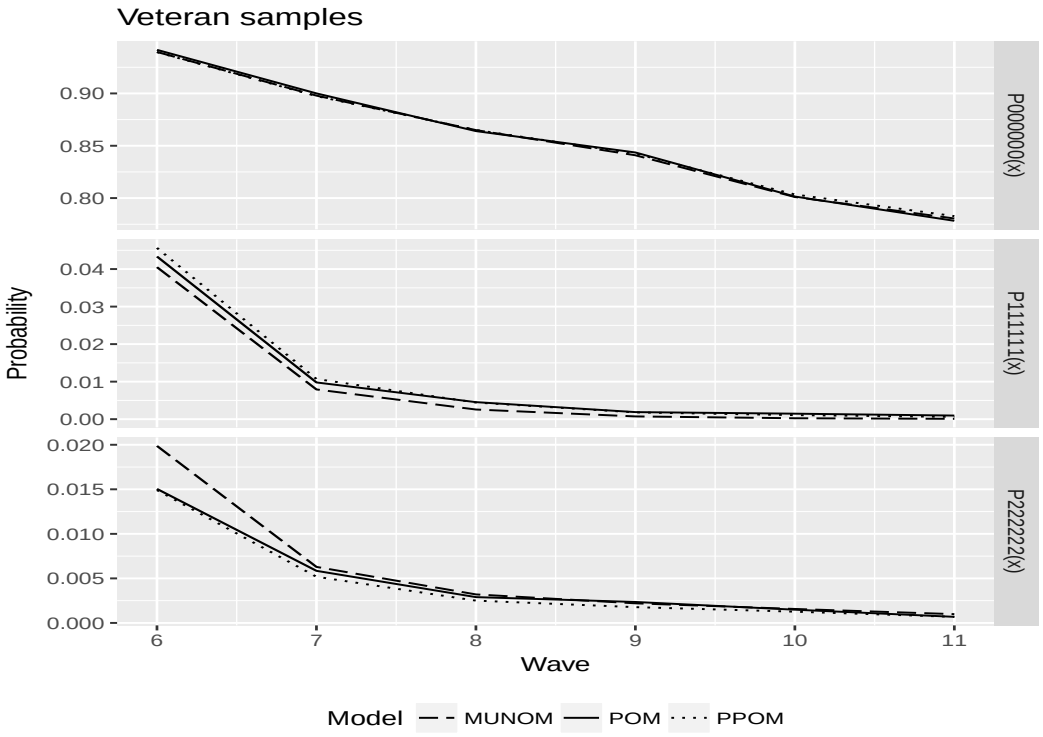
H

FIGURE 5.4: Predicted joint probability by age from three models.



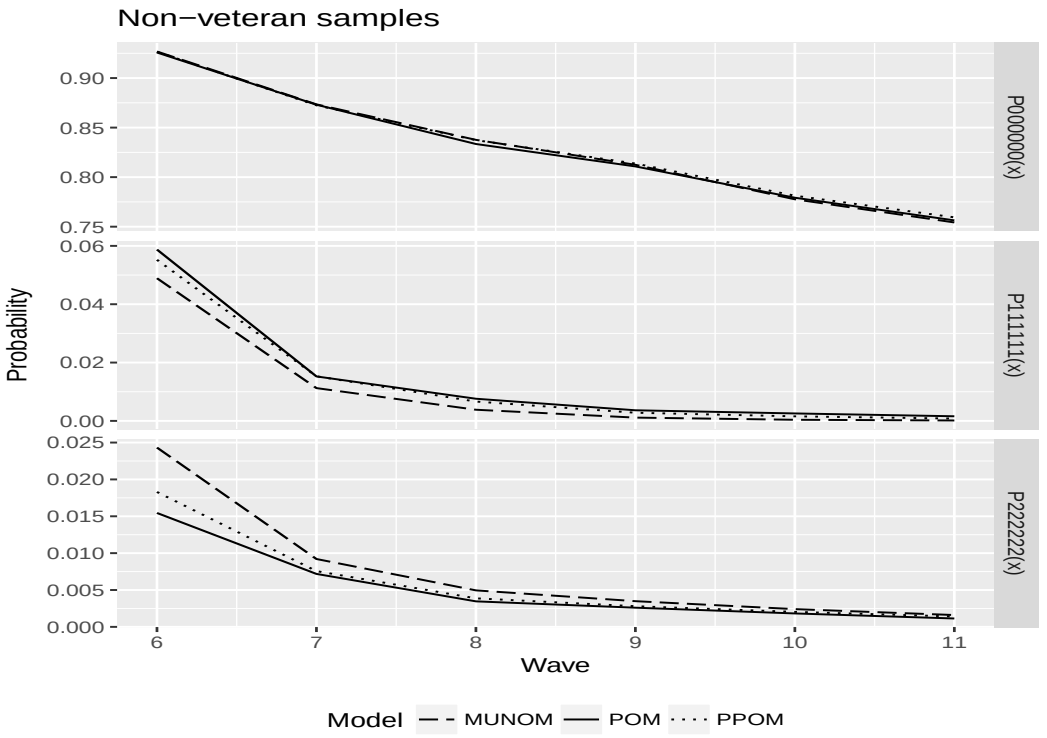
H

FIGURE 5.5: Predicted joint probability by age from three models.



H

FIGURE 5.6: Predicted joint probability vateran from three models.



H

FIGURE 5.7: Predicted joint probability non-vateran from three models.

models. Then Markov chain is used to link the marginal and conditional probabilities to obtain the joint probabilities of a sequence of events with a specified covariate vector. However, the number of conditional models need to fit increases with the increased number of follow-ups which is restricted due to over-parameterization and small sample size. In this paper, we propose three regressive models for ordinal outcomes from repeated measures (i) regressive proportional odds model, (ii) regressive partial proportional odds model and (iii) regressive multinomial logistic model. Also, we have shown multivariate model (joint model) for ordinal outcomes. Then a framework is shown to predict joint probabilities for a sequence of ordinal outcomes. The proposed model and the risk prediction framework is a new development. The major improvement of the proposed model is that only one model is needed for each repeated outcome compared to the sequence of conditional models such as Markov models. The proposed models provide the estimates for each stage in the process conditionally and using Equation (5.1) the joint model is obtained for any order to predict the risk of a sequence of events. The outcome variable at the baseline is used to estimates the parameters of the marginal model and the regressive models at the subsequent follow-ups provide the estimates of the parameters of the conditional models. The proposed framework links the marginal and conditional process and obtains predictive outcome based on the whole process through all possible trajectories.

In the proposed modeling approach interaction among previous outcomes and predictors can easily be incorporated. The interaction terms may provide a better understanding of the underlying process and the relationships between outcomes and risk factors. The

likelihood ratio test for the goodness of fit and AIC for the marginal and regressive models are shown in this paper. The prediction accuracy of POM, PPOM and MNOM for all marginal and first-order regressive models was around 0.90 which is reasonably high. This accuracy reduces with the increased order of the regressive models which was around 0.87. Also, two goodness-of-fit statistics (Pearson χ^2 and Likelihood ratio χ^2) for the proposed joint models are shown. Partial proportional odds model showed good fit for fifth order joint model whereas for MNOM fourth order joint model showed a good fit. One can easily fit proposed regressive models and predict the risk of a sequence of events using the available statistical software.

Predicted risks of a sequence of events for three selected trajectories for specified covariates vector are presented in graphs as an example. The predicted risk of outcomes from all six waves remaining in category one $P_{111111}(\boldsymbol{x})$ and remaining in category two $P_{222222}(\boldsymbol{x})$ for female by the number of previous conditions were noticeably different for a higher number of previous conditions. Similar patterns are found for a male subject. It may be noted that the number of previous conditions were significant for marginal and most of the first and higher order regressive PPOM and MNOM, but not for POM. The proposed tests for the joint model also suggested good fit for PPOM and MNOM for fifth and fourth order joint models, respectively. Partial proportional odds followed by multinomial logistic regression showed better prediction results in the case of violation of proportional odds assumption. Other available models for the ordinal outcome can easily be used in the proposed modeling framework. The proposed methods can be applied for analyzing and risk prediction for a sequence of events in many fields of studies such as epidemiology, public health, survival analysis, genetics, reliability, environmental studies, etc. This model would be very useful for analyzing big data where a large number of repeated outcomes are observed.

Chapter 6

Conclusions and future research directions

The first objective of this dissertation has been to simplify the models for risk prediction of a sequence of events by proposing multistage modeling approach for continuous time data those are observed from longitudinal studies. Proposing new regressive models for multinomial and ordinal outcomes and a framework for prediction of the joint probability of a sequence of events for longitudinally measured discrete time data has been the second objective. In addition, proposing a test for goodness-of-fit of the joint model and a test for independence of repeated outcomes was under consideration. The study objectives also included a simulation to investigate the model performance as well as use of real-life data to demonstrate the usefulness of the proposed models and risk prediction framework.

Klein *et al.* (1994) first showed existing methods of risk prediction of a sequence of events for continuous time data in terms of hazards for the transition from the multi-state model based on the work of Arjas and Eerola (1993). Later, Putter *et al.* (2006) and Putter *et al.* (2007) presented a comprehensive illustration of prediction of the probability of a sequence of events using the multi-state model. According to Aalen *et al.* (2008) this predicted probability is a reasonable estimate of the risk of a sequence of events. However, existing theories were demonstrated on the basis of specific problems and had several drawbacks. In the previous attempts, the events were not considered in the multistage framework. Hence the underlying theory remained complex. The main challenge is the simplification, and generalization of the existing method for a large sequence of events occurring at different stages. Also, the existing framework involves deriving multiple complex integrals for a specific problem and special computer skills are required to use these methods. Because of the complexity of existing methods their applications remain limited.

In predicting the sequence of events, we need to link the likely transitions at different stages of the process through potential trajectories (paths). The proposed alternative multistage approach simplifies the transition model for the underlying paths for risk prediction

and provides the estimates for each stage in the process conditionally. Then the conditional estimates are linked based on marginal-conditional models that provide the joint probabilities required for predicting the risk of a sequence of events based on the potential risk factors. The proposed method for risk prediction is a simple approach, compared to the existing ones, and this approach readily generalizes to any number of events in the process from the beginning to the endpoints. The same integral derived for a trajectory can be used for other trajectories. Predicted risk of a sequence of events for different trajectories using the real-life data set from the proposed approach produced very similar results to that from the existing method. The proposed approach can be used for risk prediction for the different disease process.

In many studies, multinomial responses are observed longitudinally during an interval or the exact time of events are unknown which produces discrete time data. One needs to deal with transitions to a number of states over time generating a large number of trajectories from beginning to the end of the study. Most of the available methods of risk prediction are for a single outcome (Yu, 2003). One approach is to use the Markov chain, where marginal and conditional models of different order are linked to obtaining the joint model required for risk prediction of a sequence of events (Islam and Chowdhury, 2010). However, this approach is restricted due to over-parameterization. Also, for repeated outcomes from a large number of follow-ups, required number of conditional models such as the Markov models needed to be fitted grows rapidly. Fitting a large number of conditional models would make the existing method inflexible and computationally infeasible (Wen *et al.*, 2016).

To overcome the problem of over-parameterization and to develop an efficient and simple method of risk prediction, following the work of Islam and Chowdhury (2010) we have proposed the regressive models for multinomial outcomes from repeated measures. The motivation for this model comes from the need for generalization of competing risks models at different stages for discrete time data and prediction of a sequence of events. The Markov chain is used to link the marginal and conditional probabilities to estimate the joint probabilities for a sequence of events. The marginal probability is estimated using a marginal model based on the outcome of the first follow-up, and conditional probabilities are estimated from the proposed regressive model for subsequent follow-ups. The main improvement made in the newly proposed method is that one needs to fit only a single model for first or higher order conditional models. This formulation can easily handle a large number of states emerging from different follow-ups from longitudinal data. This model allows to include interaction between previous outcomes and covariates in the model. The real data application shows the advantage of the model while the simulation study reveals minimal estimation bias associated with the model. These, together indicate the decent performance of the proposed regressive model. Results from training and test data showed no indication of overfitting or underfitting and showed impressive prediction accuracy.

For a robust and accurate risk prediction model, it is necessary to measure the model performance which needs to be rigorously validated (Calster *et al.*, 2017; Wehberg and Schumacher, 2004). Also, test for independence among the repeated measures is important. If the outcomes are independent, simpler models, such as marginal models for each repeated outcomes can be used instead of conditional models for risk prediction of a sequence of events. Most of the available tests for the goodness-of-fit are for a single binary or multinomial outcome and are not directly appropriate for the multinomial outcomes from repeated measures.

We proposed tests to evaluate the goodness-of-fit for a joint model for multinomial outcomes from repeated measures. One is based on Pearson chi-square, and another one is based on likelihood ratio. We used the marginal and regressive models to obtain the estimates of marginal and conditional probabilities. Then the joint probabilities are estimated linking the marginal and conditional probabilities which are used to estimate the expected frequencies. Using observed and expected frequencies Pearson chi-square test is shown. For the test of independence marginal probabilities are used similarly. Findings from the real data application and simulation study demonstrated better performance of the proposed tests.

In various discipline, ordinal outcomes are observed longitudinally producing a sequence of events over discrete time. Proportional odds model is a popular choice to model a single ordinal outcome as a function of risk factors provided that the required proportional odds assumption is fulfilled. In the violation of this assumption, partial proportional odds model is used although other alternatives are available. For predicting the risk of a sequence of ordinal outcomes, it is necessary to examine the events during subsequent follow-ups using a multivariate model. However, a multivariate approach is often complicated and would be difficult to develop for a large number of repeated ordinal outcomes. One approach for risk prediction of a sequence of events is to use marginal and conditional models to obtain a joint model for risk prediction which is limited due to over-parameterization.

Proportional odds regressive model and partial proportional odds regressive models are proposed for repeated ordinal outcomes. Then the estimates of the conditional probabilities are obtained from the proposed regressive models for ordinal outcomes. Estimates of marginal probability and conditional probability are linked to obtain the joint probability which is the risk of a sequence of events based on specified covariate vector. It poses all the advantages as in the regressive models for multinomial outcomes. Suggested re-parameterization in the proposed regressive models reduces the number of parameter sets needs to be estimated. Ordinal outcomes along with selected risk factors from HRS data are used for the application. Estimates from the proposed approach and simulation confirm the utility of the proposed models. The prediction accuracy is also reasonably high along with the absence of overfitting and underfitting.

Some important implications of the proposed methods for risk prediction of a sequence of events from the repeated measures data are:

Proposed method for continuous time data is a simple one.

The use of marginal and conditional models reduces the multivariate problem into a univariate problem. Each marginal, conditional and proposed regressive models are a univariate case.

Proposed approach allows to include interaction among previous outcomes and predictors which may provide a better understanding of the underlying disease.

The predicted risk would allow health care providers to screen individuals that would help them to suggest necessary therapy and prevention

The proposed method allows generalization to any number of stages without making the process complex.

Using existing statistical software one can predict the risk of a sequence of events with minimal programming knowledge.

This research suggests some future development for risk prediction for continuous time data where the Markov models are used. It may be possible to use regressive models to reduce the over-parameterization for continuous time data for risk prediction of a sequence of events. For ordinal outcomes, there are other alternative models (e.g., continuation ratio model, stereotype model, etc.) in the case of the violation of the proportional odds assumption. These models could easily be adopted in the proposed framework for risk prediction which will allow comparing the results from various models for ordinal outcomes.

Bibliography

Aalen, O. O., Borgan, O. and Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. Springer Science+Business Media, LLC.

Agresti, A. (2013). *Categorical data analysis, Third edition*. John Wiley & Sons Inc. Hoboken, New Jersey.

Akhter, H. H., Chowdhury, M. E. E. K. and Sen, A. (1996). *A cross-sectional study on maternal morbidity in Bangladesh*. BIRPERHT, Dhaka, Bangladesh.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* **26**, 1323–1333.

Andersen, P. K. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research* **11**, 203–215.

Andersen, P. K., Hensen, L. S. and Keiding, N. (1991). Non and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics* **18**, 153–167.

Andersen, P. K. and Perme, M. P. (2008). Inference for outcome probabilities in multistate models. *Statistical Methods in Medical Research* **11**, 91–115.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of Royal Statistical Society, Series B* **46**, 1–30.

Arjas, E. and Eerola, M. (1993). On predictive causality in longitudinal studies. *Journal of Statistical Planning and Inference* **34**, 361–368.

Austin, P. C. and Steyerberg, E. W. (2014). Graphical assessment of interal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine* **33**, 517–535.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81**, 767–775.

Barnes, D. E., Mehta, K. M., Boscardin, W. J., Fortinsky, R. H., Palmer, R. M., A, K. K. and Landefeld, C. S. (2013). Prediction of recovery, dependence or death in elders who become disabled during hospitalization. *Journal of General Internal Medicine* **28**, 261–268.

Barnett, A. G., Batra, R., Graves, N., Edgeworth, J., Robotham, J. and Cooper, B. (2009). Using a longitudinal model to estimate the effect of methicillin-resistant staphylococcus aureus infection on length of stay in an intensive care unit. *American Journal of Epidemiology* **170**, 1186–1194.

Beddoes-Ley, L., Khaw, D., Duke, M. and M, B. (2016). A profile of four patterns of vulneralibility to functional decline in older general medicine patients in victoria, australia: a cross sectional survey. *BMC Geriatrics* **16**, 1–12.

Beyersmann, J., Schumacher, M. and Allignol, A. (2012). *Competing risks and multistate models with R*. Springer, New York.

Bodilsen, A. C., Klausen, H. H., Petersen, J., Nina, B., Andersen, O., Jorgensen, L. M., Juul-Larsen, H. G. and Bandholm, T. (2016). Prediction of mobility limitations after hospitalization in older medical patients by simple measures of physical performance obtained at admission to the emergency department. *PLOS ONE* **11**, 1–19.

Bonney, G. E. (1986). Regressive logistic models for familial disease and other binary trials. *Biometrics* **42**, 611–625.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics* **43**, 951–973.

Bovelstad, H. M., Nygard, S. and Borgan, O. (2009). Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics* **10**, 1–9.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**, 1171–1178.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability (pdf). *Monthly Weather Review* **78**, 1–3.

Calster, B. V., Hoorde, K. V., Vergouwe, Y., Bobdiwala, S., Condous, G., Kirk, E., Bourne, T. and Steyerberg, E. W. (2017). Validation and updating of risk models based on multinomial logistic regression. *Diagnostic and Prognostic Research* **1**, 1–14.

Calster, B. V., Vergouwe, Y., Belle, V., Condous, G., Looman, C., Timmerman, D. and Steyerberg, E. W. (2012). Assessing the discriminative ability of risk models for more than two outcome categories: a perspective. *European Journal of Epidemiology* **27**, 761–770.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis* **5**, 315–327.

Commenges, D. (2002). Inference for multi-state models for interval-censored data. *Statistical Methods in Medical Research* **11**, 167–182.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B* **34**, 187–220.

Dabrowska, D. M., Sun, G. and Horowitz, M. M. (1994). Cox regression in a Markov renewal model: An application to the analysis of bone marrow transplant data. *Journal of the American Statistical Association* **89**, 867–877.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K. and Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham heart study. *Statistics in Medicine* **9**, 1501–1515.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall, New York.

Efron, B. and Tibshirani, R. J. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Fagerland, M. W. and Hosmer, D. W. (2013). A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine* **32**, 2235–2249.

Fagerland, M. W., Hosmer, D. W. and Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine* **27**, 4238–4253.

Farewell, V. T. (1979). An application of Cox's proportional hazard model to multiple infection data. *Applied Statistics* **28**, 136–143.

Filippi, V., Ganaba, R., Baggaley, R. F., Marshall, T., Storeng, K. T., Sombie, I., Ouattara, F., Ouedraogo, T., Akoum, M. and Meda, N. (2007). Health of women after severe obstetric complications in Burkina Faso: A longitudinal study. *Lancet* **370**, 1329–1337.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

Fox, E. R., Samdarshi, T. E. and Musani, S. K. e. a. (2016). Development and validation of risk prediction models for cardiovascular events in black adults. *JAMA Cardiology* **1**, 15–25.

Gail, M. A. (1975). A review and critique of some models used in competing risk analysis. *Biometrics* **31**, 209–222.

Goeman, J. J. and le Cessie, S. (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics* **62**, 980–985.

Gottschau, A. (1994). Markov chain model for multivariate binary panel data. *Scandinavian Journal of Statistics* **21**, 57–71.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**, 1141–1154.

Gundersen, K., Kvaløy, J. T., Kramer-Johansen, J., Steen, P. A. and Eftestøl, T. (2009). Development of the probability of return of spontaneous circulation in intervals without

chest compressions during out-of-hospital cardiac arrest: an observational study. *BMC Medicine* **6**, 1–9.

Hand, D. J. and Till, R. J. (2001). A simple generalization of the area under the roc curve for multiple class classification problems. *Machine Learning* **45**, 171–186.

Harrell, F. E. (2001). *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis*. Springer, New York.

Holt, J. D. (1978). Competing risk analyses with special reference to matched pair experiments. *Biometrika* **65**, 159–165.

Hosmer, D. W., Hosmer, T., le Cessie, S. and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* **16**, 965–980.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **10**, 1043–1069.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression. 2nd Ed.* Wiley and Sons, New Jersey.

Hosmer, D. W. and Lemeshow, S. (2013). *Applied logistic regression (Third edition)*. Wiley, New Jersey.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis* **5**, 239–264.

HRS (2014). *Health And Retirement Study, (Wave [1-10]/Year[1992-2014]) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIAU01AG09740)*. Ann Arbor, MI.

Islam, M. A. (1994). Multistate survival models for transitions and reverse transitions: An application to contraceptive use data. *Journal of Royal Statistical Society, Series A* **157**, 441–455.

Islam, M. A., Alzaid, A., Chowdhury, R. I. and Sultan, K. S. (2013). A generalized bivariate bernoulli model with covariate dependence. *Journal of Applied Statistics* **40**, 1064–1075.

Islam, M. A. and Chowdhury, R. I. (2006). A higher order markov model for analyzing covariate dependence. *Applied Mathematical Modelling* **30**, 477–488.

Islam, M. A. and Chowdhury, R. I. (2010). Prediction of disease status: A regressive model approach for repeated measures. *Statistical Methodology* **7**, 520–540.

Islam, M. A. and Chowdhury, R. I. (2017). *Analysis of repeated measures data*. Springer, Singapore.

Islam, M. A., Chowdhury, R. I., Bae, S. and Singh, K. P. (2014). Assessing the association in repeated measures of depression. *Advances and Applications in Statistics* **42**, 89–83.

Islam, M. A., Chowdhury, R. I. and Briollais, L. (2012). A bivariate binary model for testing depedence in outcomes. *Bulletin of Malaysian Mathematical Sciences Society (2)* **35**, 845–858.

Islam, M. A., Chowdhury, R. I., Chakraborty, N. and Bari, W. (2004). A multistage model for maternal morbidity during antenatal, delivery and postpartum periods. *Statistics in Medicine* **23**, 137–158.

Islam, M. A., Chowdhury, R. I. and Huda, S. (2009). *Markov models with covariate dependence for repeated measures*. Nova Science, New York.

Islam, M. A., Chowdhury, R. I. and Huda, S. (2013). A multistate transition model for analyzing longitudinal depression data. *The Bulletin of the Malaysian Mathematical Sciences Society (2)* **36**, 637–655.

Islam, M. A., Chowdhury, R. I. and Singh, K. P. (2012). A markov model for analyzing polytomous outcome data. *Pakistan Journal of Statistics and Operation Research* **8**, 593–603.

Islam, M. A., Khalaf, S. S. and Chowdhury, R. I. (2009). Estimation and tests for a longitudinal regression model based on markov chain. *Statistical Methodology* **6**, 478–489.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.

Johnson, R. A. and Wichern, D. W. (2008). *Applied multivariate statistical analysis, (6th edition)*. Prentice Hall, New Jersey.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. Wiley: New York.

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

Kay, R. (1982). The analysis of transition times in multistate stochastic processes using proportional hazard regression models. *Communications in Statistics - Theory and Methods* **11**, 1743–1756.

Klein, J. P., Keiding, N. and Copelan, E. A. (1994). Plotting sumary predictions in multistate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine* **13**, 2315–2332.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated Data*. Springer-Verlag New York.

Kleinbaum, D. G. and Klein, M. (2012). *Survival analysis: A self-learning text (Third edition)*. Springer, New York.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC curves for continuous data*. CRC Press.

Lall, R., Campbell, M. J., Walters, S. J. and Morgan, K. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research* **11**, 49–67.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley: New York.

le Cessie, S. and van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests via random effects models. *Biometrics* **51**, 600–614.

Lee, K. and Daniels, M. J. (2007). A class of markov models for longitudinal ordinal data. *Biometrics* **63**, 1063–1067.

Lipsitz, S. R., Fitzmaurice, G. M. and Molenberghs, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society, Series C* **45**, 175–190.

Liski, E. P. and Nummi, T. (1996). Prediction in repeated-measures models with engineering applications. *Technometrics* **38**, 25–26.

Long, S. J. (1997). *Regression models for categorical and limited dependent variables*. SAGE Publications, London.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Chapman and Hall, London/New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, Second edition*. Chapman and Hall/CRC.

Meira-Machado, L., Una-Alvarez, J., Cadarso-Suarez, C. and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research* **18**, 195–222.

Miller, E. M., Thomas, R., Have, T., Reboussin, B. A., Lohman, K. K. and Rejeski, W. J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple-cause nonresponse. *Journal of the American Statistical Associations* **96**, 844–857.

Moons, K., Royston, P., Vergouwe, Y., Grobbee, D. and Altman, D. (2009). Prognosis and prognostic research: what, why, and how? *BMJ* **338**, b375.

Moreira, A. and Meira-Machado, L. (2012). survivalbiv: estimation of the bivariate distribution function for sequentially ordered events under univariate censoring. *Journal of Statistical Software* **46**, 1–16.

Muenz, L. R. and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequence. *Biometrics* **41**, 91–101.

Peek, N., Arts, D., Bosman, R., van der Voort, P. and de Keizer, N. (2007). External validation of prognostic models for critically ill patients required substantial sample sizes. *Journal of Clinical Epidemiology* **60**, 491–501.

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society, Series C* **39**, 205–217.

Pierce, D. A., Stewart, W. H. and Kopecky, K. J. (1979). Distribution free regression analysis of grouped survival data. *Biometrics* **35**, 785–793.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**, 57–67.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541–554.

Putter, H. (2011). *Tutorial in biostatistics: Competing risks and multi-state models Analyses using the mstate package*. http://cran.r-project.org/web/packages/mstate/vignettes/Tutorial.pdf.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* **26**, 2389–2430.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R. and van de Velde, C. J. H. (2006). Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal* **3**, 366–380.

Rothman, K. J. (2002). *Epidemiology: An introduction*. Oxford University Press: New York.

Royston, P. (1992). The use of customs and other techniques in modeling continuous covariates in logistic regression. *Statistics in Medicine* **11**, 1115–1129.

Steyerberg, E. W. (2009). *Clinical prediction models. A practical approach to development, validation, and updating*. Springer, New York.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association* **83**, 426–431.

Sun, J. (2006). *Statistical analysis of interval censored failure time data*. Springer.

Toledano, A. Y. and Gatsonis, C. (1996). Ordinal regression methodology for roc curve derived from correlated data. *Statistics in Medicine* **15**, 1807–1826.

Tripepi, G., Heinze, G. K., Jager, J., Stel, V. S., Dekker, F. W. F. and Zoccali, C. (2013). Risk prediction models. *Nephrology Dialysis Transplant* **28**, 1975–1980.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67**, 250–251.

Vergouwe, Y., Steyerberg, E., Eijkemans, M. and Habbema, J. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology* **58**, 475–483.

Wallace, E., Stuart, E., Vaughan, N., Bennett, K., Fahey, T. and Smith, S. M. (2014). Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. *Medical Care* **52**, 751–765.

Walters, S. J., Campbell, M. J. and Lall, R. (2001). Design and analysis of trials with quality of life as an utcome: a practical guide. *Journal of Biopharmaceutical Statistics* **11**, 155–176.

Wehberg, S. and Schumacher, M. (2004). A comparison of nonparametric error rate estimation methods in classification problems. *Biometrical Journal* **46**, 35–47.

Wen, Y., He, Z., Li, M. and Lu, Q. (2016). Risk prediction modeling of sequencing data using a forward random field method. *Scientific Reports* **6**, 21120.

Yu, F. (2003). Use of a markov transition model to analyse longitudinal low-back pain data. *Statistical Methods in Medical Research* **12**, 321–331.