

ANALYSIS OF LONGITUDINAL DATA FROM THE BETASERON
MULTIPLE SCLEROSIS CLINICAL TRIAL

By

Yulia D'yachkova

B. Sc. (Mathematics) Moscow State University

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

May 1997

© Yulia D'yachkova, 1997

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date 28 May '97

Abstract

Longitudinal data sets consist of repeated observations for each subject over time; and often a corresponding set of covariates is available. Analysis of longitudinal data is often based on summaries over time. Summarizing the data allows one to use simple techniques for analysis but does not allow analysis of the patterns over time and does not take advantage of the within subject information. In many fields, repeated measures analysis of variance and multivariate analysis of variance are commonly used to analyze longitudinal data on continuous responses. Such analyses are appropriate only when the responses for each subject are multivariate Gaussian with a common covariance matrix for all subjects. In addition, all subjects are required to have measurements at exactly the same times, and no missing values may be present. In many cases, however, the longitudinal response does not satisfy these assumptions. Therefore, application of the traditional methods of analysis is limited even for continuous responses.

This thesis discusses and compares several more recently developed methods for the analysis of longitudinal data. One method, the generalized estimating equations approach, requires only minimal assumptions about the true correlation structure in the data for each subject to yield consistent estimates of regression parameters and their standard errors. The method can be applied to binary and count data as well as to continuous data. Another method, the random effects regression model, is limited to the analysis of continuous responses. An advantage of this method is that in addition to estimating population average parameters it also allows estimation of individual parameters for each subject. Finally, the modification of the random effects regression approach for the analysis of ordinal responses, the mixed effects ordinal logistic regression model, is

presented.

The methods are extensively illustrated using the data from the Betaseron clinical trial in relapsing-remitting multiple sclerosis (MS). These methods facilitated the examination of patterns over time, therefore they not only identified the presence of treatment effect, but also indicated the nature of the effect. Hence, these methods enable much more information to be extracted from the MS data set than the traditional ANOVA-based methods, and therefore provide useful and powerful tools for researchers in this subject area.

Table of Contents

Abstract	ii
List of Tables	vii
List of Figures	x
Acknowledgement	xii
1 Introduction	1
1.1 Background to the Study	1
1.2 Description of the Interferon Beta-1b Clinical Trial	4
1.3 Outline of the Thesis	6
2 Description of the Data	9
2.1 Covariates	11
3 Methodology	15
3.1 Overview	15
3.2 GEE Approach	18
3.3 Random Effects Regression Models	21
3.4 Random Effects Ordinal Regression Models	23
4 Number of Active Lesions	27
4.1 Description of UBC frequent MRI sub-study	27
4.1.1 Covariates	28

4.2	Preliminary Analysis	31
4.3	GEE Analysis	40
4.4	Summary	44
5	Exacerbations	47
5.1	Data Set Description	47
5.2	Preliminary Analysis	49
5.3	GEE Analysis	58
5.4	Analysis on Reduced Data Set	64
5.5	Summary	72
6	EDSS Scores	75
6.1	Data Set Description	75
6.2	Preliminary Data Analysis	78
6.2.1	Repeated Measures ANOVA	84
6.2.2	MANOVA	92
6.3	GEE Analysis	95
6.4	Random Effects Regression Model	100
6.5	Mixed Effects Ordinal Logistic Regression Analysis	104
6.5.1	MIXOR Implementation	105
6.5.2	Results of Analyses	110
6.6	Summary	114
7	Discussion	116
7.1	Conclusions about Data	116
7.2	Methods of Analyses	117
	Bibliography	120

List of Tables

2.1	Number of Patients in the Centers	9
2.2	Descriptive Statistics for Age, Duration and Length on Study	12
2.3	Counts for Initial EDSS, Gender and Indicator of Drop-out	14
4.4	Descriptive Statistics for Age, Duration and Initial EDSS	29
4.5	Counts for Initial EDSS, Origin and Gender	31
4.6	Descriptive Statistics for Average Number of Active Lesions	33
4.7	P-values from ANOVA for Average Number of Active Lesions	33
4.8	Square Root of Response versus Baseline Covariates	35
4.9	Age and Treatment Effects: Average Number of Active Lesions	36
4.10	Correlations Among Responses on Scans 1-6, 8, 11, 12	38
4.11	Effects for Age and Treatment Groups: Log Scale	39
4.12	Effects for Age and Treatment Group: GEE Approach	40
4.13	Linear Effects for Time: GEE Approach	43
5.14	Descriptive Statistics for Percentage of Exacerbations	48
5.15	P-values from ANOVA for Percentage of Exacerbations	50
5.16	ANCOVA of Ranked Response	52
5.17	ANCOVA on Percentage of Exacerbations	53
5.18	Effects Estimates: Percent Scale	54
5.19	Correlations Among First 10 Responses	56
5.20	Estimates of Effects: Logit Scale	57
5.21	Effects for Treatment Group, Age and Dropout: GEE	59

5.22	Linear and Quadratic Effects for Time: GEE	62
5.23	Common Linear and Quadratic Effects for Time: GEE	63
5.24	Effects for Treatment Group, Age and Dropout: GEE	64
5.25	Descriptive Statistics for Percentage of Exacerbations, Completers	65
5.26	Number of Completers by Centers	66
5.27	ANCOVA on Ranked Response, Completers	67
5.28	Effects of Treatment Group, Age and Gender for Completers: GEE	68
5.29	Linear and Quadratic Time Effects for Completers: GEE	71
5.30	Common Linear Time Effect for Completers: GEE	71
6.31	Break-down of Patients by Center	76
6.32	Original and Re-expressed EDSS Scores	78
6.33	Descriptive Statistics for Rate of Change of REDSS	80
6.34	P-values from ANOVA for Rate of Change on REDSS	80
6.35	ANCOVA on Ranked Response	81
6.36	Estimates of Effect from ANCOVA: raw scale	84
6.37	Break-down of Missing EDSS Values by Treatment Arm	86
6.38	Repeated Measures ANOVA on REDSS Change without Covariates	88
6.39	Standard Deviations of REDSS Change at Each Time Point	89
6.40	Correlation Matrices for Change in REDSS	90
6.41	Repeated Measures ANOVA on REDSS Change with Covariates	91
6.42	Estimates for Continuous Covariates: Repeated Measures ANOVA	91
6.43	MANOVA on REDSS Change with Age and Initial EDSS	94
6.44	GEE Model for REDSS Change with Separate Linear Trends	96
6.45	GEE Model for REDSS Change with Separate Linear and Quadratic Trends	98
6.46	GEE Model for REDSS Change with Covariates and Linear Trends	99

6.47	Random Effects Model for REDSS Change with Separate Linear Trends .	101
6.48	Random Effects Model for REDSS Change with Separate Linear and Quadratic Trends	102
6.49	Random Effects Model for REDSS Change with Time Trends and Covariates	103
6.50	New Categories of EDSS scores	107
6.51	Number of Non-Varying Responses and Iterations for New Categories . .	108
6.52	Fixed Effects and Thresholds Estimates from MIXOR for Data with 8 Categories	110
6.53	Results from MIXOR for Data with 8 Categories and Initial EDSS Score as Categorical	112
6.54	Results from MIXOR for data with 8 Categories and Baseline Covariates	113
A.55	Placebo Group: Number of Active Lesions Data	123
A.56	Low Dose Group: Number of Active Lesions Data	124
A.57	High Dose Group: Number of Active Lesions Data	125
A.58	UBC cohort: Baseline Covariates	126
A.59	Placebo Group (first 18 patients): Exacerbation Data	127
A.60	Low Dose Group (first 18 patients): Exacerbation Data	128
A.61	High Dose Group (first 18 patients): Exacerbation Data	129
A.62	Placebo Group (first 35 patients): EDSS Data	130
A.63	Low Dose Group (first 35 patients): EDSS Data	131
A.64	High Dose Group (first 35 patients): EDSS Data	132

List of Figures

2.1	Histogram for Length on Study (5-week bars)	10
2.2	Kaplan-Meier Survival Curves for Length on Study	11
2.3	Boxplots for Age, Duration, and Length on Study	13
4.4	Boxplots for Age, Duration and Initial EDSS	30
4.5	Boxplots for Average Number of Active Lesions	32
4.6	Average Number of Active Lesions versus Baseline Covariates	34
4.7	Plot of Residuals versus Fitted Values	37
4.8	Plot of Average Number of Active Lesions versus Time	41
4.9	Smoothed Plot of Average Number of Active Lesions versus Time	42
5.10	Boxplots for Percentage of Exacerbations	49
5.11	Percentage of Exacerbations versus Baseline Covariates	51
5.12	Proportion of Drop-outs by Center	53
5.13	Plot of Residuals versus Fitted Values	55
5.14	Percentage of Exacerbations versus Time	60
5.15	Smoothed Percentage of Exacerbations versus Time	61
5.16	Boxplots for Percentage of Exacerbations, Completers	65
5.17	Percentage of Exacerbations in Reduced Data Set	69
5.18	Smoothed Percentage of Exacerbations in Reduced Data Set	69
6.19	Boxplots for Rate of Change of REDSS	79
6.20	Rate of REDSS Change versus Baseline Covariates and Dropout	82

6.21 Treatment by Center Interaction Plot	83
6.22 Average Change in REDSS over Time	85
6.23 Smoothed Plot of Change in REDSS over Time	85
6.24 Quantile-Quantile Plots for Three Treatment Arms	94

Acknowledgement

I would like to thank my supervisor, Professor John Petkau, for his guidance throughout the development of this thesis. His hard work served as a very much needed inspiration for my efforts. The project he suggested was not only interesting and rewarding, but also remarkably useful for my present job.

I am also grateful to Michael Schulzer for his careful reading of the manuscript and his subsequent suggestions.

Most of all, I would like to thank my husband, Misha, for his continued help, patience and support.

Chapter 1

Introduction

1.1 Background to the Study

Multiple sclerosis (MS) is a common neurologic disease that is a major cause of disability, especially in young adults. Extensive research has failed to identify the possible causes of MS, but the present consensus is that damage to the nervous system results from immunologic processes. The disease is characterized by the appearance of scattered plaques or lesions in central nervous system (CNS). These lesions may slow down or completely block the transmission of impulses through the system.

The most common site from which symptoms are produced during an initial MS attack is the spinal cord. A common mode of onset of MS is a sensation of numbness or loss of feeling in the feet, ascending in the course of a few days to the waist. Rather more disabling is an initial attack in which the control over movement in a part of the body is lost. When this occurs it usually affects one of the upper limbs, resulting in loss of sense of position and of all the essential information from the muscles and joints. Such attacks make coordinated movements impossible. In approximately 15% of patients the initial symptom is optic neuritis, which typically is characterized by blurred vision and pain in one eye. Vision usually continues to decline for about a week, then the pain subsides and a week or two later, vision begins to improve. This recovery is a characteristic example of *remission* in MS; a term that means substantial or complete recovery from the effects of an initial attack or from a subsequent relapse of the disease (*exacerbation*). Double

vision that persists for several days is another common symptom of MS.

The symptoms of MS are rare in childhood. The frequency of onset of symptoms of MS begins to increase around the age of 17 and reaches a peak in early 30's. Thereafter the onset becomes increasingly uncommon but new cases continue to occur into the 60's. Women are more frequently affected than men, the ratio being close to three women to two men. The prevalence of MS varies markedly according to geography, and the most obvious factor concerned is distance from the equator. In tropical countries MS is either extremely rare or does not occur at all. In contrast, in north-west Europe, in Great Britain, in the northern states of the USA, in Canada, in southern Australia and in New Zealand prevalence is high. But MS is comparatively rare in Japan (though the severity is greater) and there are other anomalies showing that simple distance from the equator or some secondary effect of this, such as temperature, cannot be the only factors concerned. MS is, for example, said to be rare in Eskimos. There is also some evidence that MS is more common among the close relatives of those with the disease than in the general population. Two separate studies in Great Britain have shown that MS is relatively more common in those of higher social and economic standing.

One of the most remarkable facts about MS is the astonishing variability in its course and severity. The course of MS varies from that of an obviously grave disease to a mild benign form. In a moderately severe case, after an initial attack such as optic neuritis, numbness, weakness, double vision, or any of a great number of other less frequently encountered symptoms, a complete recovery follows within a few weeks. At this stage the disease becomes latent. After an interval from several months to several years new symptoms develop; and again a complete recovery follows. A further relapse, usually more severe, occurs within a year or two; but this time recovery is not complete – there are permanent residual symptoms and permanent slight disability. This pattern of successive exacerbations once or twice a year may persist for a further three or four years,

each time with less complete recovery and increasing residual symptoms. Eventually a stage of distressing disability, such as failure of normal bladder control is reached.

After the active stage of relapse and remission lasting for about five years the pattern of MS changes to more or less stable degree of disability condition. In the severe form of MS, increasing damage of CNS leads to bedridden and helpless state, and eventually to death from different kinds of infection. In another form of MS, its benign form, even after many relapses, remission is complete and no form of disability is ever developed. In 5 percent or less of cases the disease takes on a particularly severe form in which death may result within 5 years from the onset of initial symptoms of MS. Overall the average duration of life after the onset is at least 25 years.

The diagnosis of MS is extremely difficult, not only because of its relapsing and remitting nature, but also because other diseases with similar symptoms can easily be confused with MS [12]. Cranial magnetic resonance imaging (MRI) is a powerful procedure for diagnosing MS, for delineating its natural history and, potentially, as an objective quantitative outcome measure for assessing the response of MS patients to experimental therapy. In very early disease, multiple abnormalities in the CNS can often be seen with MRI, even if no corresponding symptoms are observed. However, these abnormalities are seen not only in MS but also in a number of other diseases that might be mistaken for MS.

Up to now no cure for MS has been found. Although many therapies have been used, until a few years ago none had been demonstrated to be efficacious in reducing the rate of exacerbations, the accumulation of disability, or the increase in lesion burden as judged by MRI. A few years ago, interferon beta-1b was offered as a treatment for MS [4]. Some of the data collected in this clinical trial, which was designed to test the effectiveness of interferon beta-1b in the treatment of MS, will be analyzed in this thesis.

1.2 Description of the Interferon Beta-1b Clinical Trial

The ultimate goal of any treatment is a complete cure of disease, but no treatment has been offered that could reverse all the effects of long-term disability caused by MS. The current treatments are aimed at the lesser goal of reducing disability, either by diminishing the severity, duration or frequency of exacerbations or by relieving the symptoms.

Any investigator wishing to examine an effect of treatment on MS is faced with the problem of evaluation of the effect. If, for example, it is decided to count the number of exacerbations, then one has to judge whether an increase in already existing weakness or numbness should also be counted as an exacerbation. Another problem is measuring the degree of disability. A number of schemes have been proposed for assessing abnormalities in strength, co-ordination, vision, bladder control, and many other functions and for combining these assessments into an overall score or disability rating. Unfortunately, the scoring methods are often either too sensitive, resulting in minute changes being recorded, or too coarse, so that mild but indisputable exacerbations hardly affect the score. All methods are highly susceptible to error or influence by the observer. The method most frequently used for rating disability in MS is the Kurtzke Expanded Disability Status Scale (EDSS) [6], which rates patients on a scale from 0 to 10 with half point steps (except for one step between 0 and 1), where 0 represents a healthy person and 10 corresponds to death from MS. MRI seems to offer a more objective method of assessment. But each lesion on MRI scan has to be identified, in order to be counted or measured. This brings discrepancy and subjectivity into the measurements. On the other hand, even MRI lesions frequently regress without treatment.

In spite of the difficulty of assessment, interferon beta-1b seemed to show an effect on MS. A double-blind, dose-finding pilot study in patients with MS showed that interferon beta-1b (IFNB) could be administered safely at a dose of 8 million international units

(MIU) every other day, and suggested that treatment decreased the risk of exacerbations. But as the usual rate of relapse is approximately once a year, at least two years were needed to demonstrate effectiveness.

This thesis presents a variety of analyses of data from a multicenter, randomized, double-blind, placebo-controlled study, with three parallel treatment groups (placebo, and 1.6 MIU and 8 MIU interferon beta-1b). The interferon beta-1b was manufactured by Chiron Corporation, Emeryville, CA, and supplied to the doctors as Betaseron by Berlex Laboratories, Richmond, CA.

Three hundred and seventy-two patients, all of whom had had MS for more than 1 year, were entered into the study in 11 different medical centers in the United States and Canada. All patients were between the ages of 18 and 50 years, were ambulatory with EDSS scores of 5.5 or less, and had had at least two acute exacerbations during the previous 2 years. All had been clinically stable for at least 30 days before entry and during this period had received no ACTH (adrenocorticotrophic hormone) or prednisone, medications which are used to speed recovery from relapse.

All personnel at each study site were blinded to treatment categories. Two physicians at each site were designated: one neurologist who was not aware of drug side effects to do the periodic examinations, and another "treating" neurologist who knew about side effects and injection site reactions, reviewed laboratory findings for toxicity and was responsible for overall patient care.

After randomization, each patient was instructed in self-administration of study medication. The first three injections were given at the study center under observation; thereafter, patients injected themselves subcutaneously every other day. After the first few months of more frequent visits, patients were evaluated every 12 weeks, or more often if symptoms occurred suggesting the possibility of MS exacerbations. Each evaluation included a standard neurologic examination, and a Kurtzke EDSS score was determined.

For all patients on study, the beginning and end dates of all exacerbations and EDSS scores obtained at each visit were recorded. Each patient had a baseline cranial MRI, and this was repeated yearly. A cohort of 52 patients at the University of British Columbia also had cranial MRIs repeated at 6-week intervals for the first 2 years. For this cohort the additional response variables were: the number of active (new, recurrent, or enlarging) lesions at each scan when the scan was compared to its immediate predecessor; the classification of each scan as active or not according to whether any active lesions were identified; and the burden of disease as measured by the total area of lesions on each scan. For more detail concerning description of the study, see the IFNB Multiple Sclerosis Study Group [4]. For details about the frequent MRI sub-study, see Paty, Li, the UBC MS/MRI Study Group, and the IFNB Multiple Sclerosis Study Group [17].

1.3 Outline of the Thesis

This thesis discusses several approaches to the analysis of longitudinal data, including binary, count and ordinal responses. Three distinct outcome variables correspond to these responses: the number of active lesions on each scan for the cohort in the frequent MRI UBC substudy; the presence of exacerbations for the full cohort of patients from the 11 different centers; and the EDSS scores for the full cohort of patients on study. The next chapter gives a brief description of the data sets used in the analyses.

Several methods were used to analyze each of the response variables. As a preliminary step, each of the variables was summarized over time for each patient. In particular, the counts of active lesions were summarized by the average number of active lesions per scan, the sequences of 0's and 1's characterizing whether an exacerbation began in successive 6-week periods were summarized by the percentage of periods in which exacerbations began, and the change in EDSS score from the baseline score was summarized by the average

change from baseline. For these summaries standard methods of statistical analysis, such as analysis of covariance, were used to assess treatment effects. Summarizing the data in this way results in the loss of part of the information available in the data; in particular, analyses of these summaries do not allow examination of possible patterns of change over time. These patterns may give a better understanding of treatment effects. Therefore the final conclusions about the treatment should be based on analyses which allow examination of patterns over time. Chapter 3 outlines the methodology used for such longitudinal analyses: the GEE approach, random effects regression models and random effects ordinal regression models. This chapter also presents the notation which is used throughout the thesis for precise formulation of the models used in the data analyses.

The GEE approach was used for the analysis of each of the response variables. It was utilized for the count data corresponding to the number of active lesions in Chapter 4, for the binary data corresponding to the patterns of exacerbations in Chapter 5, and as one of several approaches used for the ordinal responses corresponding to the EDSS scores in Chapter 6. To facilitate this last analysis, the EDSS scores were first transformed into a new set of scores which could more appropriately be treated as continuous data.

In addition to the GEE approach, two different approaches based on random effects models were used for the analysis of the ordinal responses in Chapter 6. Random effects regression models were used to analyze the transformed EDSS scores as continuous data and random effects ordinal regression models were employed to analyze the original EDSS scores. Because MIXOR, the program implementing the random effects ordinal regression models, had difficulty handling the many categories of the EDSS scores, these scores had to be collapsed into fewer categories for this analysis.

In Chapter 7, the thesis ends with the summary of the important conclusions about the IFNB treatment effect on the number of active lesions, probability of an exacerbation

beginning, and EDSS scores. The methods employed in the longitudinal analysis and their implementation and limitations are also discussed in this chapter.

Chapter 2

Description of the Data

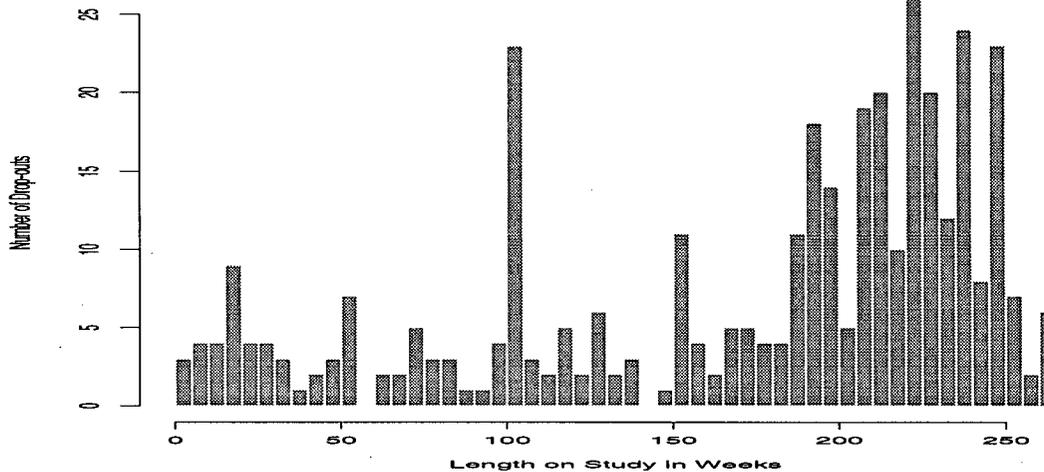
Overall, the study consisted of a total of 372 patients, randomized to either a placebo (PL), a low dose (LD) or a high dose (HD) arm. The patients were treated in 11 medical centers; the break-down of the number of patients by center is given in Table 2.1. The number of patients in different centers is roughly the same except for a high number of 52 and 48 patients in centers #259 (UBC) and #255 respectively, which will give these centers more weight in the analysis; and the very low number of 12 patients in center #266, where only 4 patients were randomized to each of the treatments. In each center, the patients divided almost evenly between the three treatment arms.

The initial plan for the study was to collect two years of data for each patient. When the majority of patients had reached the end of the second year (because of different starting dates some patients had been on study almost three years by that time), it was decided to extend the study to three years (except for frequent MRI scanning, which was terminated after two years). By the end of the third year of study therapeutic effect of interferon beta-1b became obvious and patients from all three treatment arms were

Table 2.1: Number of Patients in the Centers

Center ID	125	183	185	255	256	257	259	261	265	266	286
Total	30	30	30	48	36	24	52	45	30	12	35
PL patients	10	10	10	16	12	8	17	14	10	4	12
LD patients	10	10	10	16	12	8	18	15	10	4	12
HD patients	10	10	10	16	12	8	17	16	10	4	11

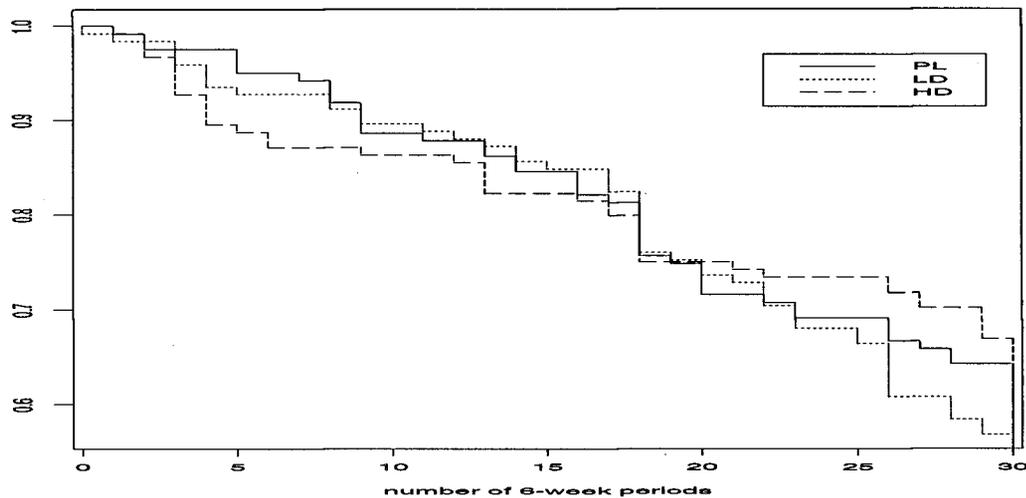
Figure 2.1: Histogram for Length on Study (5-week bars)



offered the opportunity to receive 8 MIU (the high dose) of interferon beta-1b for another two years.

The entire study continued for over 5 years, but very few patients participated that long. The histogram of the length on study for all patients in Figure 2.1 indicates a roughly constant rate of drop-out during the first three years, except for the large number of drop-outs at the end of the second year (initially intended as the end of the study). The rate of drop-out increased after the third year and it was decided to limit the analysis reported here to the data collected during the first 180 weeks (thirty 6-week periods) of the study. The Kaplan-Meier survival curves in Figure 2.2 indicate similar patterns of drop-out over time in all three arms; after the first 180 weeks about 64% of placebo patients, 57% of low dose patients and 69% of high dose patients remained on study.

Figure 2.2: Kaplan-Meier Survival Curves for Length on Study



2.1 Covariates

Several baseline covariates are available on the patients in the study. These include:

- age (in years),
- duration of disease (in years),
- initial EDSS score,
- center ID,
- gender.

The only time-varying covariate which will be investigated in the analyses is time itself. The primary focus of these analyses will be on the possible effects of treatment and the patterns over time, but the possible effects of these baseline covariates and length on study must also be considered.

Table 2.2: Descriptive Statistics for Age, Duration and Length on Study

Variable	Statistic	Placebo	Low Dose	High Dose
Age (years)	median	36.0	36.0	35.0
	mean	36.0	35.3	35.2
	SD	6.8	7.6	7.0
Duration (years)	median	5.6	5.9	7.1
	mean	7.7	8.2	8.3
	SD	6.4	6.5	5.7
Length on Study (if < 30) (6-week periods)	median	17.5	17.5	17.0
	mean	16.41	15.48	14.45
	SD	7.50	8.41	9.48
	patients	44	54	41

The median, mean and standard deviation for age and duration of disease for all study patients, and for length on study for those patients with less than 180 weeks on study are presented in Table 2.2; corresponding boxplots are provided in Figure 2.3. The distribution of age appears to be roughly the same across the three treatment groups. On average, the duration of disease is highest for the high dose arm, and lowest for the placebo arm. A number of outliers, high measurements in the range from 20 to 32 years, are indicated in the plots. But the differences in distribution for duration of disease between the treatment arms are rather small. Among the patients with less than 180 weeks on study, high dose patients tend to terminate earlier than placebo and low dose patients.

Initial EDSS and gender are summarized by counts in Table 2.3. Some lack of balance across the treatment groups is apparent within the different levels of initial EDSS. For example, 7 of the 13 patients with an initial EDSS score of 0.0 received placebo. Of course, with a relatively small number of patients at each particular level, such imbalances are not particularly surprising. There are more than twice as many females as males in the study, but the genders are balanced across the treatment groups. The low dose arm had

Figure 2.3: Boxplots for Age, Duration, and Length on Study

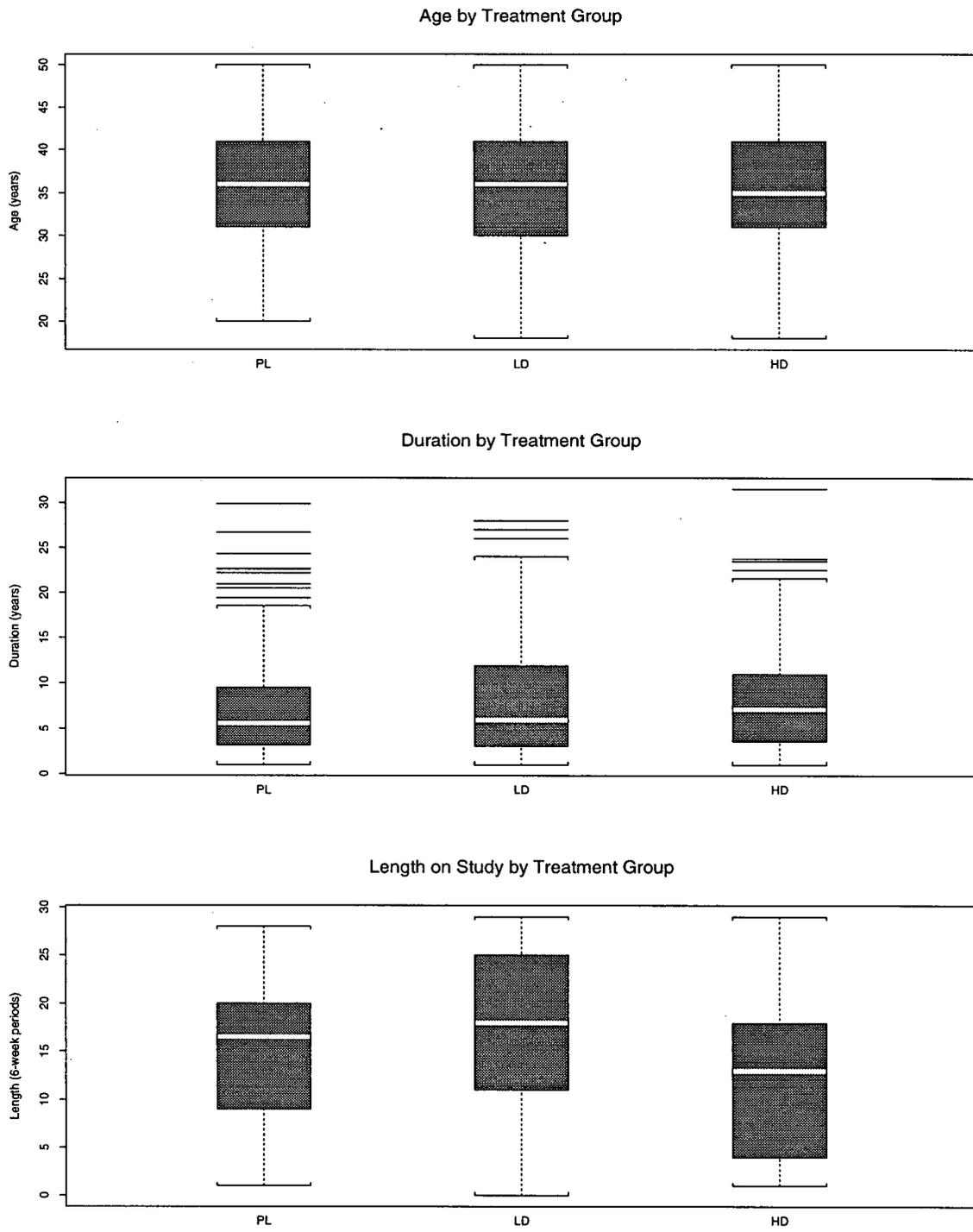


Table 2.3: Counts for Initial EDSS, Gender and Indicator of Drop-out

Variable	Value	Placebo	Low Dose	High Dose
Initial EDSS	0.0	7	4	2
	1.0	5	10	7
	1.5	13	10	11
	2.0	18	25	21
	2.5	15	10	14
	3.0	17	13	18
	3.5	23	18	15
	4.0	10	18	12
	4.5	4	5	7
	5.0	7	5	9
Gender	5.5	4	7	8
	Male	35	40	38
	Female	88	85	86
Indicator	≥ 30	79	71	83
	< 30	44	54	41

slightly more drop-outs during the first thirty 6-weeks periods than the placebo and high dose arms.

Although the covariates do not differ substantially across the treatment groups, even modest imbalances on covariates that are predictive of the response can have a substantial impact; see Senn [19]. Hence, the analyses will have to account for any possible relationships of these baseline covariates with the response variables.

Chapter 3

Methodology

3.1 Overview

Longitudinal data sets consist of repeated observations for each subject over time, and often a corresponding set of covariates for each of these subjects is also available. Typically the data sets on different subjects can be assumed to be independent of one another, but the repeated responses within each subject are usually correlated.

For the analysis of longitudinal data involving continuous responses, perhaps after some transformations, linear statistical models can be used. But classical methods, such as repeated measures ANOVA or multivariate ANOVA (MANOVA), are rather limited in their range of applicability, since they require a complete and balanced data set (where all subjects are observed at the same times) and homogeneous covariance structures for the responses. These conditions on the data structure can be relaxed, but then the analysis loses its main advantage – simplicity. So, while ANOVA methods are useful for continuous responses, they do not constitute a generally viable approach to longitudinal data analysis.

Laird and Ware [8] initiated a systematic study of an approach based on a general class of random effects models. Their development is based on a combination of empirical Bayes and maximum likelihood estimation (or restricted maximum likelihood estimation) of unknown parameters. This approach represented a major advance due to the greatly increased flexibility of the modelling, though the class of models is still limited to the

fitting of linear models for the mean response. In their development, the EM algorithm is used for iterative estimation of unknown parameters. Despite the fact that the EM algorithm can be painfully slow to converge even for the case of conditionally independent responses, this general approach has been quite widely used; see [1], [21], [7], and [22].

Lindstrom and Bates [10] provided improved computational procedures for random effects regression models based on Newton-Raphson estimation. Later [11] they developed generalizations to allow nonlinear models for the mean response, as well as associated software (*lme* and *nlme* in S-plus). Their models have flexible covariance structure, which allows for nonconstant correlation among the observations and unbalanced data. The S-plus *lme* function is used in the random effects regression analysis of the transformed EDSS score data in Chapter 6. But application of these methods is limited to longitudinal data involving continuous responses.

Although not as well developed as random effects models for continuous responses, an increasing amount of work has focused on random effects models for non-continuous response data. Such models have recently been proposed for both binary [20] and ordinal [3] outcomes. MIXOR (mixed-effects ordinal regression), a program for the analysis of repeated or clustered responses which are either binary or ordinal, was developed based on the work of Hedeker and Gibbons [3]. The program can accommodate multiple random effects, and allows for time-dependent covariates. This model is used for the analysis of the ordinal responses corresponding to the original EDSS scores in Chapter 6.

A different methodology has been developed based on modelling the marginal mean and covariance structure of the observations separately. One advantage of such methods is that they do not require specification of the joint likelihood of the data. Zeger, Liang, and Self [26] suggested one such approach for the analysis of binary longitudinal data where the marginal (rather than the conditional) probabilities are expressed as logistic functions of the covariates. They proposed the use of a “working likelihood”, rather

than the actual likelihood, for generating estimating equations for the regression parameters. Under rather weak assumptions concerning the true correlation structure for the repeated observations over time within each subject, these estimating equations lead to asymptotically valid inference.

The generalized estimating equations (GEE) approach developed by Zeger and Liang [25] and Liang and Zeger [9] applies equally well to binary, count and continuous response variables and provides asymptotically valid inference even when the correlation structure of the repeated observations is misspecified. Thus, the GEE approach provides a broadly applicable unified method for the analysis of longitudinal data. Along with other methods, the GEE approach will be used for longitudinal analyses of all the responses investigated in this thesis.

For the formulation of the statistical models the following notation is used throughout the thesis: index i labels the three treatment arms ($i = 0$ for PL, $i = 1$ for LD, and $i = 2$ for HD), j labels patients within the treatment arms ($j = 1, 2, \dots, n_i$, where $n_0 + n_1 + n_2 = N$ is the total number of patients), and t ($t = t_{j_1}, t_{j_2}, \dots, t_{j_{m_j}}$) labels the time of the observations. Note, that for the GEE approach, as well as for random effects modelling, the times at which subjects are observed are allowed to differ. Y_{ijt} , the response at time t for patient j on treatment arm i , will be modelled as a function of time and any potentially important covariates X_{ijt} . Initially all aspects of the model structure will be allowed to differ across the three treatment groups; model reductions to simplify the structure will subsequently be examined. As already indicated, the focus of the analyses will be on the effects of treatment and the patterns over time. Before proceeding with these analyses, brief descriptions of the details of the modelling approaches are provided in the remaining sections of this chapter.

3.2 GEE Approach

In the GEE approach, the regression of the response on explanatory variables is modelled separately from the within-person correlation. The average response over a subpopulation (such as a group of patients from one treatment arm) is modelled as a function of explanatory variables. Therefore, the treatment group indicators appear in the model as explanatory variables, and, for simplicity of notation, we can omit the index for the treatment groups and denote the response from the j th subject at time t as Y_{jt} , where $j = 1, 2, \dots, N$.

The GEE approach to handling the within-person correlation is based upon generalized linear models (GLM) for independent data. If there was only one time of observation for each subject, a GLM (see McCullagh and Nelder [13]) could be used to model continuous as well as discrete response variables. But for longitudinal data, each subject is observed repeatedly over time, and those observations are correlated. This correlation among the repeated observations for each subject must be taken into account in the analysis. Modelling the correlation is straightforward for Gaussian data, but there are few natural models for the joint distribution of the repeated observations over time for non-Gaussian responses. The GEE approach by-passes the difficulty of modelling the correlation for non-Gaussian responses by relying only on models for the marginal distributions.

If μ_{jt} and v_{jt} denote the marginal expectation and variance of Y_{jt} , then the expectation is assumed to depend on explanatory variables, x_{jt} , through the relationship $h(\mu_{jt}) = x_{jt}^T \beta$, where β is a $p \times 1$ vector of regression parameters and h is a known *link function* such as the logit for binary responses (such as an indicator variable for the beginning of an exacerbation) or the log for count responses (such as the number of active lesions). In addition, the marginal variance is assumed to depend on the expectation according

to $v_{jt} = \phi g(\mu_{jt})$, where g is a known function and ϕ is a scale parameter which may be unknown.

To incorporate the anticipated correlation over time between responses Y_{jk} and Y_{jl} for j th individual, $R_j(\alpha)$, a $m_j \times m_j$ "working correlation matrix" is specified for the response vector, $Y_j = (Y_{jt_1}, Y_{jt_2}, \dots, Y_{jt_{m_j}})^T$, of this individual. The working correlation matrices are assumed to be a function of an additional $s \times 1$ vector of unknown parameters, α . The working covariance matrix for Y_j is then

$$V_j = \phi A_j^{1/2} R_j(\alpha) A_j^{1/2},$$

where A_j is an $m_j \times m_j$ diagonal matrix with $g(\mu_{jt})$ as the t -th diagonal element. The working correlation matrices can differ across individuals.

The primary goal of the GEE approach is inference for the regression parameters β . The generalized estimating equations for β are given by

$$\sum_{j=1}^N P_j^T V_j^{-1} (y_j - \mu_j) = 0,$$

where $\mu_j = E(Y_j)$ and $P_j = \partial \mu_j / \partial \beta$ is a matrix of partial derivatives. These equations depend not only on β but, through V_j , also on the nuisance parameters α . The GEE estimate, $\hat{\beta}$, is obtained by iterating between solving the GEE's for β and consistently estimating α and ϕ using updated residuals.

Under mild regularity conditions, Liang and Zeger [9] show that $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically ($N \rightarrow \infty$) multivariate normal with zero mean and covariance matrix given by V , where

$$V = N \left(\sum_{j=1}^N P_j^T V_j^{-1} P_j \right)^{-1} \left[\sum_{j=1}^N P_j^T V_j^{-1} \text{Var}(Y_j) V_j^{-1} P_j \right] \left(\sum_{j=1}^N P_j^T V_j^{-1} P_j \right)^{-1};$$

here $\text{Var}(Y_j)$ is the true (unknown) covariance matrix for the response vector of the j th subject. Note that V depends upon the choice of $R_j(\alpha)$ as well as the true covariance matrix $\text{Var}(Y_j)$, but does not depend upon how α and ϕ are estimated (as

long as they are consistently estimated). A consistent estimate, \hat{V} , of the covariance matrix V can be obtained by replacing $\text{Var}(Y_j)$ by $(y_j - \hat{\mu}_j)(y_j - \hat{\mu}_j)^T$, where $\hat{\mu}_j = (h(x_{jt_1}^T \hat{\beta}), \dots, h(x_{jt_m}^T \hat{\beta}))^T$, β by $\hat{\beta}$, and α and ϕ by their consistent estimates. Details, particularly on consistent estimation of α and ϕ using the residuals, are provided by Liang and Zeger [9].

If the working correlation structure specified correctly, so that $V_j = \text{Var}(Y_j)$, the expression for V simplifies to

$$V_{naive} = N \left(\sum_{j=1}^N P_j^T V_j^{-1} P_j \right)^{-1},$$

and a consistent estimate \hat{V}_{naive} is readily available. This covariance matrix is referred to as the “naive” (or model-based) estimate of the covariance matrix of the regression coefficients since it is asymptotically valid only if the working correlation structure is equal to the true correlation structure. In contrast, because \hat{V} is asymptotically valid irrespective of the choice of working correlation structure, it is often referred to as the “robust” (or, due to its form, the “sandwich”) estimate of the covariance matrix.

An attractive property of the GEE approach is that it provides an asymptotically valid estimate of the covariance matrix of the regression parameters for any working correlation matrix used in the estimation, no matter what the true correlation structure of the data. The true correlation structure of longitudinal data is rarely known, but even when the simplest data structure, independence of the repeated measurements over time, is chosen as a working assumption, the GEE analysis leads to the asymptotically valid inferences. Nevertheless, choosing the working correlation structure to be close to the true structure increases the efficiency of the GEE procedure. The sensitivity of the inferences about β can be checked by fitting a model using different working correlation structures and comparing the different estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance model may be necessary [2].

The S-plus *gee* function was employed for the implementation of the GEE approach to the data analyses. This function allows a choice between the identity, logarithm, logit, reciprocal and probit link functions; Gaussian, Poisson, binomial or gamma variance functions; and identity, fixed, stationary-M-dependence, non-stationary-M-dependence, exchangeable, AR(M) and unstructured working correlation matrices. The logarithm link function and Poisson variance function were used to model the count response data corresponding to the number of active lesions. For the binary responses corresponding to whether a new exacerbation began during a 6-week period, the natural choices were the logit link and the binomial variance function. For the transformed EDSS scores, the identity link and Gaussian variance function were employed. Each of the independence, AR(1) and exchangeable working correlation structures were employed.

In the case of the identity link function, it is possible to formulate the models used in the GEE approach to have coefficients with the same interpretation as in the random effects regression models [2]. To compare the results obtained by applying these two models, the analysis of the re-expressed EDSS scores was repeated using the random effects regression model, which is described in the next section.

3.3 Random Effects Regression Models

The basic idea underlying a random effects model is that there is natural heterogeneity among subjects. This heterogeneity in the response is captured by the model in a subset of the regression coefficients. These regression coefficients are allowed to vary across individuals and this variability gives rise to correlation among repeated responses. Often correlation among observations for one individual arises from an unobservable or at least an unmeasured shared characteristic.

The random effects regression model for the j th individual's $m_j \times 1$ response vector

$y_j = (y_{jt_1}, \dots, y_{jt_{m_j}})$ can be written as:

$$y_j = X_j\beta + W_j\eta_j + e_j, \quad j = 1, \dots, N,$$

where X_j is a $m_j \times p$ design matrix for the fixed population effects, β is the $p \times 1$ vector of unknown regression parameters, W_j is a known $m_j \times r$ design matrix for the random individual effects, η_j is the $r \times 1$ vector of unknown individual effects, and e_j is the $m_j \times 1$ error vector. The distribution of the random effects is assumed to be multivariate normal with mean vector 0 and covariance matrix D and these effects are assumed to be mutually independent. The error vectors are assumed to be independent from the random effects, and to be mutually independently distributed as multivariate normal with mean vector 0 and covariance matrix $\sigma^2\Omega_j$. When the errors corresponding to the responses at different time points on the same individual are assumed to be independent, Ω_j is equal to the identity matrix. No restrictions are placed on the design matrices W_j and X_j . In particular, unbalanced designs and missing data are allowed.

Currently available software (for example the S-plus *lme* function or the MIXREG program by D. Hedeker and R.D. Gibbons) allows a variety of choices for the random effects variance-covariance structure, D , and for the correlation structure of the errors corresponding to the responses on the same individual, Ω_j . These choices include compound symmetry, first and second-order autoregressive, first and second-order moving average, first-order mixed autoregressive-moving average, and a general autocorrelation structure.

It follows from the model assumptions that, marginally, the y_j are independent multivariate normal with mean $X_j\beta$ and covariance matrix $\Sigma_j = \sigma^2\Omega_j + W_jDW_j^T$. To obtain the maximum marginal likelihood (MML) solution for the model parameters D , β , σ^2 , and Ω_j the marginal density of the data y_j , $h(y_j)$, has to be maximized. The MML solution is then derived by maximizing the log-marginal likelihood of the data from the

N subjects,

$$\log L = \sum_{j=1}^N \log[h(y_j)]$$

with respect to all parameters.

The EM solution proceeds by assigning starting values for the population parameters in order to estimate the individual parameters. These individual parameters are then used to obtain improved population parameter estimates. This process is repeated until convergence, which can be very slow. To improve the convergence properties, Lindstrom and Bates [10] suggested incorporating the Newton-Raphson algorithm for estimating the model parameters. Their suggestion is implemented in the S-plus *lme* function, which was used for the analysis of the re-expressed EDSS scores.

Whereas the GEE approach to longitudinal data analysis estimates only average parameters in a population, the random effects regression approach can also estimate individual parameters for each subject. This can be particularly useful in the medical setting where a proportion of subjects may respond to therapy in quite different ways from the average response.

The idea of the random effects linear regression models extends naturally to nonlinear regression models for continuous responses. Such models can be implemented through the S-plus *nlme* function. Another extension of this idea is used in combination with the idea of regression models for ordinal data in the random effects ordinal regression modelling approach described in the next section.

3.4 Random Effects Ordinal Regression Models

Following McCullagh's idea [14] of regression models for ordinal data, Hedeker and Gibbons [3] describe a random effects ordinal regression model, a statistical model for ordinal

responses in longitudinal study. In probit and logistic regression models it is often assumed that there is an unobservable latent variable which is related to the actual response through the “threshold concept” [13]. For dichotomous data, one threshold value is assumed, while for ordinal data with K ordered categories, a series of threshold values $\gamma_0 = -\infty, \gamma_1, \gamma_2, \dots, \gamma_{K-1}, \gamma_K = \infty$ is assumed. Without loss of generality, any value can be assigned to γ_1 ; for convenience, the common choice for γ_1 is 0. Then, a response occurs in category k ($Y = k$) if the latent response Z exceeds the threshold value γ_{k-1} , but does not exceed the threshold value γ_k .

The basic idea of the random effects regression model for an ordinal response is to use random effect regression to model the latent variable. Following the notation of the previous section, the random effects regression model for the j -th individual’s latent response vector z_j can be written as:

$$z_j = X_j\beta + W_j\eta_j + e_j, \quad j = 1, \dots, N.$$

With this model for the underlying latent variable Z_j , the log odds of the event $Y_{jt} \leq k$, conditional on β and η_j are modelled as

$$\log \left[\frac{P(Y_{jt} \leq k | \beta, \eta_j)}{P(Y_{jt} > k | \beta, \eta_j)} \right] = \frac{\gamma_k - \bar{z}_{jt}}{\sigma},$$

where $\bar{z}_{jt} = x_{jt}^T\beta + w_{jt}^T\eta_j$. Equivalently, conditional on β and η_j , the probability that $Y_{jt} = k$ (subject j ’s response at time t occurs in category k), is given by:

$$P(Y_{jt} = k | \beta, \eta_j) = \Psi[(\gamma_k - \bar{z}_{jt})/\sigma] - \Psi[(\gamma_{k-1} - \bar{z}_{jt})/\sigma],$$

where $\Psi(\cdot)$ represents the logistic function

$$\Psi(u) = \frac{1}{1 + \exp(-u)}.$$

Letting y_j denote the $m_j \times 1$ vector of ordinal responses on subject j , and assuming that the latent variables corresponding to different time points are conditionally independent, the probability of any pattern y_j , given β and η_j , is equal to the product of the

probabilities of the m_j responses:

$$L(y_j|\beta, \eta_j) = \prod_{t=1}^{m_j} \prod_{k=1}^K [\Psi(\gamma_k - \bar{z}_{jt}) - \Psi(\gamma_{k-1} - \bar{z}_{jt})]^{d_{jtk}},$$

where

$$d_{jtk} = \begin{cases} 1 & \text{if } Y_{jt} = k \\ 0 & \text{if } Y_{jt} \neq k \end{cases}$$

The marginal density of the $m_j \times 1$ vector of ordinal responses for subject j , y_j , in the population is given by the integral of the likelihood

$$h(y_j) = \int L(y_j|\beta, \eta_j) \varphi(\eta_j) d\eta_j,$$

where $\varphi(\eta_j)$ represents the multivariate normal distribution of η_j in the population.

For the estimation of the regression coefficients β and the threshold values γ_k ($k = 2, \dots, K - 1$), the marginal log-likelihood for the patterns from the N subjects is maximized.

MIXOR, a computer program for random effects ordinal, probit and logistic regression analysis, which was employed to carry out the analysis of the ordinal EDSS scores, uses Fisher's method of scoring to provide the solution to the maximum likelihood equations. On iteration m , the estimates for the vector of parameters Θ are improved by

$$\Theta_{m+1} = \Theta_m - E \left[\frac{\partial^2 \log L}{\partial \Theta_m \partial \Theta'_m} \right]^{-1} \frac{\partial \log L}{\partial \Theta_m}.$$

At convergence, the inverse of the information matrix provides the large-sample variances and covariances of the MML estimators which can be used to construct confidence intervals and tests of hypotheses.

To solve the likelihood equations, MIXOR performs numerical integration over the multidimensional parameter space for the random effects η_j . Gauss-Hermite quadrature is used to approximate the integrals to any desired degree of accuracy. In Gaussian

quadrature, the integration is approximated by a summation on a specified number of quadrature points (usually set to 10 or 20) for each dimension of the integration. As the number of random effects in the model for the latent variable increases, the number of quadrature points increases exponentially. Therefore specifying many random effects can prevent MIXOR from running. Specifying too many ordered outcome categories (18 original EDSS scores) or too many covariates increases the number of parameters to be estimated and can also cause problems. Collapsing some of the outcome categories and specifying fewer random effects may be necessary to fit the model.

Chapter 4

Number of Active Lesions

4.1 Description of UBC frequent MRI sub-study

As mentioned earlier, for each patient on study, in addition to the baseline scan, cranial MRI scans were obtained yearly. A cohort of 52 patients at the University of British Columbia also had cranial MRIs repeated at 6-week intervals for the first 2 years and provided a rich variety of information about changes in the CNS. The counts of number of active lesions on these 6-weekly scans for the UBC cohort will be the focus of the analyses in this chapter.

These 52 UBC patients were randomized to either a placebo (PL), a low dose (LD) or a high dose (HD) arm. Each patient was scheduled to have 17 scans in addition to the baseline scan, but several patients missed one or more of their scheduled MRI scans. Most of these missing scans resulted from patients dropping out of the study. Two of the total of 52 patients dropped out very early in the study; one LD patient (patient #423) dropped out between weeks 19 and 25 (so provided data only for the baseline plus 3 of the 17 additional visits) and the other HD patient (patient #447) dropped out between weeks 25 and 31 (so provided data only for the baseline plus 4 of the 17 additional visits). These two patients contribute a very small amount of data to the overall sub-study, so their data has been withheld from the analyses presented in this chapter. Thus, all analyses reported here are based on a total of only 50 (PL=17, LD=17, HD=16) patients.

Deviations from the target dates of this 6-weekly schedule occurred due to the practical difficulties of maintaining such a schedule. The most notable of these deviations was for the final scan which was delayed by a period of up to about two weeks for most patients. Three patients had even longer delays to their final scans (approximately 3, 4 and 5 weeks respectively) and two patients had final scans approximately 2 weeks before the target date according to the 6-weekly schedule. Most other deviations from the 6-weekly schedule were minor, though occasional more substantial deviations also occurred.

Although not essential for the methodological approach used, for the sake of simplicity, in the analyses presented here each scan was assigned to the target date of the 6-week period for which it was intended, irrespective of the deviation of the actual date of the scan from that target date. Thus, in these analyses, the “times” (since baseline) of the scans are taken to be same for all patients — as the 6-weekly schedule prescribed. These times will be referred to according to the corresponding 6-week period; thus, time 1 refers to the scan intended to be taken 6 weeks after the baseline scan time 2 refers to the scan intended to be taken 12 weeks after the baseline scan, . . . , time 17 refers to the scan intended to be taken at the end of the sub-study. Unless otherwise indicated, this definition of “time” will be utilized throughout this chapter.

Each MRI scan was compared to its immediate predecessor and the number of new, recurrent, or enlarging lesions was recorded. The *number of active lesions* results from summing the number of new, recurrent, and enlarging lesions at each scan. The derived sequence of counts for each of the 50 UBC patients is analyzed in this chapter.

4.1.1 Covariates

Several baseline covariates are available on the patients in the sub-study. These include:

- age (in years),
- duration of disease (in years),

Table 4.4: Descriptive Statistics for Age, Duration and Initial EDSS

Variable	Statistic	Placebo	Low Dose	High Dose
Age (years)	median	35.0	36.0	36.0
	mean	34.6	37.5	37.2
	SD	4.5	8.8	8.5
Duration (years)	median	6.9	6.1	10.1
	mean	7.4	9.5	11.0
	SD	5.0	7.2	6.4
Initial EDSS score	median	1.5	2.0	2.5
	mean	1.9	2.1	2.5
	SD	0.9	1.2	1.1

- initial EDSS score,
- origin (B.C. or Washington State),
- gender.

The only time-varying covariate which will be investigated in the analyses is time itself.

The median, mean and standard deviation for age, duration of disease and initial EDSS score (for the moment, ignoring the fact it is an ordinal variable) are presented in Table 4.4. These variables are also summarized by the boxplots in Figure 4.4. The distribution of age for the patients on low dose and high dose appears to have more spread than for those on placebo. However, the average age appears to be roughly comparable across the three treatment groups. The patients on high dose seem to have a longer duration of disease than the patients on placebo. The duration of disease for the patients on low dose is quite variable, covering the range of both the placebo and high dose groups. The initial EDSS score appears to be fairly similar within the three groups, although both the means and medians exhibit an increasing pattern with increasing dose. These covariates suggest that, on average, the placebo patients are slightly younger, have had the disease for a shorter period of time, and have lower initial EDSS scores.

Figure 4.4: Boxplots for Age, Duration and Initial EDSS

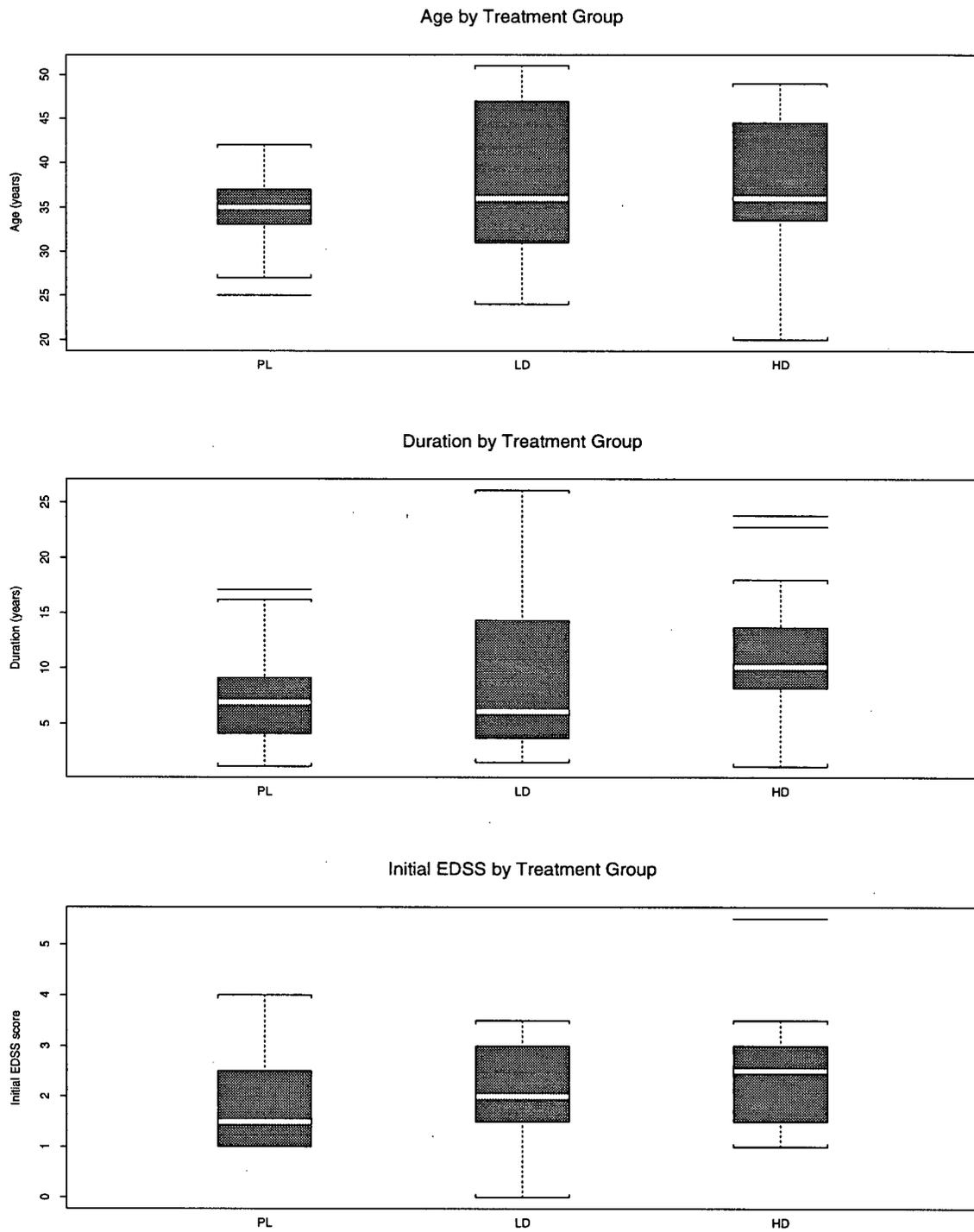


Table 4.5: Counts for Initial EDSS, Origin and Gender

Variable	Value	Placebo	Low Dose	High Dose
Initial EDSS	0.0	0	2	0
	1.0	6	1	1
	1.5	3	4	5
	2.0	3	2	0
	2.5	3	1	3
	3.0	0	3	4
	3.5	1	4	2
	4.0	1	0	0
	5.5	0	0	1
Origin	B.C.	15	11	14
	Washington	2	6	2
Gender	Male	7	2	3
	Female	10	15	13

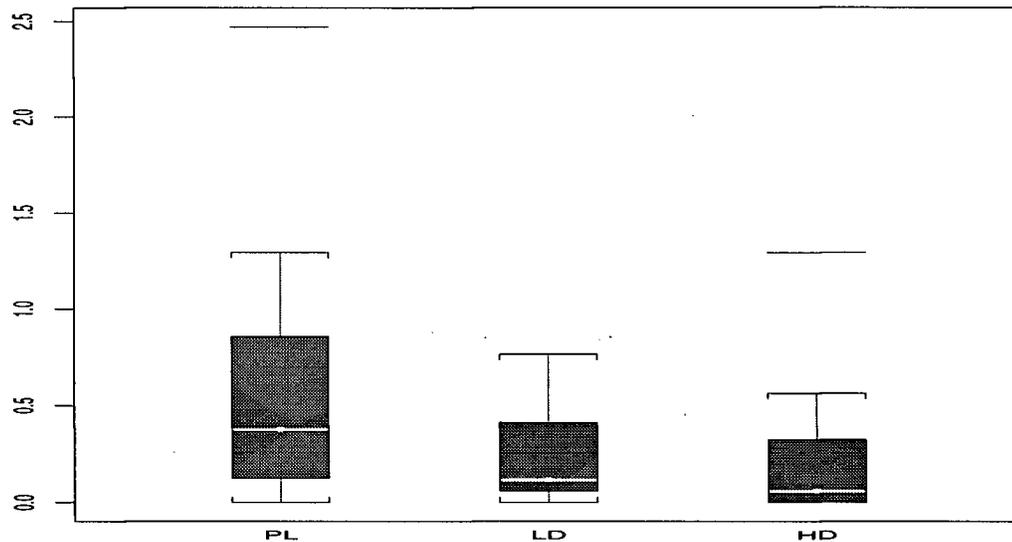
The categorical covariates are summarized by counts in Table 4.5. The most noticeable feature is the considerable lack of balance across the treatment groups within the different levels of these variables. For example, 6 of the 10 patients from Washington received low dose, 7 of the 12 males received placebo, and 6 of the 8 patients with an initial EDSS score of 1.0 received placebo. Of course, with a relatively small number of patients at each particular level, such imbalances are not particularly surprising.

The above descriptions provide a clear indication that these baseline covariates tend to differ across the treatment groups. Since they may have a substantial impact on the response variables (see Senn [19]), the analyses of the treatment groups effect has to be adjusted for all important (predictive) covariates.

4.2 Preliminary Analysis

As an initial step, the data for each patient was summarized over time by computing the average number of active lesions per scan over the period of the sub-study. For example,

Figure 4.5: Boxplots for Average Number of Active Lesions



if a patient had 12 scans subsequent to the baseline scan and had a total of 8 active lesions on these scans, then his average number of active lesions will be 0.333 lesions per scan. Boxplots of these summaries by treatment group, presented in Figure 4.5, clearly indicate that the average number of active lesions as well as their variability tends to be lower for the two treated groups. In addition, there is some suggestion of a dose response relationship. Corresponding descriptive statistics are presented in Table 4.6. Both the medians and the means suggest a difference between the two treated groups and placebo; the medians suggest a fairly clear dose response relationship, but the means are equal for the two treatment arms. The boxplots of the average number of active lesions indicate two outliers – a high value of 2.5 for one of the patients in placebo group and a value of 1.3 for a patient in the high dose group.

Table 4.6: Descriptive Statistics for Average Number of Active Lesions

Statistic	Placebo	Low Dose	High Dose
Number of patients	17	17	16
Median	0.44	0.12	0.06
Mean	0.58	0.22	0.22
SD	0.62	0.22	0.34

Table 4.7: P-values from ANOVA for Average Number of Active Lesions

Comparison	Ranks	Square Roots	Raw Averages
Overall	0.023	0.016	0.026
Placebo vs. High Dose	0.005	0.004	0.011
Placebo vs. Low Dose	0.027	0.015	0.010
Low Dose vs. High Dose	0.218	0.285	0.488

The ANOVAs on the ranks, the square roots (intended to stabilize the variance) and on the raw data of the average number of active lesions per scan resulted in the p-values listed in Table 4.7. Qualitatively the same conclusions result from the three analyses; both treatment arms appear to differ from the placebo arm, but there is no suggestion of a difference between the two treatment arms. Due to the effects of the outliers, the analysis on raw averages is less appropriate than those on the ranks or square roots of the average number of active lesions.

Next, the baseline covariates are examined as potential predictors. Plots of the response versus these covariates (see Figure 4.6) do not indicate any particularly strong relationships, although overall, the response appears to decrease somewhat with both age and duration of disease. The plots indicate the thirty year old female patient from BC with duration of disease of about 9 years and initial EDSS score of 1 as having the high average number of active lesions per scan in the placebo group and the twenty year old patient from BC with duration of disease of about 0 and initial EDSS score of 3.5 as

Figure 4.6: Average Number of Active Lesions versus Baseline Covariates

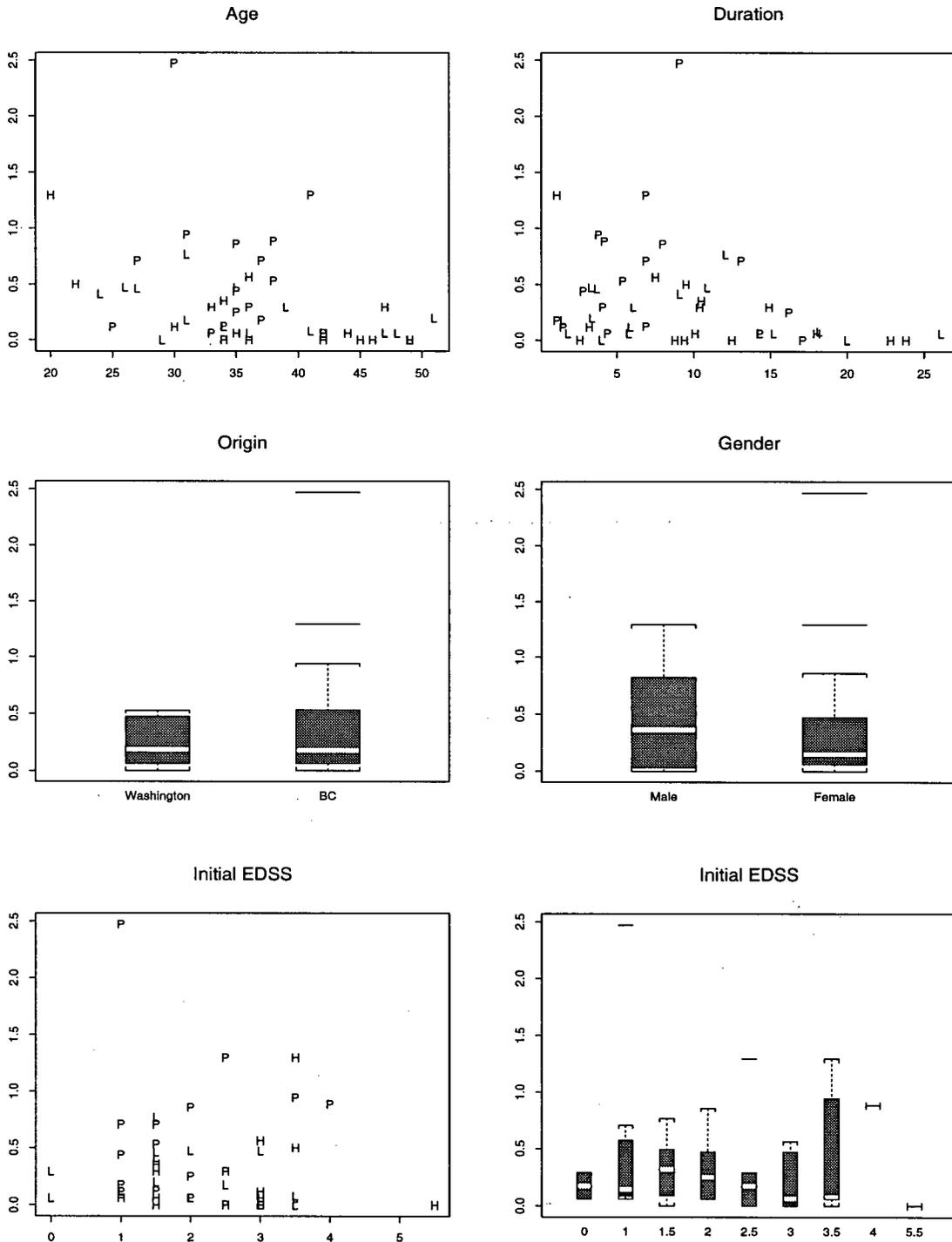


Table 4.8: Square Root of Response versus Baseline Covariates

Variable	df	F	p-value
Age	1,48	12.55	0.001
Duration	1,48	6.41	0.015
Origin	1,48	0.19	0.67
Gender	1,48	0.38	0.54
EDSS (continuous)	1,48	0.74	0.40
EDSS (ordinal)	8,41	0.74	0.65

having the high average number of active lesions per scan in the high dose group

The results of analyses, using the square roots of the average number of active lesions as the response variable and including the simplest possible effect for these covariates, are presented in Table 4.8. The p-values support a relationship with both age and duration. When a model that fits both age and duration simultaneously is considered then the importance of duration is considerably reduced (p-value = 0.22), but age continues to exhibit an effect (p-value = 0.011). These two predictors are positively correlated ($r = 0.44$), so it is not particularly surprising that only one of age and duration is important.

Keeping age in the model, effects for the treatment groups are added to evaluate their contribution after an adjustment for age has been incorporated; the overall p-value for treatment groups decreases (0.009 now, compared to the earlier 0.016). Qualitatively the same results are obtained using the average number of active lesions as the response (instead of square roots); this overall p-value decreases to 0.019 (from the earlier 0.026). The additional inclusion of duration as a covariate in the model for raw averages has very little effect on these results (p-value = 0.82) and therefore will not be incorporated in what follows.

Use of the raw data allows the magnitude of the effects of age and treatment group on the average number of active lesions to be meaningfully estimated (see Table 4.9)

Table 4.9: Age and Treatment Effects: Average Number of Active Lesions

Parameter	Estimate	SE	t	P($T > t $)
Intercept (α)	1.273	0.290		
Age (γ)	-0.020	0.008	-2.53	0.015
Low Dose (τ_1)	-0.305	0.141	-2.16	0.036
High Dose (τ_2)	-0.306	0.143	-2.14	0.037

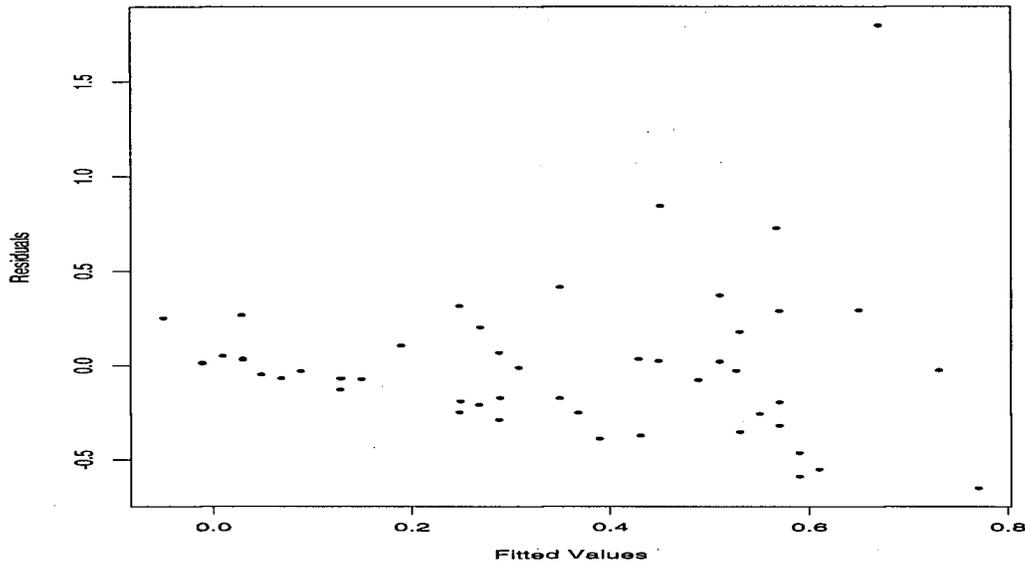
through the fitted ANCOVA model:

$$\hat{y}_{ij} = \hat{\alpha} + \hat{\tau}_i + \hat{\gamma}x_{ij},$$

where $i = 0$ (PL), 1 (LD), 2 (HD) indicates the treatment group, $j = 1, \dots, n_i$ indicates the patient within the treatment group, x_{ij} is the patient's age at the time of entry to the study, and \hat{y}_{ij} is the fitted average number of active lesions. (In fitting the model, τ_0 , the effect for the placebo group, is arbitrarily taken to be identically equal to 0.) The individual comparisons of low dose to placebo (p-value = 0.036) and high dose to placebo (p-value = 0.037) provided in Table 4.9 are quite similar to those in the original unadjusted analysis. The estimated effects indicate that, at any fixed age, low and high dose patients have, on average, about 0.3 fewer active lesions per scan than placebo patients. Of course, the results of the analysis should be interpreted with some caution since the presence of outliers in the placebo and high dose groups influences the estimated treatment groups effects and the estimate of the age effect as indicated by the plot of the average number of active lesions versus age (see Figure 4.6).

After a model is fit, the residuals, $y_{ij} - \hat{y}_{ij}$, are analyzed to help assess that fit. A plot of the residuals versus the fitted values (Figure 4.7) indicates that the spread of the residuals tends to increase as the fitted values increase. This is not too surprising because if each patient's data was uncorrelated over time then the total number of active lesions might be reasonably modelled as a Poisson response and the variance would increase

Figure 4.7: Plot of Residuals versus Fitted Values



with the mean level. Under the Poisson assumption the rate of the growth should be the same for mean and variance, but there is one extreme observation for a patient on the placebo arm, whose average number of active lesions is 2.5 (see Figure 4.6), and the model fit is poor for this observation. Additional plots (not included here) of the residuals versus treatment group, age and the other available baseline covariates, did not reveal obvious relationships, except for the outlier already noted. The variance of the residuals appears to be larger for the placebo group than the two treated groups. The residuals for the patients from B.C. also appear to more variable than the residuals for the patients from Washington. This may be related to the fact that only 2 of the 10 patients from Washington are in the placebo group whereas 15 of the 40 patients from B.C. are on placebo. Overall, the residual plots do not reveal any serious difficulties with

Table 4.10: Correlations Among Responses on Scans 1-6, 8, 11, 12

	1	2	3	4	5	6	8	11	12
1	1.00								
2	0.56	1.00							
3	0.31	0.31	1.00						
4	0.07	0.20	0.11	1.00					
5	0.55	0.59	0.10	0.37	1.00				
6	0.37	0.43	-0.01	0.39	0.73	1.00			
8	0.13	-0.06	-0.02	0.20	0.10	0.11	1.00		
11	0.33	0.52	0.07	0.25	0.58	0.64	0.13	1.00	
12	0.13	0.24	0.10	0.49	0.45	0.48	0.12	0.60	1.00

the fit, except for the outlier.

To assess the correlation structure in the original data, the 17 responses over time for each patient corresponding to the different scans are treated as separate response variables. The presence of a small amount of missing data (due either to early termination or to missing appointments for scans) complicates the calculation of correlations. The correlations among the responses from scans 1-6, 8, 11 and 12 (on which there was no missing data) are presented in Table 4.10. It is clear that the responses at different scans are positively correlated, in general, as would have been expected a priori. Further, although there are no obvious patterns in these correlations (e.g., m -dependence), it is clear that the responses on scans even quite widely separated in time can be moderately associated. Thus, the structure appears to be more exchangeable than autoregressive in character, for example. Clearly any analysis based on these original responses (rather than summaries over time, such as the average number of active lesions) must take this correlation into account.

Such analyses, using the GEE approach with Poisson regression, will be presented in the next subsection. To allow direct comparison to the GEE results, the analysis leading

Table 4.11: Effects for Age and Treatment Groups: Log Scale

Parameter	Estimate	SE	<i>z</i>
Intercept (α)	1.954	0.757	
Age (γ)	-0.074	0.022	-3.32
Low Dose (τ_1)	-0.898	0.355	-2.53
High Dose (τ_2)	-0.904	0.362	-2.50

to Table 4.9 was repeated using Poisson regression for each patient's total number of active lesions, with the log of the expected number of active lesions per scan expressed as the same linear model used to obtain the results in Table 6 (the fact that different patients had different numbers of scans is incorporated through the use of a fixed offset in the Poisson regression model).

This analysis yielded highly significant effects of the high and low doses as well as of age ($z = -5.9, -6.0, -7.9$ respectively). The Pearson X^2 statistic ($X^2 = 254$ with 46 df) indicated that the totals are overdispersed, so the results should be adjusted to allow for such overdispersion. The scale parameter is estimated as $\hat{\phi} = 5.64$ and the adjusted results are presented in Table 4.11. The significance of the age, low dose and high dose effects decreases after scaling; the z -scores are now qualitatively similar to those in Table 6. These estimated effects on the log scale indicate that, at any fixed age, the expected number of active lesions per scan for low and high dose patients is $\exp(\hat{\tau}_1) \approx \exp(\hat{\tau}_2) \approx 0.4$ times as large as for placebo patients (the corresponding approximate 95% confidence intervals contain all ratios from 0.20 to 0.82 for both groups). The conclusions to be drawn from this table are qualitatively the same as from Table 4.9, though the Poisson regression indicates a slightly higher significance of each of the effects.

Table 4.12: Effects for Age and Treatment Group: GEE Approach

Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation				
Intercept (α)	1.954	0.874	0.369	
Age (γ)	-0.074	0.024	0.011	-3.08
Low Dose (τ_1)	-0.898	0.328	0.173	-2.74
High Dose (τ_2)	-0.904	0.368	0.177	-2.45
AR(1) Working Correlation				
Intercept (α)	1.953	0.871	0.445	
Age (γ)	-0.073	0.024	0.013	-3.07
Low Dose (τ_1)	-0.918	0.326	0.210	-2.81
High Dose (τ_2)	-0.905	0.367	0.213	-2.47
Exchangeable Working Correlation				
Intercept (α)	1.952	0.851	0.718	
Age (γ)	-0.074	0.023	0.021	-3.18
Low Dose (τ_1)	-0.896	0.323	0.337	-2.77
High Dose (τ_2)	-0.907	0.362	0.344	-2.51

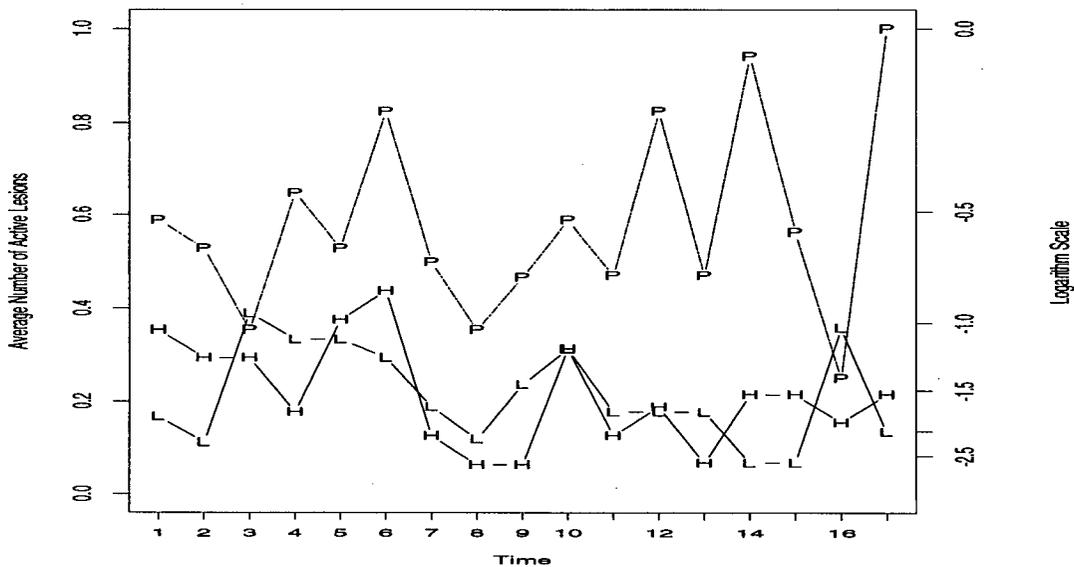
4.3 GEE Analysis

Since the activity data is count data, a GEE analysis based on Poisson regression is used to repeat the analysis reported in Table 4.11; here the fitted model is given by:

$$\log \hat{\mu}_{ijt} = \hat{\alpha} + \hat{\tau}_i + \hat{\gamma}x_{ij},$$

where μ_{ijt} is the expected count of active lesions for the j th patient in the i th treatment group at time t . Several working correlation structures are considered but as seen in Table 4.12, this has little effect on the conclusions; neither the parameter estimates nor the robust SE's are much affected. On the other hand, the naive SE's change quite considerably with the choice of working correlation (as should be expected, given the pattern of correlations in Table 4.10). Note that the robust and naive SE's agree most closely for the exchangeable correlation structure thus providing a further indication that

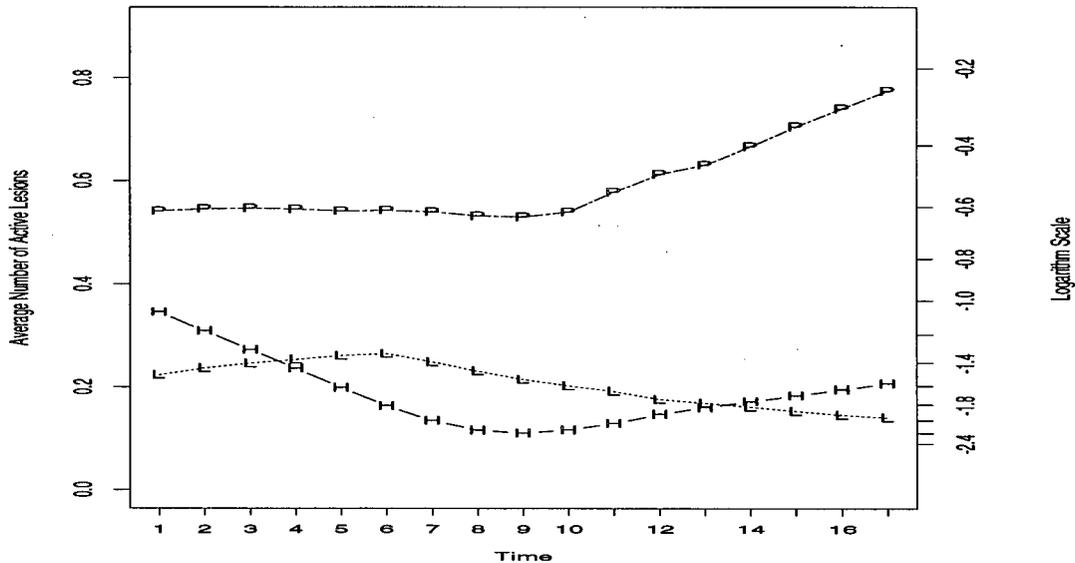
Figure 4.8: Plot of Average Number of Active Lesions versus Time



this may be a reasonable approximation to the true correlation structure in the data. The estimates of the effects obtained using the GEE approach for the independence working correlation structure are the same as those from the Poisson regression analysis on the total number of active lesions (see Table 4.11).

Since the GEE approach allows the data from each subject to be examined at each point in time, time itself can be included as a predictor; this will allow examination of the patterns over time. To examine plausible forms for these patterns, the average number of active lesions at each time of scanning for each treatment group is plotted versus time in Figure 4.8. The plot indicates that the number of active lesions tends to be higher, on average, in the placebo group than in either of the treated groups throughout the period of the sub-study but it does not suggest a strong relationship between the number of active lesions and time. However, this plot is quite noisy. Smoothing the plot using lowess (a robust, local smoother for scatterplot data) yields Figure 4.9 which suggests

Figure 4.9: Smoothed Plot of Average Number of Active Lesions versus Time



the expected number of active lesions per scan may increase over time for the placebo group and perhaps decrease slightly over time for the two treated groups.

Based on these plots, a model containing age, treatment group effects and separate linear time effects for each treatment group was fit; the estimated linear time effects are summarized in Table 4.13. The signs of the estimated time effects are as suggested by Figure 4.9 (positive for placebo and negative for the two treatment groups), but none of the effects is very large; in particular, none is significantly different from zero.

The correlation between any two of these estimated time effects is negligible for all three working correlation structures so the table provides the necessary inputs for pairwise comparisons of the estimated time effects. The separate linear time effects for the treatment groups are rather similar so it is reasonable to consider reduction of this model to one with a linear time effect which is common to the two treated groups. Comparison of $\hat{\beta}_1$ to $\hat{\beta}_2$ shows that this reduction is permissible (p-values for the test of $\beta_1 = \beta_2$

Table 4.13: Linear Effects for Time: GEE Approach

Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation				
Placebo (β_0)	0.017	0.020	0.021	0.84
Low Dose (β_1)	-0.035	0.032	0.028	-1.07
High Dose (β_2)	-0.046	0.044	0.032	-1.04
AR(1) Working Correlation				
Placebo (β_0)	0.020	0.023	0.027	0.88
Low Dose (β_1)	-0.035	0.032	0.029	-1.09
High Dose (β_2)	-0.046	0.046	0.035	-1.00
Exchangeable Working Correlation				
Placebo (β_0)	0.019	0.020	0.017	0.95
Low Dose (β_1)	-0.037	0.033	0.027	-1.11
High Dose (β_2)	-0.051	0.049	0.034	-1.03

are 0.42, 0.43 and 0.41 for the independence, AR(1) and exchangeable working correlations respectively). Fitting this reduced model to the data the estimated value of the common linear time effect for the two treated groups is -0.040, -0.041 and -0.038 with robust SE equal to 0.027, 0.028 and 0.028 for the independence, AR(1) and exchangeable working correlation structures respectively. The subsequent pairwise comparison of the linear time effects in the placebo and treatment arms leads to p-values of 0.051, 0.049 and 0.051 for the independence, AR(1) and exchangeable working correlation structures respectively, thus providing reasonable evidence of a difference between the linear time effects in the placebo and treatment arms.

On the other hand, since all the z-scores in Table 4.13 are small (less than 2 in magnitude), none of the linear time effects is significantly different from zero. This suggests that a model with a linear time effect which is common to all three treatment groups may provide an adequate representation of this data. This turns out to be the case; the overall test for the equality of these three linear time effects leads to a p-values

of 0.49, 0.48 and 0.45 for the independence, AR(1) and exchangeable working correlation structures respectively.

Given that the separate linear time effects are all negligible, the additional fact that the time effect for the placebo group is positive in Table 4.13 while those for the treated groups are negative suggests that such a common linear time effect will also be negligible. Indeed, this turns out to be the case; fitting models with a common linear time effect leads to estimates of the common linear time effect of -0.005, -0.003, and -0.006 and corresponding z-scores of -0.27, -0.15 and -0.32 for the independence, AR(1) and exchangeable working correlation structures respectively.

The smoothed plot of average number of active lesions per scan versus time (Figure 4.9) indicates slight curvature of the fitted curves for each group. A model incorporating separate quadratic trends over time led to small estimates for all the quadratic effects. Since the Wald test for simultaneous equality of all these effects to 0 resulted in p-values greater than 0.6 for the independence, AR(1) and exchangeable working correlation structures it is reasonable to exclude quadratic effects from the model.

4.4 Summary

Preliminary analyses not incorporating the trends over time indicate that both treatment arms differ from the placebo arm, but there is no difference between the two treatment arms. Age is identified as an important baseline covariate. The effects corresponding to interactions between the linear effect of age and treatment groups are negligible (p-value = 0.42 for the fit reported in Table 4.11); hence the linear effect of age can be taken to be common across the different treatment groups. The fact that the number of active lesions per scan tends to decrease with age (Table 4.9 indicates the mean number of active lesions per scan decreases by about 0.02 lesions per year of age, while Table 4.11

and Table 4.12 indicate that the mean number of active lesions per scan decreases by about 7.1% per year of age) may be of scientific interest.

Preliminary plots suggest there may be time trends in the number of active lesions per scan which differ across the treatment groups. Indeed, as Table 4.13 indicates, the estimated linear time effect is positive for the placebo arm but negative for the two treatment arms.

The data indicate that the linear time effects for the high and low dose treatment groups can be taken to be the same. If this model reduction is implemented, then there is reasonable evidence of a difference in the time trends for placebo and treated patients. The resulting model indicates that the number of active lesions per scan increases by about 2% every 6-week period on the study for the patients in placebo arm and decreases by roughly 4% every 6-week period for the patients in the two treatment arms (the corresponding approximate 95% confidence interval includes all values from -2% to 6% for placebo patients and from -9.5% to 1.5% for treated patients).

On the other hand, none of the linear time effects is significantly different from zero. This suggests an alternative (and preferable) simplification of the model to a linear time effect which is common to all three treatment groups. This leads to a negligible common linear time effect resulting in a reduced model with treatment and age effects, but no time trends. The estimated effects reported for the exchangeable working correlation structure in Table 4.12 indicate that the number of active lesions per scan for treated patients (either low dose or high dose) is only about 40% as large as for placebo patients (the corresponding approximate 95% confidence interval ranges from 21% to 77% for the low dose group and from 20% to 82% for the high dose group).

The simple analyses reported in Table 4.9 and Table 4.11 are reasonable for the preliminary analysis of the number of active lesions per scan. The results of these analysis may be considered as conclusive only if the number of active lesions per scan does not

change over time within each treatment group. But to verify this condition one has to carry out a longitudinal analysis such as GEE analysis reported here. Further, because the number of scans varies across patients in the sub-study, the final conclusions about the treatment effect should be based on the analysis summarized in Table 4.12 as only this analysis takes the detailed structure of the data directly into account.

Chapter 5

Exacerbations

5.1 Data Set Description

The beginning and end dates of any exacerbations (the appearance of a new symptom attributable to MS, or the worsening of an old symptom) the patients experienced during the study were recorded. So, for each patient, each time period can be classified as one in which an exacerbation either began or not. The derived sequence of 0's (no exacerbation began in this period) and 1's (an exacerbation began) can be used to characterize the pattern over time of *exacerbations* for each patient.

To match an earlier analysis of exacerbations for the UBC 6-weekly frequent MRI sub-study (Petkau and White [18]), exacerbations in successive 6-week periods throughout the study are considered in the analyses reported here. The same definition of "times" (since baseline) as for the analysis of number of active lesions is used for the analysis of exacerbations. Again, "times" are taken to be same for all patients and will be referred to according to the corresponding 6-week period; thus, time 0 refers to baseline, time 1 refers to the first 6-week period after baseline, time 2 refers to the second 6-week period after baseline, and so on.

Again, the primary focus of the following analyses will be on the effects of treatment and the patterns over time, but the possible effects of covariates will be considered. As seen from Figure 2.1 and Figure 2.2, a number of patients dropped out of the study very early. For these early drop-outs, the estimated rate of exacerbations can be very

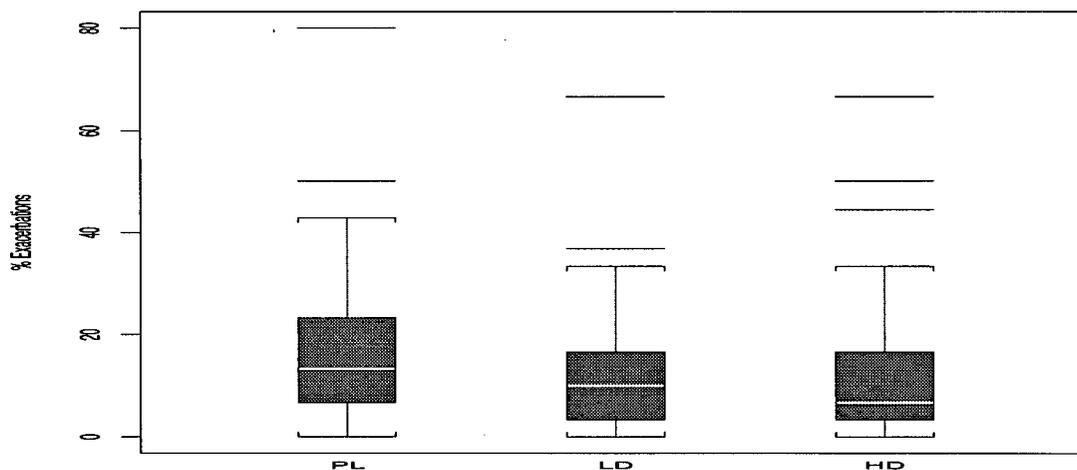
Table 5.14: Descriptive Statistics for Percentage of Exacerbations

Statistic	Placebo	Low Dose	High Dose
Number of patients	123	125	124
Median	13.33	10.00	6.67
Mean	15.57	11.44	10.85
SD	13.27	10.44	12.19

high if the patients had exacerbations during those few weeks on study, or the rate of exacerbations can be equal to 0 if they didn't have any exacerbations during those few weeks. Hence, a patient's length on study influences his rate of exacerbations and will be included in the following analysis as a covariate (in addition to the baseline covariates: age, duration of disease, initial EDSS score, center ID and gender, described in Chapter 1). Since the analysis of exacerbations is restricted to the data collected during the first thirty 6-week periods, each patient's length on study will be described by two variables: an indicator of whether a patient remained on study for less than thirty 6-week periods (this variable will be called *Dropout* throughout this chapter), and the number of 6-week periods on study for patients with less than thirty 6-week periods on study.

Some patients had two exacerbations during a 6-week period: there were 5 such periods in each of the treatment arms. In the high and low dose arms, 5 different patients had repeated exacerbations during a single 6-week period. In the placebo arm, one patient had two 6-week periods with repeated exacerbations and 3 other patients had a single such period. Since the number of periods with multiple exacerbations is balanced across treatment arms, for simplicity of analysis, we will treat these periods as if the patients had only one exacerbation beginning during these periods.

Figure 5.10: Boxplots for Percentage of Exacerbations



5.2 Preliminary Analysis

As an initial step, the data for each patient was summarized over time by computing the percentage of 6-week periods in which exacerbations began (hereafter referred to simply as percentage of exacerbations). Boxplots of these summaries by treatment group, presented in Figure 5.10, clearly indicate that the percentage of exacerbations tends to be lower for the two treated groups. In addition, there is some suggestion of a dose response relationship. Corresponding descriptive statistics are presented in Table 5.14. Both the medians and the means suggest a difference between the two treated groups and placebo and a fairly clear dose response relationship.

The boxplots of the percentage of exacerbations indicate several outliers in each treatment group. These include an extremely high value of 80% of exacerbations for one patient in the placebo group and a value of about 70% for one patient in each of the high and low dose groups. The next outliers correspond to 50% of exacerbations for patients

Table 5.15: P-values from ANOVA for Percentage of Exacerbations

Comparison	Ranks	Raw Percentages	Square Roots
Overall	0.0014	0.0039	0.0038
Placebo vs. High Dose	0.0005	0.0040	0.0018
Placebo vs. Low Dose	0.018	0.0069	0.014
Low Dose vs. High Dose	0.21	0.68	0.43

in the placebo and high dose group. Except for the patient in the low dose arm, these extremely high values correspond to early drop-outs from the study: the patient with 80% of exacerbations dropped out after about 30 weeks, the patient with 70% on the high dose arm after 36 weeks, 3 patients with 50% from the high dose arm after 12, 24 and 36 weeks respectively, and one patient with 50% from the placebo arm after the first 12 weeks. The correspondence of the high percentages of exacerbations and early terminations confirm the importance of including the length on study in the following analysis. Due to the potential effects of these outliers, an analysis on the raw percentages would be less appropriate than those on the ranks or square roots of the percentage of exacerbations.

An ANOVA on the ranks, on the square roots, and on the raw percentages of exacerbations resulted in p-values listed in Table 5.15. Qualitatively the same conclusions result from the three analyses: both treatment arms appear to differ from the placebo arm, but there is no suggestion of a difference between the two treatment arms.

Plots of the response versus the baseline covariates and Dropout are presented in Figure 5.11. The plot of percentage of exacerbations versus duration of disease indicates that the response appears to decrease with duration. The percentages for males look somewhat lower overall than the percentages for females. The boxplot for the percentage of exacerbations versus Dropout shows that, on average, the patients dropping out before

Figure 5.11: Percentage of Exacerbations versus Baseline Covariates

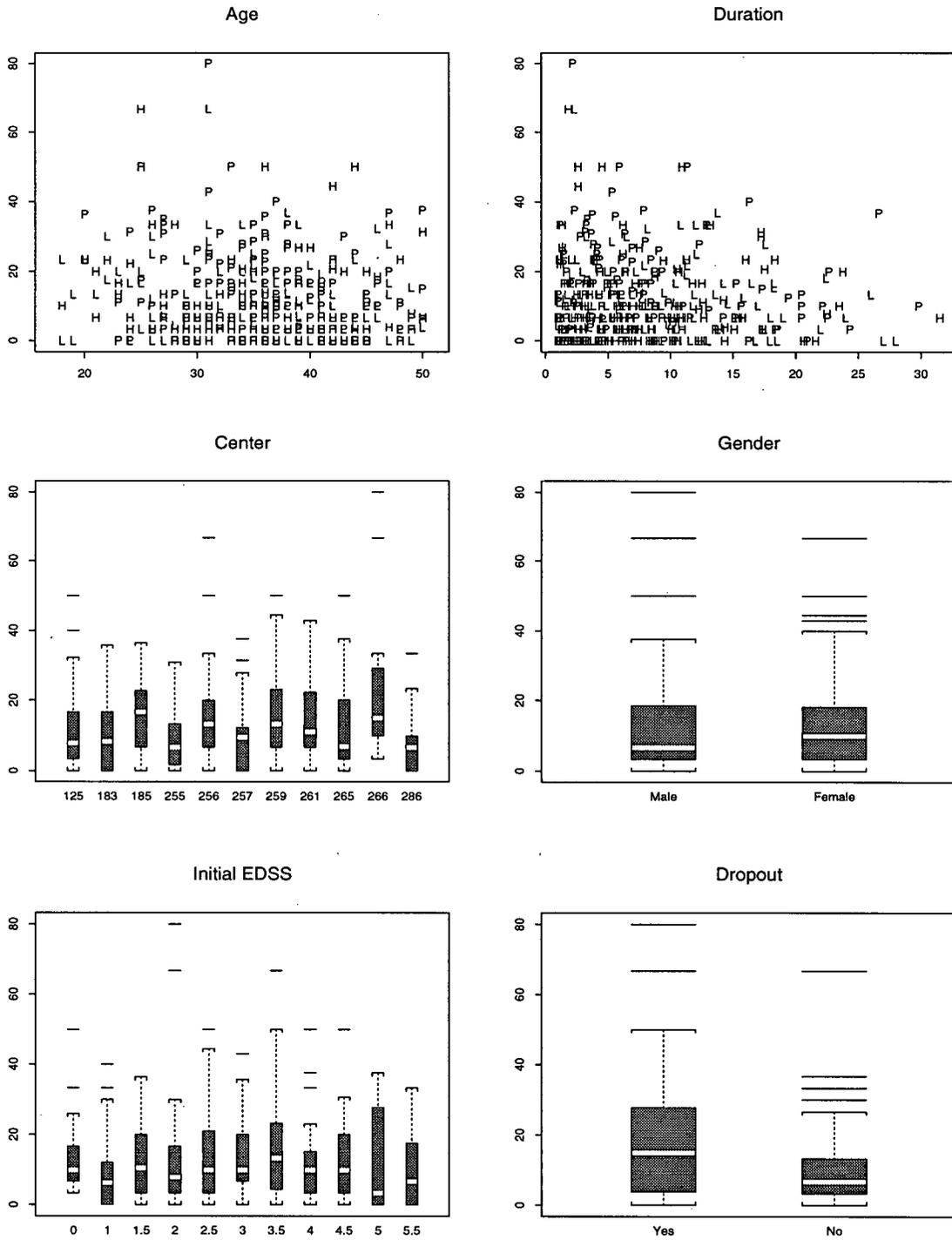


Table 5.16: ANCOVA of Ranked Response

Variable	df	F	p-value
Treatment Group	2	6.84	0.001
Age	1	5.95	0.015
Duration	1	0.20	0.66
Initial EDSS	10	1.29	0.23
Center	10	2.57	0.005
Gender	1	1.59	0.21
Dropout	1	18.41	< 0.001
Residuals	345		

thirty 6-week periods had higher percentages of exacerbations. The results of the analysis using the ranked summaries as the response variable and including all main effects are presented in Table 5.16. The p-values indicate strong treatment group effects, an even stronger effect of Dropout, as well as relationships with age and center.

In a model that fits treatment groups, age, center, Dropout and all two- and three-way interactions between these predictor variables, the strength of the main effects remains essentially unchanged. Among the interactions in the model, only the center by Dropout interaction is strong ($p = 0.03$), suggesting that centers differ in their rate of drop-out. Indeed, the display of the proportion of drop-outs by centers in Figure 5.12 confirms this suggestion. Excluding all interaction effects from the model does not change the conclusions about the main effects, hence a reduced model with only main effects for treatment, age, center and Dropout will provide a reasonable representation of this data.

Switching to the percentage of exacerbations as a response (instead of their ranks) and fitting the model with all main effects again indicates the presence of effects for treatment, age, center, and Dropout (see Table 5.17). Therefore these four variables will be included in the models which follow.

Figure 5.12: Proportion of Drop-outs by Center

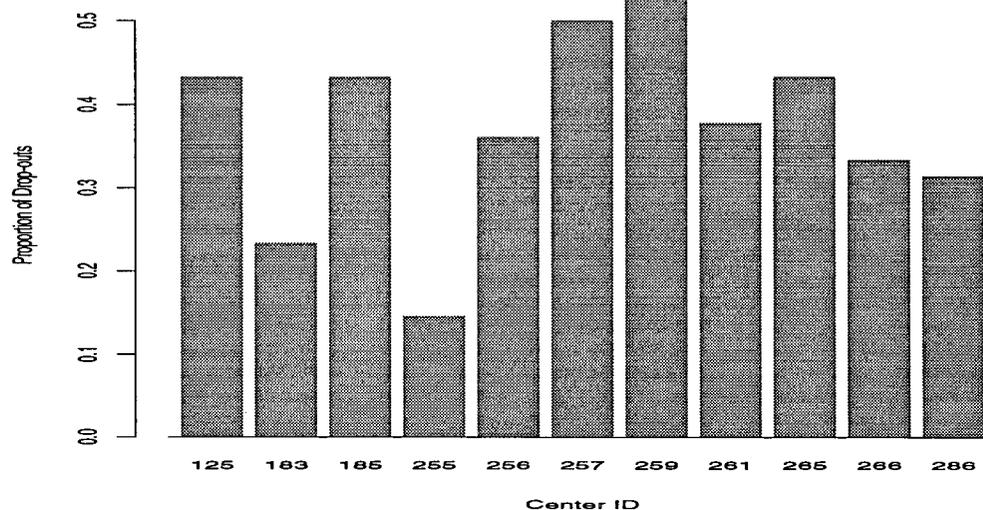


Table 5.17: ANCOVA on Percentage of Exacerbations

Variable	df	F	p-value
Treatment Group	2	6.58	0.002
Age	1	6.80	0.01
Duration	1	0.11	0.74
Initial EDSS	10	1.26	0.25
Center	10	2.88	0.002
Gender	1	0.02	0.90
Dropout	1	2.83	<0.001
Residuals	345		

Table 5.18: Effects Estimates: Percent Scale

Parameter	Estimate	SE	t	P($T > t $)
Intercept (α)	15.47	3.74	4.13	
Low Dose (τ_1)	-4.81	1.41	-3.41	0.0007
High Dose (τ_2)	-4.74	1.41	-3.36	0.0009
Age (β)	-0.20	0.08	-2.38	0.018
Dropout (δ)	7.19	1.23	5.83	<0.0001
125 (γ_1)	3.52	2.76	1.27	0.20
183 (γ_2)	3.24	2.76	1.17	0.24
185 (γ_3)	6.03	2.78	2.17	0.031
255 (γ_4)	2.65	2.48	1.07	0.28
256 (γ_5)	6.44	2.65	2.43	0.016
257 (γ_6)	0.57	2.96	0.19	0.85
259 (γ_7)	6.12	2.44	2.50	0.013
261 (γ_8)	6.49	2.51	2.59	0.010
265 (γ_9)	4.85	2.76	1.76	0.08
266 (γ_{10})	16.37	3.74	4.38	<0.0001

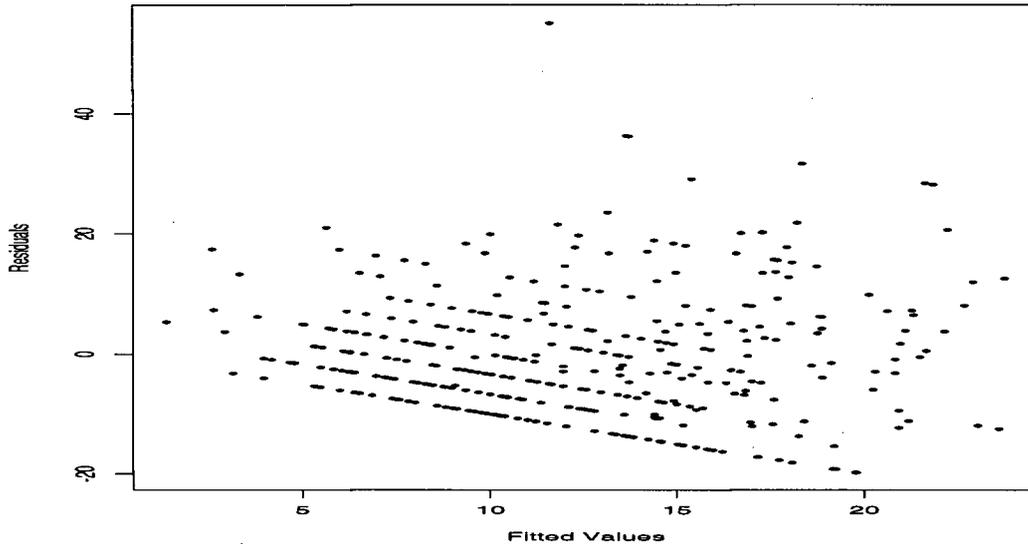
The fitted model can be written as:

$$\hat{y}_{ikj} = \hat{\alpha} + \hat{\tau}_i + \hat{\gamma}_k + \hat{\delta} * I_{ikj} + \hat{\beta} * z_{ikj},$$

where \hat{y}_{ikj} is the fitted percentage of exacerbations for the j th patient in the k th center on the i th treatment group, $\hat{\tau}_i$ is the fitted effect of the i th treatment, $\hat{\gamma}_k$ is the fitted effect of the k th center, I_{ikj} is an indicator of drop-out ($I_{ikj} = 1$ for patients with less than thirty 6-week periods on study), and z_{ikj} is the age at the time of entry to the study. The estimated effects and their standard errors are given in Table 5.18, where the effects corresponding to the placebo group and to center #286 were arbitrarily taken to be identically equal to 0 in this fitting.

The analysis shows that with respect to the percent of exacerbations the high and low dose arms are very different from the placebo arm, though the magnitudes of the high and low dose effects are very similar to each other. As indicated by the estimate

Figure 5.13: Plot of Residuals versus Fitted Values



$\hat{\delta} = 7.19$, the patients who dropped-out before the thirtieth 6-week period had on average much higher percentage of exacerbations (p-value < 0.0001) than patients who remained on study for at least thirty periods. Center #266, with only 12 patients, has the highest overall rate of exacerbations. This is partially due to the early drop-out of 4 patients, two of whom had very high rates of exacerbations (80% and 67%). Indeed, refitting the model when withholding those two patients from the analysis reduces the effect of center #266 to 7.92 (t-value = 2.08 and p-value = 0.04). At the same time the estimate for the Dropout effect $\hat{\delta}$ decreases from 7.19 to 6.29 and the p-value for the Dropout by center interaction effect increases from 0.03 to 0.04.

After a model is fit, the residuals, $y_{ikj} - \hat{y}_{ikj}$, are analyzed to help assess that fit. A plot of the residuals versus the fitted values (Figure 5.13) indicates that the spread of the residuals tends to increase somewhat as the fitted values increase. This is not very surprising because if each patient's data was uncorrelated over time then the percent

Table 5.19: Correlations Among First 10 Responses

	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	-0.05	1.00								
3	0.13	0.02	1.00							
4	0.03	0.04	-0.11	1.00						
5	0.23	0.03	0.14	0.01	1.00					
6	0.13	0.02	0.02	0.08	0.00	1.00				
7	0.06	0.01	0.01	0.07	0.06	-0.08	1.00			
8	0.21	0.03	0.16	-0.02	0.22	0.04	-0.10	1.00		
9	0.11	-0.01	-0.01	0.22	0.10	0.04	0.10	-0.02	1.00	
10	0.04	0.12	-0.04	0.05	0.07	0.06	-0.03	0.23	-0.04	1.00

of exacerbations would be a binomial response and the variance would vary with the mean level, taking on its maximum value when the percent of exacerbations was 50%. Additional plots (not included here) of the residuals versus treatment group, length on study and baseline covariates do not reveal obvious relationships. Nor do these residual plots reveal any serious difficulties with the fit.

To assess the correlation structure in the original binary data, the 30 responses over time for each patient corresponding to the successive 6-week periods are treated as separate response variables. The presence of a small amount of missing data (due to early termination) complicates the calculation of correlations. The correlations among the first 10 responses ignoring the data for the 44 patients with length on study of less than 10 6-week periods (14 on placebo, 13 on low dose and 17 on high dose) are presented in Table 5.19.

There are no obvious patterns in these correlations and most are quite modest, but some responses even quite widely separated in time are moderately associated, which suggests that an exchangeable correlation structure might be considered as the most appropriate working correlation structure for the data. Any analysis based on these original

Table 5.20: Estimates of Effects: Logit Scale

Parameter	Estimate	SE	t
Intercept (α)	-1.599	0.280	
Low Dose (τ_1)	-0.373	0.100	-3.72
High Dose (τ_2)	-0.300	0.101	-2.99
Age (β)	-0.015	0.006	-2.70
Dropout (δ)	0.790	0.088	8.96
125 (γ_1)	0.144	0.231	0.62
183 (γ_2)	0.095	0.246	0.38
185 (γ_3)	0.316	0.219	1.44
255 (γ_4)	0.073	0.230	0.32
256 (γ_5)	0.357	0.213	1.68
257 (γ_6)	-0.022	0.252	-0.09
259 (γ_7)	0.241	0.202	1.19
261 (γ_8)	0.381	0.205	1.86
265 (γ_9)	0.400	0.217	1.84
266 (γ_{10})	1.035	0.231	4.47

binary data (rather than summaries over time, such as the percentage of exacerbations) must take this correlation into account.

Such analyses, using the GEE approach with logistic regression, will be presented in the next subsection. To allow direct comparison to the GEE results, the analysis leading to Table 5.18 was repeated with the proportion of 6-week periods on study in which exacerbations began re-expressed on the logit scale. A minor difficulty arises because several patients had no exacerbations; the counts were: placebo = 16, low dose = 23, and high dose = 25. This difficulty was by-passed in the usual way through the use of empirical logits, with the data for each patient summarized as $\text{logit}[(x + c)/(m + 2c)]$, where m is the number of periods on study, x is the number of these in which an exacerbation began, and c is a fixed constant. The results for the common choice $c = 1/2$ are presented in Table 5.20. Analysis with another common choice of $c = 1/6$ gave qualitatively similar results.

The conclusions to be drawn from this table are qualitatively similar to those from Table 5.18, although re-expression on the logit scale has resulted in a slightly weaker high dose effect. The estimated treatment effects on the logit scale indicate that, at any fixed age, the odds of an exacerbation beginning for the low dose patients are only $\exp(\hat{\tau}_1) \approx 0.69$ times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.57 to 0.84), while the odds for the high dose patients are only $\exp(\hat{\tau}_2) \approx 0.74$ times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.61 to 0.90). The odds of an exacerbation beginning decrease by about 1.5% for every additional year of age at entry to the study. The patients in center #266 have the highest odds of beginning of exacerbations.

5.3 GEE Analysis

Since the original responses are binary, a GEE analysis based on logistic regression is used to repeat the analysis; the fitted model is given by:

$$\text{logit } \hat{\pi}_{ikjt} = \hat{\alpha} + \hat{\tau}_i + \hat{\gamma}_j + \hat{\delta} * I_{ikj} + \hat{\beta} * z_{ikj},$$

where π_{ikjt} is the probability of an exacerbation beginning in period t , for the j th patient in the k th center on the i th treatment group. In this analysis the effect for center #268 was again arbitrarily taken to be equal to 0, as was the placebo effect. To allow the fitting of the AR(1) correlation structure, patients with only one 6-week period on study were excluded from the analysis (3 patients were excluded: 1 from each of centers #256 (PL), #261 (LD) and #265 (HD)).

As mentioned earlier, for the sake of simplicity, in the analyses presented in this report the “times” (since baseline) are taken to be six week periods – as in the 6-weekly schedule prescribed for the UBC 6-weekly frequent MRI sub-study. In all models fit, time

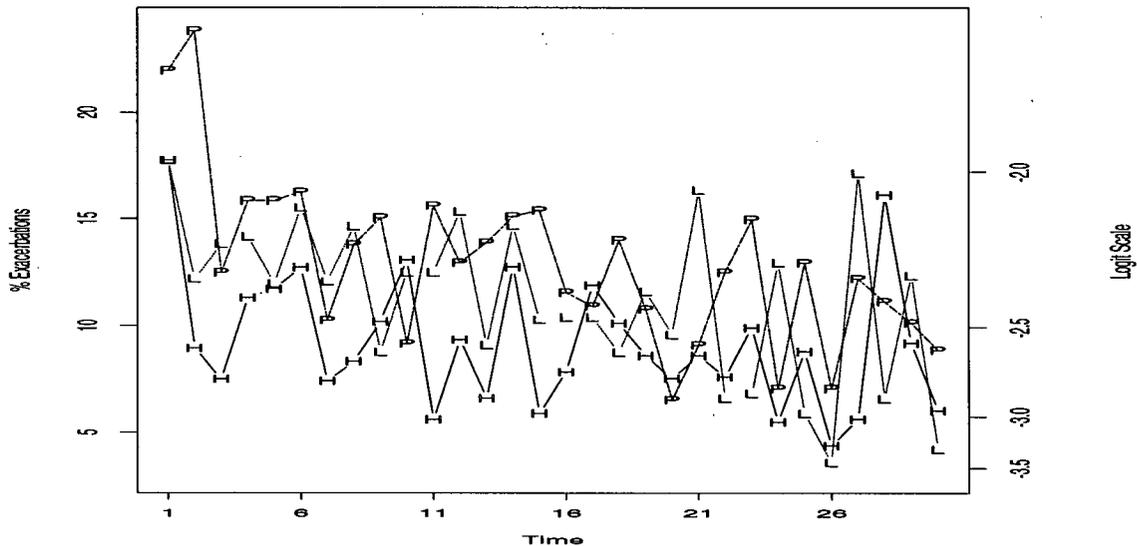
Table 5.21: Effects for Treatment Group, Age and Dropout: GEE

Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation				
Intercept (α)	-2.389	0.188	0.146	
Low Dose (τ_1)	-0.237	0.113	0.080	-2.10
High Dose (τ_2)	-0.408	0.118	0.084	-3.47
Age (β)	-0.013	0.007	0.005	-1.82
Dropout (δ)	0.534	0.104	0.077	5.14
AR(1) Working Correlation				
Intercept (α)	-2.388	0.188	0.140	
Low Dose (τ_1)	-0.236	0.113	0.077	-2.09
High Dose (τ_2)	-0.409	0.118	0.081	-3.48
Age (β)	-0.013	0.007	0.005	-1.80
Dropout (δ)	0.533	0.104	0.074	5.12
Exchangeable Working Correlation				
Intercept (α)	-1.987	0.326	0.310	
Low Dose (τ_1)	-0.274	0.112	0.106	-2.45
High Dose (τ_2)	-0.406	0.117	0.111	-3.48
Age (β)	-0.013	0.007	0.007	-1.95
Dropout (δ)	0.568	0.104	0.098	5.49

was incorporated through the centered variable $t = \text{time} - 15$ and age was incorporated through the centered variable $z_{ikj} - 30$.

Several working correlation structures are considered but, as seen in Table 5.21, this has little effect on the conclusions; neither the treatment parameter estimates nor the corresponding robust SE's are much affected. On the other hand, the naive SE's change somewhat with the choice of working correlation (as should be expected, given the pattern of mostly positive correlations in Table 5.19). Note that the robust and naive SE's agree most closely for the exchangeable correlation structure thus providing another indication that this may be a reasonable approximation to the true correlation structure in the data. The qualitative conclusions based on the GEE approach are quite similar to those

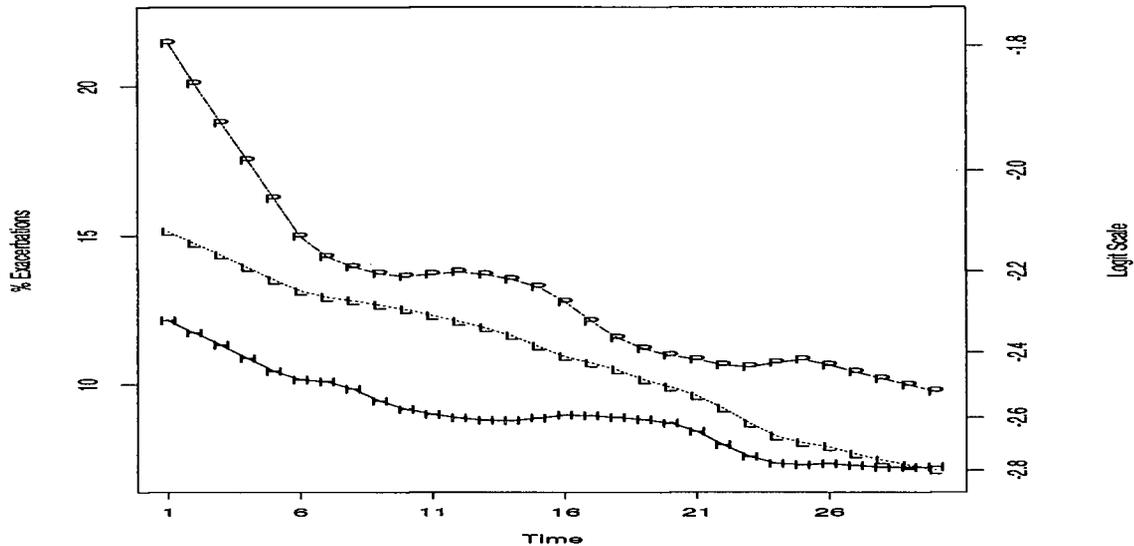
Figure 5.14: Percentage of Exacerbations versus Time



based on Table 5.20: there are strong treatment effects, as well as a modest age and a very strong Dropout effect on the probability of exacerbation. The low dose effect is weaker, but the high dose effect is stronger and the effect of age not nearly as striking in Table 5.21 as in Table 5.20.

Since the GEE approach allows the data from each subject to be examined at each point in time, time itself can be included as a predictor; this will allow examination of the patterns over time. To examine plausible forms for these patterns, the percentage of exacerbations in each period for each treatment group is plotted versus time in Figure 5.14. The plot clearly indicates that the percentage of exacerbations, on average, tends to decrease over time. However, this plot is quite noisy. Smoothing the plot using lowess (a robust, local smoother for scatterplot data) yields Figure 5.15 which suggests that the percentage of exacerbations, on average, is higher in the placebo group than in the treatment groups throughout the study, and that the high dose group has the lowest

Figure 5.15: Smoothed Percentage of Exacerbations versus Time



percentage at each point in time. The plot confirms the downwards trend which seems to be the strongest in the placebo group and the weakest in the high dose arm.

Based on these plots, a model containing treatment group, age, center, Dropout and separate linear time effects for each treatment group was fit. The signs of the estimated time effects are all negative as suggested by Figure 5.15. The time effects are strong for the placebo and low dose arms (the absolute values of all corresponding z-scores are greater than 2.3). The time effect for the high dose arm is weaker (the corresponding z-scores equal to -1.45, -1.41 and -1.27 for the independence, AR(1) and exchangeable working correlation structures respectively).

A model containing separate quadratic time trends for each treatment arm (in addition to the linear trend) was fitted next. Inclusion of the quadratic time trends in the model didn't change the significance of the linear time effects. All the estimated quadratic effects were positive as shown in Table 5.22, but none was strong. The Wald

Table 5.22: Linear and Quadratic Effects for Time: GEE

Trend	Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation					
Linear	placebo (ρ_0)	-0.019	0.006	0.007	-2.89
	low dose (ρ_1)	-0.017	0.007	0.007	-2.32
	high dose (ρ_2)	-0.010	0.007	0.007	-1.46
Quadratic	placebo (ϱ_0)	0.0012	0.0009	0.0008	1.42
	low dose (ϱ_1)	0.0004	0.0008	0.0009	0.45
	high dose (ϱ_2)	0.0011	0.0008	0.0009	1.36
AR(1) Working Correlation					
Linear	placebo (ρ_0)	-0.018	0.006	0.006	-2.87
	low dose (ρ_1)	-0.017	0.007	0.007	-2.28
	high dose (ρ_2)	-0.012	0.007	0.007	-1.42
Quadratic	placebo (ϱ_0)	0.0012	0.0009	0.0008	1.39
	low dose (ϱ_1)	0.0004	0.0008	0.0009	0.46
	high dose (ϱ_2)	0.0011	0.0008	0.0009	1.34
Exchangeable Working Correlation					
Linear	placebo (ρ_0)	-0.017	0.006	0.007	-2.70
	low dose (ρ_1)	-0.018	0.007	0.007	-2.51
	high dose (ρ_2)	-0.008	0.006	0.007	-1.28
Quadratic	placebo (ϱ_0)	0.0012	0.0009	0.0008	1.42
	low dose (ϱ_1)	0.0003	0.0008	0.0009	0.39
	high dose (ϱ_2)	0.0011	0.0008	0.0009	1.34

test for equality of these quadratic effects in the three treatment arms resulted in p-values larger than 0.7 for all three working correlation structures indicating no difference in the effect across the three arms. A model containing a common quadratic time trend resulted in an estimate of 0.0009, with a robust standard error of 0.0005 for all three correlation structures. This suggests that including a common quadratic time trend in the model is necessary.

The Wald test for equality of the linear time trends in the three treatment arms resulted in p-values close to 0.5 for all three working correlation structures indicating no

Table 5.23: Common Linear and Quadratic Effects for Time: GEE

Trend	Working Correlation	Estimate	Robust SE	Naive SE	z
Linear	Independence (ρ)	-0.015	0.004	0.004	-3.89
	AR(1) (ρ)	-0.015	0.004	0.004	-3.84
	Exchangeable (ρ)	-0.015	0.004	0.004	-3.77
Quadratic	Independence (ρ)	0.0009	0.0005	0.0005	1.91
	AR(1) (ρ)	0.0009	0.0005	0.0005	1.90
	Exchangeable (ρ)	0.0009	0.0005	0.0005	1.87

difference in these linear trends. This suggests that a model with a common linear time effect for all three treatment arms will provide an adequate representation of this data. Fitting models with common linear and quadratic time trends leads to the estimates in Table 5.23. Since the estimate of the standard error is smaller than in Table 5.21, this table indicates even stronger dependence of the probability of an exacerbation beginning on time. The estimates obtained for the treatment group, age and Dropout effects for this model with an exchangeable working correlation structure are given in Table 5.24. The p-values for these estimates are very close to the p-values from the model without time trends in Table 5.21.

The above analysis shows that at any particular time point LD and HD patients have, on average, lower odds of beginning an exacerbation than PL patients, but the data does not demonstrate evidence of differences in the patterns over time in the three treatment arms. According to the estimated Dropout effect, patients with less than thirty 6-week periods on study tend to have higher odds of beginning of exacerbation than patients who stayed on study longer. To examine how much the results of the above analyses change if only patients with at least thirty 6-week periods on study are considered, these analyses were repeated on the corresponding reduced data set.

Table 5.24: Effects for Treatment Group, Age and Dropout: GEE

Parameter	Estimate	Robust SE	Naive SE	z
Intercept (α)	-2.466	0.191	0.200	
Low Dose (τ_1)	-0.264	0.111	0.106	-2.10
High Dose (τ_2)	-0.411	0.116	0.111	-3.47
Age (β)	-0.013	0.007	0.007	-1.82
Dropout (δ)	0.489	0.104	0.100	5.14

5.4 Analysis on Reduced Data Set

The previous analysis indicated that the probability of beginning an exacerbation is closely related to the Dropout predictor variable. But including the length on study or Dropout as a covariate in the analysis of the treatment effects is somewhat questionable, because the length on study can itself be considered to be an outcome variable and could be used to assess the quality of the treatments. To avoid including Dropout we will repeat the previous analysis using only the data for the 233 patients who stayed on the program for at least thirty 6-weeks periods (call these patients *completers*).

The descriptive statistics for percentage of exacerbations for these patients are presented in Table 5.25. The number of completers in each of the treatment groups is roughly the same, with a few more drop-outs in the low dose arm. The medians for the low and high dose arms in the reduced data set are the same as in Table 5.14. The median for the placebo arm as well as the means for all three treatment arms are lower than in the full data set, suggesting that, on average, the patients who dropped out of the study before the 180-th week had higher percentage of exacerbations. Both the medians and the means for the placebo and low dose arms are now very similar; those for the high dose arm continue to be lower.

The boxplots for the percentage of exacerbations (Figure 5.16) indicate several outliers

Figure 5.16: Boxplots for Percentage of Exacerbations, Completers

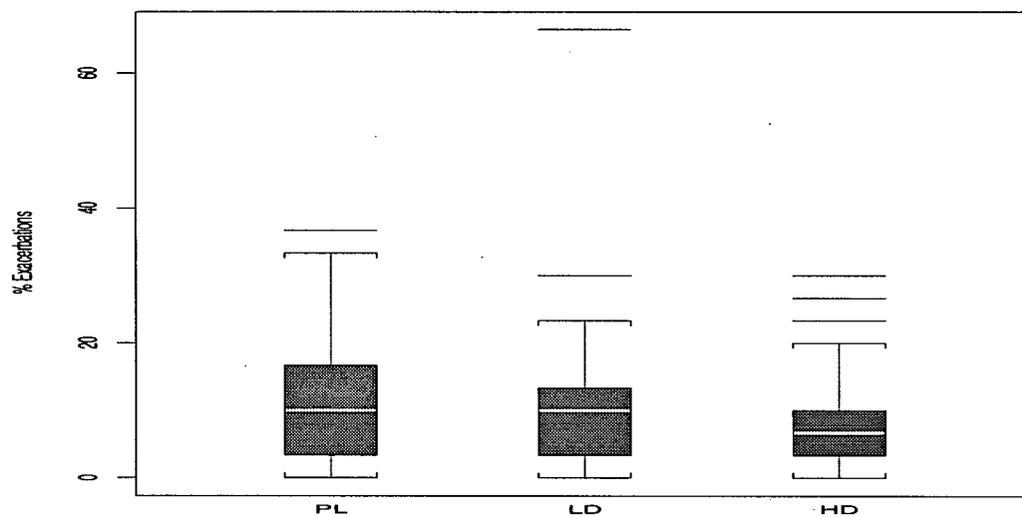


Table 5.25: Descriptive Statistics for Percentage of Exacerbations, Completers

Statistic	Placebo	Low Dose	High Dose
Number of drop-outs	44	54	41
Number of completers	79	71	83
Median	10.00	10.00	6.67
Mean	11.01	10.56	8.15
SD	8.82	9.90	7.35

Table 5.26: Number of Completers by Centers

Center ID	125	183	185	255	256	257	259	261	265	266	286
Total	17	23	17	41	23	12	23	28	17	8	24
PL patients	6	5	4	14	7	5	9	8	7	3	11
LD patients	6	9	4	14	7	1	6	9	6	3	6
HD patients	5	9	9	13	9	6	8	11	4	2	7

in each treatment group, with an extremely high value of 67% of exacerbations for a patient in the low dose arm. The patient is a 31 year old male from center #256 with 2.3 years history of MS and an initial EDSS score of 3.5.

Table 5.26 shows that the counts of completers differ considerably across the 11 centers, which gives different weights to the centers in the following analysis. The ANCOVA on the ranks of the percentages, restricted to the main effects of the baseline covariates and treatment group, is summarized in Table 5.27. The table suggests the presence of treatment group effects, a weak effect of the baseline EDSS, differences across centers and also between genders. The dependence on age is weaker in the reduced data set than in the full set. If a model also including all interactions of the baseline covariates with treatment group is fitted to the ranked response, then none of the interaction terms is shown to be important and the strength of the baseline covariate effects remain essentially unchanged. Dropping duration of disease from the model doesn't have a noticeable influence on the conclusions about the rest of the variables. Therefore, for further analysis the treatment effect will be adjusted for all baseline covariates, except for the duration of disease.

For comparison with the results obtained for the full data set, the GEE approach was used to fit the logit of the probability of exacerbation at each time point as a linear function of treatment group, age, initial EDSS, gender and center effects; the results

Table 5.27: ANCOVA on Ranked Response, Completers

Variable	df	F	p-value
Treatment Group	2	3.45	0.032
Age	1	0.97	0.33
Duration	1	0.00	0.98
Initial EDSS	10	1.68	0.086
Center	10	1.72	0.078
Gender	1	5.27	0.023
Residuals	207		

are presented in Table 5.28. The absolute values of the estimates for the low and high dose effect are smaller than the estimates for the full data set reported in Table 5.21. In contrast to the results for the full data set, only the high dose treatment effect is different from the placebo effect.

None of the estimated effects for the baseline covariates, except for four centers, is very large. The estimate of center #256 is the highest among all the centers. This high estimate is partially due to the presence at this center of the LD patient with 67% of exacerbations who is indicated as an outlier in Figure 5.16.

To examine plausible forms for the patterns over time in the reduced data set, the percentage of exacerbations in each 6-week period for each treatment arm is plotted versus time in Figure 5.17. The plot suggests a downward trend in the patterns. The smoothed version of the plot in Figure 5.18 indicates that at every time point the percentage of exacerbations in the high dose arm is lower, on average, than in two other arms. The patterns in the placebo and low dose arms seem to be quite similar. A model containing treatment group, age, gender, initial EDSS, center and separate linear time effects for each treatment group was fit to the reduced data set. The estimates for the linear trends in the PL and LD arms are almost the same as for the fit of the corresponding model (with Dropout instead of gender and initial EDSS effects) for the full data set, but the

Table 5.28: Effects of Treatment Group, Age and Gender for Completers: GEE

Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation				
Intercept (α)	-2.439	0.453	0.293	
Low Dose (τ_1)	-0.080	0.148	0.096	-0.54
High Dose (τ_2)	-0.384	0.148	0.098	-2.60
Age (β)	-0.008	0.009	0.006	-0.93
Gender	0.174	0.153	0.091	1.14
Initial EDSS	0.072	0.047	0.033	1.52
AR(1) Working Correlation				
Intercept (α)	-2.439	0.453	0.293	
Low Dose (τ_1)	-0.080	0.148	0.096	-0.54
High Dose (τ_2)	-0.384	0.148	0.098	-2.60
Age (β)	-0.008	0.009	0.006	-0.93
Gender	0.174	0.153	0.088	1.14
Initial EDSS	0.072	0.047	0.032	1.52
Exchangeable Working Correlation				
Intercept (α)	-2.439	0.454	0.442	
Low Dose (τ_1)	-0.079	0.148	0.144	-0.54
High Dose (τ_2)	-0.383	0.148	0.147	-2.59
Age (β)	-0.008	0.009	0.009	-0.94
Gender	0.174	0.153	0.133	1.14
Initial EDSS	0.072	0.047	0.048	1.52

Figure 5.17: Percentage of Exacerbations in Reduced Data Set

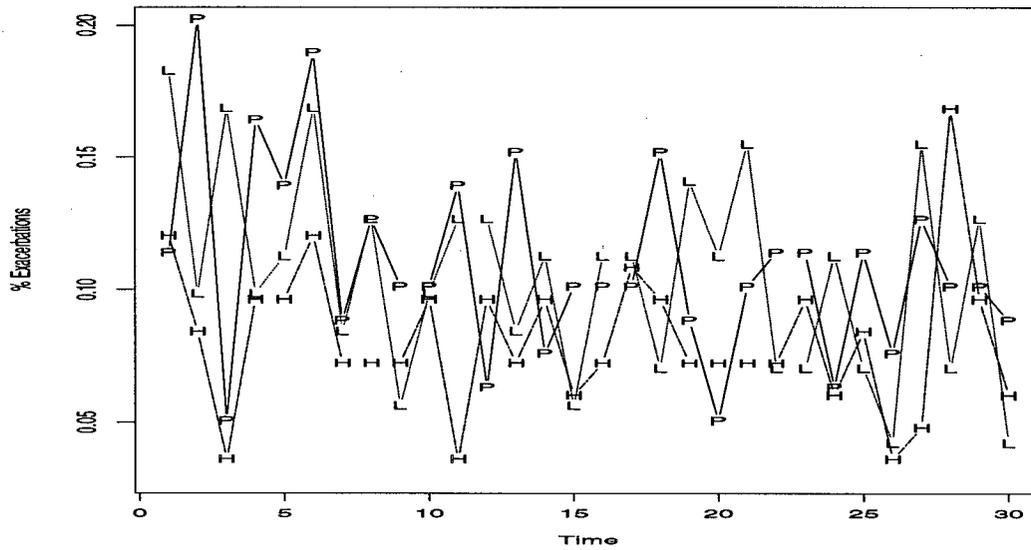
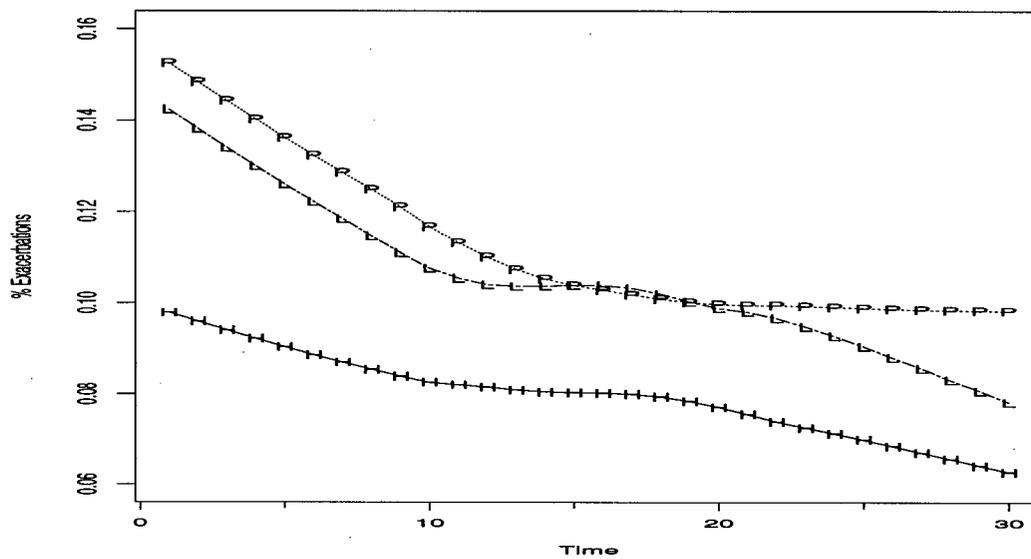


Figure 5.18: Smoothed Percentage of Exacerbations in Reduced Data Set



estimate for the linear trend in the HD arm is about half as large.

The smoothed plot for the percentage of exacerbation versus time in Figure 5.18 indicates slight curvature of the fitted curves, especially for the placebo group. The estimated linear and quadratic trends for the above model with separate quadratic time trends also included are summarized in Table 5.29. Including quadratic trends in the model didn't change the estimates for the linear effects and all the quadratic effects are small. The Wald test for the simultaneous equality of all quadratic effects to 0 resulted in p-values of 0.62 for all working correlation structures, so it is reasonable to exclude such quadratic effects from the model.

As suggested by Figure 5.18 the estimates of the linear time effect for the PL and LD arms are very close to each other, but somewhat different from the estimate for the HD. Nevertheless, as for the full data set, the reduced data set does not provide convincing evidence of differences between the HD linear time effect and those for the two other arms (the p-values are above 0.1 for the pairwise comparisons with the placebo and low dose). The Wald test for equality of the linear time effects resulted in p-values above 0.3 for all three working correlation structures; hence a reduced model with a common linear time effect seems adequate for this data set.

Fitting a model with treatment group, age, gender, initial EDSS, center and common linear time effects to the reduced data set resulted in the linear time effect estimates in Table 5.30. The estimates of the treatment group, age, gender and initial EDSS effects for this model are almost identical to the estimates in Table 5.28.

Consequently, the analysis on the reduced data set indicates that only the high dose treatment effect is different from placebo effect on completers; but the strength of this treatment effect is smaller than that for the full data set (z-score of -2.6 compared to the earlier of -3.5). The odds of exacerbations tends to decrease with time; the rate of decrease is about 1.3% per 6-week period (the corresponding approximate 95% confidence

Table 5.29: Linear and Quadratic Time Effects for Completers: GEE

Trend	Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation					
Linear	placebo (ρ_0)	-0.016	0.007	0.008	-2.23
	low dose (ρ_1)	-0.018	0.008	0.008	-2.17
	high dose (ρ_2)	-0.004	0.008	0.008	-0.46
Quadratic	placebo (ϱ_0)	0.0008	0.0010	0.0010	0.79
	low dose (ϱ_1)	0.0006	0.0010	0.0011	0.60
	high dose (ϱ_2)	0.0008	0.0009	0.0011	0.87
AR(1) Working Correlation					
Linear	placebo (ρ_0)	-0.016	0.007	0.007	-2.24
	low dose (ρ_1)	-0.018	0.008	0.008	-2.16
	high dose (ρ_2)	-0.003	0.008	0.008	-0.44
Quadratic	placebo (ϱ_0)	0.0008	0.0010	0.0010	0.80
	low dose (ϱ_1)	0.0006	0.0010	0.0010	0.60
	high dose (ϱ_2)	0.0008	0.0009	0.0011	0.87
Exchangeable Working Correlation					
Linear	placebo (ρ_0)	-0.016	0.007	0.007	-2.23
	low dose (ρ_1)	-0.018	0.008	0.008	-2.17
	high dose (ρ_2)	-0.004	0.008	0.008	-0.46
Quadratic	placebo (ϱ_0)	0.0008	0.0010	0.0010	0.79
	low dose (ϱ_1)	0.0006	0.0010	0.0010	0.60
	high dose (ϱ_2)	0.0008	0.0009	0.0011	0.87

Table 5.30: Common Linear Time Effect for Completers: GEE

Working Correlation	Estimate	Robust SE	Naive SE	z
Independence	-0.013	0.005	0.005	-2.74
AR(1)	-0.012	0.005	0.005	-2.71
Exchangeable	-0.013	0.005	0.005	-2.73

interval is from 0.3% to 2.2%) for all treatment arms.

5.5 Summary

As preliminary plots suggest, the GEE analysis on the full data set shows that there is a difference between placebo and treatment arms; the difference is overwhelming for the high dose treatment arm, but rather weak for the low dose arm. The estimated effects for the exchangeable working correlation structure reported in Table 5.24 indicate that the odds of an exacerbation beginning for the low dose patients are only 0.76 times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.61 to 0.95), while the odds for the high dose patients are only 0.66 times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.53 to 0.83).

On the other hand, for the patients who stayed on study for at least 180 weeks, there is little low dose effect. From the results for the exchangeable working correlation structure in Table 5.28 the odds of an exacerbation beginning for the low dose patients are estimated to be 0.92 times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.69 to 1.23) while the odds for the high dose patients are only 0.68 times as large as for placebo patients (the corresponding approximate 95% confidence interval contains all ratios from 0.51 to 0.92)

Smoothed plots of percentages of exacerbation versus time (Figure 5.15) suggest a decreasing trend over time in each treatment group. The GEE analysis on the full data set (see Table 5.22) shows that the trends are very strong in the placebo and low dose groups, and somewhat weaker in the high dose group. Even though the time trend for the high dose arm seems to be somewhat different from the time trends in the placebo and low dose arm, the data does not provide enough evidence to reject the equality of

the linear time effects. In addition to the linear time effect, the model identifies the presence of quadratic time trends. Even though the curve for the LD in Figure 5.15 is less bended than the PL and HD curves and the corresponding estimate is smaller, the data does not provide enough evidence to reject the equality of the quadratic time effects. In addition to the linear time trend, a common quadratic trend is detected in the pattern of exacerbations, suggesting that the rate of exacerbation decreases faster at the beginning of the trial and slower towards the end of the 3-year period as indicated by Figure 5.14 and Figure 5.15.

As for the full data set, there is a decreasing trend in the rate of exacerbations over time for patients with at least 180 weeks on study. Again, the trends are similar for the placebo and low dose arms, and somewhat smaller for the high dose arm. The smoothed plots of percentage of exacerbations versus time in Figure 5.18 indicate slight curvature of the curve for the placebo arm. But the estimates of all the quadratic effects were negligible in the reduced data set. The estimated common linear time effect is weaker than that for the full data set: the odds of exacerbation beginning decrease by about 1.3% every 6-week period (the corresponding approximate 95% confidence interval based on the results for the exchangeable working correlation structure contains all percentage reductions from 0.3% to 2.3%).

Center is identified as an important baseline covariate, with the average odds of exacerbation beginning for patients in center #266 2.3 times as large as for patients in center #268. Length on study is shown to be very important. As indicated by the GEE analysis (Table 5.21), the odds of an exacerbation beginning for the patients with less than thirty 6-week periods on study are estimated to be 1.7 times as large as for the patients remaining on study for at least thirty 6-week periods (the corresponding approximate 95% confidence interval based on the results for the exchangeable working correlation structure contains all ratios from 1.4 to 2.1). In addition, the comparison

of GEE results for the full and reduced data sets indicated that the low and high dose treatment effects are lower and linear time trend is weaker for the patients with at least 180 weeks on study.

Finally, all analyses presented in this chapter agree that, in comparison with the PL arm, there is an instantaneous reduction in the rate of exacerbations in the HD arm. On the other hand, the GEE analyses of the full and reduced data sets (since, among all the methods presented in this chapter, only the GEE approach allows the time trends to be incorporated) show that there is no difference in the time trends across the treatment arms. In addition to the common negative linear trend, a common quadratic trend identified for both data sets. The positive estimate for the common quadratic trend indicates that the odds of an exacerbation beginning decreases faster at the earlier phase of the trial than in the later phase.

Chapter 6

EDSS Scores

6.1 Data Set Description

The initial schedule was to evaluate the EDSS two weeks before the beginning of treatment (call this score *baseline* or *initial EDSS*), 6 weeks and 12 weeks after the beginning of treatment, and every 12 weeks thereafter for the first two years. After two years it was decided to continue the study. The average interval between the EDSS scores at the end of the second year was only 8 weeks; subsequent EDSS evaluations were again 12 weeks apart.

In addition to these scheduled evaluations, EDSS scores were also recorded at every non-scheduled visit, such as those caused by the worsening of patient's condition. The EDSS scores recorded when patients had exacerbations are often much higher than those assessed during exacerbation-free periods. These EDSS scores are omitted from the analysis, because our objective is to compare patients in their remission (stable condition) state.

The entire study continued for over five years and patients who completed the entire five years of the study had 23 scheduled EDSS evaluations; but due to drop-out the number of evaluations is lower for the majority of patients. The histogram of the length on study in Figure 2.1 indicates a roughly constant rate of drop-out during the first three years (up to the 14-th scheduled EDSS evaluation). The rate of drop-out increased after the third year of study and it was decided to limit the analysis reported here to the data

Table 6.31: Break-down of Patients by Center

Center ID	125	183	185	255	256	257	259	261	265	266	286
Total	28	29	26	47	34	22	51	42	21	9	31
PL patients	10	9	8	15	10	7	17	12	9	3	12
LD patients	10	10	9	16	12	7	18	14	8	4	10
HD patients	8	10	9	16	12	8	16	16	4	2	9

collected during the first three years: to the first 14 scheduled EDSS evaluations (including the baseline EDSS and the evaluation 6 weeks after the beginning of treatment). The data for 32 patients who contributed less than 5 EDSS evaluations to the data set (11 in the placebo arm, 7 in the low dose arm and 14 in the high dose arm) were also excluded from the analysis. The analyses which follow are based on data for 340 patients: 112 in the placebo arm, 118 in the low dose arm and 110 in the high dose arm. The break-down of the number of patients by center is given in Table 6.31. The number of patients in different centers are roughly the same, except for the very low number of 9 patients in center #266.

Deviations from the target dates of EDSS evaluations occurred due to the difficulties of maintaining the schedule. For the sake of simplicity, in the analysis to follow, all EDSS scores from scheduled visits were assigned to the corresponding target dates. Therefore in these analyses, the “times” (since baseline) of the EDSS evaluations the same for all patients. These times will be referred to according to the corresponding 12-week period; thus, time 0.5 refers to the EDSS score intended to be evaluated 6 weeks after the beginning of treatment, time 1 refers to the score intended to be evaluated 12 weeks after the beginning of the treatment, time 2 refers to the score intended to be evaluated 24 weeks after the beginning of the treatment, . . . , time 14 refers to the score intended to be evaluated at the end of the third year. Even though the average interval between the

EDSS scores at the end of the second year was only 8 weeks, for simplicity, we consider the time between these scores to be equal to 1. Unless otherwise indicated, this definition of “time” will be utilized throughout this chapter.

As in the analysis of exacerbations, we stratify the patients into two categories according to their length on study. For this stratification we again use a *Dropout* variable, which will be equal to 1 for the patients with less than thirteen periods on study (twelve and a half 12-week intervals, since the first period is only six weeks long) and equal to 0 for the patients with at least thirteen periods on study.

The EDSS score is an attempt to express the multiple facets of the neurologic examination as a single score. The lowest EDSS score of 0 corresponds to a normal neurologic exam; the next lowest value is 1; then scores follow 0.5 units apart up to 10, which corresponds to death due to MS (9 is the highest score in our data). The scores are ordered according to increasing deterioration, but the extent of deterioration from the score of 0.0 to the score of 1.0 may differ from the extent of deterioration from the score of 5.0 to the score of 6.0. Later in this chapter the EDSS categories will be used directly for the mixed effects ordinal regression analysis. But as a first step, we will use simpler analyses, for which we need to re-express the original scores to introduce a meaningful distance between them.

For each original EDSS score, we can regard the relative frequency of that score as a slice out of some distribution. The logistic distribution is most commonly used, because explicit formulas for re-expression are then readily available. We assign each score the numerical value of the corresponding center of gravity for the logistic distribution: every new score is calculated by the formula:

$$\text{New Score} = \frac{P * \ln(P) - (1 - P) * \ln(1 - P) - [p * \ln(p) - (1 - p) * \ln(1 - p)]}{P - p},$$

where p is a fraction of scores beyond the original EDSS score, and P is a fraction of

Table 6.32: Original and Re-expressed EDSS Scores

Original score	0	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
Re-expressed	-4.12	-2.49	-1.63	-0.89	-0.34	0.05	0.50	0.96	1.27
Original score	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0
Re-expressed	1.54	1.81	2.29	3.20	4.26	5.24	6.31	7.50	8.77

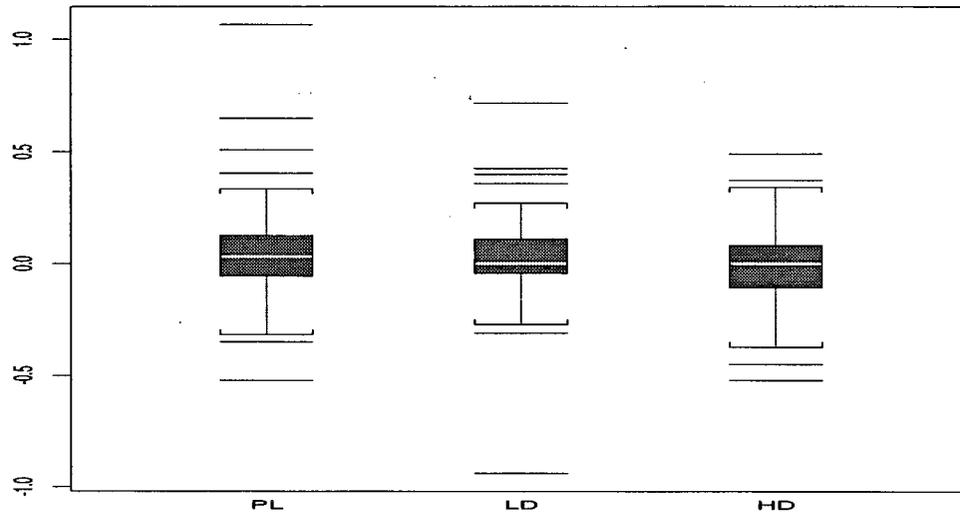
scores beyond and including the score [16]. The scores from all 14 evaluations for the 340 patients were used for the re-expression. The re-expressed EDSS scores (we'll call them REDSS) corresponding to the original scores for this data set are listed in Table 6.32. Comparing adjacent REDSS scores reveals that the distances between these scores vary: the distances between adjacent scores on the low and high ends of the scale are larger than the distances between adjacent central scores. For example, the distance between the REDSS scores corresponding to the EDSS scores of 0.0 and 1.0 is about twice as large as the distance between the REDSS scores corresponding to 5.0 and 6.0.

6.2 Preliminary Data Analysis

At the time of entry to the study all patients had EDSS scores of 5.5 or less and, as shown in Table 2.3, the three treatment arms are roughly balanced with respect to each baseline EDSS score. It is not meaningful to subtract the arbitrary scores attached to ordinal categories, but, unless otherwise indicated, we will be using REDSS as the response in the analysis to follow. The comparison of the treatment arms will be based on the change in REDSS from baseline.

As an initial step, the data for each patient was summarized over time by the average rate of change in REDSS over a 12-week period. This average rate of change was computed as the difference between the REDSS score corresponding to the last available non-relapse EDSS score evaluated by the end of the third year (the fourteenth scheduled

Figure 6.19: Boxplots for Rate of Change of REDSS



EDSS evaluation for patients who remained on study for three years and did not have a relapse at the time of this evaluation) and the baseline REDSS score divided by the number of intervening 12-week periods.

The boxplots in Figure 6.19 suggests that the rate of change in the high dose arm is slightly lower than in the other two arms. The boxplots also indicate roughly similar spreads of these rates in the three treatment arms, except for two very high rates: an average increase of 1 unit of REDSS every 12 weeks on study for a 32 year old female PL patient from center #225 with 3.7 years of MS history and initial EDSS score of 3, and an average decrease of 1 unit of REDSS every 12 weeks on study for a 45 year old male LD patient from center #261 with 9.25 years of MS history and initial EDSS score of 4.5. Due to the potential effects of these outliers, an analysis on the raw rates of change in REDSS would be less appropriate than those on the ranks of these rates.

The corresponding descriptive statistics are presented in Table 6.33. The medians

Table 6.33: Descriptive Statistics for Rate of Change of REDSS

Statistic	Placebo	Low Dose	High Dose
Number of patients	112	118	110
Median	0.032	0.000	0.000
Mean	0.044	0.029	-0.008
SD	0.191	0.163	0.165

Table 6.34: P-values from ANOVA for Rate of Change on REDSS

Comparison	Ranks of Rates	Raw Rates
Overall	0.041	0.074
Placebo vs. Low Dose	0.78	0.50
Placebo vs. High Dose	0.027	0.031
Low Dose vs. High Dose	0.031	0.098

for the treatment arms are slightly lower than for the placebo arm, but only the means suggest a fairly clear dose response relationship. ANOVA on the ranks of the rates and on the raw rates of change by treatment groups resulted in the p-values listed in Table 6.34; both suggest treatment group effects. In particular, in the pairwise comparisons based on the ranks, the HD arm is shown to be different from both the PL and LD arms, but there is no evidence of difference between the PL and LD arms.

Next, the baseline covariates are examined as potential predictors. Plots of the rate of change versus the baseline covariates and Dropout are presented in Figure 6.20. These plots do not indicate a particularly strong relationship with any of the covariates. The results of the ANCOVA on the ranks of the rates of change (including only main effects) are presented in Table 6.35. The p-values indicate modest treatment group differences and relationships of the rate of change of REDSS with age, initial REDSS, gender and Dropout. If the initial REDSS score is incorporated in this model as a continuous (rather than categorical as in Table 6.35) variable, the conclusions about the rest of the effects

Table 6.35: ANCOVA on Ranked Response

Variable	df	F	p-value
Treatment Group	2	2.21	0.11
Age	1	1.99	0.16
Duration	1	0.52	0.47
Initial REDSS	10	1.65	0.09
Center	10	1.12	0.35
Gender	1	3.22	0.07
Dropout	1	3.38	0.07
Residuals	313		

remains unchanged, but the p-value for the initial REDSS effect decreases to 0.02, suggesting a linear relationship of the rate of change with initial REDSS score.

In a model also incorporating interactions of all the covariates with treatment groups, only the treatment by center interaction was substantial ($p = 0.04$). The interaction plot of average rates of change for each treatment group in each center is presented in Figure 6.21. The plot indicates that the relative position of the three treatment group averages vary from center to center, reflecting the suggestion from the fit that the treatment group effects differ across centers. Including the treatment by center interactions in the model did not change the conclusions about the main effects; therefore, for simplicity of analysis, these interaction effects are not incorporated in the following analyses.

Switching to the raw rates as the response results in a considerable increase of all p-values, except for initial REDSS. Such increases may be associated with the presence of several outliers in the data. To minimize the effect of the outliers, the data for the two patients with the highest rates of change (described earlier) is withheld from the ANCOVA. The qualitative results of the analysis on this reduced data set are consistent with the conclusions from the analysis based on the ranked data. The treatment group,

Figure 6.20: Rate of REDSS Change versus Baseline Covariates and Dropout

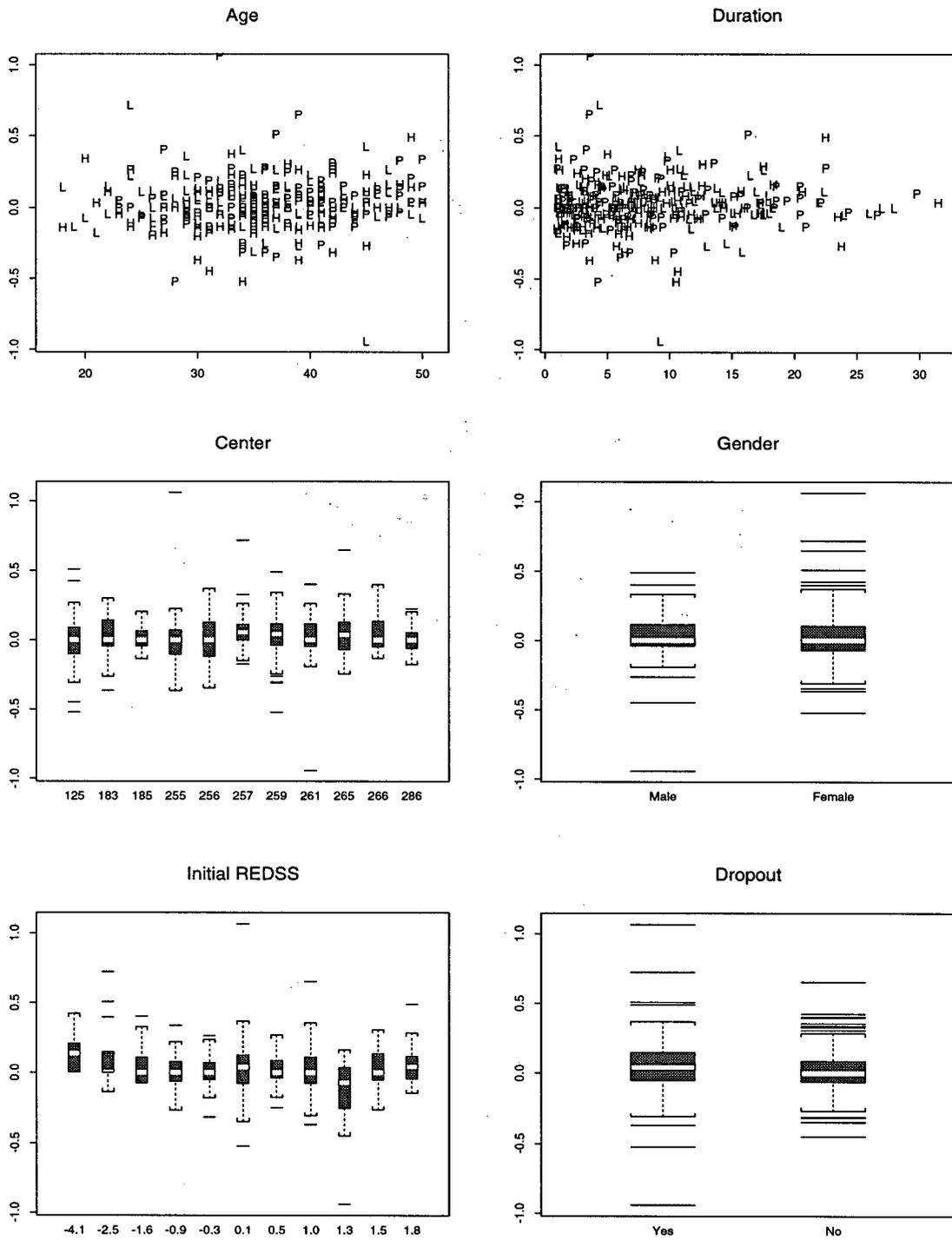
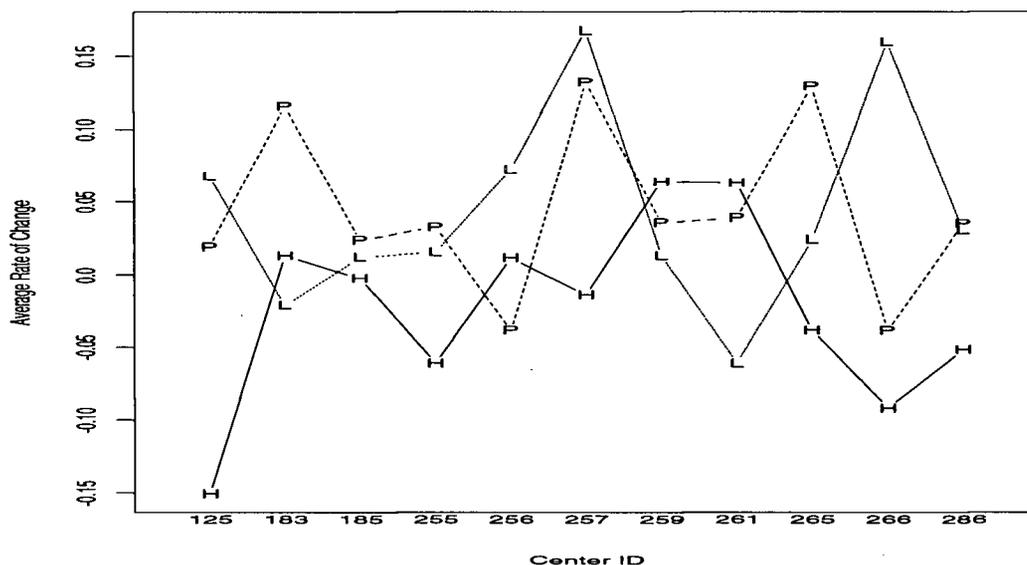


Figure 6.21: Treatment by Center Interaction Plot



age, duration, gender and Dropout effect estimates from this fit (with initial REDSS treated as continuous) are listed in Table 6.36.

These estimated effects indicate that the average rate of change of REDSS per 12-week period is 0.008 units higher in the LD arm than in the PL arm (the corresponding approximate 95% confidence interval contains all differences in rates from -0.013 to 0.029), while the average rate of change is 0.021 units lower in the HD arm than in the PL arm (the corresponding approximate 95% confidence interval contains all differences in rates from -0.034 to -0.007). The positive estimate of the age effect suggests that the average rate of increase is somewhat higher for older than for younger patients. The negative estimates for the gender and initial REDSS effects indicate that females had, on average, slower increase of score and the patients with lower baseline scores tend to have higher rates of increase. Duration of disease and center are not indicated as having an influence on the rate of change of REDSS.

Table 6.36: Estimates of Effect from ANCOVA: raw scale

Parameter	Estimate	SE	t	p-value
Low Dose	0.0075	0.0109	0.69	0.49
High Dose	-0.0207	0.0068	-3.04	0.003
Age	0.0019	0.0014	1.35	0.18
Duration	0.0005	0.0016	0.33	0.74
Initial REDSS	-0.0177	0.0069	-2.56	0.01
Gender	-0.0169	0.0096	-1.76	0.08
Dropout	0.0353	0.0206	1.71	0.09

The average rate of change is a good summary of the change in REDSS only if the rate of change is reasonably constant over this three year period. To examine the patterns of change over time, the change from baseline at each time point was averaged for patients in each treatment arm and plotted against time in Figure 6.22. The plot clearly indicates that in the PL and LD arms the change tends to increase over time, whereas in the HD arm the change remains at roughly the same level throughout the three years. The average changes for the three arms are at about the same (negative) level at the early stages of the trial, but after approximately one year the average change is consistently smaller in the HD arm than in the other two arms. Smoothing the plot using lowess yields Figure 6.23, which suggests that the average rate of change is not constant, especially for the high dose arm. Hence an analysis, which takes into account the patterns over time is required. A commonly used procedure for modelling such continuous longitudinal data is the repeated measures ANOVA described in the next subsection.

6.2.1 Repeated Measures ANOVA

About 16% of the first 14 EDSS scores for the 340 patients included in the data set for these analyses are missing. The break-down of the number of missing values by treatment arm (Table 6.37) indicates similar percentages of missing values in the three

Figure 6.22: Average Change in REDSS over Time

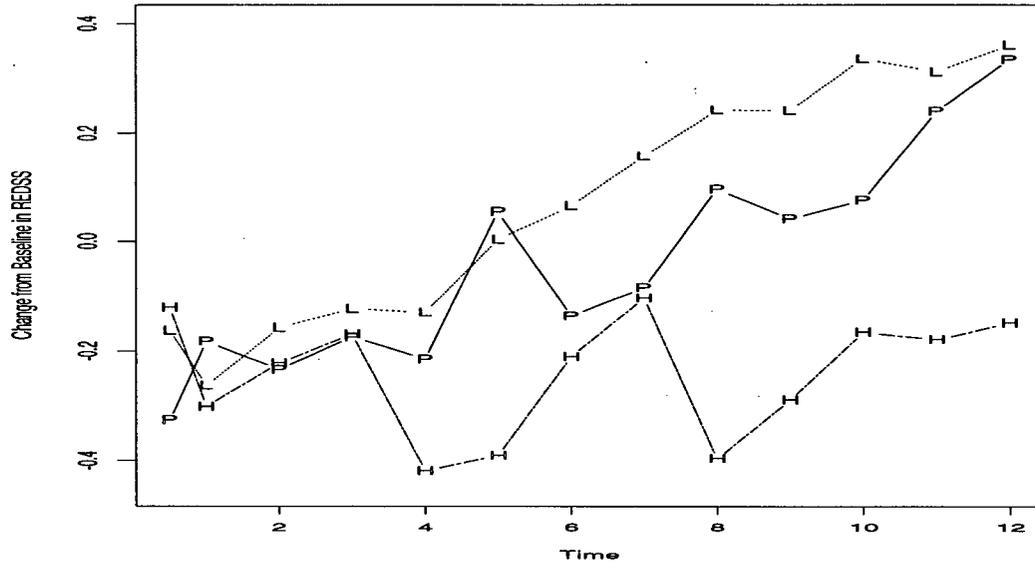


Figure 6.23: Smoothed Plot of Change in REDSS over Time

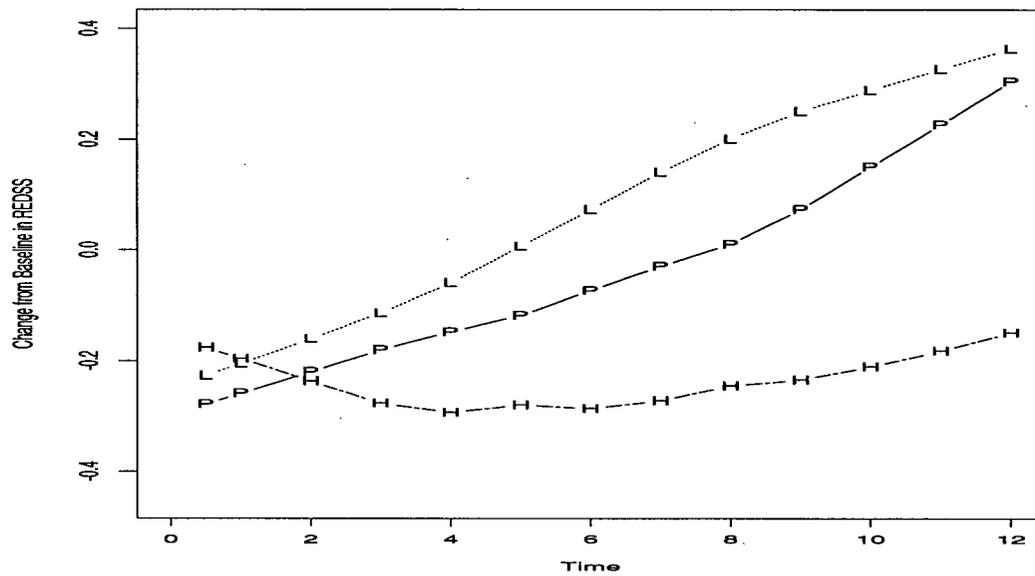


Table 6.37: Break-down of Missing EDSS Values by Treatment Arm

Treatment Arm	Patients	Missing Values	Percent Missing
Placebo	112	252	16.1%
Low Dose	118	278	16.8%
High Dose	110	210	14.0%

arms. Imputing all missing data points is a common way of obtaining a complete data set suitable for applying the repeated measures ANOVA. Missing values were estimated by the expectation-maximization (EM) algorithm as described by Johnson and Wichern [5]. To preserve possible treatment group differences, estimation of the missing values were done separately for each treatment group. For a data set with a few randomly missing observations, imputing the missing data would not be expected to have a substantial effect on the conclusions of a statistical analysis. But for a data set with a high percentage of missing values or with data missing due to the treatment side-effects, to relapses, or to other treatment-related conditions and hence nonrandom, such imputation of data may influence the conclusion about treatment effects. Because of the relatively high percentage of imputed data, a large fraction of which was missing due to relapses (as mentioned before, we are using only EDSS scores evaluated during exacerbation-free periods; scores recorded at scheduled visits when patients had exacerbations are treated as missing), conclusions from the analysis on this data set should be taken with some caution.

The simplest linear model upon which the analysis will be based has the following form:

$$Y_{ijt} = \mu + \alpha_i + \pi_{j(i)} + \beta_t + \alpha\beta_{it} + \epsilon_{ijt},$$

where

Y_{ijt} is the change in REDSS score at time t for patient j in treatment arm i ;

α_i is the effect of treatment arm i ;

$\pi_{j(i)}$ is the (random) effect of patient j in treatment arm i ;

β_t is the effect of time t ;

$\alpha\beta_{it}$ is the treatment by time interaction;

ϵ_{ijt} is the error term in the model.

In the repeated measures ANOVA model, patient-specific effects are treated as random factors nested within grouping factors such as treatment arm and categorical covariates (sex, center, initial REDSS). The patient-specific effects are assumed to be mutually independent normally distributed with mean 0 and variance ν^2 , and the random errors assumed to be mutually independent normally distributed with mean 0 and variance σ^2 . Patient-specific effects are random, since they represent deviations from the average trend in the population. Conversely, effects of time and treatment arm are treated as fixed factors, since there is interest in making inferences about the specific time points and treatments included in the study. The repeated measures ANOVA tests of time and treatment effects for the above model (ignoring the baseline covariates) are summarized in Table 6.38.

As suggested by the patterns in Figure 6.22, where there is no difference between the treatment arms at the outset but a rather large difference by the end of the time period, the analysis identifies the presence of treatment by time interactions. Partitioning the treatment by time interactions into the components for treatment by linear, treatment by quadratic and treatment by higher order components for interactions indicates a strong effect of the treatment by linear component of these interactions (p-value = 0.02) and no effects of the treatment by quadratic component (p-value = 0.36) or the treatment by higher order component (p-value = 0.99). This suggests that the treatment effects change linearly over time.

For this analysis to be valid, the variance-covariance matrix of the data has to satisfy

Table 6.38: Repeated Measures ANOVA on REDSS Change without Covariates

Source of variation	df	MS	F	p-value
<i>Between patients</i>	339			
Treatment groups	2	30.69	2.21	0.112
Patients within groups	337	13.92		
<i>Within patients</i>	4080			
Time	12	8.50	12.90	< 0.001
Treatment by time	24	1.46	2.22	0.001
Residuals	4044	0.66		

the so-called symmetry conditions [23]. If the conditions are not tenable, the F ratios may not follow the F -distribution, and the test results may be misleading. The standard deviations for each treatment arm at each time point are presented in Table 6.39 and the correlation matrices in Table 6.40. Since only subjects with baseline EDSS score below 6 were eligible for the study, the variability is relatively low for the first few EDSS evaluations. As the EDSS scores change over time the range of scores grows, so it is not surprising that the standard deviations increase over time, as indicated in Table 6.39. In addition, the rate of increase of the standard deviations is somewhat lower in the low dose arm than in the high dose and placebo arms. Table 6.40 indicates that the correlations between REDSS scores evaluated at 2 time points in the initial phases of the trial are lower than between EDSS scores evaluated at 2 time points in the later phases. These observed patterns in the standard deviation and correlation structures suggest that the symmetry conditions are violated in the REDSS data.

The symmetry conditions can be examined using the multivariate generalization of Box's M test and Mauchly's test of sphericity [15]. Both Box's M test of the homogeneity of variances across the treatment arms and Mauchly's test of sphericity yielded $p < 0.001$, hence the symmetry conditions appear to be violated. One approach to dealing with this situation is to adjust the numerator and denominator degrees of freedom of the critical

Table 6.39: Standard Deviations of REDSS Change at Each Time Point

Arm	1	2	3	4	5	6	7	8	9	10	11	12	13
PL	0.71	0.99	1.09	1.17	1.02	1.30	1.37	1.72	1.68	1.57	1.73	1.72	1.89
LD	0.85	0.75	0.93	1.11	1.12	1.20	1.06	1.29	1.34	1.26	1.35	1.40	1.32
HD	0.69	0.74	1.10	0.97	1.00	1.11	1.31	1.73	1.40	1.48	1.55	1.60	1.71

value for the test to compensate for the extent of the violation. Two estimates of this adjustment, called ϵ , are available. The Huynh-Feldt ϵ is an attempt to correct the Greenhouse-Geisser ϵ , which tends to be overly conservative [23]. Adjusting the F-ratio for the time by treatment interactions by the Huynh-Feldt ϵ yielded a p-value of 0.012, which does not change the conclusion about the existence of the treatment by time interactions. So, all we can conclude from this analysis is that the treatment group effects change across time.

Including all the baseline covariates, Dropout and treatment by center interaction terms in the repeated measures model leads to the results summarized in Table 6.41. Adjusting the degrees of freedom of the F-ratios for the treatment by center interaction and for gender effects by the Huynh-Feldt ϵ yielded p-values of 0.14 and 0.13 respectively, which does not change the conclusion about the existence of modest effects. But the treatment by center interaction effect on the change of REDSS score is somewhat weaker than the corresponding effect on the average rates of this change per 12-week period. Unlike the analysis on the average rates of change in Table 6.35, the results in Table 6.41 do not suggest a Dropout effect.

“Regression” in Table 6.41 corresponds to the continuous covariates: age, duration of disease and initial REDSS; the estimated effects and their standard errors are listed in Table 6.42. The conclusions about these covariates are similar to those from Table 6.35. The table shows a very strong linear effect for initial REDSS and suggests the possibility

Table 6.40: Correlation Matrices for Change in REDSS

	1	2	3	4	5	6	7	8	9	10	11	12	13
Placebo Arm													
2	0.43	1.00											
3	0.38	0.64	1.00										
4	0.33	0.58	0.73	1.00									
5	0.36	0.59	0.67	0.76	1.00								
6	0.44	0.55	0.62	0.67	0.73	1.00							
7	0.35	0.58	0.67	0.63	0.76	0.76	1.00						
8	0.34	0.46	0.52	0.57	0.67	0.80	0.81	1.00					
9	0.33	0.55	0.58	0.62	0.68	0.75	0.79	0.88	1.00				
10	0.36	0.50	0.67	0.66	0.69	0.73	0.72	0.77	0.85	1.00			
11	0.47	0.47	0.53	0.66	0.59	0.75	0.73	0.80	0.80	0.84	1.00		
12	0.22	0.41	0.54	0.53	0.56	0.71	0.69	0.76	0.78	0.80	0.83	1.00	
13	0.30	0.43	0.53	0.57	0.61	0.76	0.77	0.83	0.85	0.84	0.86	0.86	1.00
Low Dose Arm													
2	0.49	1.00											
3	0.72	0.48	1.00										
4	0.51	0.37	0.75	1.00									
5	0.49	0.45	0.68	0.69	1.00								
6	0.33	0.41	0.59	0.74	0.82	1.00							
7	0.49	0.41	0.62	0.58	0.74	0.76	1.00						
8	0.45	0.29	0.63	0.62	0.70	0.68	0.75	1.00					
9	0.49	0.25	0.71	0.70	0.77	0.79	0.83	0.83	1.00				
10	0.42	0.25	0.47	0.46	0.57	0.63	0.77	0.78	0.80	1.00			
11	0.30	0.24	0.42	0.52	0.52	0.58	0.56	0.64	0.62	0.71	1.00		
12	0.53	0.45	0.56	0.50	0.59	0.62	0.71	0.67	0.71	0.75	0.75	1.00	
13	0.37	0.27	0.41	0.30	0.52	0.58	0.62	0.67	0.60	0.75	0.75	0.78	1.00
High Dose Arm													
2	0.55	1.00											
3	0.37	0.54	1.00										
4	0.30	0.48	0.66	1.00									
5	0.21	0.53	0.51	0.71	1.00								
6	0.51	0.52	0.49	0.65	0.75	1.00							
7	0.35	0.59	0.60	0.60	0.65	0.78	1.00						
8	0.34	0.54	0.60	0.56	0.63	0.79	0.87	1.00					
9	0.17	0.40	0.42	0.52	0.70	0.71	0.69	0.79	1.00				
10	0.27	0.42	0.63	0.62	0.68	0.71	0.73	0.77	0.80	1.00			
11	0.26	0.51	0.52	0.59	0.65	0.68	0.72	0.81	0.85	0.80	1.00		
12	0.29	0.46	0.54	0.40	0.57	0.70	0.80	0.79	0.77	0.83	0.77	1.00	
13	0.20	0.40	0.44	0.48	0.60	0.66	0.73	0.77	0.79	0.80	0.86	0.81	1.00

Table 6.41: Repeated Measures ANOVA on REDSS Change with Covariates

Variable	df	F	p-value
Treatment*Center	20,314	1.43	0.11
Regression	3,314	6.59	< 0.001
Gender	1,314	2.44	0.12
Dropout	1,314	0.01	0.91
Time	12,3804	9.04	< 0.001
Gender*Time	12,3804	1.10	0.35
Dropout*Time	12,3804	0.30	0.99
Treatment*Center*Time	240,3804	0.95	0.70

Table 6.42: Estimates for Continuous Covariates: Repeated Measures ANOVA

Variable	Estimate	SE	t	p-value
Age	0.012	0.009	1.36	0.18
Duration	0.006	0.010	0.62	0.54
Initial REDSS	-0.189	0.044	-4.26	< 0.001

of an age effect on change in REDSS score. There is no indication of an effect due to duration of MS.

Excluding Dropout, duration and the treatment by center interaction terms from the model results in p-value of 0.04 for treatment by time interactions and p-value of 0.06 for center by time interactions. The estimates of gender, linear initial REDSS score and linear age effects remain similar to the estimates from the previous model.

The conclusions about the effects of the baseline covariates drawn from this analysis are similar to the conclusions from the preliminary analysis based on the rate of REDSS change in Table 6.36. The repeated measures analysis identifies the existence of treatment by time interaction effects, which confirms that using all the data collected over time instead of only the overall rate of change for the analysis of the REDSS data supplements conclusions about treatment effect with information about the time dependence.

Since the symmetry assumptions of the repeated measures ANOVA cannot be justified for the REDSS data, we will use the multivariate ANOVA (MANOVA) model, which allows a more general form for the variance-covariance matrix.

6.2.2 MANOVA

In the MANOVA model, each patient's repeated measurements are treated as a multivariate response vector. For the MANOVA, as for the repeated measures analysis, we use the data with missing values imputed by the EM algorithm. The multivariate analysis of variance model for treatment group effects has the form:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where

Y_{ij} is the 13-dimensional vector of change in REDSS scores for patient j in treatment arm i ;

α_i is the vector of effects associated with treatment arm i ;

ϵ_{ij} is the vector of errors.

The vector of errors is assumed to have a multivariate normal distribution with mean vector 0 and covariance matrix Σ that is common to the three treatment arms. The test for treatment group main effects gives a Wilks' Lambda of 0.9 which corresponds to an F-statistic of 1.47 with 26 and 650 degrees of freedom and p-value of 0.064. Hence this multivariate test provides some indication of treatment differences. Multivariate pairwise comparisons of the treatment groups resulted in p-values of 0.44 for the PL vs. LD comparison, 0.004 for the PL vs. HD comparison, and 0.22 for the LD vs. HD comparison. These p-values from the comparison of the treated group to placebo are very close to those obtained for the treatment group effect estimates on the rate of change in

Table 6.36. So, the analysis identifies only the high dose treatment effect, as suggested by Figure 6.23.

Since treatment differences are indicated, it may be informative to examine the univariate test results to get some idea of the time points at which the differences occurred. As suggested by Figure 6.22 the univariate F-tests (with 2 and 337 degrees of freedom) for treatment effects at each of the 13 time points do not indicate any treatment effects in the early portion of the study and show treatment differences more frequently towards the end of the third year of the study.

The MANOVA adjusted for all baseline covariates, Dropout and treatment by center interaction didn't indicate treatment by center interaction effects ($p = 0.45$), a Dropout effect ($p = 0.96$), a gender effect ($p = 0.80$) or a duration of MS effect ($p=0.71$) on the change in REDSS score.

The results of the analysis based on a model reduced to account for the treatment group, age, linear initial REDSS and center main effects (F-statistics corresponding to Wilks' Lambdas) are summarized in Table 6.43. As for the preliminary analysis on the ranks of the rates of change (Table 6.34), the MANOVA indicates treatment group effects on the change in REDSS. The MANOVA allows examination the patterns of dependence of the change in REDSS scores on the continuous covariates over time. The effect of the initial REDSS score is shown to be very strong (p -values < 0.001) up to time 7 (84 weeks on study) and gradually decreasing in the later phase (with p -values above 0.10 at time 11, time 12 and time 13).

The MANOVA relies on the assumption of multivariate normality for the data. To check this assumption, the quantities

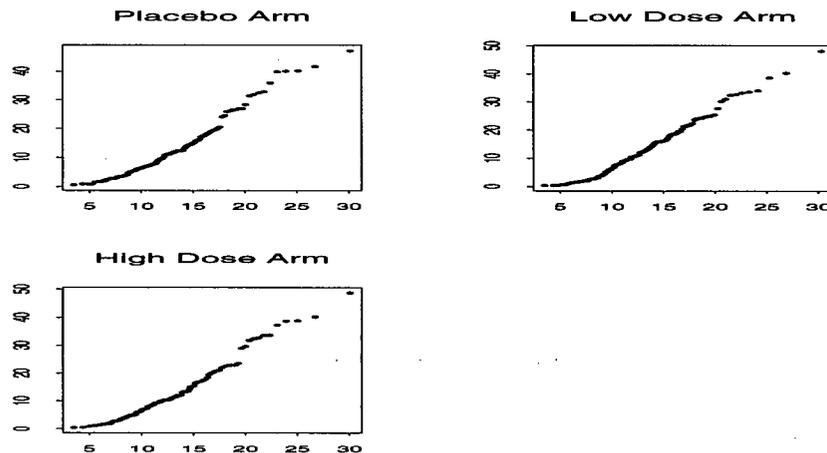
$$(Y_{ij} - \hat{\mu}_{ij})' \hat{\Sigma}^{-1} (Y_{ij} - \hat{\mu}_{ij}),$$

where Y_{ij} is the 13-dimensional vector of observations for a patient, $\hat{\mu}_{ij}$ is the estimate

Table 6.43: MANOVA on REDSS Change with Age and Initial EDSS

Variable	df	F	p-value
Treatment Groups	26,626	1.55	0.042
Regression	26,626	4.00	< 0.001
Center	130,2520	1.19	0.074

Figure 6.24: Quantile-Quantile Plots for Three Treatment Arms



of the mean vector for this patient, and $\hat{\Sigma}$ is the estimate of the variance-covariance matrix (common for the three treatment arms), were compared to a χ^2 distribution with 13 degrees of freedom. Quantile-quantile plots for each treatment arm separately (see Figure 6.24) do not indicate strong departures from multivariate normality.

In addition to the assumption of multivariate normality for the data, the MANOVA relies on a critical assumption of equality of the variance-covariance matrices across treatment arms. Though the standard deviations and correlation matrices for the three arms look reasonably similar at most time points (Table 6.39), Box's M test yielded p-value < 0.001 indicating that the homogeneity of variance-covariance matrices across treatment arms appears to be violated. Because the MANOVA assumptions may be unreasonable

for the change in REDSS scores data, the conclusions from the MANOVA may be incorrect. Methods which do not require these assumptions might be more reliable for analysis of the REDSS score data. Results for two such methods, the GEE approach and random effects regression models, will be presented in the following sections.

6.3 GEE Analysis

The GEE approach allows for different numbers of measurements for patients and the presence of missing values in the data. Therefore the data set with missing values (not imputed by the EM algorithm as for the repeated measures ANOVA or MANOVA) will be used for the GEE analysis.

Treating REDSS as a continuous response, the GEE analysis of the change in REDSS scores can be based on the linear regression model,

$$Y_{ijt} = \mu + \alpha_i + \beta_i * t + \epsilon_{ijt},$$

where μ is the intercept for the placebo arm, α_1 is the difference in the intercepts due to the low dose effect, α_2 is the difference in the intercepts due to the high dose effect. The time variable was centered in this fit, hence the fitted intercepts correspond to time 6 (66 weeks on study).

Table 6.44 lists the parameter estimates and the robust and naive estimates of the standard errors for above model. The choice of the working correlation structure has little effect on the conclusions as neither the parameter estimates nor the robust standard errors are much affected. On the other hand, the naive SE's change with the choice of the working correlation structure (as should be expected, given the pattern of consistently positive correlation in Table 6.40). The fitted slopes for all treatment arms are positive, indicating an increase in scores over time, and those for the PL and LD arms are highly

Table 6.44: GEE Model for REDSS Change with Separate Linear Trends

Parameter	Estimate	Robust SE	Naive SE	<i>z</i>
Independence Working Correlation				
Intercept (μ)	-0.041	0.110	0.038	
LD (α_1)	0.107	0.138	0.053	0.77
HD (α_2)	-0.199	0.143	0.053	-1.40
PL slope (β_0)	0.047	0.015	0.010	3.07
LD slope (β_1)	0.056	0.012	0.010	4.58
HD slope (β_2)	0.004	0.014	0.010	0.28
AR(1) Working Correlation				
Intercept (μ)	0.003	0.108	0.075	
LD (α_1)	0.072	0.135	0.105	0.53
HD (α_2)	-0.204	0.140	0.106	-1.46
PL slope (β_0)	0.061	0.015	0.015	4.13
LD slope (β_1)	0.057	0.012	0.014	4.84
HD slope (β_2)	0.011	0.014	0.015	0.76
Exchangeable Working Correlation				
Intercept (μ)	0.010	0.108	0.096	
LD (α_1)	0.062	0.134	0.134	0.45
HD (α_2)	-0.211	0.144	0.136	-1.46
PL slope (β_0)	0.042	0.014	0.007	2.92
LD slope (β_1)	0.049	0.011	0.007	4.36
HD slope (β_2)	0.008	0.013	0.007	0.62

significant. But, the estimate for the HD arm is very close to zero, which suggests rather small average change in the REDSS scores over the three years of treatment.

The smoothed plot of the average change in REDSS score over time in Figure 6.23 indicates possible curvature in the patterns, especially in the high dose arm. A model also incorporating separate quadratic trends over time led to the estimates listed in Table 6.45. The table indicates negligible quadratic trends in the PL and LD arms and only a weak trend in the HD arm. The positive sign of the quadratic effect suggests that in the high dose arm the REDSS scores tend to decrease during the first part of study and increase later. But the Wald test for simultaneous equality of all quadratic effects to zero yielded p-values of 0.40, 0.39 and 0.38 for the three working correlation structures, suggesting that it may be reasonable to exclude such quadratic effects from the model. The Wald test for equality of all linear time effects (from the model without quadratic effects) yielded p-values of 0.01, 0.02 and 0.04 for the independence, AR(1) and exchangeable working correlation structures respectively, suggesting that the rate of REDSS change over time differs across treatment groups.

Including all the baseline covariates and a Dropout main effect in the model with the separate linear trends over time leads to low z-scores (absolute values below 1.07) for the gender and duration effects suggesting a possibility of model reduction by excluding these two covariates. Refitting the model without gender and duration leaves the conclusions about all other main effects essentially unchanged. The estimates and standard errors for all effects, except centers, from this model fit are listed in Table 6.46.

As suggested by Figure 6.23, the Wald test indicates difference in the slopes for the three fitted lines. In particular the estimate of the slope for the HD arm is much smaller than for the PL arm, which indicates that, on average, the change in the REDSS score from baseline increases slower over the three year period in the HD arm (only about

Table 6.45: GEE Model for REDSS Change with Separate Linear and Quadratic Trends

Time Effect	Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation					
	Intercept (μ)	-0.072	0.126	0.058	
	LD (α_1)	0.140	0.157	0.081	0.89
	HD (α_2)	-0.221	0.163	0.081	-1.36
Linear	PL(β_0)	0.046	0.015	0.010	3.05
	LD (β_1)	0.056	0.012	0.010	4.60
	HD (β_2)	0.003	0.014	0.010	0.24
Quadratic	PL(ϱ_0)	0.002	0.004	0.003	0.82
	LD (ϱ_1)	-0.0001	0.003	0.003	-0.06
	HD (ϱ_2)	0.004	0.003	0.003	1.50
AR(1) Working Correlation					
	Intercept (μ)	-0.032	0.125	0.100	
	LD (α_1)	0.074	0.160	0.140	0.46
	HD (α_2)	-0.245	0.167	0.143	-1.47
Linear	PL(β_0)	0.061	0.015	0.015	4.10
	LD (β_1)	0.056	0.012	0.014	4.84
	HD (β_2)	0.010	0.014	0.015	0.68
Quadratic	PL(ϱ_0)	0.002	0.003	0.004	0.73
	LD (ϱ_1)	0.002	0.003	0.004	0.64
	HD (ϱ_2)	0.004	0.003	0.004	1.43
Exchangeable Working Correlation					
	Intercept (μ)	-0.020	0.120	0.100	
	LD (α_1)	0.100	0.156	0.140	0.64
	HD (α_2)	-0.227	0.162	0.142	-1.40
Linear	PL (β_0)	0.041	0.014	0.007	2.89
	LD (β_1)	0.049	0.011	0.007	4.36
	HD (β_2)	0.007	0.013	0.007	0.58
Quadratic	PL(ϱ_0)	0.002	0.003	0.002	0.91
	LD (ϱ_1)	-0.0005	0.003	0.002	-0.21
	HD (ϱ_2)	0.003	0.002	0.002	1.50

Table 6.46: GEE Model for REDSS Change with Covariates and Linear Trends

Parameter	Estimate	Robust SE	Naive SE	z
Independence Working Correlation				
Intercept (μ)	-1.005	0.340	0.141	
Low Dose (α_1)	0.123	0.131	0.051	0.94
High Dose (α_2)	-0.144	0.132	0.052	-1.10
PL slope (β_0)	0.048	0.015	0.010	3.19
LD slope (β_1)	0.058	0.012	0.010	4.73
HD slope (β_2)	0.006	0.013	0.010	0.44
Initial EDSS	-0.173	0.043	0.017	-4.00
Age	0.019	0.008	0.003	2.42
Dropout	0.205	0.119	0.055	1.72
AR(1) Working Correlation				
Intercept (μ)	-0.929	0.333	0.278	
Low Dose (α_1)	0.093	0.128	0.101	0.73
High Dose (α_2)	-0.147	0.130	0.103	-1.13
PL slope (β_0)	0.061	0.015	0.014	4.15
LD slope (β_1)	0.058	0.012	0.014	4.93
HD slope (β_2)	0.012	0.014	0.015	0.80
Initial EDSS	-0.180	0.041	0.033	-4.37
Age	0.019	0.008	0.006	2.46
Dropout	0.198	0.119	0.105	1.67
Exchangeable Working Correlation				
Intercept (μ)	-0.907	0.354	0.356	
Low Dose (α_1)	0.080	0.131	0.128	0.61
High Dose (α_2)	-0.156	0.134	0.131	-1.16
PL slope (β_0)	0.042	0.014	0.007	2.93
LD slope (β_1)	0.049	0.011	0.007	4.37
HD slope (β_2)	0.008	0.013	0.007	0.63
Initial EDSS	-0.181	0.043	0.042	-4.26
Age	0.017	0.008	0.008	2.12
Dropout	0.185	0.122	0.127	1.52

0.008 units of REDSS score per 12-week period with corresponding 95% confidence interval from -0.018 to 0.036 units as estimated for the exchangeable working correlation structure) than in the PL arm (about 0.042 units of REDSS score per 12-week period with corresponding 95% confidence interval from 0.014 to 0.070 units as estimated for the exchangeable working correlation structure). Given that at the beginning of the treatment the average change from the baseline of the REDSS scores in the three treatment arms was similar, this analysis indicates that the high dose treatment effect on the change of REDSS scores is superior to the placebo effect.

As the GEE approach, the random effects regression model allows for different numbers and times of measurements for different patients. Hence the data with missing values (not imputed by the EM algorithm as for the repeated measures analysis and MANOVA) will be used in the random effects regression analyses presented in the next section.

6.4 Random Effects Regression Model

The main question of interest is whether fitted curves show differences across the three treatment arms. Therefore, in addition to the overall intercept, the deviation from this intercept due to the treatment group effects will be included in the model. As well as for the GEE approach separate linear time trends for each group will be incorporated into the initial model.

One way to express the corresponding simplest linear random effects model is as

$$Y_{ijt} = \mu + (\alpha_i + a_{ij}) + \beta_i * t + \epsilon_{ijt},$$

where

Y_{ijt} is the change in REDSS score at time t for patient j in treatment arm i ;

μ is the intercept for placebo group;

Table 6.47: Random Effects Model for REDSS Change with Separate Linear Trends

Parameter	Estimate	SE	z
Intercept (μ)	-0.035	0.058	
Low Dose (α_1)	0.017	0.098	0.24
High Dose (α_2)	-0.081	0.041	-1.97
PL slope (β_0)	0.053	0.013	3.90
LD slope (β_1)	0.049	0.013	3.72
HD slope (β_2)	0.011	0.014	0.78

α_i is the deviation from the placebo intercept due to the treatment arm i effect (fixed effect);

a_{ij} is the deviation from the population-average intercept for patient j in treatment arm i (random effect);

β_i is the slope for treatment arm i (fixed effect);

ϵ_{ijt} is the error term in the model.

It is assumed that for all patients the a_{ij} are independent and identically distributed with a $N(0, D)$ distribution and the vectors of errors for different patients ϵ_{ij} are mutually independent and distributed with a $N(0, \sigma^2 \Omega_{ij})$ distribution, independent of the a_{ij} . For this analysis, as for the GEE approach, the time variable was centered so that the intercept corresponds to time 6 (sixty six weeks on study). In addition, the deviation for the PL arm, α_0 , is arbitrarily taken to be equal to zero and the errors ϵ_{ijt} are assumed to be independent (Ω_{ij} is the identity matrix) with common variance σ^2 . This model corresponds to fitting a different overall line for each treatment arm. The lines for the individual patients are assumed to be parallel to the overall line for the corresponding treatment arm.

The estimates of the slopes and intercepts for the three groups are listed in Table 6.47. The fitted slopes for all treatment arms are positive indicating that in every treatment

Table 6.48: Random Effects Model for REDSS Change with Separate Linear and Quadratic Trends

Time Effect	Parameter	Estimate	SE	z
	Intercept (μ)	-0.060	0.059	
	LD (α_1)	0.030	0.072	0.42
	HD (α_2)	-0.096	0.042	-2.27
Linear	PL (β_0)	0.053	0.013	3.90
	LD (β_1)	0.049	0.013	3.72
	HD (β_2)	0.010	0.014	0.77
Quadratic	PL (γ_0)	0.002	0.002	0.96
	LD (γ_1)	-0.0003	0.002	-0.16
	HD (γ_2)	0.004	0.002	2.28

arm the change in REDSS scores increases over the three year period. The rate of increase is relatively high and similar in the placebo and low dose arms but different from the rate in the high dose arm (since the correlations between the estimates are negligible, the table provides the necessary inputs for assessing the significance of these differences). The low z-score corresponding to the estimate of the slope in the high dose group suggests that this slope can be considered as negligible. If the linear model provides an adequate fit to the data in the high dose arm, this indicates that, on average, the REDSS scores for the HD arm remained essentially unchanged during the three years of treatment. The conclusions about the group effects drawn from this table are qualitatively similar to those from Table 6.44.

Because the smoothed plot for the change of REDSS score versus time in Figure 6.23 indicates slight curvature, the model was expanded by also including separate quadratic effect for each treatment arm. The estimates for this model and the corresponding standard errors are listed in Table 6.48. The table indicates rather small quadratic effects in the PL and LD arms and a strong quadratic effect in the HD arm. The positive sign of the quadratic effect suggests that in the HD arm REDSS scores tend to decrease

Table 6.49: Random Effects Model for REDSS Change with Time Trends and Covariates

Parameter	Estimate	SE	z
Intercept (μ)	-0.700	0.217	
Low Dose (α_1)	0.030	0.069	0.43
High Dose (α_2)	-0.060	0.041	-1.48
PL slope (β_0)	0.053	0.014	3.90
LD slope (β_1)	0.050	0.013	3.75
HD slope (β_2)	0.011	0.014	0.84
Quadratic	0.002	0.001	1.76
Initial EDSS	-0.249	0.031	-8.07
Age	0.017	0.006	2.90
Center #125	0.119	0.096	1.24
Center #183	0.108	0.057	1.88
Center #185	0.025	0.033	0.76
Center #255	0.032	0.028	1.14
Center #256	0.007	0.027	0.27
Center #257	-0.013	0.016	-0.79
Center #259	-0.028	0.015	-1.85
Center #261	-0.012	0.018	-0.64
Center #265	0.014	0.025	0.55
Center #266	-0.015	0.012	-1.20

during the first part of study but increase during the later phase. The Wald test for the equality of all quadratic effects yielded a p-value of 0.22; the test for the simultaneous equality of all quadratic effects to zero resulted in p-value of 0.10. These p-values indicate that it is possible to reduce the model to a common quadratic effect for the three groups, but it may not be reasonable to further reduce the model to the linear trend only. Fitting the reduced model yields the estimated common quadratic effect of 0.002 with z-score of 1.76. The results for the other effects remained essentially the same as in the fit with the separate quadratic effects.

Including fixed effects for the baseline covariates and Dropout in the model resulted in z-scores of -0.22, 0.78 and 0.23 for the gender, duration and Dropout effects respectively.

Excluding these variables from the model did not change the conclusions about the other effects, hence such model reduction seems reasonable. The fit of this model is summarized in Table 6.49.

The estimates of the slopes and their standard errors from this analysis are similar to the corresponding estimates in Table 6.46. Both analyses indicate that, on average, the rate of REDSS increase is slower in the HD than in two other arms. But unlike the analysis based on the GEE approach, the random effects regression analyses identify the presence of a quadratic trend in the pattern of REDSS change over time. Though the estimates for the other effects listed in Table 6.49 are somewhat different from the estimates in Table 6.46, the qualitative conclusions concerning these effects are very similar. Both analyses suggest a strong age effect. The positive sign of its estimate indicates that, on average, older patients tend to have a larger change from baseline than younger patients. Both analyses indicate an extremely strong influence of the initial REDSS score on the change in REDSS. The negative sign of the estimate suggests that, at each time point, among patients of the same age, the patients with the higher initial REDSS score tend to have smaller change from the baseline score.

All the analyses to this point were done using re-expressed scores and treating those scores as continuous, none of them used the actual ordinal EDSS categories. The next section presents results from mixed effects ordinal logistic regression analysis which utilizes the original data with ordinal EDSS categories.

6.5 Mixed Effects Ordinal Logistic Regression Analysis

The mixed effects ordinal logistic regression analysis is a preferable way to analyze the ordinal EDSS scores because it uses the original data and does not rely on a somewhat arbitrary rescaling.

Because subtracting the EDSS scores associated with the ordinal categories is not meaningful, the analysis in this section, unlike all other analyses in this chapter, is performed on the EDSS categories themselves, not on their difference with the baseline score. Nevertheless, in order to reduce the number of parameters to be estimated, at first, the initial EDSS score is incorporated in the model as a continuous covariate. As in the models for the GEE approach and mixed effects regression analysis, the time variable is centered in this analysis, so that time $t = 0$ corresponds to sixth period after the beginning of the treatment.

Using the notations of the random effects regression model for REDSS change, the initial mixed effects regression model for the latent response can be expressed as:

$$z_{ijt} = \mu + (\alpha_i + a_{ij}) + \beta_i * t + \theta * \xi_{ij} + \epsilon_{ijt},$$

where θ is a fixed initial EDSS effect (which was indicated by every performed analysis as having a very strong effect on the change in EDSS scores). The same assumptions are used in this model as in the random effects regression model for REDSS change.

MIXOR, the computer program written by D. Hedeker and R.D. Gibbons [3] for the mixed effects ordinal logistic regression analysis was employed to fit the model. Implementation of the program is described in the following subsection.

6.5.1 MIXOR Implementation

MIXOR can estimate the starting values for the parameter and threshold estimates or can incorporate user-defined starting values. For almost all runs of the MIXOR, the number of iterations increased (or the program failed to converge) whenever user-defined starting values were specified. The number of iterations increased substantially when the common convergence criterion of 0.001, corresponding to termination when corrections for all parameters became less than 0.001 [3], was replaced by 0.0001.

Unfortunately the number of categories that can be handled by MIXOR is rather limited. Even for the simple model for the latent response described above, using all 18 categories of the original EDSS scores resulted in very many iterations (120 iterations with the convergence criterion of 0.001 and automatic starting values). To increase the likelihood of convergence, MIXOR makes an incremental adjustment to the diagonal elements of the information matrix when the program encounters a non-increasing likelihood or other numerical difficulty during the iterations. This adjustment, called the ridge, starts at 0.0, and increases by 0.1 each time difficulties are encountered. Therefore the ridge value can be used as an indicator of computational difficulty that the program encountered. The lowest value of the ridge was equal to 0.2 for the data with all original EDSS categories.

More serious estimation difficulties also occurred: the program failed to converge every time user-defined starting values were used for the data with 18 categories. The choice of these sets of starting values was based on the parameter estimates obtained from the fit with the automatic starting values and on the automatic starting values themselves. One of the sets of starting values was equal to the parameter estimates rounded to the first decimal place and few other sets of starting values were chosen from the interval between the automatic starting values and the corresponding parameter estimates. All these indications of estimation difficulties suggest that the estimates obtained from the analysis on the original 18 categories of the EDSS scores may be unreliable.

These problems were by-passed by collapsing some of the outcome categories. Forming new categories by combining adjacent EDSS scores reduces the information available from the EDSS scores themselves; but the form of the conclusions about the treatment effects should stay unaffected by such combining. Hence the parameters of interest will not depend for their interpretation on the actual response categories involved although the estimates will in general be affected [14].

Table 6.50: New Categories of EDSS scores

EDSS	Frequency	Number of Categories							
		4 (Freq.)	6 (Freq.)	7	8	9	10	12	
0	183	1 (975)	1 (975)		0	0	0	0	
1	318			1	1	1	1	1	
1.5	474			2	2	2	2	2	
2	621	2 (996)	2 (621)	3	3	3	3	3	
2.5	375						4	4	
3	383	3 (793)	3 (758)	4	4	4	5	5	
3.5	410						5	6	6
4	278	4 (910)	4 (688)	5	5	6	7	7	
4.5	111								8
5	155			5 (345)	6	6	7	8	9
5.5	79								10
6	156								
6.5	78								
7	29								
7.5	17		6 (287)	7	7	8	9	11	
8	4								
8.5	1								
9	2								

Except for a rather coarse stratification into four categories, which was done based on the frequencies alone, two guidelines were used to guide the most appropriate stratification of the EDSS scores. The first was to choose meaningful subdivisions of scores from the severity of MS point of view. The second was to choose a stratification with an approximately equal number of observations in each category. As the result of these two considerations, all the EDSS scores above 5.5 form one of the new categories. Even though there are relatively few observations above 5.5, because these scores are considered to be very high, they had to be separated from the rest of the scores. The scores of 4.5, 5, and 5.5 are also somewhat critical and formed another category. The scores below 4.5 were stratified according to their frequency.

Table 6.51: Number of Non-Varying Responses and Iterations for New Categories

Categories	4	6	7	8	9	10	12	18 (EDSS)
Non-Varying Responses	56	30	19	15	9	9	6	5
Number of Iterations	61	32	30	30	32	35	37	120

With EDSS scores stratified into 6 categories, MIXOR did not report any estimation difficulties and the estimates were consistent for all sets of specified starting values. After successful estimation of the model parameters for 6 categories finer stratifications were attempted. The new stratifications had 7, 8, 9, 10, and 12 categories. The break-down of the original EDSS scores into these strata is given in Table 6.50.

With 12 categories different estimates were obtained with different sets of starting values; therefore no finer stratification was attempted. The estimates of the model parameters for all smaller numbers of categories seemed to be rather consistent, both for different specified starting values and for different numbers of categories. The value of the ridge was 0.0 for all these fits.

Coarser stratification resulted in an increasing number of patients with non-varying responses; these are patients whose repeated responses all fell into the same stratum. As the number of such patients increases, the computational difficulty in fitting also increases. For data with almost all subjects having constant responses, reducing each patient's vector of responses to a scalar summary and using non-longitudinal methods of analysis may be more efficient. The number of patients (out of total 340) with non-varying responses with different numbers of categories is listed in Table 6.51. The table also contains the number of iterations MIXOR required to reach the convergence criterion of 0.001 when automatic starting values were used. The smallest number of iterations required was 30 with the EDSS scores stratified into 7 and 8 categories. For that reason, we decided to focus the analysis on these numbers of categories.

The required number of iterations seems to increase not only for the data with more than 8 categories, but also for the data with less than 7 categories. The latter increase may be associated with the increasing number of patients with non-varying responses. To assess the influence of the non-varying responses, the data for the 19 patients with non-varying responses was withheld from the data set with 7 categories. Withholding this data from the analysis gave an unexpected result – the number of iterations increased from 30 to 38. On the other hand, withholding the data for the 9 patients with non-varying responses from the data set with 9 categories, reduced the number of iterations from 32 to 29. This suggests that other factors in addition to the number of response categories, the convergence criterion, the specification of user-defined starting values, and the number of non-varying responses affect the speed of convergence of the MIXOR program.

To evaluate the effect of an increasing number of the fixed effect parameters on MIXOR performance, the initial EDSS score was reparametrized as a categorical covariate and age, gender, duration, Dropout, and center fixed effects were included in the model. Including these covariates in the model increased the number of iterations from 30 to 230 and from 30 to 76 for the data sets with 7 and 8 categories respectively (with corresponding ridge values of 0.8 and 0.2). Including more covariates in the MIXOR model also noticeably increases the time for each iteration. Increasing the number of parameters increased the sensitivity of the fit to the choice of starting values. The parameter estimates changed substantially for different sets of starting values. The choice of these sets of starting values was again based on the parameter estimates obtained from the fit with the automatic starting values and on the automatic starting values themselves. These observations suggest that the MIXOR may have difficulties fitting models with very many covariates.

On the other hand, the estimates for the fixed linear time effects were much more

Table 6.52: Fixed Effects and Thresholds Estimates from MIXOR for Data with 8 Categories

Parameter	Estimate	SE	z	p-value
Intercept (μ)	1.182	0.157		
LD (α_1)	0.087	0.129	0.67	0.50
HD (α_2)	-0.215	0.125	-1.72	0.085
PL slope (β_0)	0.069	0.009	7.76	< 0.001
LD slope (β_1)	0.111	0.009	11.83	< 0.001
HD slope (β_2)	0.015	0.009	1.63	0.10
Initial EDSS (θ)	1.898	0.045	42.37	< 0.001
Threshold 1 (γ_1)	0.000			
Threshold 2 (γ_2)	1.954	0.069	28.52	< 0.001
Threshold 3 (γ_3)	3.649	0.074	49.03	< 0.001
Threshold 4 (γ_4)	5.409	0.086	62.77	< 0.001
Threshold 5 (γ_5)	7.537	0.098	76.67	< 0.001
Threshold 6 (γ_6)	9.932	0.112	88.60	< 0.001
Threshold 7 (γ_7)	11.722	0.121	96.94	< 0.001

consistent than the estimates for the intercepts: their values remained unchanged to the second decimal not only for all sets of specified starting values but also across the fits with different numbers of categories. The results of the analyses based on the EDSS scores stratified into 8 categories are presented in the next subsection.

6.5.2 Results of Analyses

For the initial mixed effects regression model for the latent response specified at the beginning of this section, the estimates for the parameters and thresholds obtained with the EDSS scores stratified into 8 categories are listed in Table 6.52. In terms of the continuous latent variable assumed in the model, this table indicates (recall that in this chapter we are analyzing the EDSS scores, not changes from baseline) that at six periods after the beginning of the treatment the average disability level in the HD is somewhat lower than in the PL. The positive estimates for the slopes in each of the three treatment

arms suggest that the latent variable for the EDSS scores tends to increase over time. The Wald test of the equality of the linear time trends in the three treatment arms resulted in a p-value < 0.001 ; subsequent pairwise comparisons of the trends also indicated strong differences between each pair. Therefore we can conclude that the analysis indicates a high dose treatment effect, in the sense that the disability level tends to increase over the three years of study more slowly in the HD arm than in the PL arm (or in the LD arm). The analysis indicates a huge (z-score of 42) influence of the initial EDSS score. The estimate of the effect is positive, indicating that the patients with a higher score at the beginning of the study tend to remain in a higher category than the patients with a lower initial EDSS score. Qualitatively the same results were obtained from the analysis with 7 strata.

Including fixed effects for quadratic time trends in the model for the data with 8 categories resulted in very small estimates for quadratic trends with p-values of 0.32, 0.99 and 0.18 for the PL, LD and HD arms respectively. This suggests that a model without quadratic time trends will provide an adequate representation of this data. This turns out to be the case; the Wald test of the simultaneous equality of the quadratic time trends to zero resulted in a p-value of 0.42. Excluding the quadratic time trends from the model left the conclusions about the rest of the effects essentially unaffected.

At the next step, in order to get a better understanding of the effect of the initial EDSS score, we reparametrized the initial EDSS variable in the model. Instead of treating it as a continuous covariate, now we treat it as categorical and look at the effect of each initial EDSS score separately (the effect of the initial score of 0 was arbitrary taken to be zero in this fitting). The resulting effect estimates are listed in Table 6.53. The effect estimates for initial EDSS increase with the category, except for the initial scores of 4.5 and 5.5; this departure from the overall pattern may be associated with the relatively low frequency of these baseline scores. The linear trend in effect estimates for the categories

Table 6.53: Results from MIXOR for Data with 8 Categories and Initial EDSS Score as Categorical

Parameter	Estimate	SE	z	p-value
Intercept (μ)	0.333	0.267		
LD (α_1)	0.197	0.123	1.61	0.11
HD (α_2)	0.071	0.124	0.57	0.57
PL slope (β_0)	0.068	0.009	7.72	< 0.001
LD slope (β_1)	0.112	0.010	11.39	< 0.001
HD slope (β_2)	0.016	0.009	1.80	0.071
Initial EDSS (Freq.):				
0 (11)	0.000			
1 (21)	1.002	0.342	2.93	0.003
1.5 (29)	3.508	0.309	11.35	< 0.001
2 (59)	4.370	0.289	15.13	< 0.001
2.5 (38)	4.998	0.289	17.26	< 0.001
3 (47)	6.799	0.299	22.74	< 0.001
3.5 (50)	7.415	0.296	25.08	< 0.001
4 (35)	9.080	0.314	28.94	< 0.001
4.5 (12)	5.787	0.402	14.40	< 0.001
5 (20)	9.987	0.346	28.89	< 0.001
5.5 (18)	9.532	0.346	27.53	< 0.001

of the initial EDSS score suggests it is reasonable to use the initial EDSS as a continuous covariate in the following analysis.

Next, to assess the effect of covariates, fixed age, gender, duration, Dropout, and center effects were included in the model. Including the baseline covariate and Dropout effects in the model resulted in the estimates listed in Table 6.54. The table indicates that the average value of the latent variable for EDSS score differs across centers (the effect for center #268 was arbitrary taken to be equal to 0), it is higher for older patients, for patients with less than 66 weeks on study, and for females. The random effect variance term for this fit was estimated as 2.038 with a standard error of 0.056.

For all described models and attempted stratifications of the original EDSS score,

Table 6.54: Results from MIXOR for data with 8 Categories and Baseline Covariates

Parameter	Estimate	SE	z	p-value
Intercept (μ)	-2.277	0.397		
LD (α_1)	0.725	0.120	6.07	< 0.001
HD (α_2)	0.114	0.125	0.91	0.36
PL slope (β_0)	0.067	0.009	7.24	< 0.001
LD slope (β_1)	0.111	0.009	11.90	< 0.001
HD slope (β_2)	0.018	0.009	2.01	0.04
Initial EDSS (θ)	1.749	0.044	40.10	< 0.001
Age	0.075	0.009	8.10	< 0.001
Gender	-0.509	0.105	-4.84	< 0.001
Duration	0.015	0.010	1.50	0.13
Dropout	0.534	0.145	3.68	< 0.001
Center #125	-0.770	0.274	-2.81	0.005
Center #183	0.731	0.261	2.80	0.005
Center #185	0.374	0.226	1.65	0.10
Center #255	0.883	0.215	4.10	< 0.001
Center #256	0.484	0.235	2.06	0.040
Center #257	1.067	0.271	3.93	< 0.001
Center #259	0.650	0.223	2.91	0.004
Center #261	0.089	0.218	0.41	0.68
Center #265	-0.230	0.243	-1.23	0.22
Center #266	1.989	0.399	4.99	< 0.001

MIXOR was relatively consistent in estimating the time trends for the three treatment arms. The analyses indicated that the latent variable for EDSS score increases over the three year period in all treatment arms, but the pairwise comparison suggested that the rate of increase is much lower in the HD than in the two other arms. The estimates from Table 6.54 suggest that for any category, the odds of having a score in this category or below decrease by 4.6% per 12-week period on study for placebo patients (the corresponding approximate 95% confidence interval contains all percentages from 3.7 to 5.5), by 7.5% per 12-week period on study for low dose patients (the corresponding approximate 95% confidence interval contains all percentages from 6.6 to 8.3), while the odds for high dose patients decrease by only 1.2% per 12-week period on study (the corresponding approximate 95% confidence interval contains all percentages from 0.4 to 2.1).

6.6 Summary

As a first attempt to analyze the ordinal EDSS scores, this response was re-expressed to treat it as a continuous variable. Five different methods were used to analyze the change from baseline of these re-expressed EDSS scores. First, the vector of repeated observations for each patient was summarized as a scalar variable, representing the average rate of change of REDSS score from per 12-week period. The analyses based on these summaries indicated a modest high dose treatment effect.

Next, two ANOVA methods, repeated measures ANOVA and multivariate ANOVA, were employed. Unlike the analyses on the average rates of change, these methods account for differences across treatment groups in the patterns of the re-expressed scores over time. In order to obtain an equal number of measurements for each subject, for these analyses the missing data was imputed using EM algorithm. As suggested by Figure 6.23, both analyses indicated the presence of treatment by time interaction. In addition, both

analyses identified a very strong effect of the re-expressed baseline EDSS score and the possibility of a weak age effect. But the validity of both analyses relies upon restrictive assumptions about the variance-covariance structure of the repeated observations on individual patients in the three treatment arms. The assumptions appear to be violated, casting some doubt on the validity of the results.

The next two methods, the GEE approach and random regression models, allow a more general approach to the problem of modeling repeated measurements. Both models allow for missing observations, hence the original data set (with the rescaled EDSS scores) was used. As suggested by Figure 6.23 the rate of REDSS change fitted by these analyses was noticeably lower in the HD arm than in the PL arm. A slight curvature in the fitted curves was suggested by the random effects model approach to the analysis.

Finally a direct analysis of the original EDSS scores was attempted using MIXOR. This analysis might be viewed as the most reliable, because it does not require re-expression of the data, or pretending that EDSS scores can be treated as continuous responses. Unfortunately, due to complexity of the estimation procedure for the mixed effects ordinal regression analysis, the number of categories and the number of covariates which can be reliably modeled by the program for a data set of this size is rather limited. The analysis based on the mixed effects ordinal regression approach agrees with the GEE and random regression modelling approach that the HD patients tend to have, on average, slower increase of the EDSS score, than patients from the PL and LD arms.

Chapter 7

Discussion

7.1 Conclusions about Data

The main focus of all the analyses on the three outcomes of the interferon beta-1b clinical trial considered in this thesis has been on the effects of treatment and the patterns over time. Preliminary plots for the number of active lesions, rate of exacerbations and EDSS scores over time suggest differences between treatment arms. Indeed, the high dose treatment group effects are clearly identified by all approaches for all variables considered in this thesis.

Apparent treatment effects are detected for the number of active lesions per scan and for the probability of beginning an exacerbation by both attempted analyses, the analysis of covariance on summaries over time and by the GEE approach. But only longitudinal analysis, such as the GEE approach, can be used to assess differences in patterns over time. Preliminary plots for the number of active lesions for the cohort in the frequent MRI UBC substudy suggest there may be time trends in the number of active lesions per scan which differ across the treatment groups. Indeed, the estimated linear time effect is positive for the placebo arm but negative for the two treatment arms. On the other hand, none of the estimated linear time effects is significantly different from zero. Thus, the longitudinal analysis has established that, for this sub-study, the data do not provide convincing evidence of trends over time in the logarithm of the expected number of active lesions.

Preliminary plots for the exacerbations for the full cohort of patients from the 11 different centers suggest a decreasing trend in the rate of exacerbations over time in each treatment group. Even though the time trend for the high dose arm seems to be somewhat different from the time trends in the placebo and low dose arm, the data does not provide convincing evidence of differences in the linear time effects. In addition to the linear time trend, the model identifies the presence of a common quadratic time trend. The positive sign of the estimate for the quadratic time trend along with the negative estimates for the linear trends suggests that the rate of exacerbation decreases faster at the beginning of the trial and slower towards the end of the 3-year period.

Since the time trends in the three treatment arms appear to be similar, these analyses suggest an early (within the first 6-week period) effect of the high dose of Betaseron on both the number of active lesions and the probability of exacerbations. On the other hand, the nature of the treatment effect on the EDSS scores appears to differ. All attempted longitudinal analyses, the repeated measures ANOVA, MANOVA, GEE approach or random effects regression approach on the re-expressed EDSS scores as well as mixed effects ordinal regression analysis on the original EDSS scores, indicated differences in patterns over time in three treatment arms. In particular, these analyses suggested that, even though there is no evidence of an early Betaseron effect on the EDSS scores, the EDSS scores increase more slowly in the high dose arm than in the placebo arm.

7.2 Methods of Analyses

The previous chapters have illustrated application of several methods to longitudinal analysis for binary, count and ordinal responses. As a preliminary step the repeated observations for each patient were summarized over time to create a univariate response. Summarizing the data in this way allows one to use simple techniques for the analysis

but does not allow analysis of the patterns over time. Therefore longitudinal analysis techniques are essential for extracting maximum information about the treatment effects from the available repeated observations.

Repeated-measures analysis of variance and multivariate analysis of variance are commonly used to analyze longitudinal data on continuous responses. These two methods of analysis were applied to the EDSS data with the original scores re-expressed to treat them as continuous responses. These analysis of variance methods are appropriate only when the responses for each subject are multivariate Gaussian with a common covariance matrix for all subjects. In addition, all subjects are required to have measurements at exactly the same times, and no missing values may be present. In our case, however, the re-expressed EDSS scores did not satisfy these assumptions. A large percentage of the data was missing due to the removal from the data set of EDSS scores obtained during periods of relapse and the covariance structure did not satisfy either the sphericity or the homogeneity assumptions. This example illustrated that application of these traditional methods of longitudinal analysis is rather limited. One of the purposes of this thesis was to explore the use of several more recently developed methods of longitudinal analysis.

One of these methods, the GEE approach, applies equally well to binary, count and continuous response variables and provides asymptotically valid inference even when the correlation structure of the repeated observations is misspecified. Thus, the GEE approach provides a broadly applicable and unified method for the analysis of longitudinal data. One advantage of the GEE approach is that it does not require specification of the joint likelihood of the data. Another advantage of this approach is that it can incorporate time-dependent covariates and different times of observations for different subjects (including missing data). But the GEE approach has some drawbacks, namely that the lack of full distributional assumptions limits the availability of diagnostic tools and can lead to an over-reliance on estimated regression coefficients and their standard

errors [24].

Random effects regression models were also applied to the analysis of the continuous response. One of the disadvantages of these models is the requirement of more restrictive assumptions about the data structure. The approach allows time-dependent covariates and different times of observations for different subjects. Whereas the GEE approach to longitudinal data analysis estimates only average parameters in a population, the random effects regression approach can also estimate individual parameters for each subject. This can be particularly useful in the medical setting where a proportion of subjects may respond to therapy in quite different ways from the average response.

The mixed effects ordinal regression model can be employed for the analysis of repeated or clustered responses which are either binary or ordinal. It allows for time-dependent covariates and different times of observations. This approach has the same advantage of accounting for individual differences for each subject while estimating population parameters as the random effects regression model. Due to the complexity of the estimation processes, the approach, at least as implemented in the MIXOR software, is rather limited to data with relatively few response categories and simple models with few random effects.

Bibliography

- [1] Cook, N.R. and Ware, J.H. (1983). Design and analysis methods for longitudinal research. *Annual Review of Public Health* **4**, 1-23.
- [2] Diggle, P.J., Liang, K.Y. and Zeger S.L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.
- [3] Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933-944.
- [4] The IFNB Multiple Sclerosis Study Group (1993). Interferon beta-1b is effective in relapsing-remitting multiple sclerosis: clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* **44**, 655-661.
- [5] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis (3rd edition)*. Prentice Hall, New Jersey.
- [6] Kurtzke, J.F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* **33**, 1444-1452.
- [7] Laird, N.M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305-316.
- [8] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.
- [9] Liang, K.Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- [10] Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association* **83**, 1014-1023.
- [11] Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673-687.
- [12] Matthews, B. (1993). *Multiple Sclerosis THE FACTS (3RD EDITION)*. Oxford University Press.
- [13] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (2nd edition)*. Chapman and Hall, London.

- [14] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**, 109-142.
- [15] Morrison, D.F. (1967). *Multivariate Statistical Methods*. McGraw-Hill.
- [16] Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression : A Second Course in Statistics*. Addison-Wesley Publishing Co.
- [17] Paty, D.W., Li, D.K.B., the UBC MS/MRI Study Group, and the IFNB Multiple Sclerosis Study Group (1993). Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* **43**, 662-667.
- [18] Petkau, J. and White, R. (1995). Longitudinal analyses for the UBC 6-weekly frequent MRI sub-study of the Betaseron multiple sclerosis clinical trial. *Biostatistics Research Report #10*, Biostatistics Research Group, University of British Columbia.
- [19] Senn, S.J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**, 467-475.
- [20] Stiratelli, R., Laird, N.M. and Ware, J.H. (1984). Random-effects models for serial observations with binary responses. *Biometrics* **40**, 961-971.
- [21] Ware, J.H. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *The American Statistician* **39**, 95-101.
- [22] Waternaux, C., Laird, N.M. and Ware, J.H. (1989). Methods for the analysis of longitudinal data: Blood lead concentrations and cognitive development. *Journal of the American Statistical Association* **84**, 33-41.
- [23] Winer, B.J. (1971). *Statistical Principles in Experimental Design*. McGraw - Hill, Inc.
- [24] Xu, J.J. (1996). *Statistical Modelling and Inference for Multivariate and Longitudinal Discrete Response Data*. Ph.D. Thesis, Department of Statistics, University of British Columbia.
- [25] Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.
- [26] Zeger, S.L., Liang, K.Y., and Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72**, 31-78.

Appendix A

Listings of Data

Table A.55: Placebo Group: Number of Active Lesions Data

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
420	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
421	1	1	1	1	1	1		0	1	0	0	0	0	0	0	0	0
443	0	0	0	0	0	2	2	1	0	4	0	0	0	0	0	1	2
446	0	0	0	0	0	0	0	0		0	0	0	0	0	1	0	1
449	0	0	0	1	0	2	1	1	0	1	1	2	0	1	1	0	4
451	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
498	2	2	1	1	0	0	0	0	0	0	2	2	0	2	2	2	0
501	0	0	0	0	0	0	0	1		0	0	1	0	1	0	0	1
504	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
507	0	0	1	3	1	0	3	1	0	2	0	2	2	5	0	0	2
522	0	0	0	0	0	2	0	0	0	0	1	0	0	2	0	0	0
523	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
539	2	1	1	1	2	1	1	2	0	0	1	0	0	0			
540	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
545	1	2	0	2	5	5	1	0	3	3	3	5	3	2	3	1	3
550	0	2	0	0	0	0	0	0	1	0	0	0	2	2	0	0	2
565	3	0	2	1	0	1	0	0	2	0	0	1	0	1	1	0	0

Table A.56: Low Dose Group: Number of Active Lesions Data

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
419	0	0	0	3	1	1	1	0	0	2	0	1	1	1	0	1	1
424	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
448	1	0	0	0	0	0	0	1	0		0	1	0	0	0		0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
452	0	0	2	0	0	0	0	0	1	0	0	0	0	0	1	0	1
499	0	0	0	0	2	0	1	0	0	1	0	1	0	0	0	2	0
502	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
505	0	0	1	0	0	0		0	0	0	0	0	0	0	0	0	0
508	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
521	1	0	0	0	0	0	0	1	2	1	0	0	1	0	0	2	0
525	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
542	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
544	0	1	2	0	1	0	0	0	1	0	1	0	0				
547	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
548	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
564	0	0	0	2	2	2	1	0	0	1	0	0	0	0	0	0	0
568	1	0	0	0	0	0	0	0	0	0	0	0	0				

Table A.58: UBC cohort: Baseline Covariates

ID	age	edss	dur	dose	orig	gen	ID	age	edss	dur	dose	orig	gen
420	42	2.0	4.4	0	1	0	508	31	2.5	1.5	1	1	1
421	35	1.0	2.8	0	1	0	521	35	3.0	10.9	1	0	1
443	27	1.5	6.9	0	1	1	525	36	3.5	15.2	1	0	1
446	34	1.5	6.9	0	1	1	542	34	1.0	5.9	1	1	1
449	38	4.0	4.2	0	1	0	544	27	1.5	3.7	1	1	0
451	34	2.5	17.1	0	1	0	547	42	2.0	1.8	1	0	1
498	31	3.5	3.8	0	1	0	548	47	3.5	26.1	1	0	1
501	35	2.0	16.2	0	1	1	564	26	2.0	3.3	1	0	1
504	33	1.0	14.3	0	1	1	568	41	3.5	18.2	1	0	1
507	41	2.5	6.9	0	1	0	422	34	1.5	12.5	2	1	1
522	36	2.5	4.1	0	0	0	444	44	3.0	18.0	2	1	0
523	37	1.0	1.1	0	1	1	445	47	2.5	14.9	2	1	1
539	35	2.0	8.0	0	1	1	453	35	1.0	10.1	2	1	1
540	25	1.0	1.5	0	1	1	454	42	1.5	10.1	2	1	1
545	30	1.0	9.1	0	1	1	497	20	3.5	1.1	2	1	1
550	38	1.5	5.4	0	0	1	500	33	1.5	10.4	2	1	1
565	37	1.0	13.1	0	1	1	503	30	3.0	3.2	2	1	1
419	31	1.5	12.1	1	1	0	506	22	3.5	9.5	2	1	1
424	29	3.0	4.0	1	1	1	524	36	3.0	7.5	2	1	1
448	51	1.5	3.4	1	1	1	526	46	2.5	9.4	2	1	0
450	49	3.5	20.0	1	1	1	541	36	2.5	8.8	2	1	1
452	39	0.0	6.1	1	1	1	543	45	1.5	23.8	2	1	1
499	24	1.5	9.1	1	1	1	546	49	5.5	22.8	2	0	0
502	48	0.0	5.8	1	1	1	549	34	1.5	10.5	2	0	1
505	47	3.0	14.3	1	1	1	566	42	3.0	2.6	2	1	1

ID: patient study identification number

age: age (in years)

edss: EDSS score

dur: duration of disease (in years)

dose: treatment group: 0=Placebo; 1=Low Dose; 2=High Dose

orig: origin: 0=Washington State; 1=B.C.

gen: gender: 0=male; 1=female

Table A.59: Placebo Group (first 18 patients): Exacerbation Data

ID	1 16	2 17	3 18	4 19	5 20	6 21	7 22	8 23	9 24	10 25	11 26	12 27	13 28	14 29	15 30
202	0 0	0 0	0 0	0 0	0 0	0 0									
206	0 0	0 0	0 0	0 0	0 0	0 0									
210	1 0	0 0	0 1	0 0	1 0	0 0	1 0	0 0	0 0	0 0	0 0	0 1	0 0	1 1	0 0
212	0 0	0 0	0 0	0 0	0 1	0 0	0 0	1 0	0 1	0 0	0 0	0 0	0 0	0 0	0 0
214	0 1	0 0	0 0	0 0	0 0	1 0	0 0	0 0	0 1	0 0	0 1	0 0	0 0	0 0	1 0
215	1 0	0 0	0 0	1 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	1 1	0 0	0 0	0 0
220	0 1	0 0	0 1	0 0	0 0	0 0	0 1	0 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0
222	0 0	1 0	0 1	0 0	0 0	0 1	0 0	0 1	0 0						
226	0 0	0 0	0 0	0 0	0 0	0 0									
228	0	0	1	0	1										
232	0 0	0 0	0 1	0 0	0 0	0 0	1 0	0 0	1 0	0 0	1 0	0 0	0 0	0 0	1 0
233	0 0	0 1	0 0	0 0	0	0	0	0	0	1	0	1	0	1	0
239	0 0	1 0	0 0	1	0	0	0	0	1	0	1	0	0	1	0
240	1	1	1	0	0	1	0	0	0	0	0				
244	0 0	0 0	0 1	0 0	0 0	0 0	0 0	0 0	0 0						
246	0	0	0	0	0	0	0	0	0						
249	0	0	0	0	0	1	0	0							
250	0 0	0 0	0 1	0 0	0 0	0 0									

Table A.62: Placebo Group (first 35 patients): EDSS Data

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
202	4	4	3	4.5	4	4.5	5	5	5	5	4.5	4.5	4.5	5
206	5	5	5	5	3.5	4.5	2	4	2	2	3	2.5	2	1.5
210	2.5	1.5	1.5	1.5	2.5	2	2.5	2	1.5	2.5		2	2	1.5
212	2.5	2.5	2.5	2	1.5	2	2	2.5	2	2.5	2.5	2.5	2.5	2.5
214	1	1	1	1		1	2	1	1.5		1.5	1	1.5	
215	3.5		2.5		2.5	3.5	3		5	5.5	6	6	3.5	6
220	1.5	1	1	0	0	1	1	1	0		0		0	1
222	0	0	0	0	0		2	2.5	2.5	0	0	2	2	2
226	1	1	1	1	1	1.5	1	1.5	1	1	0	0	0	0
228	1	1.5	2	3.5	2.5	2	3							
232	3.5	3.5	2.5	3.5	3	3.5	3.5		3			3	3	3
233	2.5		2.5	3	3	3.5				4.5		4.5		
239	1.5	1		1.5	2	1	1	2.5		2	2	1	1	2
244	4	3	3.5	3.5	3	4	4	6	6	6		2.5	4.5	5
249	4	3.5	5	5			4.5	4.5				3	4	5
250	3.5	3.5	3.5	2.5	2.5	3	2	2.5	3.5	3	2	1.5	3	3.5
256	1	1.5	1	0	1	1	1		1.5		1	1	1	2
259	3	3	3.5	3	3	2	2		2	5	3	3	3	3.5
262	2	2	3	2	2	1.5	2		2.5	3	3	2.5	2.5	2.5
266	1.5	2	2	1.5	1.5	2	1.5	1	2	1	2	1.5	1	1
268	3	3.5	1.5	3	1.5	2	3	4			5		5.5	5.5
272	3.5	3	2		3	3		3.5				4	6	4.5
273	4	4	4	5.5		6								
277	1.5	1	1	1	1	1	1	1.5	1	1	1	1	2	1
279	1.5	1.5				2	2.5	1.5	1.5	2	1.5	2	1.5	1.5
282	0	0	0		0		0	0	0	0	0	0		
288	4.5		2	3.5	3.5	3		3	2.5		3	2		3.5
289	2.5	2	2	1.5	1.5	2	2	1	2	1.5	1		2	
291	3	3	2	2	1	1	1	1	1	0				
296	1.5	1	1	2	2	1.5	2.5				2			
298	2	2		1	1.5	3	3	2		2.5	2	3	3.5	
303	4.5	4.5		3.5	3.5	3.5	3	3	3	3	3		2	2.5
307	3.5	3.5	3.5	2.5	2.5	3	3	3	3.5	3	3	3	3.5	3.5
309	2		2.5	2.5			3.5	3		4	3	3.5	3	
312	3	3.5	3.5	3.5	3.5	3	3.5	3.5			4	4	4	

Table A.63: Low Dose Group (first 35 patients): EDSS Data

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
204	2	2	2	3	3.5	3.5	3.5		3.5	3.5	3	3.5	3.5	3.5
205	5.5	3.5		4	3	3.5	4	3.5	3	3	3.5	3.5		3.5
207	2	2	2	3		3	2	1.5	1.5	1.5	1.5	3	3	3
211	2			2	2.5	2	2	2	1.5	1.5	1.5	2	0	1.5
213	4	4	3.5	4		4	4	4	4	4	4	4.5	4.5	4.5
218	1	2.5		2	2	2			1.5	1.5	1.5	1.5	2	1.5
219	1.5	1.5	0		1	1	1	1.5						
221	2	2	2	2	2	2	2	2	2	2	2	2	2	2
225	3.5	3.5	3	3	3.5	6	4	3	3.5	3.5		3		3.5
229	2.5	2.5	2.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	4	4	4.5	
234	3.5	3.5	3.5	4	3.5	4	4	4	4	4	4	4	4	4
236	4	4	4	4	4	4	4			4	4	4	4	4
237	3.5	3.5		3.5	3.5			4.5	3.5			4.5		4
238	2.5	2.5	2.5	2.5	2.5	2.5	1.5	1.5	1.5		1.5	1.5	2	2
243	4	3	2.5	3	3		2	3	2		3	2		3
245	3		2			1.5	1.5	2		1.5	2	3	2	3
253	2.5	2.5	1.5	3	1.5	2	1.5	2				3		2.5
254	1	1	1.5		0	1		1.5	0	1	0	0	2	1
257	2.5	2	2	2		3.5	3.5	2.5	2.5	2.5	2.5	2.5	2	2
258	3.5		3.5	3.5	4.5		4		4	5	4.5	4.5	5	
261	4	4	3	3.5	3.5									
265	4.5	4	4	4	4	4	4	4	4	4	4	4.5	4	4.5
267	2.5	2	2	2	2	2	1.5	2	2.5	1.5	2	1.5	1.5	1.5
270	2	1.5	1.5	1	1.5	0	0	0	0	0	0	0	1.5	0
274	2	2	2	2		3	4.5	3	3.5	3	3	6	5.5	3.5
275	4.5	2	3.5	2	3.5			5.5	5.5	5.5	5.5	5	5	5.5
280	1.5	1.5		3					2	2.5	2.5	2	2	3
284	3.5	4	2	3.5	3.5	2.5	2		3	3	2.5	3.5	3	3
286	3	3	2.5	2.5	1.5	2	2	2	1.5	1.5	1.5	1.5		2
290	0	0	0	0		0	1	0	0	0	0	0		4
293	5	5	5	5	5		5			4	5			
294	0	0	1	0	0	0	1	0	0	0			0	1
306	2	2	2	2	1.5	1.5	2	2	2	2	1.5		2	2
311	2	2	3		2.5	2.5	2		3.5	1.5		5	2	
313	3	1.5	1.5	2	2	1.5	1.5	1.5	3	2.5	3.5	1.5	1	1.5

Table A.64: High Dose Group (first 35 patients): EDSS Data

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
201	3		2.5	1.5	3.5	3.5	2	1	1.5	3			1.5	
203	3	3	3	3	3	3	3	3	3.5	3		3	3.5	3.5
208	4.5	4.5	5	4.5	4.5	4.5	4.5	5	4.5		5		5.5	5
209	5.5	4.5	3.5	4.5	4.5	6	4	3.5	4	6	5	5	5	4
216	2	2.5	2	2	1.5	1.5	2	3	3.5	2.5	2.5	2.5	2.5	2.5
217	3	4	4	4	4	4	4.5		4	4	4.5	3	3	
224	2	2	2.5	2.5		2	2	2	2		0	1.5	1.5	1
227	1	1	0	0	1	1	1.5	0	0	2		1.5	0	0
230	4.5	3.5	3	1	5	4	3	2.5		1	0	0	0	1
231	2.5	2	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2	1.5	1.5	1.5
235	2	2	2	1.5		1.5	1.5		2	1.5	1.5	1.5	1.5	1.5
241	3	2.5	3		3.5	4	4	4		4	4	6	6	
242	3		3	3		3.5		4	3.5		3.5	4	4	
247	4	3.5		3.5	3.5			2	1.5					
248	1.5	2.5	1.5	1.5	2	1.5	3	3	1.5	1	3	1.5	2.5	3
251	5					4.5	4.5		4	3	4.5	3	4	4
252	1	2	2	1.5		1	1.5	1	2		1	1.5	1	1.5
255	4.5		3	3		1.5	1.5	1		1	1	1	1	0
263	3.5	3.5	3.5	4	4	4	4	3.5	4	4	3.5	4	4	3.5
264	2.5	2.5	2.5	3	3	1.5	3	3	2.5	2.5	2	3	2.5	2.5
269	4	4	3.5		6	4		6.5	6		6	6	7	
271	5	5	4	4	4.5	5.5	4.5	4.5	4.5	5	5	5	5	5
283	2.5		1.5	1.5		2.5	2.5	2.5	3.5	2	2.5	3.5	2	
285	5	5	5	5	5	5	5	5	5	5	5	5	5	5
287	4.5	4.5	3	3		2.5		4	4.5	4	4	3.5		
292	2	1.5	1.5	1.5	1	1	1		1	1	0	0	1	1
295	3	2	2.5	2	2.5	2	2			3	3			3.5
304	3	2.5	2	1.5	1.5	2	2.5	2	1.5	1.5	1.5	1.5	1.5	1.5
310	5	3.5	2.5	2.5	2.5	1.5	2.5	2.5	3	3	1.5	3	2	2
314	2	2	1.5	1.5	1.5		2	2	2.5	2	2	0	2	0
317	3	3	3	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
321	4	2.5	3	3	4	4	2.5	2		2.5	2		2	2
326	3.5	2	3	3.5	3.5		2	2	2.5	3		2.5	3	3
327	2.5	2.5	3	6	2.5			6	6	2				
331	2	1.5	2	3	4.5	4	2	2	2		4		1	