A comparison of statistical techniques used in the longitudinal analysis of the modified Rankin scale in stroke randomised controlled trials

Jessica Emma Potts

A thesis submitted for the degree of Doctor of Philosophy in Statistics

Keele University

December 2018

Contents

List of tables

List of figures

Acknowledgements

There are several people that I would like extend my thanks to that have helped make this research project and the writing of this thesis possible.

Firstly, to my supervisors, Dr John Belcher, Professor Julius Sim, Professor Christine Roffe and Professor Anand Pandyan for their support, encouragement, and advice at different points throughout this journey.

Specifically, I would like to thank Dr John Belcher, for your patient guidance throughout my time as your student, and for answering my questions – not matter how small and trivial they were. Secondly, I would like to thank Professor Christine Roffe for providing the SO_2S trial data, as well as Dr Tracy Nevatte, for the effort you put in to obtain the data in a suitable format – it may have taken longer than anticipated but your efforts were appreciated.

I would like to thank Keele University for providing me with ACORN funding for a studentship throughout my time as a postgraduate student here. As well as the School of Computing and Mathematics, for providing me a space to work, keeping my spirits up and lending an ear when needed.

I also extend my deepest thanks to Becky for many years of friendship both in and out of work and for reading and commenting on my thesis – I will return the favour!

Finally, I would like to thank my friends and family, for their support in getting to the end of this thesis. My final thanks go to my husband James, and my parents who have kept me sane and supported me throughout this project.

Abstract

Most analyses in stroke clinical trials assess the primary outcome at a single time point and do not consider conducting a longitudinal analysis, even when the outcome has been recorded at several time points. Efforts to improve the quality of statistical analyses have previously been made, but have been directed at techniques that compare groups at a predetermined fixed time point. The aim of this thesis was to specifically investigate the use of longitudinal techniques, such as generalised linear mixed models and latent variable methods, to model serially correlated stroke outcome data, and to consider the challenge of having a score for death within the scale. The chosen outcome was the modified Rankin scale, which is a functional assessment outcome after stroke and is one of the most popular endpoints in clinical trials; often recorded repeatedly during follow-up.

Initial chapters describe the epidemiology of stroke and various stroke outcomes, with a review of what appears in the literature regarding current longitudinal modelling of mRS data. Subsequent chapters report how proportional-odds models are fitted to longitudinal repeated measures trial data, with the effect estimates compared to an analysis at a single time point. Simulation techniques assess how different patterns of drop-out affect treatment estimates derived using all the data. These are then compared to a Markov analysis that treats death as an absorbing state. Finally, clustering methods are considered in order to try and describe trajectories of the mRS, but it is highlighted that this is challenging given limited movement between states. Each of the methods have their own merits, and trials should be encouraged to consider longitudinal analyses when there is repeated measures data.

List of abbreviations

ADL – Activities of daily living

AIC – Akaike's information criterion

ANOVA – Analysis of variance

AvPP – Average posterior probability

BFGS – Broyden-Fletcher-Goldfarb-Shanno

BI – Barthel Index

BIC – Bayesian information criterion

BLRT – Bootstrap likelihood ratio test

CT – Computerised tomography

EQ-5D – EuroQol five dimensions questionnaire

IC – Information criterion

GCS – Glasgow Coma Scale

GOS – Glasgow Outcome Score

KPS – Karnofsky performance scale

LCA – Latent class analysis

LCGA - Latent class growth analysis

LLCA – Longitudinal latent class analysis

LTA – Latent transition analysis

LOC – Level of consciousness

LRT – Likelihood ratio test

LMM – Linear mixed models

MAR – Missing at random

MCAR – Missing completely at random

MRI – Magnetic resonance imaging

mRS – Modified Rankin Scale

NIHSS – National Institutes of Health Stroke Scale

NINDS – National Institute of Neurological Disorders and Stroke

NEADL – Nottingham extended activities of daily living

OAST – Optimising analysis of stroke trials

OHS – Oxford Handicap Scale

RCT – Randomised controlled trial

rt-PA – recombined tissue plasmogen activator

SF-36 – 36 item short form health survey

SO_2S – Stroke oxygen study

TIA – Transient ischaemic attack

VLMR LRT – Vo-Lin-Mendel-Rubin Likelihood Ratio Test

WHO – World Health Organization

1. Introduction

1.1 Project introduction

To a patient who has had a stroke, one of the most important questions they want answering is, “when will I recover?” They want to know whether they will recover and be able to live the life they did before and how long that process will take. Every stroke is different and therefore the recovery will vary from person to person. The objective of my thesis is to investigate the recovery of patients who have suffered a stroke during the first 12 months of rehabilitation.

Very little research is conducted using longitudinal methods in stroke trials, and when they are conducted they use specific methods that do not give the best interpretation or use all the data available. It is especially important that when collecting data on several scales, these scales are used to their full potential and not dichotomised just because the interpretation is easier, as the loss of statistical power is great when these techniques are applied (Bath et al. 2012). Previous studies that have looked at a follow-up of stroke patients and conducted longitudinal analysis have mainly used repeated measures analysis of variance or a linear mixed model. Specifically, functional recovery over time is not well researched in stroke trials, this may be due to the fact that general functional recovery is hard to specify, with most studies researching the effects of functional recovery in a specific body part for example the arm or the leg), or that a lot of recovery is done in the first few weeks after the stroke which is often before follow-up interviews and questionnaires have been conducted and so they fail to capture the moment at which patients’ recovery begins. Recently, it has been realised that better statistical methods are needed in stroke trials, with research in America looking at the use of multi-state models, a method we are also considering.

Inspiration for this work came from researching different longitudinal techniques, with special concentration on the techniques that can be applied to ordinal scales – specifically, the modified Rankin Scale (mRS), which is the outcome scale that we are most interested in, as it is the most common scale recorded during stroke trials but also appears to be the scale in which the least amount of research has been conducted. Because of this, it was decided to focus the research conducted in the thesis on the longitudinal analysis of the mRS within a randomised controlled trial environment, as this was the type of data that we had available to us. It is worth noting that there will be situations other than randomised trials where these techniques may be appropriate, for example cohort studies, although attention has not been given to this throughout this project. Indeed, many of the techniques used within the project are more commonly found in the analysis of cohort studies, with the use within a clinical trial setting being relatively new.

1.2 Outline of thesis

Chapter 2 provides an initial background and introduction into the acute condition of stroke. The chapter aims to detail a brief introduction as to what a stroke is and the different types of stroke that an individual can have. It also covers information about stroke background including how a stroke is diagnosed, the different causations and how it is treated. It will provide a wider context for the aims of the thesis, and give a good understanding of stroke treatment. In addition to this, the chapter looks at the different ways that the severity of a stroke can be measured and how these stroke scales differ from each other in what they measure. Finally, this chapter introduces the two different datasets that are used in the analyses throughout the thesis, the Stroke Oxygen Study and the National Institute of Neurological Disorders and Stroke Trial. The details of the design and setting of the trials are given, along with how the data was collected and also the follow-up of the trial and the responses at each time point.

Chapter 3 reviews the current literature surrounding stroke trials and the methods used to analyse them. There are several different parts to the review of the literature. A systematic review was undertaken to investigate:

1) The methods used in recently published stroke trials when considering studies that follow individuals over a period of time and report multiple time points. The interest lies in how many trials actually conduct some sort of longitudinal analysis and how many ignore the repeated measures and conduct the analysis at a single time point.
2) The different methods that have been applied to the mRS in longitudinal studies in stroke research. The mRS is the main outcome scale of interest in this thesis and so it is important to research previous and current literature surrounding the longitudinal analysis of this, in order to identify previously applied techniques and gaps in the literature.

As well as looking at longitudinal methods for analysing the mRS, consideration was taken into looking how stroke trials analyse the mRS at a single time point in order to identify techniques that may be able to be modified and then applied to the mRS in a longitudinal analysis. This was further extended to look at the application of potential techniques had been discussed and decided upon, in areas other than stroke clinical trials. This was done in order to obtain a better understanding of whether the included methods were appropriate for use in the analysis of the mRS.

Finally, the literature was explored to look at other outcome assessment scales that, like the mRS have a category for death included in them. These scales can come from anywhere within medical assessments and may help to provide information on different methods on how to analyse the mRS.

Chapter 4 investigates the different ways in which an ordinal outcome scale can be treated and the potential methods of analysis that can be used that are appropriate for each of different ways the scale can be treated. Following this, the methodology for these methods is detailed, focusing mainly on the different regression models that can be fitted to the data in order to analyse an ordinal outcome scale considered at both at a single follow-up time point and using the longitudinal data of the follow-up of a trial, by fitting a repeated measures model. The results produced when fitting these models are presented within this chapter, the effects calculated by these regression models are then compared and how well each of the models fit the data is considered.

Chapter 5 is a simulation study that extends the analysis that was conducted in chapter 4. The simulation study aims to look at the comparison between the longitudinal repeated measures model and the single time point model and how different types of missing data can affect the comparison. The study looks at ordinal data simulated from 2 different types of distributions, comparing the effect of treatment at the first time point when using just the data at that time point alone, to the treatment effect at the first time point having considered the other time points in the follow-up. The study compares the differences in estimated effects of the two models, when different correlation values between the repeated time points are used.

Chapter 6 starts to apply more novel methods to the ordinal outcome data, the first being the multi-state model that is fitted in this chapter. This method splits the categories of the scale into different states and calculates the risk of an individual moving between different states in the model. By doing this it allows the category of death to be used in a meaningful way, as it can be included without the ordinality of the modified Rankin scale being questioned.

Chapter 7 applies a technique called latent class growth analysis, where the trajectories of an individuals modified Rankin scores are considered over time. Initially, patients included in the Stroke Oxygen Study are grouped into various numbers of clusters dependent on their mRS values over the follow-up period, with the best fitting model being chosen. After this, only individuals who have at least 1 transition between scores over time are considered in the analysis and split again into clusters. It is also investigated whether the treatment group and the standard care group have different trajectories in the clusters assigned to them. Following this, models were fitted firstly to all survivors and secondly to survivors who had a transition between scores over the follow-up time. Finally, models were fitted to all patients in the National Institute of Neurological Disorders and Stroke Trial dataset as well as the treatment and control groups separately to consider the effect of having more time points in the model.

Chapter 8 discusses all the different models that have been applied in the thesis, with a brief summary of the results and the strengths and limitations of the different longitudinal methods that have been applied to the data, as well discussing the impact of the research conducted. Finally, the chapter will discuss some of the key decisions that were made throughout the analysis and the impact of these, as well as thinking about areas in which future research can be conducted in order to extend the research and findings of the project.

1.3 Thesis aims

The overall aim of the thesis is to identify and apply appropriate longitudinal techniques for analysing stroke trials with mRS outcome data and a longitudinal follow-up. The research conducted in this project will help to identify the best possible methods for analysing stroke outcome data, in order to inform clinicians and patients about expected recovery in these patients. The main focus of the project is the change in the mRS of patients over the first 12 months following the stroke. It is important to find a method to apply to longitudinal data that produces meaningful results that are easy for clinicians to interpret. The overall aim of the thesis can be broken down into the following sub-aims:

Aim 1: To identify appropriate longitudinal methods that can be applied to the mRS for use in randomised clinical trials.

The specific objectives of this aim will identify:

- Previous longitudinal methods applied to any scale in a stroke trial.
- Previous longitudinal methods applied to the mRS.
- The use of longitudinal methods that have been identified as appropriate when applied to other scales, outside of a stroke clinical trial.

Aim 2: Explore the different treatment effects produced by using regression models at a single time point in the follow-up and a repeated measures regression model

The specific objectives of this aim are to:

- Identify appropriate regression models in order to calculate a treatment effect.
- Apply the regression models to the SO_2S data and calculate the effect of treatment using data from a single time point and data from all the included time points.
- Simulate new ordinal data to produce alternative scenarios in order to compare the differences in the two types of regression model further.

Aim 3: Model the ordinal mRS, with a multi-state model, as a series of states between which patients can transition.

The specific objectives of this aim are to:

- Identify the appropriateness of using multi-state models for ordinal data.
- Apply the multi-state model using complete data.
- Apply the multi-state model taking into account missing data.

Aim 4: Explore whether patients can be clustered into recovery trajectories over time and if these trajectories differ between the treatment and control groups.

The specific objectives of this aim are to:

- Identify appropriate latent class techniques to model the data.
- Apply latent class growth analysis models in order to identify different clusters of patient trajectories in various scenarios.
- Identify best models for each scenario with an optimal number of clusters for each scenario.

It is expected that the different models fitted to satisfy each of the project aims will produce differing results, and therefore it is important to consider how these different models compare and if there is a single method that produces the best model to be applied to the ordinal outcome data.

The initial interest of this project was to look at the longitudinal analysis of stroke trial data, as an extension of the SO_2S project, from which the initial data on this project were collected. After obtaining the data and conducting research into the types of scales available, specific focus was drawn to the mRS, due to its ordinal nature with the inclusion of the category of death. The use of a score for death within the scale was particularly interesting, along with consideration of the impact of using the full scale in longitudinal analysis, rather than just the single 90 day (3 month) time point. The work conducted in this project, although specific to the recovery of stroke patients, may be applicable in other fields with ordinal outcome scales, specifically if those scales have a category for death included.

A more comprehensive clinical understanding of the treatment effect on outcomes after stroke may be better described using repeated measures analysis, if improvement or worsening is expected beyond the 90 day outcome.

2. An overview of stroke

This chapter will give a background to the main themes of the thesis – highlighting the specific nature of the acute condition of stroke, giving a brief overview to some of the many outcome assessments that are used in order to measure the impact of a stroke, and providing a wider context for the main aims, both theoretical and clinical, of the project. This chapter will also introduce the two randomised clinical trials from which the data used in the thesis were taken. It details information about the design and setting, how the data was collected and the information collected during follow-up.

2.1 Stroke

2.1.1 Defining a stroke

A stroke or cerebrovascular accident is an acute condition that occurs when the blood supply to the brain is obstructed. The World Health Organization (WHO) defines a stroke to be a "neurological defect of cerebrovascular cause that persists greater than 24 hours, or is interrupted by death less than 24 hours" (The World Health Organisation 1978). During a stroke the restriction of the blood supply to the brain can cause brain cells to die, which can result in brain damage, causing disability or even death. The resulting disability can have varying consequences and has the potential to affect all aspects of an individual's identity – personally, socially and physiologically.

A stroke is actually a collective term for two different types of brain injury, which are defined by the different causations. The first, an ischaemic stroke, is caused by a blood clot that reduces the flow of blood to the brain; the other, a haemorrhagic stroke, is where a blood vessel bursts and bleeds into the brain. Approximately 85% of all strokes are ischaemic in nature, usually due to a cerebral infarction (Intercollegiate Stroke Working Party 2012). There are several causes of an ischaemic stroke, including an embolism, thrombolysis, and a lacunar stroke. A related condition to a stroke is a transient ischaemic attack (TIA), which can be referred to as a mini-stroke. This condition does not fall within the remit of a stroke, as it usually lasts for less than 24 hours; however it can be an indicator of an individual having an increased risk of developing a stroke (Johnston et al. 2000, Lovett et al. 2003).

2.1.2 Epidemiology

Globally, 15 million people suffer from a stroke each year; of these, 5 million will die as a result of their stroke (Mackay & Mensah 2004). A stroke was reported to be the second leading cause of death worldwide after ischaemic heart disease in 2011 (The World Health Organisation 2013). In 2010, stroke was the fourth largest cause of mortality in the United Kingdom (UK) of people under 75, after cancer, cardiovascular disease and respiratory disease (Townsend et al. 2012). A stroke accounts for almost 6% of deaths in men and 8% of deaths in women in the UK (Stroke Association 2016). However, in recent years there has been a decreasing trend in mortality in early strokes, with a 46% reduction in the number of deaths due to stroke between 1990 - 2010 (Feigin et al. 2014), as more people are surviving stroke due to better health care and medical advances.

A stroke is the leading cause of adult disability in the UK. In 2005, there were over 900,000 people who had suffered a stroke living in England, with 300,000 of these living with moderate to severe disability (Department of Health 2005). Stroke causes a greater range of disability than any other disease. Of those who survive a stroke, approximately 42% will be independent afterwards, 22% will have mild disability, 14% will have moderate disability, 10% will have severe disability and 12% will have very severe disability, resulting in them needing constant nursing attention as they are unable to do anything for themselves (Stroke Association 2013). More than half of stroke survivors are left dependent on others for help in everyday activities and as a result of this approximately 11% of stroke survivors are newly admitted into a care home after hospital discharge (Royal College of Physicians, Clinical Effectiveness and Evaluation Unit 2014).

Strokes are largely preventable if an individual with an increased risk of stroke considers a lifestyle adjustment; however, there are some risk factors that increase the chances of having a stroke that cannot be altered. Patients who are over the age of 65 have an increased risk of having a stroke. In all stroke studies, it is clear that with an increase in age there is an increase in incidence and prevalence of strokes. Men are also at a greater risk of suffering from a stroke, with approximately a 25% higher risk of having a stroke than women (Townsend et al. 2012). People who have a close relative who has suffered a stroke or have a family history of either heart disease or diabetes are also at an increased risk. Ethnicity can also affect the chances of having a stroke, with those of south Asian, African and Caribbean descent at a higher risk; however, this is partly due to an increased risk of other factors within these ethnicities such as diabetes and high blood pressure (Stroke Association 2016).

2.1.3 Symptoms

The symptoms of stroke will vary from person to person, as they will depend on several factors, including the type of stroke, the area of the brain affected, and the severity of the stroke. Regardless of what symptoms an individual has, they are likely to come on very suddenly. Common symptoms include: numbness or weakness in the face, weakness in the arms and/or legs on one side of the body, confusion, difficulty speaking, headaches, loss of vision and dizziness. Less common symptoms include: nausea, fever, fainting and loss of consciousness. In the late 1990s a group of stroke physicians developed a training tool for ambulance personnel, using the acronym FAST, which stands for (F)acial drooping, (A)rm weakness, (S)peech difficulties and (T)ime, to identify quickly if a patient was having a stroke. Since its development it has helped to reduce the number of misdiagnosed strokes (Harbison et al. 2003). Recently in the UK, the National Health

Service (NHS) has used the FAST test in a national campaign to advertise the need to identify if someone is having a stroke quickly, so that prompt treatment can be given.

2.1.4 Diagnosis

Stroke treatment differs for the different types of stroke and so being able to identify what type of stroke a person is having quickly is one of the most important steps in stroke management. The use of computerised tomography (CT) and magnetic resonance imaging (MRI) scans have revolutionised this distinction. In the UK, all suspected stroke patients should receive a brain scan within 24 hours of stroke onset (NHS Choices 2012a). It may be necessary for further tests to be undertaken in order to identify the type of stroke, including a swallow test, as difficulty swallowing can lead to aspiration, which can cause further problems such as chest infections, including pneumonia (NHS Choices 2012a). Blood tests may also be undertaken to check blood sugar and cholesterol levels and pulse measurement is also used in order to detect an irregular heartbeat. To ensure a correct diagnosis, ultrasound scans may be carried out on the neck to check arteries to the brain or echocardiograms to look for blood clots around the heart (NHS Choices 2012a).

2.1.5 Interventions and prevention

The 10-point stroke action plan identified from the UK’s National Stroke Strategy in 2007 deems that the care a patient receives in a stroke unit is a principal determinant of a person’s outcome following a stroke. All stroke patients should have prompt access to an acute stroke unit with specialist care. An acute stroke unit is a specialist unit that provides high-dependency care, including physiological and neurological monitoring, rapid treatment of stroke and the associated complications, as well as early rehabilitation and palliative care (Department of Health 2007). There are different interventions available for the different types of stoke. Ischaemic strokes can be treated by clot-busting drugs including recombined tissue plasminogen activator (rt-PA) and alteplase; however, not all patients are suitable for this thrombolysis treatment. These treatments are most effective when given early, and so the number of patients benefitting from this treatment may be small given the short therapeutic time window. This treatment is not suitable for haemorrhagic strokes, which are usually treated by finding the cause of the bleeding in the brain and stopping it; surgery may also be considered, with decompressive surgery used to reduce brain swelling or a craniotomy to repair burst blood vessels (Stroke Association 2012).

After a patient has suffered a stroke, anticoagulants or antiplatelet medication may be prescribed in order to reduce the chance of blood clots developing and thereby reduce the risk of a further stroke, usually in the form of aspirin, a low-cost, widely available and easily administered medicine. Medication such as beta blockers may also be given to control high blood pressure and statins may be prescribed as a further preventative measure to help lower cholesterol levels if they are high (NHS Choices 2012b). Medication like this may also be prescribed to people who are classed as high risk of having a stroke in order to prevent the stroke occurring. However, the primary method of preventing a stroke is to control the risk factors that increase the chances of having a stroke; this can mainly be done by changing lifestyle factors. Factors that are considered to increase the risk of a person having a stroke are hypertension, smoking, high cholesterol, diabetes, excessive alcohol intake, obesity, lack of exercise, stress and atrial fibrillation (Stroke Association 2016).

2.1.6 Effects of stroke

Stroke can cause widespread disability and recovery can be a gradual process. The rehabilitation of each stroke patient will differ depending on severity of the stroke and the symptoms the patient exhibits. After a stroke, patients may suffer from psychological problems such as depression or anxiety disorders. Patients may feel anger or frustration about the impact of the stroke, and may find it difficult to control their emotions. Strokes may also result in cognitive impairments; a patient’s memory may be affected, most commonly short term, as well as affecting the ability of the patient to concentrate on tasks. The body may have difficulty processing information, leading to apraxia (the inability to execute learned movements), difficulty in decision making and social cognition. Physically, strokes can cause muscle weakness or paralysis, which leaves patients struggling to move around and can lead to balance and coordination problems, especially if a patient also has problems with vision or hearing, as vision loss in one or both eyes can also be a common effect of stroke. Fatigue is also an issue with stroke survivors, making patients feel like they are not in control of their recovery (NHS Choices 2012c).

2.1.7 Stroke complications

After suffering a stroke, there are several complications that may occur. Patients may develop blood clots and muscle weakness after being immobile for a long period of time. Being immobile for long periods of time can increase the risk of a patient developing blood clots in the deep veins of the legs and can also lead to muscle weakness and decreased muscle flexibility. If the stroke has affected the muscles used for swallowing, a patient may find it difficult to eat and drink. There is also a greater risk of them inhaling food into the lungs (aspiration), which can lead to pneumonia. Finally, some strokes affect the muscles that are used to urinate, leading to loss of bladder control. When this happens patients may be fitted with a catheter, which can itself cause complications such as urinary tract infections (Poisson et al. 2010).

2.1.8 Measuring stroke severity

The severity of a stroke can differ greatly. Some patients may recover within a day, whereas others may die as a result of the stroke. In order to meaningfully describe the recovery in stroke survivors, more sophisticated measures are required than simple dichotomous endpoints such as mortality or stroke recurrence. There are many outcome scales used to assess the impact of a stroke including scales that measure disability, impairment and handicap. These terms are often used interchangeably but have distinct definitions from the World Health Organisation. An impairment is defined as a loss of psychological, physiological or anatomical function. Handicap is defined as the disadvantage for an individual that prevents or limits fulfilment of a role that is normal. A disability is defined as the restrictions of ability of the ability to perform an activity within a range that would be considered normal. The term disability can be used generally as an umbrella term for any impairment of body function, limitation of activities or participation restriction (Bowling 2005). It is important that the scales are reproducible and that they capture at least one of the neurological defect, functional impairment and the psychological impact of the stroke not only at onset but also during the recovery process, in order to monitor patients and to be used for prognostic purposes. At present, the mRS (Rankin 1957), Barthel Index (Mahoney & Barthel 1965) and National Institutes of Health Stroke Scale (Brott et al. 1989) are the most routinely used.

2.2 Stroke scales

Stroke scales can be classed as clinimetric scales and functional impairment or handicap scales. There are many outcomes that can be used to compare patients after a stroke, such as mortality or quality of life, however, functional recovery is the most commonly used outcome, as strokes represent the leading global cause of adult disability. Functional outcome scales are usually ordinal scales that range from the worst possible state to the best. Questions are asked or signs and symptoms observed, with the responses quantified in different categories in order to produce a score on the scale. Alternatively, some scales rely on the person assessing the patient to decide what response best describes the patient.

It is recommended that an outcome assessment is undertaken at 3 months for trials intending to demonstrate sustained benefit of acute treatment in stroke (Lees et al. 2012). It is believed that a patient will have undergone most recovery in the first 3 months after the stroke. However a patients’ recovery may continue for up to a year and even beyond. Below is a summary of the different scales used to assess a patient after they have had a stroke. It is important to consider the validity, reliability, sensitivity to change and simplicity of the scale.

The validity of the scale assesses the extent to which it accurately measures the underlying concept of interest. Validity can be measured by assessing face validity, content validity and criterion-related validity. Face validity evaluates whether the scale is unambiguous and appropriate, content validity assesses whether the scale included all relevant concepts of the attribute that is being measured, and criterion-related validity compares how well the scale measures up to the 'gold standard' (Cozby & Bates 2012). In many cases, especially when looking at quality of life, this is very hard as there is not usually a recognised gold standard, and so criterion-related validity is very hard to assess.

Reliability refers to the ability to produce consistent results on different occasions, when it is known no change has occurred. Test-retest reliability looks at whether the results are the same after applying the scale repeatedly to the same population. Other ways reliability can be assessed are: internal consistency (measurement of the same concept by different scale items), inter-rater reliability (consistency of a measure when administered by different interviewers) and intra-rater reliability (consistency of a measure when administered more than once by the same interviewers). The sensitivity of the scale is the ability of the scale to detect clinically important changes in a patient's condition. The simplicity of a scale is also important, as a complex scale will reduce reliability.

2.2.1 National Institutes of Health Stroke Scale (NIHSS)

The NIHSS is an observational scale, developed in the early 1980s to provide consistent reporting of neurological deficits in acute stroke studies. It was designed to assess the differences between interventions given in clinical trials but is now increasingly used as an initial assessment tool, due to the fact that it provides a quick and accurate initial evaluation after a stroke, often only taking around 6 minutes to perform with no additional equipment needed (Brott et al. 1989). The NIHSS is a 15-item impairment scale that standardises and quantifies measures of the key components in a standard neurological examination, with the attention centred on those components that are most relevant to a stroke. The NIHSS provides a measure of acute impairments by assigning numerical values to various aspects of neurological function (Brott et al. 1989). The scale incorporates assessment of consciousness, language, motor function, sensory loss, visual fields, coordination, extra-ocular movements, neglect and speech. The items that are assessed along with the scores received are given in Table 2.1. Scores are summed across each category to give a total value that can range from 0 to 42, with 0 representing no impairment. A severe score on the NIHSS is often described as a score greater than 21 (Harrison et al. 2013). Different components in the score have different scoring values, with some ranging between 0-4 and other only taking a value of 0 or 1. The summative score cannot be considered truly ordinal, although a higher score indicates a more severe impairment, as it is possible to obtain the same score through different combinations of the items on the scale. There is a standardised approach to assessment using the NIHSS, usually assessing state of consciousness first, and guidance for when the patient is unable to respond to the command (Harrison et al. 2013).

The validity and reliability of the overall NIHSS has been widely noted (Muir et al. 1996). However, there are criticisms for its redundancy and complexity, as it contains items with poor reliability, which can lead to low inter-rater reliability (Ghandehari 2013). These poorer items include those relating to facial palsy, ataxia, dysarthria and level of consciousness (Lyden et al.

1999). There is a modified version of the NIHSS scale that maintains a similar structure, but has the items with poor reliability removed (Lyden et al. 2001). The modified NIHSS was intended to be easier to administer as it is an 11-item scale and has a simpler grading system. The reliability and validity of the modified NIHSS scale has been shown to be higher than those of the NIHSS scale (Lyden et al. 2001). Despite the advantages of the modified NIHSS scale (less inter-rater variability and simplification) it is not frequently used, and in terms of the length of the scale it is debatable whether a shorter scale is needed as the NIHSS is already quick and easy to perform. The NIHSS also has predictive validity, as an initial score is a robust predictor of in-hospital complications and outcome at 3 months (Johnston & Wagner 2006).

Table 2.1: National Institutes of Health Stroke Scale (NIHSS)

Domain assessed	Response	Score
Level of consciousness (LOC)	Alert; keenly responsive	0
	Not alert; verbally arousable or aroused	1
	Not alert; only responsive to repeated or strong stimuli Totally	2
	unresponsive; Responds only with reflexes	3
LOC questions (state month and age)	Answers both correctly	0
	Answers one correctly	1
	Both incorrect or no reply	2
LOC commands (open and close eyes, grip and release normal hand)	Correctly performs both tasks	0
	Correctly performs one task	1
	Performs neither task correctly	2
Best gaze	Normal; able to follow pen or finger to both sides	0
	Partial gaze palsy; gaze is abnormal in one or both eyes, but gaze is not totally paralysed	1
	Total gaze paresis; gaze is fixed to one side	2
Visual fields	No visual loss (or in coma)	0
	Partial hemianopia; no visual stimulus in a specific quadrant	1
	Complete hemianopia; no visual stimulus in one half of the visual field	2
	Bilateral hemianopia (blind from any cause)	3
Facial palsy	Normal symmetrical movements	0
	Minor; asymmetry on smiling	1
	Partial; total or near total lower face paralysis	2
	Complete; absence of movements in upper/lower face	3
Best motor arm (score either right or left)	No drift; hold limb at 90 degrees for 10 seconds	0
	Drift; drifts down but does not hit bed	1
	Some effort against gravity	2
	No effort against gravity	3
	No movement	4
Best motor leg (score either right or left)	No drift; hold limb at 45 degrees for 5 seconds	0
	Drift; drifts down but does not hit bed	1
	Some effort against gravity	2
	No effort against gravity	3
	No movement	4
Limb ataxia	Absent (or in coma)	0
	Present in 1 limb	1
	Present in 2 or more limbs	2
Sensory	Normal	0
	Partial loss; patient feels the pinprick but less sharp	1
	Dense loss; unaware of being touched (or in coma)	2
Best language	No dysphasia	0
	Mild dysphasia; obvious loss of fluency or comprehension	1
	Severe dysphasia; all communication is fragmented	2
	Mute; no usable speech (or in coma)	3
Dysarthria	Normal articulation	0
	Mild dysarthria; patient understood with some difficulty	1
	Unintelligible or worse (or in coma)	2
Neglect	No neglect (or in coma)	0
	Partial neglect	1
	Complete neglect	2

2.2.2 Modified Rankin Scale (mRS)

The Rankin scale was developed by Scottish physician John Rankin in 1957, in order to describe the positive outcomes seen in his prototypic stroke unit (Rankin 1957). Although it was not originally intended to be used as an assessment for clinical trials, a modified version of the Rankin Scale (mRS) was developed in 1987 (van Swieten et al. 1988). The mRS is a simple 6-point scale that assesses the impairment of the patient due to both the limitation in activities and the changes in a patient's lifestyle. It is a hierarchical ordinal scale that ranges from a patient having no disability at all to a patient having severe disability, where a patient is unable to look after themselves, requiring constant nursing care and attention. There are 6 scores on the mRS, with no disability scored at 0 up to severe disability scored at 5. Recently, a score of 6 has been included to represent death, used in clinical trials (Uyttenboogaart et al. 2005, Banks & Marotta 2007). The mRS has grown in popularity and is now the most popular primary endpoint in clinical trials of stroke (Banks & Marotta 2007). The scale is a measure of functional independence and is usually assessed under a structured interview.

A single point change on the mRS will always be clinically relevant; however, with fewer categories the mRS may not be very responsive to change. Previously, the reliability of the mRS had been found to be satisfactory, though the descriptions of the categories of the mRS were broad and open to subjective interpretation (Wilson et al. 2002) and therefore it was left open to the raters to develop idiosyncratic criteria. This led to low levels of inter-rater reliability as there was the suggestion some raters may systematically assign higher or lower grades than others, introducing bias. The use of a structured interview conducted to ascertain the mRS score of a patient helps to reduce the variability and bias between raters assigning MRS grades to patients, and therefore could potentially improve the quality of results seen in clinical trials, substantially improving reliability (Wilson et al. 2002). Another modification to the mRS assessment calculates a pre-stroke score. However, the wording of the mRS is not suited to such an assessment and so unsurprisingly it only has moderate validity and reliability when so applied (Fearon et al. 2012). Another modification is to use proxies, usually an informant who knows the patient, to supplement or complete the interview, as stroke survivors may have physical or language problems that would complicate the standard interview (McArthur et al. 2013). Although this method is intuitive, there is suboptimal validity and reliability, and so the standard interview for the mRS is the preferred assessment. A third version of the mRS has also been developed, the Oxford Handicap Scale (OHS), which keeps the same categories but includes information on lifestyle factors to allow the scale to measure the handicap of the patient rather than the disability (Bamford et al. 1989).

Table 2.2: Modified Rankin scale (mRS)

Level	Description
0	No symptoms
1	No significant disability despite symptoms; able to carry out all usual duties and activities
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability; requiring some help, but able to walk without any assistance
4	Moderately severe disability; unable to walk without assistance, unable to attend needs without assistance
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention
6	Dead

2.2.3 Barthel Index (BI)

The Barthel Index was developed in 1965 by Dr Florence I. Mahoney and Dorothea W. Barthel by adapting the Maryland Disability Index to produce "a simple index of independence, useful in scoring improvement in rehabilitation" (Mahoney & Barthel 1965). It was developed to assist with discharge planning in long-term wards. The scale describes 10 basic aspects relating to self-care and mobility, with scores given according to the amount of time or assistance required by the patient. The scale ranges from 0–100, as it is usually summed to give a total score, with higher scores representing greater independence. There are several different definitions of a good outcome on the BI. The most common definition was determined using statistical modelling, where a good outcome is defined as a BI score of greater than 80, where patients are generally independent. An excellent outcome uses a score of 95 or greater as the optimal choice with less than 75 being scored as a poor outcome (Uyttenboogaart et al. 2005).

The BI is now the most popular activities of daily living (ADL) scale in clinical practice (Wade & Collin 1988). It is used repeatedly to assess improvement over time and the score is determined by interviews and distant observation of the patient in tasks (Quinn et al. 2011). The BI score is used to describe what a patient can do at the time of grading. It is second only to the mRS as a stroke outcome measure of choice; however, unlike the mRS the BI does not have a separate score to represent mortality, considered to be the poorest outcome after a stroke. The BI has been adopted by many disciplines other than stroke including nursing, spinal injury, burns, cardiac disease, rheumatoid arthritis and amputations (Quinn et al. 2011), and is a recommended assessment in older adult care (Harrison et al. 2013).

Validity of the scale is well described, and it is recognised as a valid prognostic tool for predicting recovery and the level of care required (Huybrechts & Caro 2007). Inter-observer reliability is good, and reasonable reliability has been observed. The BI is limited in its ability to represent changes throughout the spectrum of potential outcomes, and is susceptible to both a floor and ceiling effect, where the score does not change from the minimum or maximum score despite clinical change (Schepers et al. 2006).

Table 2.3: Barthel Index (BI)

Domain assessed	**Response**	**Score**
Bladder	Incontinent (or catheterised)	0
	Occasional accident	5
	Continent	10
Bowels	Incontinent	0
	Occasional accident	5
	Continent	10
Toilet use	Dependent	0
	Needs some help	5
	Independent	10
Grooming	Needs help	0
	Independent	5
Dressing	Dependent	0
	Needs some help	5
	Independent	10
Feeding	Unable	0
	Requires assistance	5
	Independent	10
Bathing	Dependent	0
	Independent	5
Transfer (bed to chair and back)	Unable	0
	Major help	5
	Minor help	10
	Independent	15
Mobility (on level surface)	Immobile	0
	Wheelchair independent, including corners	5
	Walks with help	10
	Independent	15
Stairs	Unable	0
	Needs help	5
	Independent	10

2.2.4 Nottingham Extended Activities of Daily Living scale (NEADL)

The Nottingham Extended Activities of Daily Living scale is an instrumental ADL scale, which assesses the level of activity actually performed by a patient (Nouri & Lincoln 1987). It was designed in 1987 by Nouri and Lincoln, an occupational therapist and a clinical psychologist respectively, at the Stroke Research Unit, Nottingham. The scale considers 22 activities, which can be divided into 4 subscales: mobility, kitchen, domestic and leisure activities. There are 4 possible responses (0-not at all, 1-with help, 2-on my own with difficulty, 3-on my own). Respondents are asked whether they actually undertake the activity rather than whether they are able to do so. The NEADL scale has been validated for postal use, and it was designed for this purpose (Nouri & Lincoln 1987).

Originally, items in the scale were designed to be a dichotomous scale, with responses 0 and 1 combined to represent patients being dependant on others [0]. Patients who scored 2 or 3 were classed as independent [1], leading to a summary score out of 22. There has recently been a movement to give scores based on the responses (1, 2, 3 and 4), looking at summative subscales and total scores. There is inconsistent use of both scoring methods, although the validity of the total summative score has been questioned. Analysis has been conducted to show that using the responses (1, 2, 3 and 4) is not recommended as the scale is not unidimensional, but Rasch analysis conducted supported the use of the 4 subscales (das Nair et al. 2011). The NEADL compares favourably to the BI and is less susceptible to the latter's ceiling effects (Sarker et al. 2012).

Table 2.4: Nottingham Extended Activities of Daily Living scale (NEADL)

Domain Assessed	Questions asked	Score
Mobility	Do you walk around outside? Do you climb stairs? Do you get in and out of the car? Do you walk on uneven ground? Do you cross roads? Do you travel on public transport?	0-not at all 1-with help 2-on my own with difficulty 3-on my own
In the kitchen	Do you manage to feed yourself? Do you manage to make a hot drink? Do you take hot drinks from one room to another? Do you do the washing up? Do you make yourself a hot snack?	0-not at all 1-with help 2-on my own with difficulty 3-on my own
Domestic tasks	Do you manage your own money when out? Do your own shopping? Do you wash small items of clothing? Do you do a full cloths wash?	0-not at all 1-with help 2-on my own with difficulty 3-on my own
Leisure activities	Do you read newspapers and books? Use the telephone? Do you write letters? Do you go out socially? Do you manage your own garden? Do you drive a car?	0-not at all 1-with help 2-on my own with difficulty 3-on my own

2.3 Comparison of stroke scales

The choice of which stroke scale to use for assessment must be made on the basis of the question of interest and the evidence base around its clinimetric properties. There is no ideal stroke measure and so the relative strengths and limitations of each assessment strategy must be considered, in order to find the most suitable. Within the three core assessment scales, the mRS would be recommended when looking at disability, BI when considering basic ADL, and NIHSS when looking at the neurological impact of stroke, with training available for each assessment scale. There are strong correlations among the three most widely used outcome measures (Lees et al. 2012). With 3-month mRS, the Spearman rank correlation was 0.94 with the BI and 0.91 with the NIHSS (Lees et al. 2012), when assessed in 9,275 patients from Virtual International Stroke Trials Archive (VISTA) (Ali et al. 2007). The correlation between NIHSS and BI was 0.85. Thus, there are strong correlations between the outcome measures, especially if examining the full range of the scales. The mRS is hierarchical in nature; however, this is not true for either the NIHSS or the BI scale, as a particular score may be attained through various combinations of sub items that are not necessarily equivalent. Improvements in reliability can be seen on the mRS with the introduction of a structured interview (Wilson et al. 2002).

2.4 Stroke Oxygen Study

The main data used in the thesis are from the Stroke Oxygen Study (SO_2S). The aims of the thesis were designed with these data in mind. Although each chapter will give a more detailed explanation of the data used from the trial for each analysis, an overview is given below of the trial design, data collection and response to follow-up.

2.4.1 Design and setting

This trial was a multi-centre randomised controlled trial of oxygen supplementation in patients with acute stroke. The main trial hypothesis was that a fixed dose of oxygen-treatment during the first 3 days after an acute stroke improves outcome, with a primary outcome of the mRS 90 days post-stroke. Further hypotheses extended the times at which the outcome was recorded to 6 months and 12 months post-stroke (Roffe et al. 2014).

This thesis considers the values of the mRS recorded during the whole follow-up period. In the trial there are values of the mRS, BI and NEADL scores recorded at 90 days (3 months), 6 months and 12 months post-stroke.

Patients in the study were randomised, based on covariates from the baseline assessment, using minimisation to one of three groups in a ratio of 1:1:1:

1. Continuous oxygen supplementation for 72 hours
2. Nocturnal oxygen supplementation for 3 nights
3. No routine oxygen supplementation

Individuals who were randomised to receive oxygen were given it at a rate of 2 to 3 l/min, depending on baseline oxygen saturation.

2.4.2 Data collection

A pilot study for the SO_2S was conducted between July 2004 and April 2008, recruiting 300 patients (Roffe et al. 2011, Ali et al. 2013). Following the implementation of minor amendments, the main study began recruitment in April 2008. Patients were recruited from multiple (>30) centres throughout the UK and worldwide. Medical centres were eligible for participation in the study if they admitted patients with acute stroke, were able to provide oxygen treatment and monitor oxygen saturation, and if there was a local researcher who was willing to act as the principal investigator for the locality. All adult patients who were admitted to one of the centres with symptoms of an acute stroke within the preceding 24 hours were eligible to be considered for participation in the study, as long as there was no indication for, or contraindication to oxygen treatment in the doctor's opinion (Roffe et al. 2014).

Patients were not eligible for inclusion in the trial if the responsible doctor considered the patient to have definite indications or contraindications to oxygen treatment. This decision was left to the responsible doctor and used to ensure best clinical practice. Patients were also excluded if the stroke was not the main health problem, or they had another serious life threatening illness that was likely to lead to death. These patients were excluded as they were unlikely to receive any benefit from the study intervention. The eligibility criteria for inclusion in the trial reflected this uncertainty about who should receive oxygen treatment and for how long. This allowed as many patients as possible to be recruited into the study.

The initial assessment was conducted at baseline by the researcher randomising the patient. It included the baseline demographics along with date and time of event, the Glasgow Coma Scale, predictors of outcome, and the NIHSS. An assessment one week after the stroke was conducted in order to confirm diagnosis. It was performed by a trained assessor seven days (± one day to allow for weekends and holidays) after enrolment. The assessment documented deaths and neurological status (NIHSS), compliance with the intervention, and complications arising during the first week.

Follow-up at 3 months, 6 months and 12 months was conducted centrally in order to ensure blinding of the assessors. The assessment was a questionnaire sent out to patients at each specific time point, after checking that the patient was still alive. Non-responders were contacted and were able to answer the questions by telephone where possible in order to reduce the amount of missing data and loss to follow-up. The follow-up questionnaire the patient received contained the assessments for three stroke scales the mRS, BI and NEADL. As well as this patients were assessed using the EuroQol five dimensions questionnaire (EQ-5D), which provides a single index value for health status as well as being asked questions regarding memory, sleep, speech and discharge status.

2.4.3 Response during the follow-up

The original sample size for the study was 6,000 participants. With allowance for a maximum of 10% loss to follow-up, the trial had a recruitment target of 6,669 (Roffe et al. 2014). The recruitment target was subsequently revised in October 2012 to 8,000 patients, to provide greater power to detect an interaction between stroke subgroups (defined by severity) and the effect of oxygen versus control.

In total there were 8,003 individuals that were recruited to the study between April 2008 and June 2013. All patients had full information recorded at baseline as this information was used in the minimisation process. There were 129 patients that had died within the first seven days of the study.

Follow-up questionnaires were sent to individuals with a series of questions that allowed the mRS, BI and NEADL to be calculated. At 3 months, 7,370 (92%) patients in the trial were still alive, of which 6,936 (94%) of these patients responded to the 3 month questionnaire. At 6 months 7,167 (89%) patients were still alive and 6,594 (92%) of these patients responded to the 6 month questionnaire. At 12 months, 6,957 (86%) recruited patients were still alive and 6,020 (87%) of these patients provided responses to the questionnaires. Unlike the other data set included in the thesis, where multiple imputations had previously been conducted before the data was obtained, no values of the SO_2S data set were imputed. This is because all methods that are used for analysis throughout the thesis are able to deal with missing data, and there are some concerns about the validity of imputing the outcome variable. One if the issues with the missing data for the outcome in this case is that there is more than one measurement for each individual, which may lead to the introduction of bias within the imputation strategy (Sterne et al. 2009). Also in order to compute the outcome, it is recommended that there is complete data for all covariates that affect the outcome, therefore caution should be taken imputing the outcome variable, and it was decided not to do this for the SO_2S data.

2.5 National Institute of Neurological Disorders and Stroke Trial (NINDS)

The second dataset that was used in the thesis was the National Institute of Neurological Disorders and Stroke Trial (NINDS) trial data. This was used because it had the outcome of interest recorded at more time points than were available in the SO_2S study dataset. This was important for the analysis in chapter 7 specifically. This information for the NINDS trials was found using papers that had previously conducted analysis of the trial data (Kwiatkowski et al. 1999, Li et al. 2010) any more information needed on the NINDS trial can be found in the original paper (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group 1995).

2.5.1 Design and setting

The NINDS trial was a clinical trial in acute stroke therapy that was conducted to compare the effect of tissue plasminogen activator (t-PA) versus placebo in patients with acute ischaemic stroke. The main hypothesis was to test whether t-PA reduced morbidity and mortality in stroke patients. The primary outcomes considered were an improvement by more than 4 points on the NIHSS or complete resolution at 24 hours and also the proportion with a favourable outcomes at 90 days after stroke.

The trial looked at 4 different outcomes: BI, mRS, Glasgow Outcome Scale (GOS) and NIHSS. The BI and GOS are recorded at 90 days, 6 months and 12 months post-stroke. The mRS was recorded at 7 days, 90 days, 6 months and 12 months post-stroke. Finally the NIHSS was recorded at baseline, 24 hours 7 days and 90 days post stoke.

Patients were randomised 1:1 to two treatment groups using a permuted-block design with blocks of various sizes used for randomisation. Patients were stratified in the randomisation according to clinical centre and time of onset of stroke and received one of the following treatments:

1. Placebo
2. A recombinant t-PA (rt-PA), in a dose of 0.9 mg per kilogram of body weight (with a maximum dose of 90 mg), 10% of which was given as a bolus followed by delivery of the remaining 90% as a constant infusion over a period of 60 minutes.

2.5.2 Data collection

The study was split into two separate parts. The first part assessed the change in neurological deficit in 24 hours and recruited 291 patients. The second part looked at favourable outcomes at 90 days and recruited 333 patients. Both parts of the trial completed the 12-month follow-up and so can be considered as one trial, which recruited 624 patients, with 312 patients randomised to t-PA and 312 patients randomised to placebo.

The study was conducted between January 1991 and October 1994. Patients were recruited from 8 centres across the USA. The inclusion criteria stated patients must have suffered an ischaemic stroke with a clearly defined time of onset (in order for treatment to be administered within 3 hours), a deficit measurable on the NIHSS, and a baseline CT scan of the brain that showed no evidence of an intracranial haemorrhage.

Some of the exclusion criteria were previous stroke within 3 months, major surgery in the previous 14 days, a seizure at the onset of stroke and the use of anticoagulants or an antithrombotic within 48 hours preceding stroke onset. A full list can be seen in the original paper (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group 1995).

The initial assessment was conducted at baseline by the researcher randomising the patient. It included the baseline demographics along with date and time of event, and an indication of severity, including a NIHSS score.

The protocol defined that individuals were followed up for 12 months post-stroke, even though the primary outcomes for both parts of the trial were assessed at 90 days post-stroke. Despite this, information regarding each individual in the trial was recorded at 24 hours, 7-10 days, 3 months, 6 months, and 12 months post-stroke. There were 3 outcome scales considered at each of these time points, the BI and Glasgow Outcome Scale are recorded at 3 months, 6 months and 12 months post-stroke. The mRS was recorded at 7-10 days, 3 months, 6 months and 12 months post-stroke.

At 24 hours post-stroke and at 3 months the outcomes of interest were determined by a certified examiner who had not been present during the initial treatment, and who also had not performed the baseline assessment on the individual.

2.5.3 Response during the follow-up

In total there were 624 individuals recruited as parts of the NINDS trial, of these 291 were recruited as part one and 333 were recruited as part two, giving the trial 624 individuals that were randomised, 312 in each treatment group. At the initial assessments and for the primary endpoint of the trial (3 months) full information is detailed for each individual. At 6 months there were 15 individuals for whom information was not available, approximately 2.4% of the data, and at 12 months there were 26 individuals (4.2%) for whom information was not available. For the individuals with missing data the trial used multiple imputations in order to give individuals complete values during the follow-up period. There were 5 individuals who had missing data recorded at 7-10 days; these values were not imputed in the dataset.

2.6 Conclusions

This chapter aimed to give an overview of the acute condition stroke and introduce the main assessment scales that can be used to assess a stroke. As well as this the two datasets that are to be used in the thesis have been introduced. Although the trials were conducted 10 years apart, the thesis is less interested in the results and conclusions and more interested in the methods that can be applied to longitudinal data and the treatment effect they produce.

From this overview of stroke and the assessment scales, it was decided to focus on the mRS as it is of particular interest having a score for death included. The next chapter goes on to conduct a systematic review to look at what research has already been conducted in longitudinal stroke trials, both generally and specifically with the mRS, in order to identify where the research conducted in this thesis can bring new knowledge.

3. Current literature surrounding stroke trials

This chapter looks at the current literature surrounding stroke trials and the methods used to analyse them. Three systematic reviews were undertaken; firstly, to investigate methods used in recently published stroke trials when considering studies that follow up individuals over a period of time and report multiple follow-up points, and secondly to look at the different methods that have previously been applied to the mRS in longitudinal studies in stroke research. A third systematic review was conducted to look at methods that had been identified, and their application in other clinical trials. The literature was also explored to investigate other outcome assessment scales, similar to the mRS, which include a category for death like the mRS.

3.1 The longitudinal analysis of stroke trials.

A longitudinal study is a study that involves measuring the same variables repeatedly over a specified period of time. Unlike cross-sectional studies that compare individuals at one time point, longitudinal studies take each individual and track him or her over a continuous period of time. The observed data from monitoring an individual are known as repeated measures data and form the basis of the analysis.

Clinical follow-up studies are longitudinal in nature; where a disease is monitored systematically in order to establish how the illness progresses over time and what influences their prognosis (Coggon et al. 2007). Most follow-up studies document the characteristics of each subject when they are recruited to the study (such as age, sex, duration and severity of symptoms) so that the association of these variables with prognosis can be examined. Longitudinal studies can have either prospective (cohort) or retrospective (case-control) designs, as well as observational and experimental designs.

In the literature, it is very common to find longitudinal data as panel data, as although the patients are being followed up over time, the points at which they assessed are usually predefined and therefore the status of each individual is only known at specific time points during the follow-up period. Panel data records cross-sectional units at repeated time points and can be either balanced (observations occur at same time periods) or unbalanced (observations occur at differing time periods.

Longitudinal analyses are often used to look at trends across a life span, or life events over generations, and therefore can mainly be found in research in psychology and sociology. There is little research conducted using longitudinal methods in stroke trials, even though the trial design may incorporate repeated assessments of patients over a follow-up in order to test the effect of an intervention.

3.1.1 Introduction

A systematic review of the literature was conducted looking at all stroke trials that had published results since 2014 to investigate whether the trial had outcomes that were repeatedly measured over a follow-up period and, if so, whether the trials used techniques that analysed these data appropriately. To do this for all stroke trials that had ever been published would be a formidable task. This review focuses specifically on trials that have been conducted in stroke research recently, with articles sourced from papers that have been published from January 2014 onwards. By looking at papers that have been published from 2014 onwards, the trials identified can be considered as using methods that are current practice. Also, recently more emphasis is being placed on statistical methods used in stroke research in order to ensure the best use of all available outcome data (Bath et al. 2012). It would therefore be expected that more recent publications would use methods appropriate for longitudinal data, and consider using the chosen outcome scale in a form that uses the most information available.

3.1.2 Methods

3.1.2.1 Search Strategy

A search was conducted in Medline and Embase in November 2015 to identify randomised controlled trials in stroke research that had been published since January 2014. In order to make sure that all possible articles were captured, the search terms in the search strategy included "Stroke", "Randomised", "Randomized" and "RCTs". As well as looking at articles that explicitly mention the search terms in the title, articles that were tagged as stroke or randomised control trials were also identified. Article tags are applied to articles in order to classify the contents and help to make them easier to identify.

Search terms were kept as broad as possible in order to try to capture all relevant articles. Once the search had been conducted, several restrictions were placed on the articles including the date published from (January 2014) and also restricting the articles to trials that were conducted in human subjects and that the text was available in English. After articles that did not fulfil these criteria were discarded, the references that remained were investigated further and an inclusion criterion was defined in order to keep only those articles that were relevant to the aim of the review. After this, the references of the chosen articles were searched in order to ensure completeness of the review.

3.1.2.2 Inclusion Criteria

Initially, the titles of all the articles found were reviewed and those that were obviously not appropriate were removed from the list. After this abstracts were reviewed followed by full texts of the articles if they were identified to be of interest. Each article was checked to see if the trial considered recorded the outcomes at multiple time points or if only a single time point was considered. If the outcome was only recorded at one specific time point, these articles were discarded as the interest of the review lies in those that have multiple time points recorded. Once the articles that had multiple time points had been identified, the type of analysis was considered as to whether the trials utilised all the time points by conducting a longitudinal analysis. The different longitudinal methods used by the different trials were noted. Trials were not considered if the study population had not had a stroke, if the studies were looking at the risk of having a stroke, or if having a stroke was an outcome these trials. Additionally, articles had to refer to a stroke trial with results; protocols and other pre-trial articles were excluded.

3.1.3 Results

3.1.3.1 Search strategy

The search conducted in Medline and Embase identified 191,187 articles that had used terms that were specifically related to stroke and 428,256 articles that used terms that were related to randomised controlled trials. After these two searches were conducted separately, they were combined to look for articles that were relevant to both searches and this produced 9,649 articles, after the removal of duplicates. Restrictions were placed on the search to exclude trials conducted on animals and to consider only those that looked at human subjects, which reduced the number of articles down to 9,577.

Finally, restrictions were placed on the search to look at articles published from January 2014 up to the date of the search. This greatly reduced the number of articles to consider and there were 1,396 articles left that were suitable to consider in more detail about the time points at which the outcomes were recorded and the methods that were used to analyse the articles. Appendix A contains a flow diagram detailing the articles that were excluded at each stage of the screening.

3.1.3.2 Articles found

Initial screening identified 499 studies that were clearly not conducted on stroke patients from the title of the articles and so they were instantly discarded. The abstracts of the remaining articles were considered in more detail. Out of the remaining studies, 180 were excluded as they did not report the results of a stroke trial. A full breakdown of what these excluded articles were looking at is given:

- Study protocol (n=79)
- Commentary/discussion (n=16)
- Study design (n=13)
- Rationale (n=11)
- Secondary analysis (n=7)
- Cost effectiveness (n=7)
- Review (n=7)
- Other study type (n=7)
- Methodology (n=7)
- Efficacy & safety (n=6)
- Feasibility (n=5)
- Meta-analysis (n=4)
- Statistical analysis plan (n=3)
- Trial design (n=3)
- Baseline characteristics (n=2)
- Study update (n=2)
- Recruitment strategy (n=1)

This left 717 studies with abstracts to be checked, of these, a further 196 were found to not be eligible, even though the word stroke featured in the title. This is because a stroke may have been an outcome or the study was looking at the risk of stroke in a certain population. As well as this, the information was unclear in the abstracts of 16 studies, and full texts were not able to be found for them as they were conference abstracts with no published article. Three studies were not able to be considered due to the main article being written in a non-English language.

This left 505 articles that fitted all the inclusion criteria. After these articles had been identified, each abstract and full text were checked to see if the study had the outcome in the study recorded only at a single time point or at multiple time points in a follow-up period; the numbers of such studies were 329 and 176 respectively. At this point, the type of analysis was not considered.

The articles recording more than one time point were then further assessed, and articles were excluded if the analyses only considered a single time point, ignoring the repeated measures. Of the remaining 176 articles, 110 conducted some type of longitudinal analysis, and 66 did not. This shows that over a 1/3 of the eligible stroke studies that followed up the patients for several months only considered one time point in their analyses. These articles had the references checked in order to identify any more potential articles for inclusion, but none were found within the considered time period.

There were eight different statistical methods that were applied to the different outcomes considered in the stroke trials. There was no restriction put on the outcome of the study and so different methods will have been used that were appropriate for the outcome of interest. The methods used included repeated measures Analysis of Variance (ANOVA), linear mixed models, Cox regression and Kaplan-Meier curves, ordinal regression, Freidman test, logistic regression, latent class growth analysis and functional principal components analysis. There were two studies that conducted both repeated measures ANOVA and linear mixed modelling; these were counted twice in the list below, which shows the number of studies using each of the statistical methods.

- Repeated measures ANOVA (n=55)
- Linear mixed models (n=21)
- Cox regression and Kaplan-Meier curves (n=31)
- Freidman test (n=3)
- Latent class growth analysis (n=1)
- Functional principal component analysis (n=1)

Repeated measures ANOVA is an extension of a one-way ANOVA, however, the groups are assumed to have a relationship between the outcomes, due to the outcome measure being repeatedly assessed for each group. The aim of the repeated measures ANOVA is to detect a change in the repeated outcome. In order to conduct this analysis, the outcome variable considered needs to be continuous and the intervention variable (within-subjects factor) nominal or ordinal. The repeated measures ANOVA looks at the mean change in the outcome variable over two or more time points, commonly it is three or more time points as if there were only 2 time points a t-test would be used. A repeated measures ANOVA cannot be applied in the case of the mRS, as the outcome variable needs to be continuous in order to compare the mean values. Therefore, this method is not appropriate for use within this project, but may be useful if considering other stroke scales such as the NIHSS, which can be treated as continuous.

Linear mixed models are an extension of generalised linear models (e.g. logistic regression). The term mixed model is derived from the fact that the models can include both fixed and random effects. They are particularly useful in a longitudinal analysis when the same outcome measure is assessed at repeated time points, as they are able to deal well with missing data, which means they are preferable to the repeated measures ANOVA. Linear mixed models are able to be applied to both binary and ordinal data. This is useful, as by using these methods we are able to consider the effect of dichotomising the mRS scale and contrast the estimates produced from the dichotomised and the full ordinal scale by comparing the linear mixed models fitted to the data. These methods are well established and aim to model the odds of a patient being in a specific category on the scale depending on the covariates that are included in the model. Linear mixed models are applied to the data within this project.

Cox regression uses time to event data, and so is less interested in the simple fact that a specific outcome has occurred, but with when the outcome occurred. The most common use of Cox regression analysis is in survival analysis when considering how long patients survive after a treatment or diagnosis, for example. One of the main interests in the mRS lies in the fact that it includes a point on the scale for death, which can therefore can be considered as part of the main analysis. Cox regression would be of interest if we were to consider death separately; however, our main outcome is an ordinal scale, not a time to event variable. It is possible to consider joint models that combine both the ordinal scale and a time to event variable, and this was considered as a potential method to be included in the project. However, the series of follow-up points is very small within the data collected, making this type of method hard to conduct. It was decided not to split the mRS up and consider death separately, but to keep the mRS as a whole scale and apply methods that are appropriate to the ordinal scale including death. Because of this Cox regression models were not fitted to the data in this instance, but the can be used in future if mortality is the outcome of interest.

Several papers used a Freidman test, which is a non-parametric statistical test that is comparable to repeated measures ANOVA. Its method is similar to a Kruskal-Wallis test, which compares two or more independent samples. As a non-parametric test, it is not assumed that the data are normally distributed and the data are ranked rather than using the raw scores. This method would be appropriate to apply to the mRS data as the Friedman test requires the data to be in an ordinal form. The Friedman test uses a ranking system and finds a critical value in order to accept or reject the null hypothesis. There need to be at least 3 observations in order for the test to be conducted, and if the result is significant then you need to run a second analysis in order to find out where the differences actually lie.

These were the most popular methods that were used by papers where the outcomes had been recorded at multiple time points. One study conducted functional principal components analysis (FPCA). FPCA demonstrates how a set of functional data varies with regard to its mean, and in terms of these modes of variability; it is commonly used for studying race walking (Dona et al. 2009). This study identified in the review was a rehabilitation study looking at stroke patients regaining independence after stroke. Although this is an interesting and novel method of analysis, we are not able to use it in this project as it cannot be used to analyse the mRS, or any of the other outcomes that are recorded in the data set.

The final method that was identified in the review was a latent class growth analysis (LCGA). This method works by clustering patients into groups that have similar profiles over time. This is a novel method for longitudinal analysis as there has only been one stroke trial published in the time period considered that used the method. It can be applied to both continuous and ordinal data and therefore is an appropriate method for use with the mRS. This method is considered later in the thesis, along with a more detailed review of other areas that have used the method within a randomised controlled trial.

3.1.4 Discussion

Although there were many articles that needed to be examined, it was found that only 505 of 1,396 found were deemed to be suitable for consideration, once non-English, non-stroke and non-trial articles had been removed. Of these, there were 176 articles that had multiple follow-up time points and over two-thirds of the articles used a method that allowed the repeated measures nature of the follow-up to be considered. This means that there were nearly 1/3 of trials that recorded measurements at several time points but failed to conduct an appropriate analysis to make the full use of the data that they had. A reason for this may be that there is no consensus on the methods that should be used for serial measurements and therefore it is easier to just consider a single time point. There are advantages to using the full longitudinal data over just the single time point; the sample size would be greatly increased when using the full data rather than just the single time point, and there would be greater statistical power with the multiple time points. There are increased costs associated with following individuals up over time; therefore, it makes sense that the analysis conducted is appropriate and makes use of all the data rather than just a single time point, for which it was not necessary to follow up the individuals frequently in the specified time period after the intervention has been given. This highlights the importance of the aims of this thesis as it will hopefully provide a consensus on the best statistical methods to use when analysing longitudinal data from studies of stroke.

3.2 Analysing the mRS in stroke trials.

Recently there has been much research conducted on how best to analyse the outcomes of stroke trials with particular emphasis on the mRS (Bath et al. 2012). However, the main interest of the majority of the research conducted has been the analysis of the mRS as the primary outcome at a single time point rather than longitudinal analysis methods.

The Optimising Analysis of Stroke Trials (OAST) Collaboration conducted work in 2007 that looked at ways in which the statistical analysis of stroke trials can be improved (Bath et al. 2007). They assessed which statistical approaches were most efficient when analysing the functional outcomes of stroke trials. The outcomes they considered were the mRS, the BI, and "Three questions". The three questions were developed to see whether simple questions could assess outcome after stroke (Lindley et al. 1994), and asked:

- Is the patient alive?
- In the last two weeks did you require help from another person for everyday activities?
- Do you feel you have made a complete recovery from your stroke?

The OAST collaboration used a variety of different approaches to re-analyse the functional data comparing two treatment groups. They obtained 55 datasets, which was comprised of 47 trials with approximately 54,000 patients between them. They found that methods that kept the functional outcome data in its ordinal form were more statistically efficient than those methods that dichotomised the data. The results were consistent across the different outcome measures considered. The trials that were analysed in this study looked at the functional outcome at a single time point, with methods including ordinal logistic regression, t-tests and robust ranks tests as the most favourable methods.

More work was conducted in 2012 as part of the European Stroke Organisation Outcomes Working Group to look at the statistical analysis of the primary outcome in acute stroke trials (Bath et al. 2012). They focused on the different methods that could be used to analyse ordinal data without the use of dichotomisation, as this can rarely be recommended using the mRS. They concluded that preferred statistical approaches to analyse the mRS could include using a sliding dichotomy on the scale or treating the scale as either ordinal or continuous. However, they conclude that there is no single best approach that will work for all stroke trials and that there will be some methods that are better than others depending on the scenario and the question that is being considered.

More recently a systematic review was conducted to look at the analysis of the mRS in acute stroke trials (Nunn et al. 2016). The primary aim of the review was to assess whether the recommendations given by the OAST collaboration in 2007 had had any influence on the methods used to analyse stroke trials since 2007.

The review identified 42 clinical trials for inclusion in the review, with approximately 32,500 patients in the incorporated studies. They found that although all reported the mRS, the primary outcome differed between published studies, with only 24 studies reporting the mRS as the primary outcome alone. Out of the 42 studies, 25 (60%) used a dichotomous analysis compared to only 8 (19%) using an ordinal analysis. There were 9 (21%) studies that did not fall into either category.

The studies that used an ordinal analysis used several different methods. Three studies used ordinal logistic regression, five studies used the Cochran-Mantel-Haenszel test, four used a test based on the normal distribution, and one used a Mann-Whitney U test. It is important to note that there were 12 additional studies that used ordinal methods as a secondary or sensitivity analysis. Once again, the main focus of the review was the mRS at a single time point, but it shows that despite the recommendations of OAST collaboration, there was not a dramatic shift in the types of analysis performed when analysing acute stroke trials. In 2007, the included studies in the OAST project paper had 18 studies that used the mRS as the main outcome, with 30% using an ordinal approach and 50% using a dichotomised approach. If we compare this to the findings of Nunn et al we see that only 19% of included studies use the mRS as an ordinal scale. It is important to note the original OAST project aimed to combine trial data from stroke RCT's and therefore there may have been other studied that satisfied the criteria of the search by Nunn et al. that were not originally included, due to the study not satisfying the inclusion criteria for the OAST project. This means that the numbers of ordinal mRS papers between the OAST project and systematic review may not be comparable.

3.3 Longitudinal analysis of stroke trials specifically using the mRS

Many stroke trials tend to look at the psychosocial changes over time rather than functional recovery. Psychosocial factors could be mental problems such as depression or social problems such as personal relationships. Those studies that do look at functional recovery tend to be focused on a particular area (i.e. recovery in the hand) rather than general recovery. This may be because recovery is so specific to the individual that it is hard to obtain a global measure of recovery. The most common scale recorded in stroke trials is the mRS and therefore this is the scale that we are most interested in. I am specifically interested in this scale because of the frequent inclusion of death in the scale, and I was interested in how this was dealt with in the longitudinal analysis. In order to identify all possible methods that could be applied to the mRS, the restriction of looking in clinical trials was removed when compared to the first search of the literature, with restrictions placed to look only for longitudinal studies or analysis. This was to give the broadest possible selection of studies that use longitudinal analyses with the mRS scale, to help inform potential methods for inclusion in this project.

3.3.1 Introduction

A review was conducted to look specifically for stroke studies that used the outcome of mRS for the longitudinal analysis that was conducted. In order to make sure that all possible papers were found we extended the search to also look for articles that used the OHS as an outcome as well, as this is just an extension of the mRS.

3.3.2 Methods

3.3.2.1 Search Strategy

A search was conducted to identify papers of longitudinal studies in which the mRS or the OHS was used as the outcome. First, articles that had been published on stroke patients were identified using search terms including ‘stroke’ and ‘cerebrovascular’. Secondly, longitudinal studies were identified using the search term ‘longitudinal’, and also those that were tagged as longitudinal studies. These two searches were then combined. Finally, a search was done to identify those papers that looked at the mRS as the outcome of interest. These searches were then combined again. Further restrictions that were placed on the search limited the articles to human patients and studies that were published in English.

3.3.3 Results

3.3.3.1 Search Strategy

By searching in both Medline and Embase, a total of 41 papers were identified using the search strategy above after the searches were combined and the restrictions of human subjects and English language publications were placed on the found articles. Appendix A contains a flow diagram detailing included studies. Between these two search engines, 9 articles that were found were duplicates and were removed. Of the 32 remaining studies, all the abstracts were read in order to identify which of the remaining studies were informative in the methods used and results of longitudinal analysis using the mRS as the primary outcome. When reading the abstracts, it became apparent that there is very little information on longitudinal studies using the mRS. There were very few of the studies that took a longitudinal approach to the analysis of the follow-up data they reported. There were 6 trials that did not use the mRS as an outcome and 13 trials that were not a longitudinal analysis. Finally, two trials were not actually undertaken in stroke patients. Only two studies conducted a longitudinal analysis using the mRS as the outcome.

3.3.3.2 Articles

The first study that conducted a longitudinal analysis using the mRS as an outcome was published in 2012 and used previously published trial data from the NINDS rt-PA trial (Li et al. 2010). The study conducted joint modelling in order to look at the longitudinal ordinal data of the mRS with competing risks survival times. The study using the NINDS rt-PA data looked at the mRS values at baseline, 7-10 days, 3 months, 6months and 12 months post-stroke. It also re-categorised the scale to: 1 = no symptoms or no significant disability despite symptoms, 2 = slight disability, 3 = moderate disability or moderately severe disability, 4 = severe disability or dead. It was found that the mRS scale has a decreasing trend over time, suggesting recovery of individuals, as a lower mRS score incicates less disability, and that conditional on other covariates and random effects, the cumulative odds ratio of a smaller mRS value was 8.33 (95% CI 5.63, 12.33) for 3 months, 9.68 (95% CI 6.54, 14.32) for 6 months and 11.59 (95% CI 7.53, 17.84) for 12 months, compared to 7-10 days. The study considered itself to be the first to consider competing risks failure times, in an attempt to deal with possible non-ignorable missing values in the longitudinal ordinal measurements. The missing data after death or dropout could be non-ignorable in this case as patients with a higher mRS score would be more likely to die or drop out of the study because of low efficacy of the treatment.

This study kept the mRS is an ordinal form, with a reduction in the number of points on the scale by merging some of the categories together, combining severe disability and death within the same severity category. This study also used a pre-stroke mRS value as the baseline. This was not available in the SO_2S data for all patients, only approximately 1000 patients, who had the information retrospectively considered. Clinicians were unsure of the accuracy of the pre-stroke Rankin data collected in the study, and therefore it was decided not to include it within the analysis conducted in the thesis. The joint model fitted by Li et al. investigates the time to drop out and the time to death/severe disability. If there were more points available in the SO_2S data then it would definitely be worth considering fitting a joint model to the SO_2S data; however, with such a short follow-up time frame, the risk of death and dropout over only 2 follow-up points (as there is no baseline value) may not provide accurate estimates.

The second study published in 2007 was a post hoc analysis of a previously published Vitamin Intervention for Stroke Prevention (VISP) trial (Spence et al. 2001), which enrolled 3,680 adults and followed them up for 2-years post-stroke (Newman et al. 2007) . With an outcome of the mRS score, linear mixed effect models (LMM) were fitted to the longitudinal data measured over the 2 years. This method allowed the correlation of repeated measures within each subject to be considered as well as the inclusion of missing data. The LMM found that there was an improvement in the mRS values each month over the 2-year follow-up, suggesting a reduction in disability over time. It was found that there was an increase in disability for older patients and in those who were black, but there were no gender effects. The main aim of the paper was to investigate the impact of diabetes on functional recovery. A linear mixed effects model is a suitable model for the analysis of longitudinal mRS, and one that will be considered in the thesis using both a dichotomised and ordinal version of the scale.

3.3.4 Discussion

The two studies both found that over time there was an improvement in the mRS score, indicating decreasing disability in a patient who is recovering from a stroke. One study used a joint model to investigate the recovery while the other used a linear mixed model. Both papers used data from a randomised clinical trial, one looking at a possible treatment for the stroke itself and the other the use of an intervention to prevent further vascular injuries. Both of the papers found were post hoc analyses of trials that had been previously published, one with a lengthy follow-up of two years, the other with a follow-up of a year.

There were no studies found that looked at the recovery of stroke patients in a natural cohort, that is those who received usual treatment and no trial intervention, and as there are only two studies that have looked at any longitudinal analysis of the mRS in stroke patients, any new research in this area will be novel and will provide new information into the recovery of patients over time. The use of just the natural cohort of the SO_2S trial was initially considered, this would have been done by using just the control arm of the trial. However, it was decided that in order to compare the different methods used an effect of treatment should be estimated, which required all data from the SO_2S trial, meaning that both the treatment group and the control group were used in the analyses conducted.

3.4 Scales that have death included

When a patient dies within a clinical trial or cohort study, they are usually treated as lost to follow-up and removed from the study. This means that they are unable to contribute any information to the study after they have died. When a scale has a category for death included like the mRS does, the patient that has died has a recorded value and therefore can still contribute to the analysis. When a patient has died in this situation, it is not appropriate to treat the value as missing and the individual has a known value and data should use the last observation carried forward to complete follow-up points after a patient has died. A search was conducted to look at other scales that have a category for death included in them.

3.4.1 Methods

No formal search strategy was conducted to identify the scales, as it was found to be very hard to define search terms that were able to identify the scales that included death and separate these trials from a trial conducted with any assessment scale where the patient had died. Initial information came from a text called *Measuring Disease* by Ann Bowling (Bowling 2001). This allowed several scales to be identified, following which more research was conducted into these scales, leading to the identification of more scales and a better understanding of the clinical areas in which scales including death may be used.

3.4.2 Results

There are two scales that, although not principally designed with stroke assessment in mind, have been applied frequently to assess brain injury in stroke patients, usually on hospital admission. These scales are mainly used to assess traumatic brain injury.

The first scale is the Glasgow Coma Score (GCS) and it is possible for a patient to be classified as dead on this scale (Teasdale & Jennett 1974). The Glasgow Coma scale was developed in 1974 and modified in 1975 by Graham Teasdale and Bryan Jennett to give a reliable, objective way of recording the conscious state of a person. It is now commonly used to assess consciousness in all acute medical and trauma patients. A patient is assessed against the criteria of the scale, and the resulting points give a patient score between 3 (indicating deep unconsciousness) and 15 (fully alert). The scale assesses the patient's ability to open their eyes (E) as well as motor (M) and verbal (V) responses; a total summative score is calculated from the 3 categories. In the scale, a person who receives a score of 3 is regarded as being in a deep coma or dead.

The second score found to include death as a category is the Glasgow Outcome Scale (GOS). This is a hierarchical ordinal scale, which describes the disability and handicap in patients with a brain injury (Jennett & Bond 1975). Developed by Bryan Jennett and Sir Michael Bond in 1975, it has 5 categories with scores ranging from 1 (death) to 5 (good recovery). It has been suggested that the categories are too broad and so an extended version was developed in 1998 (Wilson et al. 1998), which has scores 1-8, including categories with upper and lower degrees to distinguish disability better in the higher 3 categories; so, for example, good recovery becomes lower good recovery and upper good recovery. It is easy to complete and is conducted via an interview. Although not designed to measure stroke severity, these scales have been known to be used on patients who have had a stroke as an initial assessment of stroke severity.

The first scale found in oncology was the Karnofsky Performance Scale (KPS), designed in 1949 by David A Karnofsky and Joseph H Burchenal, which is a quality of life assessment scale (Karnofsky et al. 1948). Its main use is to allow clinicians to evaluate a patient's ability to survive chemotherapy treatment for cancer. However, this is not the purpose that the scale was designed for; originally it was used as a measure of nursing workload. But, its assessment of physical performance and dependence can be used in oncology. The scale contains rank-ordered decile points that range from 0 (dead) to 100 (normal). Patients are assigned to categories using a summation of scores to produce an overall score based on classifications made by health professionals, although the validity of the summations is yet to be tested and there are variations on the categories that exist. The KPS has been altered to look at paediatric patients. This alteration produced the Lansky Scale, however, within this scale the category for death has been removed (Lansky et al. 1987).

Also in the field of oncology, there is a second scale including death that looks at a cancer patient’s general well-being and activities of daily living, although it goes by several names; the WHO functional scale, the Zubrod Scale, or the Eastern Cooperative Oncology Group (ECOG) Performance Status Rating scale (Oken et al. 1982). It is effectively a condensed version of the KPS, with 6 ordered categories, that provides information on what a patient is capable of doing, rather than what they actually choose to do. In this scale, the range of scores goes from 0 (no impairment) to 6 (dead).

Within patients who suffer from multiple sclerosis (MS), an autoimmune disorder of the central nervous system, the Kurtzke Expanded Disability Status Scale (EDSS) is a scale used to quantify the disability seen as a result of MS (Kurtzke 1983). It was designed by John F Kurtzke as an extension of his 10-step Disability Status Scale (DSS) and is the most widely used assessment scale within MS. It has increments of 0.5 and can range from 0, which signifies a normal neurological exam, to 10, which signifies death due to MS. Here the death must be specifically related to MS, which has not been a stipulation on many of the other assessment scales.

There are also scales that include death within cardiovascular disease. The Olsson Ranking Scale is used to categorise the degree of cardiovascular disability (Olsson et al. 1986). It is based on the New York Heart Association (NYHA) functional classification scale, a measure used widely in clinical trials of congestive heart failure. The Olsson Ranking Scale classifies patients into seven mutually exclusive categories of health state from 1 (dead) to 7 (NYHA Class 1 without side effects or complications).

The Bulpitt's Hypertension Questionnaire is an adaptation of Fansel and Bush's Health Status Index (Bulpitt 1982). It focuses on the symptoms associated with hypertension as well as looking at the effect on everyday life. It is split into two sections; the first looks at disease-specific symptoms and the second about the effect of hypertension on daily life and functioning. The questionnaire can be self-administered, with the disability section of the questionnaire being scored on a continuum from 0 (dead) to 1 (total well-being).

There are also scales that have been altered in order to include death, in the hope of improving the interpretability or to make them comparable to other scales that are used for assessment. The 36-Item Short Form Health Survey (SF-36) scale been altered to include a defensible value for death. As well as this, the BI can be altered to have a score of -5 (one worse than worst possible outcome on the scale) and also the cog-4 subset of the NIHSS, proposed as a quick test of cognitive impairment, has a score of 10 added (one worse than the max score of 9).

3.4.3 Discussion

Outcome assessment scales are used in a large variety of clinical areas and assess many different outcomes. This being said, there are very few that, like the mRS, have a category in them that is specifically for patients who have died. The review above has found a few scales that this is appropriate for; however, it would be hard to form conclusions about how best to treat these data when the clinical areas considered are so different. A brief search was conducted in order to identify any alternative methods used for longitudinal analysis using any of the scales mentioned above, that have not previously been considered, that would be appropriate to use with the mRS. Although these scales included death as a category on the scale, the main analyses of papers using these scales were time to death analyses, and therefore the majority of analyses that were used were Cox regression methods, which are not appropriate for the mRS as an ordinal scale. There was one method identified using the EDSS scale for multiple sclerosis that fitted a Markov multi-state model to the data looking at disability progression over time (Palace et al. 2014). Although this paper did not include the score of death in the multi-state model, it is possible for this to be done, allowing the whole of the mRS scale to be used in full. This would make it an appropriate method to be used for an analysis in this project.

3.5 Clinical trials using other methods

Throughout the chapter, different methods for analysing the longitudinal data in the thesis have been considered. These include linear mixed models, whose use in clinical trials is well established. There were 21 papers identified in the review of trials published in stroke using longitudinal methods in Section 3.1. Other methods that will be considered include LCGA and Markov multi-state models, whose application in stroke trials is more novel. The use of these models is not as well reported, and a quick literature search was conducted to identify any recent clinical trial (since January 2014) that used either of these 2 methods as the main method of analysis.

3.5.1 Methods

A search was conducted in Medline and Embase in November 2017 to identify randomised controlled trials that had been published between January 2014 and December 2016. In order to make sure that all possible articles were captured, the search terms in the search strategy included "Latent class growth analysis", "Markov multistate model", "Randomised", "Randomized" and "RCTs". Article tags are applied to articles in order to classify the contents and help to make them easier to identify.

Search terms were kept as broad as possible in order to try to capture all relevant articles.

Once the search had been conducted, several restrictions were placed on the articles, including the date published (from January 2014) and also restricting the articles to trials that were conducted in human subjects, and that the full text was available in English. After articles that did not fulfil these criteria were discarded, the articles that remained were investigated further in order to identify only those articles that were relevant to the aim of the review. After this, the references of the chosen articles were searched in order to ensure completeness of the review.

3.5.2 Results

The search in Medline and Embase identified 32 articles for latent class growth analyses in RCTs and 54 articles for multi-state models in RCTs, before the removal of duplications and restrictions on dates. Initially, 13 articles about LCGA and 23 articles about multi-state models were removed due to duplication of articles. After the restriction of publication dates there were 10 LGCA and 9 multi-state articles of which the abstracts were checked for suitability.

Out of the 10 potential LCGA articles, 3 were conducted in cohort studies, with the remaining 7 articles conducting LCGA in RCTs. In these 7 articles, there were 3 that looked at trajectories in depressive and or anxious symptoms. The first in the effect of cognitive behavioural therapy in adolescence, second in overweight/obese patients with postpartum depression, and third in patients undergoing haemodialysis. There were 2 papers that looked at compliance with medication – one looking at adherence to medication in kidney transplant patients and the other looking at individuals using nicotine replacement therapy. The final 2 articles were both looking at trajectories of weight, with one looking at weight loss in patients with osteoarthritis and the other in the life course of patient BMI in Guatemalan adults who were given supplements as an infant. All the RCT's included had a follow-up of at least 12 months up to 30 years' worth of follow-up, and regardless of the number of patients included in the trials, the number of classes in a model fitted to the data ranged between 2 and 4 clusters.

Out of the 9 potential multi-state articles, 1 did not fit a multi-state model and 3 were not conducted in RCTs, which left 5 articles that conducted a multi-state model in an RCT setting. One trial looked at disability in the form of the major disability scale and looked at the impact of long-term physical activities. There were 2 cancer studies, one looking at failure type in non-small cell cancer and the other looking at risk screening for prostate cancer. The latter considered 6 potential states, similar in size to the mRS, and was the largest number of states found, with all other trials fitting a simple illness-death model, which has 3 states of healthy, ill and dead. The final 2 trials looked at mortality in coronary artery bypass graft patients compared to percutaneous transluminal coronary angioplasty, and heart failure hospitalisations in patients with diagnosed heart failure with reduced ejection fraction.

3.5.2.1 Latent Class Papers

A study published in 2016 by Ford et al. looked at life course body mass index (BMI) trajectories in Guatemalan adults, randomised to receive improved nutrition in the first 1000 days after conception in a study conducted in 1962-77. This RCT had an extended follow-up of over 30 years and this was used to fit LCGA models, in order to predict trajectories of BMI. The trajectories were derived from up to 22 possible measures of height and weight, which is a very large amount of data, much more than seen in most other studies conducting LCGA. They also used the idea of fitting different models to men and women separately, as it was agreed that there would be distinct differences in the change of BMI over the follow-up in both men and women. This meant that there were 2 models fitted, 3 different trajectories were found in men and only 2 different trajectories in women. The smallest cluster contained 15% of the data, which was one of the male trajectories. The paper describes well the process for deciding the optimal number of classes, which included using BIC, the bootstrap likelihood ratio test (BLRT), entropy and posterior probabilities. They took the interpretability of the classes into account when deciding on the optimal number of clusters.

The paper used Mplus in order to carry out the analysis. It also had to extend the models in order to account for within family correlation, as over 84% of participants had at least 1 sibling included in the trial (Ford et al. 2016). This is not something we need to include in our analysis, as we are assuming independence between each individual.

A second paper found also looked at change in weight data and was conducted by de Vos et al. in 2014. This paper focussed on the difference in weight between follow-ups, rather than calculating a BMI for the individual. The data was taken from the PROOF study, which was an RCT that investigated the effects of weight reduction on the development of knee arthritis. Over a follow-up of 2.5 years, information was recorded at 6 month intervals, so patients had a maximum of 6 follow-up points to contribute. The paper used a combination of the BIC, BLRT and entropy in order to decide on the best fitted model and aimed to fit models with between 2 and 6 clusters with linear, quadratic and cubic trajectories. They chose the 3 class linear model. There was a better fitting model, however it was decided not to choose this model due to having groups of very small numbers; the smallest group in the chosen model was 11.8%. They fitted multivariate multinomial regression models alongside the LCGA, in order to test the effect of the baseline covariates on the outcome. The paper accurately described the methods used, they fitted the LCGA models with Mplus and the regression models in SPSS (de Vos et al. 2014).

In 2014, Scherphof et al. conducted a LCGA in the use of nicotine replacement therapy in adolescents. The aim of the study was to evaluate predictors of compliance trajectories. Firstly, the compliance trajectories were identified and then the predictors of these trajectories investigated. The models included 6 questionnaires, from the 1st day after quitting until the last day after treatment, with uneven spacing in between the considered follow-up points. The models were fitted with between 1 and 4 trajectories and the models were compared using BIC, the bootstrap likelihood ratio test and entropy, in order to identify the best fitting model. The 4 class model had the best fit according to BIC but an insignificant BLRT meant that the 3 class model was chosen as the best fitting model. Following on from the identification of the clusters, multinomial regression models were fitted, which compared the groups with moderate and strong decreases in compliance over time, with the cluster in which participants were the most compliant over the study period. As with previous papers, Mplus was used in order to conduct the LCGA, and the paper provides a detailed explanation of the methods used and decisions made in order to choose the best fitting model for the data (Scherphof et al. 2014).

The group based symptom trajectories in the prevention of adolescent depression were investigated by Briere in 2015. The RCT contained 3 different interventions and aimed to create trajectories based on groups of individuals from pre-test to 2-year follow-up. Initially, each intervention was considered distinctly, to look if there were differences in the interventions and control, however because the trajectories had similar numbers and shapes across the groups, all the patients were considered together. They modelled between 1 and 5 classes, with the best model being chosen using information criterion and likelihood ratio tests. Three information criterion AIC, BIC and adjusted BIC, and 3 likelihood ration tests Lo-Mendall-Rubin, Vuong-Lo-Mendall-Rubin and BLRT were used. This is the most thorough paper in terms of testing model fit, as they used more tests than the other papers. We considered using all of these tests within the model fitting of our LCGA, as this would hopefully give us the most information for fitting the best fitting model, if they are in agreement. From this, 4 classes were identified, most having a steady change over time, with 1 showing a decrease and resurgence in the depressive symptoms. The smallest cluster in this model has 6% of the data in it. After the trajectories had been identified, they then considered the effect of the intervention in the trajectories by looking at the prevalence of class membership by condition, with no significant difference found in the trajectory membership. Finally, multinomial regression models were used to investigate difference between predictors within the trajectories (Briere et al. 2015).

3.5.2.2 Multi-state Papers

The first multi-state RCT paper identified was an analysis of chemotherapy in non-small cell lung cancer – conducted by Rotolo in 2014. The aim of the study was to use multi-state models to describe the effect of chemotherapy, for approximately 200 patients recruited between 1995 and 2001. The median length of follow-up was 7.5 years, in which patients could suffer a local replace, distant metastasis or die. The paper doesn't make it very clear how many states are included in the model as they define 2 different types of metastasis but from the text and the diagrams it is unclear whether they consider them as distinct or together. The paper provides no indication on how model fit was assessed, and although it presents whether calculated hazard ratios are significant or not significant, this information could be presented much more clearly. There is no information on what statistical package was used to fit the models or what was actually included in them, with the main results being adjusted but not described in the text (Rotolo et al. 2014).

There was another paper in the field of oncology identified that used multi-state models. In 2015, Yen et al. looked at risk stratification for prostate cancer screening and fitted a 6 state disease progression model to a Finnish RCT on cancer screening. The paper presented the way that the model was fitted well, with the transitions defined between states in the disease progression and used a 2 step approach. First, it estimated the transition rates without considering genetic markers using the Cox and Miller method, followed by the inclusion of predefined genetic markers. The models produced a 10 – year risk of progressive prostate cancer, for low, medium and high risk groups, who began screening at different ages. There is no description in the text about how model fit was assessed, or the program used in order to fit the model to the data, but the model provides odds ratios for the effect of the selected SNPs on the risk of prostate cancer (Yen et al. 2015).

Also in 2015, Zhang et al. analysed Bypass Angioplasty Revascularisation Investigation (BARI) trial data using a multi-state model. The RCT randomised patients to 2 different cardiovascular treatments, and used a primary outcome of all-cause mortality. The model fitted was a 3 state model of receipt of intervention, myocardial infarction (MI) and death, and was compared to a standard survival analysis using Cox regression and Kaplan-Meier curves. It was found that the differing effect between treatments may have been disguised without the intermediate MI state. As with other papers, there is no discussion on model fit, but the paper does report the effects of several covariates on the transitions estimated. Although it doesn't mention the software used, the paper is the only paper to recognise limitations of the multi-state model itself within the study limitations, including the complex nature of using software packages to fit the data, and the validity of the Markov assumption although, they believe that it holds in this instance.

Upshaw et al. (2016) used a multi-state model in order to predict heart failure hospitalisations and all-cause mortality in patients with reduced ejection fraction. They included 12 covariates in the analysis, which were chosen a priori. They also used a complete case scenario as there was less than 1% missing data. They fitted the multi-state model using Cox proportional hazards regression, with 3 states in the model, one of which was the absorbing state for death, which was estimated separately. The paper explains well how the included covariates were selected using a process of univariate models and comparison of the AIC. They then went on to further test the models by validating it in a new cohort. Models were fitted using R and the 'mstate' and survival packages for multi-state modelling (Upshaw et al. 2016).

The final paper is probably most similar in terms of outcomes to the analysis conducted throughout the thesis. In 2016, Gill et al. used multi-state modelling of a RCT to analyse the effect of structured exercise on transitions between states of major mobility disability (MMD). MMD was defined as the inability to walk 400m and was assessed every 6 months for up to 3 years. A 3 state model of no MMD, MMD and death was used, with missing or indeterminate data included as a censored state within the model. The model estimates the cumulative incidence for the first 3 transitions, and obtained estimates for 1 and 2 years of follow-up. The median follow-up was 2.7 years. The model produced hazard ratios for the comparison of the 2 considered treatment groups, with adjustment some baseline characteristics. The paper accurately details the observed transitions between states and the risks associated based on subgroups. The paper clearly presented the results and assessed model fit with the addition of covariates using likelihood ratio tests.

3.5.3 Discussion

From the papers that have been identified conducted in RCTs, it seems clear that the methods detailed in the papers conducting LCGA are much better defined than the papers in which multi-state models are conducted. This may be because all the papers used the same software, specifically designed for conducting LCGA, and therefore may be following guidelines defined by software developers and other research papers. Most of the papers found use continuous follow-up time rather than discrete periods as with the SO_2S, but this could be mainly due to use in areas such as oncology where the focus is event based. We also see that the papers have many more follow-up points than the 3 and 4 time points used in the data within the thesis. Within the multi-state models found, most fit a simple 3 state model with an intermediate state to death, however, there was one model that fitted 6 states to the data. LCGA models tended to only be fitted with between 2 and 5 clusters, but model fit was assessed and reported well within these papers, suggesting more model selection takes place with these models, more than multi-state models. The papers identified here form a basis on which to build, when planning to conduct the different types of analyses used throughout the thesis. They show that although the different areas in which the methods have been applied to RCTs are wildly different, it is not common for these methods to be used, and therefore the impact of using them is novel work and they should be promoted if investigations show they are suitable to be used to analyse RCT data from stroke clinical trials.

3.6 Conclusion

This chapter identified the areas in which research has already been conducted by looking at potential methods that have previously been used with the mRS and other scales in stroke trials. Although many different methods have been applied in previous trials, there are very few found for which the mRS is an appropriate outcome. It appears that there is also very little research into the longitudinal analysis of the mRS, which is the main focus of this thesis, suggesting that the work conducted in this project will be novel and will be able to provide information on which is the best method to use. It can also be seen how few scales there are in which there is a category for death, suggesting that if the statistical techniques that are used in the thesis work well with the category for death included in the scale and the results are sensible, then it would be possible to apply the methods that have been identified here in randomised controlled trials that are not just stroke related.

4. Analysis using regression models

Previous research has been conducted into the best methods for analysing stroke data, but all this research has been conducted using a single primary end point (Bath et al. 2007, Sajobi et al. 2015). Using an ordinal method was recommended from this research, but it shows that despite the recommendations of the OAST collaboration (Bath et al. 2007) , there has not been a dramatic shift in the types of analysis performed when analysing acute stroke trials.

This chapter investigates potential ways that ordinal scales can be treated, looking at dichotomising the scale, keeping it as an ordinal scale, and even the potential of using the scale as a continuous outcome measure. Next, the different regression models that could be applied to the data are discussed and the form of data needed for each of the models. The chapter also considers whether to use the full scale with the score for death, or whether it should be removed or condensed in some form.

Once all of this had been reviewed, regression models were applied to the data, including binary logistic regression and ordinal logistic regression models. There were some assumptions made about using these models, and the analysis goes on to consider how tenable the assumptions are, and what the conclusions would be if different assumptions had been made.

4.1 Ordinal outcome scales

Ordinal data are a form of categorical data whereby patients are classified into a specific category and these categories have a natural order; for example, a rating scale: “poor”, “good”, “excellent”. The scale can be ordinal in nature from the start, like the mRS, or it can be a summative score, which is a specific type of ordinal score. It is a measurement score where a series of questions are asked and rated, which can be combined to form a total score. This is then usually the ordinal score that is due to be analysed. Different methods are needed to analyse these two types of scale, as a summative score made of ordinal items may have interval properties. This chapter will focus on methods for just an ordinal score, and will not consider methods for a summative score.

Ordinal outcome scales pose a statistical challenge for those trying to model them. The nature of the ordinal scale means that although the data are in a ranked form, the relative degree of difference between points on the scale is not known. The idea of exploiting the ordinal nature of an ordered outcome scale is not a new concept in the statistical community. Nevertheless, this approach has not been applied to the analysis of clinical trials on a regular basis (Roozenbeek et al. 2011).

There are various approaches that can be applied to ordinal outcome data when analysing at a single time point. How the ordinal scale is treated will affect the different types of longitudinal methods that could be applied to the data. The different ways in which the scale can be treated are detailed below. Each different model type has different assumptions associated with it, which will need to be assessed as to whether or not the data meet these assumptions. Each method has its own advantages and disadvantages, which will be discussed. It is also important to develop a way in which fitted models can be compared as this will help to determine the most appropriate way to analyse longitudinal stroke trials in future.

4.1.1 Dichotomisation of the scale

The first option is to dichotomise the scale into two categories. This method is most commonly used in clinical trials, where the scale is converted into a binary measure, usually a favourable and an unfavourable outcome. The main advantage of using the ordinal score in a dichotomised form is that the methods used to analyse the data are considered to be simpler methods than regression models that make multiple comparisons between the data. They also provide effect estimates that are much easier to interpret because there are only two categories being compared. Using only two categories makes calculations like the number needed to treat easier, as there are only 2 options, a bad outcome and a good outcome. Dichotomisation is an accepted approach for an ordinal scale; however, the simplicity of the scale is obtained at some cost (Altman & Royston 2006).

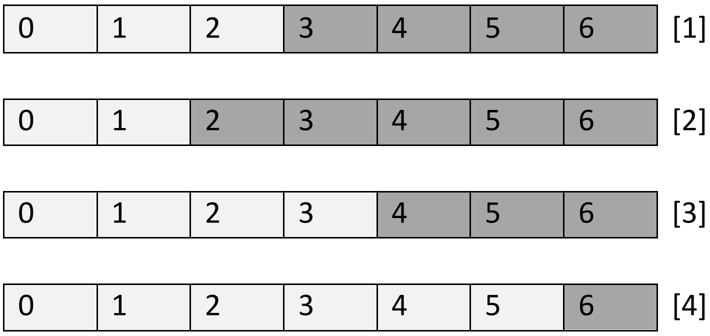
One disadvantage of dichotomising an ordinal scale is that you are reducing the statistical power of being able to detect a relationship between the variable and the patient outcome as information is lost due to several categories being combined together. It is only possible to capture the transitions of patients across a single cut-off point – other changes, either side of this cut-off, are not captured.

Secondly, there is usually not a defined cut-off for the ordinal scale. Unless extensive research has previously been conducted and the best dichotomisation defined, there are several ways in which a scale could be dichotomised, which could all lead to different results. This can lead to different trials in the same area producing very different results due to the different cut-off points applied to the data. One common method, known as the median split, takes the median value of the data as the cut-off point. This is more common when categorising continuous data. Historically, dichotomising the mRS is the most popular approach to the analysis of clinical trials at the primary end point (Bath et al. 2012).

4.1.1.1 Example of potential dichotomisations.

There are 7 points on the mRS and there are several different ways in which the scale can be sensibly dichotomised. Figure 4.1 shows the different ways that the mRS has been dichotomised in previous studies.

Figure 4.1: Potential dichotomisations of the mRS



The first part of Figure 4.1, [1] shows the most common dichotomisation of the mRS. An example of a study using this dichotomisation is the NEST-3 trial (Levine & Hill 2014), where a binary mRS score is the primary endpoint at 90 days. On this scale, a favourable outcome, defined as success, is a mRS score between 0-2 and an unfavourable outcome, defined as a failure of the treatment, is a mRS score of 3-6. The second dichotomisation of the mRS, [2], has recently been used in the NBP study (Cui et al. 2013), amongst others. In this study, the primary endpoint was also the mRS score at 90 days; however, the favourable outcome was defined as a mRS score of 0-1, with the unfavourable outcome the remaining 2-6 mRS scores.

Another possible dichotomisation of the score is to place the cut-off at a mRS value of 3, [3]. This is a less favoured dichotomisation as there have been fewer trials that have used it, however, it was used by the Controlling hypertension and hypotension immediately post stroke (CHHIPS) study in 2009 (Potter et al. 2009). Here, the researchers defined death or dependency as a score of mRS > 3. So the unfavourable outcome contained mRS scores from 4 to 6, and the favourable outcome contained mRS scores from 0 to 3.

Finally, the mRS has been dichotomised into patients who are either dead or alive [4], as was used in a subsidiary analysis of a prospective cohort study and systematic review of Interleukin-6 (Whiteley et al. 2009). Here they detailed that although a dichotomisation of 0-2 and 3-6 was used for the primary analysis, a dichotomisation of alive (mRS 0-5) and dead (mRS=6) was also used, turning the mRS into a binary outcome reflecting survival.

The practice of dichotomising ordinal scales is not just common practice in stroke trials but also in many other clinical areas. It is now well recognised that dichotomising an ordinal (or continuous scale) and analysing the resulting binary scale means discarding patient information that can only be compensated for by substantially enhancing the patient numbers recruited in clinical trials. Therefore, it is of interest to use all possible patient information in order to reduce recruitment numbers and therefore costs for clinical trials in stroke research.

4.1.2 Ordinal scale

It is possible to keep the data in its raw form and treat it as an ordinal scale. This approach keeps all the categories separate with no collapsing of the categories and assumes that all the data follow a specific order. The categories are not assumed to be fairly evenly apart; so, there is room for there to be a slight variation in the perceived change when a patient transitions from one category to the next. This method has advantages over dichotomising the scale as there is more statistical power, as information is not being thrown away. Obviously, this method also rules out the need to define specific cut points for the data as well, meaning that the results can be more comparable across different trials. Of course, if the ordinal scale in question has several categories it is possible to condense the categories together, but not into just two categories like the dichotomisation. This type of categorisation will also lead to an ordinal scale, just with fewer categories than the original ordinal scale. Even in a condensed ordinal scale, there is likely to be greater power to detect changes compared to the dichotomised scale.

The application of ordinal outcome analysis substantially increases the power of a clinical trial. Adopting and ordinal outcome analysis will allow the detection of smaller treatment effects, as when the outcome is analysed in an ordinal way, all patients can contribute to the detection of a treatment effect (Roozenbeek et al. 2011).

4.1.3 Sliding dichotomy analysis

An extension of the dichotomisation at a single cut-off point is to fit a sliding dichotomy to the data. This method can also be referred to as a responder analysis, or a prognosis-adjusted analysis. The method works by applying different dichotomies to the data and defines a favourable outcome as a function of the baseline prognosis or severity. Each severity band defined at baseline will use a different dichotomisation of the scale. By splitting the data in this way, there is less information lost, which results in a slight increase in the statistical power and more changes are able to be captured, which would otherwise be undetected in a binary analysis. Although there are benefits over using a single dichotomy, there are some disadvantages of using the sliding dichotomy. There is a lack of efficiency in the method of the analysis, particularly when there are a larger number of categories on the ordinal scale (McHugh et al. 2010). The results produced can be difficult to interpret, and in some case may not vary a lot between different dichotomies (Saver 2011), where results are produced by combining the estimates from each strata of the dichotomisation. It is also important that sliding dichotomy scale must be defined before the trials, with the calibration dependent on the type of patients likely to be recruited (Bath et al. 2012).

4.1.4 Continuous approach

It is possible to take an ordinal scale and treat it as if it was continuous. If the scale was to be treated as continuous then an assumption would be made that a patient moving one point on the scale would be a change of the same magnitude regardless of where that patient is on the scale (Knapp 1990). For example, a patient moving from 2 to 1 would have the same gain as a patient moving from 5 to 4, which may not be the case for an ordinal scale. It is possible to test whether treating an ordinal scale as continuous results in a loss of information (Long & Freese 2006). It has previously been shown that using linear regression models on an ordinal outcome can lead to incorrect conclusions (Winship & Mare 1984). Considering the mRS a continuous scale was considered in 2015 by Sajobi et al. and although they found a continuous scale to be as powerful as an ordinal scale, caution must be undertaken taking an ordinal scale and treating in a continuous manner.

4.2 Logistic regression models

Logistic regression is a regression model where the outcome variable is categorical in nature. The regression model works by estimating the probabilities of the relationship between the categorical outcome variable and one or more predictor variables considered by the logistic function. The outcome variable in logistic regression is able to be binary, ordinal or multinomial. The main interest is in how using the different types of regression model can affect the treatment effect predicted from the model.

4.2.1 Binary logistic regression

Binary logistic regression is conducted when the outcome of interest has only two possible categories. The aim of logistic regression is to predict the likelihood that the outcome variable is equal to 1 (compared to 0) given certain values and combinations of the predictor variables.

First, let us start with the probabilities π_i that depend on a vector of observed covariates x_i. It is possible to let π_i be a linear function of the covariates, where β is a vector of regression coefficients.

$$\pi_i = x_i\beta$$

This model is known as the linear probability model and is often estimated from individual data using ordinary least squares. The main issue with this model is that the left-hand side is a probability and must range from 0 to 1, whereas the right-hand side is a linear combination and can take any real value. The solution to this problem is to transform the probability in order to remove the restrictions of the range of the left-hand side of the equation. First, this is done by transforming the probability to the odds, which is defined as the ratio of the probability to its complement, as there are not restrictions on the odds.

$$odds_i = \frac{\pi_i}{1 - \pi_i}$$

Once this has been done, a second stage of the transformation is to take natural logarithms, which removes the effect of the floor restrictions. This leaves us with the log-odds, sometimes known as logit. The logit is a one to one transform that maps probabilities from the range (0, 1) onto the real number line.

$$\log (odds_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = logit\ (\pi_i)$$

Now that the logit of the probability has been defined, we are in a position to define logistic regression, and the logit, rather than the probability, follows a linear model.

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X$$

The model works by estimating the log-odds depending on the covariates included, which can then be transformed to give the odds ratio between the two categories of the outcome variable by taking the exponential. The link function of a model defines how the model is related to the response variable, in this case the model uses a logit link function, which has been defined above as the log odds.

4.2.2 Ordinal logistic regression

The previously defined model for binary logistic regression can be extended to analyse the data when the outcome variable is ordinal in nature and not binary. In order to do this, a small change is needed to define the probabilities for the outcome variable differently. Instead of considering the probability of a single event, the model now considers the probability of that event and all the events that are ordered before it. This model also uses a logit link function to define the relationship of the predictor variable and the outcome. Ordinal models have been considered with other link functions, including the probit link function (Winship & Mare 1984). Both binary and ordinal regression models can use a probit link function rather than a logit link function.

Using a probit link function assumes the errors follow a normal distribution and the effect that is calculated cannot be interpreted as easily as using the logit link function, with no odds ratio being estimated as there is when using the logit link function. It was decided that the probit link function would not be used unless there was a reason why the logit link function could not, as testing the distribution of the errors is not easily done, and there is nothing to concern us that a logistic distribution is not appropriate. In this situation there was no need to use the probit function, and so we focus on logistic regression models because the parameters estimated are conveniently interpreted as log-odds ratios.

Now we can rewrite the odds that we are considering, where k is the event of interest for the outcome variable.

$$odds_i = \theta_i = \frac{P(Y_{ij} \leq k)}{P(Y_{ij} > k)} = \frac{P(Y_{ij} \leq k)}{1 - P(Y_{ij} \leq k)}$$

Once again natural logarithms of the odds are taken in order to give the model for ordinal logistic regression.

$$\ln(\theta_i) = \alpha_j + \sum_{k=1}^{p} x_{ik}\beta_{jk}$$

Each logit calculated has its own value for α_j but the same coefficient for β. The α_j terms are known as the threshold values, although they are of little interest. Because of this, there is an assumption that the effect of the predictor variable is the same for different logit functions. This is known as the proportional odds assumption and is why ordinal logistic regression is also known as the proportional odds model. The assumption needs to be checked as if it does not hold true another model will need to be fitted. The assumption can be checked by comparing the separate binary logistic modes underlying the overall model (Winship & Mare 1984). This can be done by conducting a Brant test (Brant 1990). Due to the proportional odds assumption, we find $\beta_{1k}= \beta_{2k}=\ldots= \beta_{c-1k}$, where *c* is the total number of outcomes and so we can write the log odds as

$$\ln(\theta_i) = \alpha_j + \sum_{k=1}^{p} x_{ik}\beta_{k}$$

There are no limits to the types of predictor variable that can be included in the model; they may be continuous, categorical or ordinal themselves, although if the predictors were ordinal or categorical then they would need to be entered as a set of dummy variables. However, it is assumed that if there is more than one predictor variable, there is no multicollinearity, where two or more variables are highly correlated with each other.

4.2.3 Multinomial logistic regression

If an ordinal regression model is fitted to the data, but the assumption of proportional odds is violated, then a multinomial model can be fitted to the data. This is an extension of the binary logistic regression model, when there are more than two categories. Although the scale used is ordinal there is no assumption is made about the relationship between each category of the scale, meaning that the ranking nature of the data is lost. The model is often used to model nominal response data where there are various categories with no set order.

The simplest method of a multinomial model is to nominate one of the response categories to act as the baseline or reference category and then to calculate the log-odds for all the other categories in the scale relative to that baseline. The model used to fit a multinomial model is analogous to a binary logistic regression model, except that the probability distribution of the response in now multinomial rather than binomial and there are now multiple equations. A common choice for the baseline category is the final response category.

Let there be *c* categories for the response variable. Then the model described above can be extended so that the odds for each score compared to baseline can be written as

$$odds_i \ = \ \theta_i = \frac{P(Y_i = y_j)}{P(Y_i = y_c)} \qquad for\ j = 1,2, \dots, c-1$$

Once again natural logarithms are taken, to work in terms of the log odds, which are written as

$$\ln(\theta_i) = \ \alpha_j + \sum_{k=1}^{p} x_{ik}\beta_{jk}$$

In order for the parameters to be identified, it is important to place restraints on certain parameters as each value of β will have a fixed constant added to it – giving the same predicted probabilities. So the values α_c=0 and β_{ck} = 0 are fixed.

4.2.4 Partial proportional odds model

The partial proportional odds model is very similar to the proportional odds model (ordinal logistic regression); however, the assumption of proportional odds is relaxed for some of the variables. This is done when there is a violation of the proportional odds assumption for some but not all of the predictor variables considered in the model. There are two ways to model these variables that do not achieve the proportional odds assumption. One is to fit an unconstrained model whereby there are no constraints placed on how the odds ratios vary over different thresholds. The other version is a constrained model where there are constraints imposed for the variables for which the assumption fails on the linear (proportional) relationship for the odds ratios. These constraints are the same for each predictor variable and are derived using a priori assumptions about how the data should behave.

4.2.4.1 Unconstrained proportional odds model

An unconstrained proportional odds model takes the following form

$$\ln(\theta_i) = \alpha_j + \sum_{k=1}^{p} x_{ik}\beta_{jk} + \sum_{k=1}^{p} T_{ik}\gamma_{jk} \qquad for\, j = 1,2,\dots,c-1$$

Here the x_{ik} are the values for an individual in each group i with a full set of p explanatory variables, and β_{ik} are the regression coefficients associated with the p variables in x_k. T_{ik} are the q covariates, such that $q < p$ and contains the values of an individual in group i on the subset of p explanatory variables for which the proportional odds assumption is not assumed, or is yet to be tested. The values for γ_{ik} are the regression coefficients associated with the q variables in T_{ik} and so $T_{ik}\gamma_{ik}$ is an increment associated only with the cumulative logit with $\gamma_{1k}=0$. If all the values of $\gamma_{ik}= 0$ then the model reduces to the proportional odds model.

4.2.4.2 Constrained proportional odds model

Restrictions can be placed on the unconstrained proportional odds model in order to give the following constrained proportional odds model.

$$\ln(\theta_i) = \alpha_j + \sum_{k=1}^{p} x_{ik}\beta_{jk} + \sum_{k=1}^{p} T_{ik}\gamma_k\Gamma_j \qquad for\, j = 1,2,\dots,c-1$$

Here we see that Γ_{j} are fixed scalars that are pre-specified with Γ_1=0. The new parameters γ_k are now not dependent on j. However, they are multiplied by the fixed constant scalar Γ_j in the calculation of the j^{th} cumulative logit.

4.3 Other models for analysing an ordinal scale

The regression models that have been detailed in the previous section are not the only methods that are available to be used to analyse ordinal data. There are models that can be applied if the data were to be treated as continuous; however, this method is not really appropriate in the case of the mRS where there are distinct categories and the change between categories does not appear to be linear. We also wanted the model to be interpretable from an individual perspective; therefore it was considered whether to use generalised estimating equations (GEEs). However, the decision was taken not to use these because they take a population averaged approach, which is not what we wished to consider. Using GEEs would allow the benefit of asymptotically consistent variance-covariance estimates when data are nonexchangeable, even when the precise nature of that dependence is unknown, and would be a valid method if we had been interested in estimating effects at a population level.

As well as these, there was consideration as to whether to use robust/cluster standard errors; this is a means or empirically correcting the variance-covariance estimates in the presence of heteroscedasticity, clustering, and other forms of conditional dependence. However, it was decided that using the linear mixed models in the process, and by using random effects in order the take into account the correlation, was a better approach than using an adjustment on the standard errors.

There are, however, several other models that are available that can be fitted to ordinal data, which uses the data in its ordinal form. These are the continuation ratio model, the stereotype logistic regression models and the adjacent category model, which will be discussed below.

4.3.1 Continuation ratio model

The continuation ratio model compares the probability of being in a particular category of the ordinal scale with the probability of being in a higher category, given that the individual has already reached the category being compared. This model is similar to the proportional odds model as it assumes that the effect of treatment would be the same at each level of the outcome variable. However, it differs from the proportional odds model as the continuation ratio model calculates the probability of being in a category or higher given that the category has been obtained, which means that lower categories are discarded, whereas the proportional odds model calculates the probability of being in a category or higher compared to any category lower. This means that the model does not allow individuals to reverse the progression of the ordinal outcome being considered (Armstrong & Sloan 1989).

4.3.2 Stereotype logistic regression

The stereotype logistic regression model was developed in 1984 by Anderson and is an extension of the multinomial regression model (Anderson 1984). The model is nested within a multinomial regression model as it has fewer parameters included and can be described as a non-linear form of the constrained multinomial model (Lunt 2001). The stereotype regression model, like the multinomial model, compares each category in the scale to a baseline category. However, it differs from a multinomial model because although the model does not automatically assume an order in the outcome, it can allow for an order, unlike multinomial regression. This means that the odds ratios that the model produces may be larger for extreme categories and the model would not give a single estimate for the difference in the treatment effect. This means that the interpretation of the model can be difficult, and although the model is more flexible than previously discussed models, it may be that the model is complicated to fit and interpret.

4.3.3 Adjacent category model

The adjacent categories model compares the log odds of each possible pair of adjacent categories simultaneously (O'Connell 2006). As with the proportional odds model, the predictor variable effects are assumed to be the same over each level of the outcome. The model fits binary logistic regression for each of the adjacent category pairs in the model. Each of these models is fitted simultaneously and the regression coefficients of each model are equal due to the assumptions made. Although the model uses the scale in an ordinal form, it uses less data than previously suggested models as it only considers adjacent pairs, and it looks at the associations with moving into the next highest category for adjacent pairs (O'Connell 2006).

4.4 Dealing with the inclusion of death in the mRS

In a follow-up study, once patients have died they are usually dropped out of the study with the date of death recorded along with other information such as the cause of death. They will then make no further contribution to the study follow-up. The mRS is unusual because it has a category for death in it. This means that individuals who have died are not lost during the follow-up and will contribute a score throughout the follow-up that is attempting to be modelled. The mRS is an ordinal score with higher categories indicating more severe disability, meaning that the score of death being the highest in the scale may give it more influence.

Research has been conducted into the perceptions of individuals of the outcomes on the mRS. A study conducted in both stroke patients and non-stroke patients found that 69% of stroke patients and 82% of non-stroke patients would prefer death to severe disability (Hanger et al. 2000). This suggests that death may be deemed preferable to severe disability with severe language, cognitive and motor defects. By including a score of 6 for death in the mRS, it can be argued that the ordinality of the scale is removed and therefore to include death as a higher category on the scale may lead to misleading results as the ordinal nature of the scale could be questioned.

As well as this, the mRS may not be the only scale in which a patient is measured during follow-up. Depending on what is being assessed, a patient who is dead would contribute no score to the BI. If the two scales were to be analysed and compared, the patients included in the analysis would be different as patients who had died would be able to be analysed on the mRS scale but not on the BI.

Previous research investigated the effect of including death in an ordinal scale, mainly looking at health status over time, and there have been some suggestions as to how death can be treated. Research conducted by Diehr et al proposed several approaches for incorporating death when considering health-related variables in longitudinal studies (Diehr et al. 1995). These included adding a category for death to the variable of interest as an extreme value on the scale or dichotomising the scale into healthy and not healthy and including those individuals who have died as not healthy. Finally, they proposed a method where they transformed the scale into the probability of being healthy in the future (Diehr et al. 2005). They concluded that the scale needed to make sense in terms of having an extreme value included for death, but that analysis of a scale including death in a longitudinal setting was reasonable. They found that strategies that gave death less influence tended to show more favourable changes in health over time (Diehr et al. 1995).

There are several different things we can do with the mRS scale to accommodate the category of death. Firstly, the value of death can be removed. There are two ways to do this. Firstly, death can be treated as loss to follow-up and the values of 6 in the mRS score replaced with a missing value. This means that there will be different sample sizes at each time point in the follow-up; this is known as an available case analysis. The other way to remove death is to completely remove anyone who died from the study and therefore all values at each time point are omitted for an individual who has died; this is known as complete case analysis.

Secondly, we can recode the data to combine the categories of 5 and 6 together in order to eliminate the concern about whether the ordinality of the scale holds. This option has been done in several studies previously (Lansberg et al. 2009, Whiteley et al. 2009) . Finally, we can leave the mRS scale exactly as it has been designed and use it as a seven point ordinal scale that ranges from 0 to 6 with the category for death assigned an extreme value for the variable.

It is possible to use all these methods with the mRS scale in all the different types of regression models that were previously detailed in section 4.2. It was decided to keep as much information from the scale as possible and analyse the mRS scale as a scale that has 6 included for an individual who has died.

4.5 Appropriate choice of ordinal model

There are several issues to consider when choosing what model to apply to the ordinal data that we have. Firstly, it is important to decide if an ordinal model is necessary and whether the outcome variable being ordinal is important with respect to the predictor variable(s) being considered. It has been suggested previously that in order to model the data as ordinal, the outcome must be ordinal with respect to the predictor variables (Anderson 1984).

Secondly, it needs to be considered whether the model that is being applied is suitable for the ordinal scale that is being considered, namely a non-linear probability model (Long & Freese 2006). Finally, for the modelling that is being conducted here, it must be decided whether the model is appropriate to be used in terms of a repeated measures analysis rather than analysis at just a single time point.

A summary of the advantages and disadvantages of using each of the models that have previously been described in the chapter are given in Table 4.1. When considering the ordinal mRS scale there are several models that are not appropriate, or are less appropriate than other models. The aim of this modelling is to find the effect of treatment when using single and repeated measures models; therefore the model should produce a single estimate that can be compared between models fitted.

Both multinomial and stereotype logistic regression models compare each category on the scale to a single reference category and therefore produce several treatment effect estimates for each model fitted. This is not ideal for the comparison with models that produce a single estimate and therefore these methods will not be considered. The continuation-ratio model requires the outcome variable to be progressive and therefore will not allow an individual go in the opposite direction to the progression of the scale. It can be seen that individuals move both up and down the mRS scale between two given time points and therefore this method is not appropriate for the analysis.

For the adjacent categories model, the distances between all the adjacent categories are assumed to be equal. This may not be the case for the mRS, and is the reason why the scale cannot be considered continuous; therefore this method may not be appropriate for longitudinal modelling of the mRS scale. As well as this the model is concerned with individuals moving to the category one higher than the current category, which suggests that it does not model the scale moving both ways, just like the continuation-ratio model.

This leaves us with a binary logistic regression model, an ordinal logistic regression model (proportional odds model) and the partial proportional odds model. As the partial proportional odds model is an extension of the proportional odds model, it will only be required if the assumption for proportional odds fails, leaving us with two models to be considered that are both easy to interpret for clinicians and easy to fit within statistical packages.

Table 4.1: Advantages and disadvantages of potential statistical methods

Potential methods	Advantages	Disadvantages
Binary logistic regression	Easy to fit Simple to interpret Single effect estimate	Ignores potential relationships – due to single cut-off in outcome Reduced statistical power
Ordinal logistic regression (Proportional odds model)	Most basic ordinal model Easy to fit Single effect estimate	Strong assumptions Slightly more complex to interpret
Multinomial logistic regression	Few assumptions needed Easy to fit	Does not account for ordering Estimates many parameters Compares to reference category – no single effect
Partial proportional odds	More flexible the proportional odds model	Harder to fit than the proportional odds model – not available in all packages Can produce negative probabilities
Continuation-ratio model	Outcome variable is progressive	Harder to fit than regression models – not available in all packages Strong assumptions Does not use all data
Adjacent categories model	Good if the outcome variable increases by 1 category	Harder to fit then regression models – not available in all packages Strong assumptions Does not use all data
Stereotype logistic regression	Does not automatically assume order, but can allow an order unlike multinomial regression	Not able to fit when the outcome has questionable ordinality Does not use all the data

4.6 Application to data

It was decided to fit the different regression models to the SO_2S data and look at the effect of the model on the 3-month treatment effect. First, each model was fitted at the single 3-month time point and so only took into consideration the data at that one time point. Secondly, a repeated measures model was fitted to the data, which includes all the information at each of the follow-up points. The 3-month treatment effect can then be deduced from the model using a combination of the logits that are calculated. The 3-month time point was selected as the point of comparison over the other time points. This is because it is the most commonly recorded endpoint in clinical stroke trials. Therefore, if the analysis showed that using further time points in a repeated measures model was preferable to just the single time point, the future implications of analysing this time point would be greatest in trials. Trials would need to wait for all data to be collected over the follow-up before the analysis can take place. This analysis was only conducted in the SO_2S data, as previous publications (Li et al. 2010) have looked at analysing the NINDS data set already, and the SO_2S data are the main trial data obtained for the project.

4.6.1 Binary logistic regression

This first regression model fitted to the data was at a single time point. This model ignores where an individual was at a previous time point or where they will move to in the future. By doing this we remove the longitudinal aspect of the analysis and this is the most common type of analysis conducted in stroke trials. The equation for this model is the simple model calculating the log odds given in Equation 4.1 below

$$Ln\left[\frac{P}{1-P}\right] = \beta_o + \beta_1 Treatment_i \qquad [4.1]$$

Where β_1 represents the difference in the 2 treatment groups at whatever time point the model is estimating and β_0 represents the log odds estimates for the standard care group. The model is simple to fit as there is only one predictor considered and the results, when exponentiated, give an odds ratio for the comparison of being dependent to being independent between the oxygen treatment group and the standard care group.

The other model fitted to the data was a repeated measures logistic regression model; this allowed all follow-up points in the data to be considered for the model. The model is able to estimate the treatment effect at any point during the 12-month follow-up, by using an appropriate contrast. The model is also able to take into account the correlation between the mRS scores as they are being estimated at several time points included in the model. The model fitted is described in Equation 4.2, with the same covariate of treatment that was fitted in the first model, but also with the inclusion of time. As there were only 3 follow-up time points and they were spaced irregularly, time was included in the model as a categorical variable by using dummy variables for each time point and not as a continuous time variable. Due to the inclusion of time, an interaction term between treatment and time was also included in the model. This allows the two groups to vary differently over time, as it would be wrong to assume that the groups vary the same over time. The model was fitted with a random intercept to allow the individuals to have subject specific effects over time. The following logistic regression model was fitted to the data, in order to calculate the log odds (Equation 4.2):

$$Ln\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 Treatment_i + \beta_2 Time1_j + \beta_3 Time2_j + \beta_4\left(Treatment_i \times Time1_j\right) + \beta_5\left(Treatment_i \times Time2_j\right) + v_i \qquad [4.2]$$

Where β_0 represents the log odds estimates for the standard care group at 3 months and $Time1$ is the contrast from 6 months to 3 months and $Time2$ contrasts 12 months to 3 months. *Treatment* was coded 0 for the standard care group and 1 for the treatment group. Due to the coding of the model, we find that β_1 represents the difference in the 2 groups at 3 months and β_2, β_3 represent the difference of the effects at 6 and 12 months compared to 3 months for the standard care group. Finally, β_4 and β_5 represent the difference in the time effects between groups, which is the extra difference seen in the treatment group compared the standard group at the follow-up points and v_i is included due to the random effects included in the model. The odds ratio is calculated from the repeated measures model by taking the exponential of the estimated parameter and expresses the odds of being dependent compared to independent between the treatment group and the standard care group.

4.6.2 Ordinal logistic regression

It has already been explained why dichotomising an ordinal scale is not the best way to perform an analysis due to the power and information that is lost. The next analysis fits the same two types of model, the single time point regression model and the repeated measures model; however, the data that are being modelled is now the full ordinal mRS scale rather than the dichotomised version. This model extension means that the log odds of being in a category or lower are calculated rather than the odds of being dependent or independent.

Equation 4.3 shows the model fitted to the single time point using the ordinal scale was as follows,

$$Ln\left[\frac{P(Y_{ij} \leq k)}{1-P(Y_{ij} \leq k)}\right] = \alpha_{ok} + \beta_1 Treatment_i \qquad k=1,\ldots,k-1 \quad [4.3]$$

Where α_{ok} are the *k-1* intercept terms to model the marginal frequencies of the *k* ordered categories, which represent the log estimate for the standard care group. Equation 4.4 extends the model in order to fit the repeated measures data:

$$Ln\left[\frac{P(Y_{ij} \leq k)}{1-P(Y_{ij} \leq k)}\right] = \alpha_{ok} + \beta_1 Treatment_i + \beta_2 Time1_j + \beta_3 Time2_j + \beta_4(Treatment_i \times Time1_j) + \beta_5(Treatment_i \times Time2_j) + v_i \quad k=1,\ldots,k-1 \quad [4.4]$$

As before, α_{ok} are the k-1 intercept terms to model the marginal frequencies of the *k* ordered categories, which in the model represent the log odds estimates for the standard care group at 3 months. The covariate *Time1* contrasts 6 months to 3 months and *Time2* contrasts 12 months to 3 months and *Treatment* is coded 0 for the standard care group and 1 for the treatment group. Due to the coding of the model, we find that β_1 represents the difference in the 2 groups at

3 months and β_2, β_3 represent the time effects for the standard care group at 6 and 12 moths respectively compared to the 3 month effect. Finally, β_4 and β_5 represent the difference in the time effects between groups, which as before is the additional difference in log odds for the treatment group compared to the standard care group and v_i is included due to the random effects included in the model. Now, instead of the model calculating the odds of being dependent on the dichotomised mRS compared to independent, this model looks at the odds of being in a category or higher category combined on the mRS compared to the odds of being in any lower category combined.

4.6.3 Partial proportional odds model

It is not possible to conduct a Brant test on the repeated measures model in order to assess the assumption of proportional odds. It is possible to extend to ordinal regression model with proportional odds that has been fitted into a partial proportional odds model as detailed in section 4.2.4. The previous model included homogenous covariates of treatment, time and the treatment time interaction. This meant they were not allowed to vary for each potential cut-off point but were forced to be the same, for proportional odds. The first partial proportional odds model fitted heterogeneous treatment and time but still included a homogenous treatment by time interaction. By including heterogeneous covariates, the proportional odds assumption was relaxed and these covariates were allowed to vary for each cut off point. Equation 4.5 shows the partial heterogeneous repeated measures model fitted was (Hedeker & Mermelstein 2000):

$$Log\left[\frac{P(Y_{ij}\leq k)}{1-P(Y_{ij}\leq k)}\right] = \alpha_{ok} + \alpha_{1k}Treatment_i + \alpha_{2k}Time1_j + \alpha_{3k}Time2_j + \beta_4(Treatment_i \times Time1_j) + \beta_5(Treatment_i \times Time2_j) + v_i \qquad [4.5]$$

Where the regression covariates fitted for time and treatment are now represented by alpha rather than β, showing that they are estimated, and like the α_{ok} terms where there are the $K\text{–}1$ terms to model the marginal frequencies. The second partial proportional odds model fitted heterogeneous covariates, relaxing the assumption of proportional odds for all the covariates included in the model, Equation 4.6. The full heterogeneous repeated measures model fitted was (Hedeker &Mermelstein 2000):

$$Log\left[\frac{P(Y_{ij}\leq k)}{1-P(Y_{ij}\leq k)}\right] = \alpha_{ok} + \alpha_{1k}Treatment_i + \alpha_{2k}Time1_j + \alpha_{3k}Time2_j + \beta_{4k}\left(Treatment_i \times Time1_j\right) + \beta_{5k}\left(Treatment_i \times Time2_j\right) + v_i \qquad [4.6]$$

4.7 Results

4.7.1 Binary logistic regression

Firstly, before the analysis could be undertaken, the mRS was dichotomised into 2 categories: those who were independent and those who were dependent, Table 4.2. Patients who had a favourable outcome were classed as independent received a mRS of 0-2 as these people were able to perform activities and live the life that they were previously able to do before their stroke. If patients received a mRS of 3-6, they were classed as dependent, the unfavourable outcome, as they were no longer able to live the life they could before the stroke and required assistance, or they had died. After much discussion, this was decided to be the optimal cut-off that made clinical sense and for which there was a definite distinction between the two groups: those who could still do activities that they could before their stroke and those who required assistance to live the life they had before the stroke. The details of the number of patients who are in each of the dichotomised scores are given below for both the treatment group and the standard care group. The trial had randomised patients on a ratio of 1:1:1 to receive either constant oxygen, nightly oxygen or no oxygen (control); the two types of treatment were combined meaning that there was a ratio of 2:1 for the numbers of patients in the treatment and control groups.

Table 4.2: Number of individuals dichotomised into each outcome at each of the follow-up points

Follow-up (months)	**Standard care**			**Treatment**		
	Favourable / independent	Unfavourable/ dependent	Total	Favourable / independent	Unfavourable/ dependent	Total
3	1,317 (52%)	1,206 (48%)	2,523	2,661 (53%)	2,385 (47%)	5,046
6	1,305 (52%)	1,174 (48%)	2,479	2,618 (53%)	2,333 (47%)	4,951
12	1,178 (49%)	1,191 (51%)	2,369	2,430 (52%)	2,267 (48%)	4,697

Looking at both the treatment and standard care group we see that the numbers of individuals in each of the categories of the dichotomised outcome are quite evenly split. There appears to be more of a difference in the treatment group, but there are almost twice as many individuals in this group compared to the standard care group. The observed odds ratios can be calculated for both the oxygen treatment group and the standard care group over the three follow-up points. We can also calculate the logits for each of the time points. The observed odds ratios are given in Table 4.3 below, with the logits associated with the odds ratios given in brackets. The logistic regression model was fitted to the data in order to examine the effect of treatment more formally.

Table 4.3: Odds and logits of being dependent compared to independent calculated for each follow-up point

Follow-up	**Standard care**	**Treatment**
3	1,206/1,317 = 0.92 (-0.080)	2,385/2,661 = 0.90 (-0.100)
6	1,174/1,305 = 0.90 (-0.100)	2,333/2,618 = 0.89 (-0.110)
12	1,191/1,178 = 1.01 (0.009)	2,267/2,430 = 0.93 (-0.070)

The regression model at the single time point estimates the log odds of the model, which is given in the following table, Table 4.4, along with the standard error and confidence interval.

Table 4.4: Binary logistic regression models fitted to data at 3 months

	Estimate	Standard error	95% Confidence interval
Treatment	-0.021	0.049	-0.271, 0.033
Intercept	-0.088	0.040	

The model shows that at 3 months those in the oxygen treatment group have a reduction in the log odds of being dependent compared to the standard care group of 0.02. This can be converted into an odds ratio by taking the exponential of the parameter estimate for the treatment effect of the odds of being dependent, which for just the mRS scores recorded at 3 months is calculated to be 0.98, suggesting that there is a 2% decrease in the odds of being dependent for those individuals on the oxygen treatment. The repeated measures model is more complex than the model fitted at the single time point; it now considers the effect of time on the outcome as well as the treatment. The following table, Table 4.5, presents the coefficients for the log odds of the model, along with the standard errors.

Table 4.5: Repeated measures binary logistic regression model results

	Homogeneous effects		
	Estimate	**Standard error**	**95% Confidence interval**
Intercept	-0.173	0.129	-0.427, 0.080
Treatment	-0.100	0.143	-0.410, 0.209
6 month	-0.109	0.101	-0.308, 0.089
12 month	0.261	0.103	0.059, 0.463
Treatment x 6 month	0.020	0.124	-0.222, 0.263
Treatment x 12 months	-0.212	0.126	-0.458, 0.034
Subject SD	4.787	0.115	4.567, 5.019
Log-likelihood	-11,145.6		

The regression model fitted to the data calculates the log odds of being dependent compared to independent. We see that the treatment group has a reduction of log odds of 0.100 compared to the standard care group at 3 months. There is also a reduction in the log odds of being dependent compared to independent in the standard care group at 6 months (-0.109) but an increase at 12 months (0.261) compared to 3 months. For those in the treatment group, the change in log odds is 0.020 for 6 months compared to 3 months and an increase of -0.212 for 12 months compared to 3 months. Because of how the repeated measures model is defined, the treatment effect at 3 months is easily found by the model. The odds ratio of being dependent for the treatment group compared to the standard care group is calculated to be 0.90. This suggests that there is a 10% reduction in the odds of being independent for those individuals who received oxygen treatment. The subject standard deviation given in Table 4.5, details the standard deviation of the random effect term in the model, which looks at the heterogeneity. The value for this model suggests that there is heterogeneity between the individuals included, as if there was no heterogeneity then the standard deviation would be close to 0. This value is quite large and it suggests that there is substantial variability in the propensity to experience a good outcome. However, this is to be expected as there is only a treatment covariate in the model. The log-likelihood is also presented; it is commonly used to compare models. However, as it is a function of the sample size, in order to compare models there needs to be the same number of individuals included in each of the models.

Table 4.6 below compares the odds ratio and log odds of being dependent for the treatment group compared to the standard care group from the two different types of model fitted to the data. The standard errors and 95% confidence intervals are also presented.

Table 4.6: Comparison of model results for all binary regression models fitted

Treatment vs standard care	**Log odds**	**95% CI**	**Standard error**	**Odds ratio**	**95% CI**
Single time					
3 months	-0.021	-0.271, 0.033	0.048	0.979	0.891, 1.069
Repeated measures					
3 months	-0.100	-0.410, 0.209	0.143	0.904	0.660, 1.199

It can be seen that there is a difference of 0.079 between the log odds of an individual being dependent at 3 months post-stroke (-0.021 – (-0.10)), with both models agreeing that the oxygen treatment group shows a reduction in the log odds of being dependent compared to independent, suggesting that both models show that the oxygen treatment may be beneficial. Although it can be seen from the confidence intervals calculated that neither reduction in the log odds is statistically significant. We see that by taking the exponential of the log odds using a single time point the odds ratio is calculated to be 0.98 (95% CI 0.89, 1.07) and using a repeated measures model the odds ratio is calculated to be 0.90 (95% CI 0.66, 1.2). The standard error calculated from the repeated measures model is much larger due to the inclusion of random effects in the model in order to account for the correlation between the repeated measures. We see that the odds ratio suggests that there is a larger effect of treatment for the repeated measures effect than the effect seen at 3 months, as the odds ratio is further from the null than the crude odds ratios calculated in Table 4.3. There could be many reasons for this; however, the most probable is that there are missing data in the repeated measures models, which could affect the estimation as there will be different numbers included in both the model and the observations at each time point.

Also, the inclusion of a random intercept represents the combined effect of omitted subject-specific covariates. By including a random intercept we assume that there in a random heterogeneity in a subjects underlying risk of a favourable outcome that persists throughout. These differences in the odds ratios could also be due to a different interpretation of the parameters in the two models, as the random effects model describes the effect of treatment conditional on the specific subject.

4.7.2 Ordinal logistic regression

Table 4.7 below shows the distribution of the mRS scores at each time point for both the standard care and treatment groups using the whole mRS scale with no dichotomisation.

Table 4.7: Distribution of individuals on each mRS score at each follow-up point stratified by treatment

Follow-up (months)	**Standard care**							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	292 (12%)	710 (28%)	315 (13%)	422 (17%)	420 (17%)	168 (7%)	196 (8%)	2,523
6	315 (13%)	667 (27%)	323 (13%)	402 (16%)	374 (15%)	125 (5%)	273 (11%)	2,479
12	303 (13%)	608 (26%)	267 (11%)	373 (16%)	358 (15%)	111 (5%)	349 (15%)	2,369
	Treatment							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	649 (13%)	1,361 (27%)	651 (13%)	875 (17%)	769 (15%)	304 (6%)	437 (9%)	5,046
6	705 (14%)	1,234 (25%)	679 (14%)	819 (17%)	713 (14%)	238 (5%)	563 (11%)	4,951
12	661 (14%)	1,179 (25%)	590 (13%)	688 (15%)	681 (15%)	201 (4%)	697 (15%)	4,697

It can be seen that there are more variations in the scores that people receive at 3 months, 6 months and 12 months than in those of the dichotomised scale. This shows how the dichotomised scale loses information when the categories are condensed, as these differences would not be able to be compared. As with the binary logistic regression, it was possible to fit the ordinal regression models individually at each of the follow-up points. Table 4.8, shows the log odds for the treatment effect at 3 months.

Table 4.8: Model results for ordinal logistic regression at 3 months

	Estimate	**Standard error**	**95% CI**
Treatment	-0.028	0.041	-0.111, 0.062

It can be seen that those individuals in the oxygen treatment group have a 0.03 reduction in the log odds of being in a higher mRS category than those individuals in the standard care group. The estimate of the treatment effect can be used to calculate an odds ratio for the odds of being in a higher category on the mRS, which for the model fitted at the single time point is 0.97. So at any category on the scale an individual in the oxygen treatment group has a 3% reduction in the odds of being that category or higher on the mRS, to being in a lower category compared to the standard care group. A higher mRS category indicates higher disability and so this suggests that those in the treatment group are less likely to have a higher disability than those in the standard care group.

As mentioned before, the ordinal regression model works by assuming that there is one common odds ratio for all cut-off points on the scale. This is a substantial assumption to make and one that can be tested using a Brant test. The 3-month model produces a *p*-value of 0.162 for the Brant test. This suggests that the proportional odds assumption has not been violated and that the model fitted satisfies the underlying assumptions.

Below in Table 4.9, the coefficients for the log odds for the repeated measures model are presented along with the standard errors and 95% confidence interval. The ordinal logistic regression model has an assumption that there are proportional odds across each categorisation of the mRS; the validity of this assumption cannot be found using a Brant test for the repeated measures data.

Table 4.9: Model results for repeated measures ordinal logistic regression

	Estimate	**Standard error**	**95% confidence interval**
Treatment	- 0.073	0.137	-0.340, 0.201
6 month	- 0.048	0.059	-0.155, 0.067
12 month	0.238	0.061	0.121, 0.36
Treatment x 6 month	- 0.018	0.072	-0.161, 0.123
Treatment x 12 months	- 0.126	0.074	-0.265, 0.020
Subject SD	5.112	0.814	4.961, 5.274
Log-likelihood	- 31,911.9		

The regression model fitted to the data calculates the log odds of being in a higher mRS category compared to a lower one. We see that the treatment group has a reduction of log odds of 0.07 compared to the standard care group at 3 months. In the standard care group, there is also a reduction in the log odds of being dependent compared to independent at 6 months (-0.048) but an increase at 12 months (0.238) compared to 3 months. In the treatment group the difference in the log odds at 6 months compared to 3 months is -0.018 and an increase of 0.126 in the log odds at 12 months compared to 3 months. Once again the way that the repeated measures model is defined means the treatment effect at 3 months is easily found by the model. The estimate of the treatment effect can be used to calculate an odds ratio for the odds of being in a higher category on the mRS, which for the model fitted at the single time point is 0.93 (95% CI 0.71, 1.21). Here the subject standard deviation suggests there is more heterogeneity in the model, when the outcome is included in an ordinal manner, but once again there is only the treatment covariate in the model to explain the difference seen between individuals. The log-likelihood is much larger in the ordinal model, which is expected with a more complex model to be fitted.

Table 4.10 summarises the odds ratio and log odds of being in a higher mRS category for the treatment group compared to the standard care group from the two different types of model fitted to the data. The standard errors and 95% confidence intervals are also presented.

Table 4.10: Comparisons on ordinal regression model results

Treatment vs Standard care	**Log odds**	**95% CI**	**Standard error**	**Odds ratio**	**95% CI**
Single time					
3 months	-0.028	-0.111, 0.062	0.041	0.972	0.889, 1.050
Repeated measures					
3 months	- 0.073	-0.340, 0.201	0.137	0.929	0.708, 1.210

It can be seen that the difference in the log odds calculated at 3 months in the two models is 0.04, which is a smaller difference than the difference in the log odds from the binary model, suggesting that the single time point and repeated measures model agree more for the ordinal regression model than the binary logistic regression model. Once again the log odds calculated for both models are not statistically significant and the repeated measures model has a larger standard error due to the inclusion of random effects. The model fitted to a single time point produced an odds ratio of 0.97 (95% CI 0.89, 1.05) and for the model fitted with repeated measures, the treatment effect at 3 months gives an odds ratio of 0.93 (95% CI 0.71, 1.21). It can be seen that the odds ratio calculated for the repeated measures model gives a bigger reduction in the odds of being in a higher category on the mRS compared to the single time point model. The difference in the odds ratios seen is due to the repeated measures model taking into account what people score throughout the follow-up and not just at the 3 month value.

It had previously been decided that because there were only 3 time points in the data, and they were irregularly spaced, using time in a categorical form would be the best way to use the time variable in the analysis. The effects of including time as a continuous variable were also considered to see if it changed the conclusions drawn from the ordinal regression models. The first model fitted included time as a continuous variable, having been centred at 3 months. The second model transformed the value of time into the square root value of time, which helps to linearise the covariate and force it to be well behaved. Both models fitted used random effects to adjust for the correlation between the outcomes at each of the time points, due to there being repeated measures, as before

Table 4.11 looks at the treatment effect and the standard error for the treatment effect, specifically at the first time point in the model. This is compared with the treatment effect when categorical time was used. As well as this it was considered how well the model fits the data by looking at the value of the log-likelihood, Akaike information criterion (AIC) and Bayesian Information Criterion (BIC), full descriptions of these are given in Chapter 8.

Table 4.11: Comparison of model results with categorical, continuous and square root time

	Categorical time	Continuous time	Continuous time – square root
Treatment effect (OR)	-0.073 (0.93)	-0.063 (0.94)	-0.059 (0.94)
Standard error	0.139	0.136	0.137
Log-likelihood	-31911.91	-31918.99	-31924.93
AIC	63847.81	63857.98	63869.85
BIC	63943.84	63938	63949.87

By using continuous time instead of categorical time, it can be seen that the difference in the log odds values of the treatment and standard care group are similar to the difference in log odds between the treatment and standard care group calculated by the repeated measures model when using categorical time. All these treatment values presented in Table 4.11 are calculated at 3 months. There is a small reduction in the standard errors associated with the treatment effect when using continuous time. Both models, continuous time and square root time, calculated an odds ratio of 0.94, suggesting that in both models the treatment group has a 6% reduction in the odds of being in a higher category, compared to a 7% reduction when using categorical time. It must be noted that in the models, all the treatment effect values are not significant. The smallest value of the log-likelihood is found when the model includes categorical time, as is the smallest value of the AIC, whereas for the BIC the smallest value is found when the time variable is included in a continuous form. As the smallest log-likelihood and AIC values occur in the model with categorical time, this suggests that is may be the best fitting model when treatment, time and the treatment time interaction are included in the model. The treatment time interaction is actually not significant for any of the models, so it was removed to see how it affected the treatment effect calculated, Table 4.12.

Table 4.12: Comparison of model results after interaction removal for categorical, continuous and square root time

	Categorical time	Continuous time	Continuous time – square root
Treatment effect (OR)	-0.117(0.89)	-0.117 (0.89)	-0.118 (0.89)
Standard error	0.132	0.132	0.132
Log-likelihood	-31913.53	-31920.55	-31926.19
AIC	63847.06	63859.09	63870.37
BIC	63927.07	63931.11	63942.39

After the removal of the interaction term, which allows the effect of the treatment group to vary differently to the log odds over time of the control group over time, the difference in the log odds between treatment and control groups at the 3-month follow-up point increases. The log odds calculated by the model is still the 3-month effect, but both of the treatment groups are assumed to change the same over time rather than being allowed to vary individually. Now for all 3 models the difference in log odds is approximately -0.117, which is equivalent to an odds ratio of 0.89, suggesting that the treatment group has an 11% reduction in the odds of being in a higher category on the scale when the influence of the repeated measures values is considered. As before, the categorical time model has the smallest log-likelihood value, and also the smallest AIC and BIC values, suggesting that it is the best fitting model to those data.

The inclusion of time as a continuous variable was considered because it may be of help if the random effects model is extended to include more random effects than just allowing variation between individuals. By including a continuous time variable, the number of parameters in the model are reduced as categorical time requires dummies and therefore there are more parameters in the models. The use of continuous time was assessed graphically, and the assumption of a linear association was valid with a log odds outcome. The current model only includes a random intercept and allows each individual to vary, but they all follow the same trajectory. By including more random effects, this trajectory is able to be varied and a random slope can also be included as well as the random intercept. This is done by placing random effects on each of the covariate terms in the model. When these random effects were added to the model with categorical time, the model struggled to converge as there were challenges with the numerical derivation of the model. The magnitude of random effects that were found was very small, in the order of 10^{-36}, and did not change the log-likelihood calculated and therefore did not fit a better model. Considerations of parsimony suggest that, in this case, fewer random effects are better, as there are too many random effects that the model is trying to estimate when the time is categorical. Using continuous time, these extra random effects were also added to the model, but like the categorical time model, the inclusion did not provide a treatment estimate that was any closer to the single time point model at 3 month, or warrant the extra random effects in the model.

4.7.3 Partial proportional odds model

It is not possible to conduct a Brant test on the repeated measures model in order to assess the assumption of proportional odds. Statistical significance is defined at a 5% level, where a *p*-value <0.05 is statistically significant, so considering the repeated measures model fitted, by looking at the *p*-values of the included covariates it can be seen that the 12 months covariate is statistically significant (*p*<0.001), which suggests that the proportional odds assumption may not hold for the repeated measures model. Table 4.13 shows the results of the partial proportional odds model with heterogeneous time and treatment effects, but homogeneous interaction effects. It can be seen from Table 4.14 that the interaction between treatment and month has been kept constant across each of the cutoff points as these are the covariates that still follow the proportional odds assumption. Looking at the treatment effect we can see that there is quite a lot of variation in the log odds calculated for each cutoff point. It is important to note that the covariate is still not statistically significant for all of the cutoff points, except for the 0 vs 1-6 cut off.

Table 4.13: Model results for partial proportional odds model – mixture of homogeneous and heterogeneous covariates

Parameter (SE)	0 vs 1-6	0-1 vs 2-6	0-2 vs 3-6	0-3 vs 4-6	0-4 vs 5-6	0-5 vs 6
Treatment	-0.343 (0.162)*	0.033 (0.1427)	-0.06 (0.142)	0.004 (0.145)	0.067 (0.163)	0.143 (0.181)
6 month	-0.26 (0.091)	0.002 (0.074)	-0.142 (0.074)	-0.118 (0.078)	0.013 (0.096)	0.701 (0.119)
12 month	-0.259 (0.094)	0.033 (0.077)	0.066 (0.076)	0.31 (0.081)	0.601 (0.097)	1.814 (0.121)
Treatment x 6 month	-0.019 (0.073)*	-0.019 (0.073)	-0.019 (0.073)	-0.019 (0.073)	-0.019 (0.073)	-0.019 (0.073)*
Treatment x12 month	-0.123 (0.073)*	-0.123 (0.073)	-0.123 (0.073)	-0.123 (0.073)*	-0.123 (0.073)*	-0.123 (0.073)*

* Indicates statistical significance (p<0.05)

Looking at the time covariate, there is also a lot of variation between the log odds estimated for each of the cutoff points. There are statistically significant values for 0 vs 1-6 and 0-5 vs 6 for the 6 month time point and 0 vs 1-6, 0-3 vs 4-6, 0-4 vs 5-6 and 0-5 vs 6 for the 12-month time point.

Table 4.14: Model results for partial proportional odds model – all heterogeneous covariates

Parameter (SE)	0 vs 1-6	0-1 vs 2-6	0-2 vs 3-6	0-3 vs 4-6	0-4 vs 5-6	0-5 vs 6
Treatment	-0.437 (0.185)*	-0.005 (0.153)	0.008 (0.159)	-0.099 (0.158)	-0.001 (0.189)	0.232 (0.239)
6 month	-0.257 (0.138)	-0.018 (0.098)	-0.140 (0.097)	-0.151 (0.107)	0.006 (0.143)	0.827 (0.192)*
12 month	-0.359 (0.139)*	0.064 (0.101)*	0.175 (0.099)	0.228 (0.109)	0.562 (0.143)*	1.897 (0.192)*
Treatment x 6 month	-0.023 (0.166)	0.014 (0.119)	-0.023 (0.118)	0.033 (0.131)	-0.014 (0.177)	0.024 (0.169)
Treatment x12 month	0.025 (0.169)	-0.166 (0.123)	-0.287 (0.122)*	0.005 (0.134)	-0.075 (0.176)	-0.265 (0.232)

* Indicates statistical significance (p<0.05)

When the heterogeneous model is applied to the data, the estimates for treatment vary slightly but are still similar to the partial heterogeneous model, the standard errors have increased a little, and only the 0 vs 1-6 cut off has a statistically significant estimate. After the interaction term is included, only the 0-5 vs 6 cut off is statistically significant for the 6 month covariate, but the statistically significant estimates for 12 months are the same. The interaction terms obviously now vary and are found to be not statistically significant for all the treatment x 6 months estimates, and only statistically significant for the estimate at the 0-2 vs 3-6 cut off for the treatment x 12 month interaction.

The goodness of fit of the model can be explored by looking at the -2 log-likelihood value as well as the AIC and BIC values calculated from each model. The -2 log-likelihood was considered here, as that was what the partial proportional odds regression model produced. Table 4.15 gives these values for the ordinal regression model with homogeneous covariates (proportional odds model), the ordinal regression model with partial homogeneous covariates, and the ordinal regression model with all heterogeneous covariates, calculated using the code given in Appendix B (Carriere & Bouyer 2006).

Table 4.15: Goodness of fit statistics for proportional and partial proportional odds models

	Proportional odds model	**Partial proportional odds model - treatment and time**	**Partial proportional odds model – all covariates**
-2 log-likelihood	63822	64024	64017
AIC	63847	64078	64091
BIC	63944	64266	64349

The smallest values for all 3 criteria are found in the proportional odds model, which includes a random effect to account for the repeated measures but includes all covariates as satisfying the proportional odds assumption. This suggests that relaxing the proportional odds assumption does not improve the fit of the model and the spread of variation seen shows us that these models are still not able to estimate the single time point treatment effect.

Looking at the partial proportional odds models has shown that there is no improvement in fit over the proportional odds model; however, we are still unsure how well this model fits the data due to the lack of a Brant test. It is possible to check how well the proportional odds model fits by comparing the numbers of estimated individuals in each treatment group for each of the mRS scores with the numbers observed. Therefore, plots have been produced that show both the expected and the observed proportion of individuals in both the treatment and controls groups at each of the follow-up time points. The following plots show the predicted proportions of individuals in the treatment and standard care groups for the repeated measures ordinal model at 3, 6 and 12 months in the follow- up. This is done separately for each value of the mRS scale.

Figure 4.2: Observed and expected proportions of individuals during follow-up when mRS=0

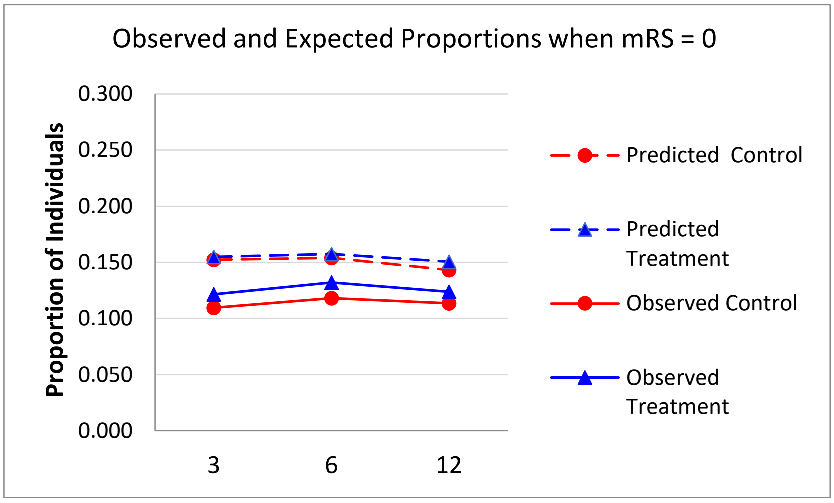


Figure 4.3: Observed and expected proportions of individuals during follow-up when mRS=1

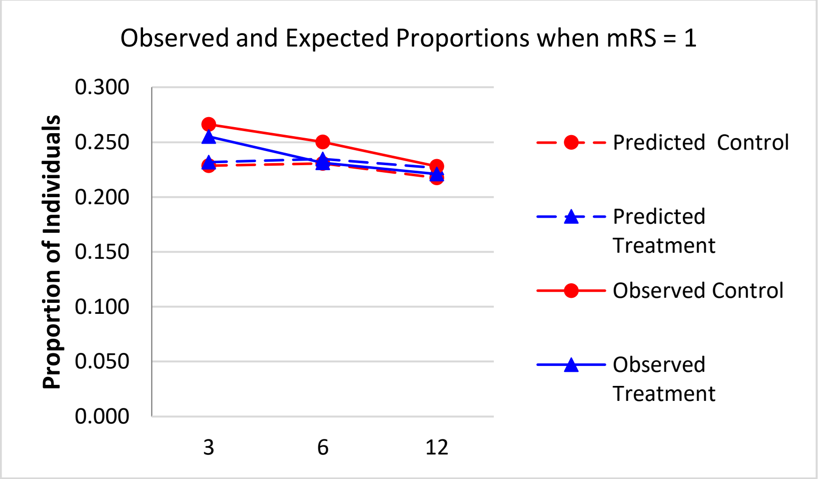


Figure 4.4: Observed and expected proportions of individuals during follow-up when mRS=2

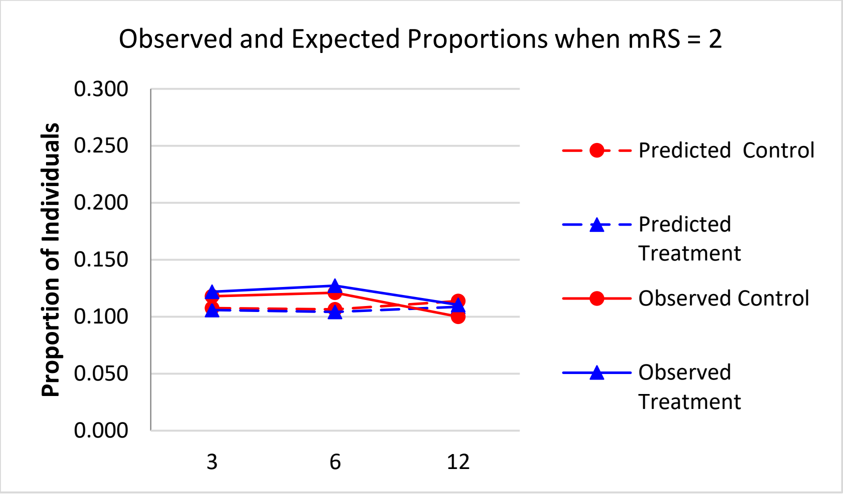


Figure 4.5: Observed and expected proportions of individuals during follow-up when mRS=3

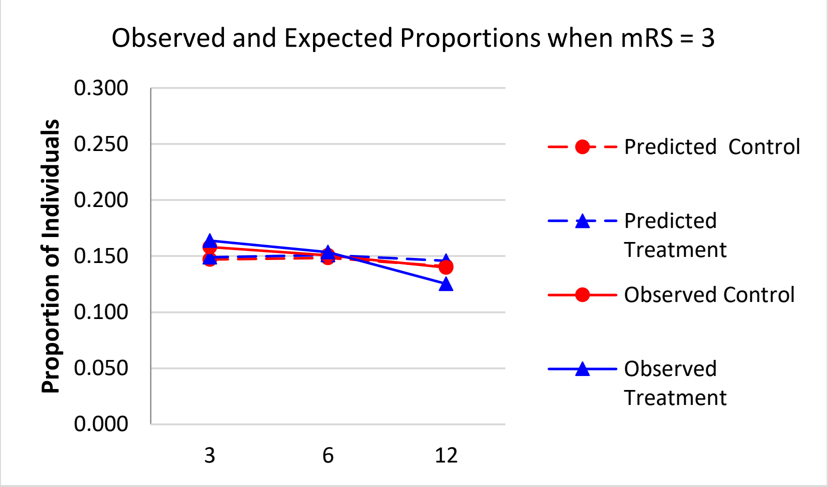


Figure 4.6: Observed and expected proportions of individuals during follow-up when mRS=4

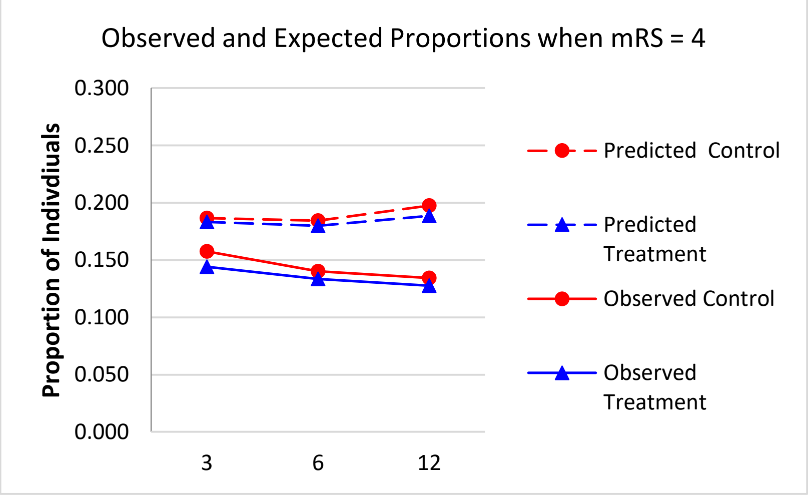


Figure 4.7: Observed and expected proportions of individuals during follow-up when mRS=5

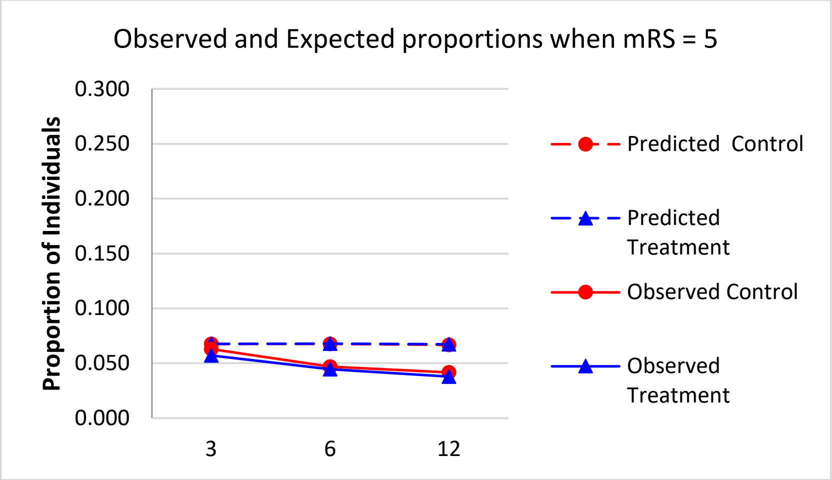
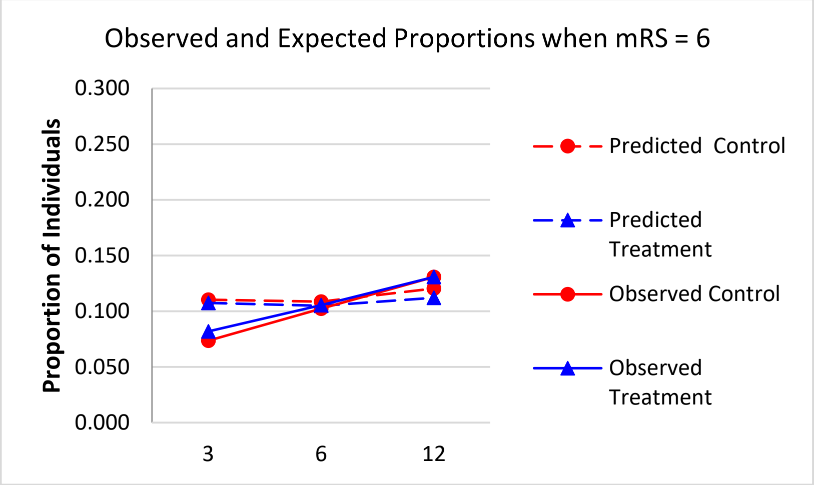


Figure 4.8: Observed and expected proportions of individuals during follow-up when mRS=6



Figures 4.2 to 4.8 show the proportion of individuals observed and expected from the ordinal regression model with proportional odds. It can be seen that the random effects repeated measures model shows some variation between the observed and expected values, with some categories fitting better than others. The random effect included in the model has allowed for a random intercept but a fixed slope so it assumes that everyone has the same change in the mRS score over time. When the mRS is 0 the model looks like it may over predict the proportions of individuals in the treatment and control groups. When the mRS is 1, the model slightly underestimates the proportion in each group at 3 months but is very similar at 12 months. When

the mRS is 2 once again the model underestimates the proportions in each group at 3 months but is well fitted at 12 months. When the mRS is 3 the model fits well at 3 months and at 12 months for the standard care group but overestimates the oxygen treatment group, which has a sudden drop in the observed numbers compared to the standard care group. When the mRS is 4 the model over predicts the proportion of the expected numbers in each of the treatment group compared to the observed numbers. When the mRS is 5 the model estimates the proportions in each group well at 3 months, but slightly over-estimates the proportions at 12 months. Finally, when the mRS is 6 the model overestimates the proportions seen at 3 months but estimates the proportion in each group well at the 12 months follow-up time point. In most of the graphs the numbers of observed individuals in the oxygen treatment groups and the standard care group are fairly similar. It is unclear why the model fits some of the categories on the mRS better than others, although looking at the graphs it seems that where there are larger changes over time in the observed number in each of the group, the poorer the fit is between the observed and expected values.

4.8 Discussion

This chapter aimed to assess potential regression models for the analysis of the mRS scale. Several different models were considered, in the end only 2 different types of model were used. A logistic regression model was used, with the scale dichotomised into a binary variable and an ordinal regression model. Both models were fitted to a single time point and all the time points in the data. It was found that the treatment effects calculated were not statistically significant but that there was variation in the effects calculated dependent on the type of model fitted.

These models fitted calculate a treatment effect by assessing its effect on the value of the mRS. These methods have been applied before in clinical stroke trials and therefore are an obvious choice for analysing the longitudinal clinical trial data using the mRS. It is possible that these methods can be extended to analyse data that differs from the data included in this analysis. For example, the models are able to handle data with more time points in the model, and there would be no effect if these extra points were regularly spaced or irregularly spaced time points, as has been shown in the current analyses. The model would be able to handle the inclusion of a baseline measurement, indeed this would be of benefit to the analysis and is a limitation of using the mRS scale as the main outcome. Missing data has been suggested to affect the treatment effect calculated in the repeated measures model, but the model handles some missing data well, although little to no missing data is preferable. Also if the data shows little or no change over time the models can also deal with this, as well as modelling data that shows large changes over time.

4.8.1 Model comparison

Both binary and ordinal logistic regression models were fitted to the SO_2S data. There were two different models fitted, the first looked at the mRS values at just the 3-month time point and the second fitted a repeated measures model, including all three time points in the model. From this model it is possible to estimate the effects at 3 months as well and compare to the previous model.

It is very hard to compare the single time point models and the repeated measures models using the usual goodness of fit statistics, AIC, BIC or log-likelihood. This is due to there being different numbers of observations in the models due to the there being one or three time points.

In the binary regression model, the repeated measures model reduced the odds ratio from 0.98 in the single time point model to 0.90 in the repeated measures model. This model included a dichotomised mRS with 0-2 representing independent individuals, the favourable outcome and 3-6 representing dependent individuals. Both of these odds ratios show that the treatment is favourable as there is a reduction in the odds of being dependent compared to independent for the treatment group in comparison to the standard care group at 3 months. This reduction increases from 2% for the single time point model to 10% in the repeated measures model. Both of the odds ratios calculated are not statistically significant, as the confidence intervals contain 1, which would indicate no difference between treatments. The standard errors calculated from the repeated measures model were much larger due to the random effects in the model in order to account for the correlation between the repeated measures. By dichotomising the mRS scale, information from the scale is lost and may exaggerate (or reduce) the differences seen in comparing the two treatment groups due to categories in the scale being combined and only compared across one value of the mRS.

Because of this an ordinal regression model was fitted to the data using the same data but keeping it in the form of a seven point ordinal scale rather than dichotomising it. The model fitted to a single time point produced an odds ratio of 0.97 and the model fitted with repeated measures gave an odds ratio of 0.93 at 3 months. It can be seen that the odds ratio calculated for the repeated measures model gave a bigger reduction in the odds of being in a higher category on the mRS (for the treatment group compared to standard care) than the single time point model. The ordinal regression model calculates the odds of being in a higher category on the mRS, which has a reduction of 3% for the treatment group compared to the standard care group in the single time point model and a 7% reduction in the odds for the treatment group compared to the control group for the repeated measures model. The difference in the odds ratios produced by the ordinal models at 3 months are smaller than the difference in the odds ratios produced by the binary models at 3 months. Once again the log odds calculated for both models were not statistically significant and the repeated measures model had a larger standard error due to the inclusion of random effects. The difference in the odds ratios seen was due to the repeated measures model taking into account what people score throughout the follow-up and not just at the 3-month value.

The proportional odds assumption was tested for the single time point model using a Brant test; this calculated a *p*-value of 0.162 for the treatment variable, suggesting that the proportional odds assumption is valid. The proportional odds assumption could not be tested for the repeated measures model. However, fitting a model where part of the proportional odds assumption was relaxed showed no improved fit or reduced log-likelihood compared to the proportional odds model, neither when the assumption was relaxed only for treatment and time nor when relaxed for all of the covariates in the model.

The use of categorical time was also investigated, comparing it with using continuous time in the model. The use of continuous time was considered using time centred on 3 months and also using a square root transformation of time. Neither improved the fit of the model from the categorical time repeated measures model. The fit of the model was also considered by plotting the observed and expected values of individuals with each mRS score at each time point. These graphs show that the fit of the model is not too bad; for some scores the observed and predicted values are very close together, for others the fit of the model is worse.

The differences seen between the single time point model and the repeated measures model are apparent and they occur due to the difference in the modelling properties, whereby the repeated measures model borrows an effect from all the considered time points; the values are therefore influenced by either future or past observations. This is introduced as a random effect, which adds uncertainty into the model, reflected in the size of the confidence intervals, which are wider than those of the single time point model.

4.8.2 Further work

The main analysis in this chapter looked only at the inclusion of the treatment variable, as well as a time covariate in the repeated measures model. An analysis was run with basic covariates included (age, gender and baseline NIHSS), and although statistically significant themselves, they had no effect on the treatment estimate, which is the only variable of interest in the models fitted here. It may also be of interest to fit and undertake a full model selection with available covariates now that it has been decided that the ordinal regression models are better than binary regression models when considering the ordinal longitudinal analysis of the mRS.

The systematic review of mRS longitudinal analysis of the mRS identified a paper for which a joint model of the mRS and survival had been conducted. Due to restrictions with the SO_2S data and because this had already been conducted within the NINDs data and shown to be appropriate, it was decided to focus just on the effects of longitudinal models in the treatment effect and to not consider survival as well. This however could be a good extension to the regression models fitted, but may require a longer follow-up period to obtain more repeated measures to model as 3 time points is unable to give a reasonable profile if an individual was to die.

As well as this, more consideration could be taken into methods that compare the goodness of fit of a model when the AIC/BIC cannot be compared in this situation. Roozenbeck et al. (2011) calculated the ratio of the Wald statistics, which can be interpreted as the gain in information density, and therefore, could be considered a suitable measure for the efficiency of the different approaches and allow two different types of model to be compared (Roozenbeek et al. 2011).

Also, the use of interaction terms within the model could be considered. When considering the results of the repeated measures model it may be of interest to consider the effect of treatment independent of time of the follow-up point. But looking for such a statistic in the context of a model with interaction terms doesn't actually make any sense. By adopting a model with interaction terms, you are explicitly claiming that the effect of group differs from one time to the next. A pooled effect can be nothing other than some sort of weighted average of those, and it is more an artefact of the weights you choose for that averaging process than anything else. If you believe that the effect varies over time, then you need an interaction model and it is a contradiction in terms to talk about a pooled effect. However, if you were unsure about the necessity of the interaction terms then it would be possible to use techniques to create a pooled effect, which would give the effect of treatment independent of time.

4.8.3 Conclusions

When the data that are being considered for the analysis are ordinal, it is important to make the best use of the data and fit an appropriate model. Although it is commonly claimed a binary regression model is easier to interpret, most ordinal models are an extension of a binary model and therefore the interpretation is not much harder. It is usually suggested that a binary model is easier to fit, but ordinal methods are found in several statistical packages now and they are therefore readily accessible. It may be that there is not a lot of difference in the results of a binary and an ordinal regression model, but it is important to use an ordinal model in order to get the best effect estimate. Here it was found that an ordinal regression model was the best model fit to the data. The model needs the random effects included to account for the repeated measures, but allowing the covariates in the model to have random effects assigned to them does not improve the fit of the model. There was no evidence that relaxing the proportional odds assumption made the data fit the model any better, which would need to be assessed for every application as it is an important assumption to test.

Having fitted an ordinal regression model with repeated measures, it was found that the repeated measures model was not able to estimate the 3-month treatment effect very well, overestimating the odds, it was considered whether there was any situation in which the two models would agree. A simulation study was conducted with ordinal data generated to investigate whether the repeated measures model was able to estimate the first time point effect similar to the single time point model and also if there was missing data in the follow-up, that the model was able to adjust for this and still produce similar estimates.

5 Simulation

The previous chapter identified that a random effects proportional odds logistic regression model, is the best type of model to apply to the stroke data when conducting a longitudinal analysis of an ordinal variable. This being said, this repeated measures model produces a different treatment effect than if the model was fitted at the single time point. Our data has a high correlation between the repeated measurements and also loss to follow-up is fairly small. If the data had a lower correlation value, would similar results occur? It could be expected that the higher the value of correlation between the repeated measures, the more similar value are throughout follow-up and therefore the more similar the results of the two different model types.

The aim of this chapter is to conduct a simulation study, where ordinal data are randomly generated and then the ordinal logistic regression models applied to the data, with comparisons made between the treatment effect estimated at the first time point using both a single time point and repeated measures model. It was also decided to add different dropout patterns and magnitudes of drop out to the follow-up data generated, in order to see if the results were affected by the type of dropout, as in a longitudinal analysis dropout over time due to loss to follow-up or withdrawal will normally occur.

5.1 Introduction

Having conducted regression models in the previous chapter, this chapter now focuses further on the comparison between fitting a regression model at a single time point and a repeated measures regression model that estimates the results at the same time point as the single time point model. It was decided to focus specifically on the ordinal regression model with the proportional odds assumption, as this was favoured in the previous chapter.

The primary analysis in a longitudinal randomized controlled trial is sometimes a comparison of arms at a single time point. While a two-sample *t*-test is often used, missing data are common in longitudinal studies and decreases power by reducing sample size. Mixed models for repeated measures (MMRM) can test treatment effects at specific time points by using an appropriate contrast matrix. This chapter highlights a simulation study to compare the performance of an ordinal regression model fitted to a single time point with that of a mixed effects one applied to all available data. In this way two odds ratios assessing any potential treatment effect are generated. The impact of within and between-person variance, dropout mechanism, and variance-covariance structure were all considered.

The main aim of the chapter is to look at how the different missing data mechanisms affect the results for the treatment effect at the first time point calculated from the repeated measures model and the single point treatment effect. In order to investigate this, a simulation study was conducted applying different dropout mechanisms to the generated data and the treatment effects were compared. This chapter presents a detailed introduction to simulation studies as well as an account of the design of the study conducted in the thesis. The different dropout patterns were applied to datasets that had been generated with different correlation structures. The correlation structure suggests how correlated the values of the repeated measures are across each of the individuals generated in the simulation study. By varying the correlation structures, the data is generated so that if the correlation value is high, the ordinal data is more likely to be more similar across time points that a lower correlation value.

Dataset 1: 20% correlation structure, between outcome correlations of 0.2

Dataset 2: 40% correlation structure, between outcome correlations of 0.4

Dataset 3: 60% correlation structure, between outcome correlations of 0.6

5.2 Background

A simulation study, also known as Monte Carlo simulation, is a computerised procedure that uses numerical techniques to help evaluate the relative effectiveness of statistical methods. A simulation study involves random sampling either from probability distributions or from a population with known parameters. The goal of a simulation study is to use a statistical method to make inferences about the population from the random sample, by generating a dataset with specific characteristics for the statistical methods to be applied to. Using Monte Carlo simulation techniques, the simulation is completed hundreds of times and the results compared. Simulation studies are very useful when a theoretical explanation is difficult, or cannot be obtained. Simulation studies may be used to compare the performance of two or more models that are analysing the same problem, as is the case with the simulation in this thesis.

A simulation study evaluating different missing data methods takes random samples of the same size from the population, with complete data throughout the sample. An ordinal regression model estimates the treatment effect at the first time point in this model. Following this, missing data mechanisms impose different structures of missing values on these random samples. Once the missing data mechanisms have been applied to the samples, a second ordinal regression model is fitted to the data and the parameters estimated from this model are compared with the previously estimated parameters. A detailed description of the method is given in the following section.

5.3 Methods

Ordinal data for the simulation study were generated using SAS 9.3, and used methods developed for generating correlated discrete ordinal data (Ibrahim & Suliadi 2011). It was decided to use this method as this way it is possible to model different correlation structures. This allows for a comparison to be made between differing correlation values, knowing that the data have been generated using the same methods so variation seen will be due to the differing values of the correlation structure chosen. It is possible to use the trial data to inform the generation of the simulation data, however this would lead to only a single dataset with the same properties of the trial. We have learnt how the two types of model compare in such a highly correlated dataset and therefore this is of less interest to us and the comparison of how the models work with differing correlation values is of most interest to us.

The analysis was conducted using Stata 14, using the same regression models fitted in Chapter 4, both a single time point model and a repeated measures ordinal regression model with the proportional odds assumption. Although this assumption cannot be tested easily it was decided not to use partial proportional odds models with the simulated data, in order to provide a clear and concise message from the results. A detailed method for the simulation procedure is explained below.

5.3.1 Step 1: Generating complete datasets

There were several different complete data sets generated for the simulation study. Using the SAS macro developed by Ibrahim and Sulialdi (Ibrahim & Suliadi 2011) it was possible to generate correlated ordinal response data. The basics of this model include a binary predictor for the group effect, and assessment time *(T)* and an interaction term between the group and time. This produces a proportional odds model written as:

$$Logit\left[P\left(Y_{ij} \leq k\right)|\, x_i, t_j\right] = \beta_{0k} + \beta_x x_i + \beta_t t_j + \beta_{tx} x_i t_j \qquad [5.1]$$

Where x is the group covariate, *t* is the time covariate *and i =1 …, N; j= 1 … T; k=1, …, k-1.* There are only 3 required elements to Ibrahim's macro, the marginal probabilities at each time point, the correlation structure, and the sample size.

5.3.1.1 Marginal probabilities

In order to generate the longitudinal data for the two groups, the macro was applied twice, once with *x=0* and once with *x=1* and then the data was combined to form the whole dataset. The marginal probabilities were calculated for both of the groups in order to be used in each separate run of the macro. The marginal probabilities were calculated using theoretical values of the model parameters displayed in Table 5.1.

In order to generate the ordinal data a value of k must be chosen, which represents the number of categories of the ordinal scale being generated. In our case, we picked K=7 in Table 5.1, as this is the number of categories that is in the mRS scale, this was done in order to make the generated data comparable to the mRS to a degree. However, it would have been possible to pick any of the values of K in order to generate an ordinal scale with between 2-7 categories. In the table the β values are those in Equation 5.1, and are used in order to calculate the log odds used in the data generation.

Table 5.1: Theoretical values of coefficients for calculating log odds used in generation of ordinal data

Distribution	K	β_{01}	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_x	β_t	β_{tx}
Well-balanced	2	-0.25						0.10	0.10	-0.15
	3	-0.71	0.66					0.10	0.10	-0.15
	4	-1.10	0.00	1.10				0.10	0.10	-0.15
	5	-1.39	-0.41	0.41	1.39			0.10	0.10	-0.15
	7	-1.79	-0.92	-0.29	0.29	0.92	1.79	0.10	0.10	-0.15
Skewed	2	1.00						0.80	0.10	-0.25
	3	-2.20	-0.85					0.80	0.10	-0.25
	4	-0.41	0.00	0.41				0.80	0.10	-0.25
	5	-0.85	-0.20	0.20	0.85			0.80	0.10	-0.25
	7	-1.39	-0.66	-0.16	0.16	0.66	1.39	0.80	0.10	-0.25

The marginal probabilities are calculated by substituting the values of Table 5.1 into Equation 5.1, once setting *x=0* and working out the marginal probabilities for that group at each time point in the model and once setting *x=1* and working out the marginal probabilities for each time point in the model. The different values of k indicate the number of ordinal responses that are being generated. In the above table, the data generated can have between 2 and 7 responses. It was chosen that the simulated data generated would have 7 ordinal responses and both well-balanced data and skewed data was considered. Well-balanced data generated data from marginal probabilities that were of a similar size across each outcome score, whereas the skewed data had marginal probabilities that were skewed to favour lower outcome scores.

The following marginal probabilities (Tables 5.2 and 5.3) were calculated in order to generate the complete datasets that were used in the simulation studies. The ordinal response variable generated had 7 categories included in it, the same number as in the mRS. The datasets were simulated to contain 3 time points, just like the SO_2S dataset. For the well-balanced distribution of the data, the marginal probabilities in Table 5.2 were calculated for each of the two groups.

Table 5.2: Marginal probabilities for generation of well-balanced data

	Control			**Treatment**		
Outcome score	**Time point 1**	**Time point 2**	**Time point 3**	**Time point 1**	**Time point 2**	**Time point 3**
1	0.149	0.143	0.137	0.156	0.169	0.184
2	0.146	0.142	0.138	0.150	0.158	0.166
3	0.145	0.143	0.141	0.147	0.150	0.153
4	0.144	0.144	0.144	0.144	0.143	0.141
5	0.141	0.143	0.145	0.139	0.134	0.129
6	0.138	0.142	0.146	0.134	0.126	0.118
7	0.137	0.143	0.149	0.131	0.120	0.110

For the data generated from the skewed distribution of the data, the marginal probabilities in Table 5.3 were calculated for each of the two groups.

Table 5.3: Marginal probabilities for generation of skewed data

	Control			Treat		
Outcome score	**Time point 1**	**Time point 2**	**Time point 3**	**Time point 1**	**Time point 2**	**Time point 3**
1	0.216	0.233	0.252	0.208	0.199	0.192
2	0.148	0.154	0.159	0.145	0.141	0.138
3	0.121	0.123	0.124	0.120	0.119	0.118
4	0.080	0.079	0.078	0.080	0.080	0.080
5	0.117	0.114	0.110	0.118	0.119	0.120
6	0.135	0.128	0.121	0.138	0.141	0.145
7	0.184	0.169	0.156	0.192	0.199	0.208

5.3.1.2 Correlation structure

The correlation structure defines how the outcomes that are generated at each time point are related to each other. For the correlation structure, a simple exchangeable correlation structure was used, *(i =1 … n…; j, k=1, …, T):*

$$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ x & j \neq k \end{cases}$$

The values of *x* differed for each of the different scenarios that were used, so *x* took the value of 0.2, 0.4 and 0.6. These values are smaller than the correlations seen between the mRS outcomes at each time point, which are all in the region of 0.8. It was attempted to include a correlation structure of 0.8 for the simulation study, however the model failed to converge for these models. An exchangeable correlation structure was used as this allows for values to be equally correlated for each individual over time. This is the most appropriate choice for a correlation structure, as we know the observations are not unstructured or independent and they are from the same individual, therefore there must be some correlation observed. An auto-regressive correlation structure could have ne chosen, however when looking at the data we would expect values to be more correlated over time as recovery slows, down rather that less correlated, hence the exchangeable structure being the most appropriate correlation matrix of choice.

5.3.1.3 Sample size and time points

The simulated dataset also contained 3 time points like in the SO_2S dataset, and it was decided to generate 1000 individuals for inclusion in the data. There were 500 generated in the treatment group and 500 generated in the control, giving the whole dataset 1000 individuals. It was thought that this was a sufficiently large number, but it would still be computationally manageable even though the numbers were not of the magnitude of the SO_2S trial itself. There was one dataset for which the data was well balanced and one where the data was drawn from a skewed distribution.

A master data set was generated for each of the complete dataset options; this was used to work out the treatment effect at the first time point, using an ordinal regression model. Following this, the data generating mechanism was replicated 500 times in order to create the final dataset on which the repeated measures analysis was conducted. This figure was conducted based on the analysis conducted looking at multiple imputation methods for incomplete longitudinal ordinal data (Donneau et al. 2015).

5.3.2 Step 2: Generating missing data

From the complete datasets, incomplete datasets were created by using pre-defined deletion criteria. The missing data mechanisms were based on a simulation study looking at the effect of missing data in randomised clinical trials (Joseph 2015). It was decided that the first value in the data set would remain complete for all individuals. This is because in most longitudinal trials the majority of data are caused by individuals who drop out during the follow-up time increases as the time between recruitment and follow-up increases. The type of missing data mechanism applied to the datasets follows a monotone pattern; this means that individuals will have missing data for all subsequent visits after they have one missing observation. Obviously, in practice, there may be a small number of individuals who have non-monotone missing data (where an individual has missed one assessment, but returns for subsequent assessments). For simplicity, it was assumed that a monotone missing pattern is observed for the dropout rates and mechanisms applied in the simulation study. This is reflected in the SO_2S trial data, where at the first time point, which is 3 months post-stroke, the value of the mRS is known for 98% of the individuals. This reduces to 87% by 12 months. Both the dropout rate for the time points and the choice of dropout mechanism used in the simulation study were predefined.

In a randomised trial there are usually 2 causes for missing data, one is loss to follow-up and the second is death. In the mRS scale with the category for death included missing data in the follow-up of a study can only occur due to loss to follow-up. Consideration was made as to whether an absorbing score, such as that of death in the mRS could be included in the setup of the simulation study. Because of the missingness mechanisms used (with a monotone pattern), including a score where data is never missing in the simulation is complicated. This is because the missingness mechanisms are added by generating a random variable, the value would be required to be re-added, if the value representing death was omitted. This would only be possible if the value was known before, therefore could only be useful for those who died at the first time point as it would be possible to carry forward the observation. But by doing this there would be an effect on the amount of missing data, as although the different rates defined below would have been applied, values would have been re-added and therefore the rates would not be the same across the different datasets used in the analysis. It was decided that as the data generated in not specifically from the SO_2S trial and is randomly generated data that an absorbing category would not be considered and that dropout may occur form any point on the scale. Although this is different to the outcome analysed in the previous chapter, there is still merit in the work as it is looking at what happens in different correlation structures and in different data (well-balanced or skewed), and understanding whether missing data, and the type of missing data, affect the conclusions drawn.

The two main types of mechanisms used are described below, missing completely at random and missing at random. The first, as the name suggests, has data missing completely at random, whereas the second relies on some noticeable pattern to inform drop-out. As there are no baseline characteristics, apart from treatment group included in the model, it was decided that to inform the dropout the missing at random pattern would restrict dropout to those individuals who had a generated values larger than the median value of the whole dataset for that time point. This method was chosen as in the mRS scale, those with higher scores are more disabled and therefore may be more likely to be lost to follow-up as they are unable to complete follow-up questionnaires themselves. This method satisfies the criteria for missing at random data when there are no observed variables in order to inform the pattern of missingness.

5.3.2.1 Choice of dropout rate

After deciding that the first time point would have complete data, it then needed to be decided how many individuals would drop out at the remaining points in the data set. It was decided that two of the dropout patterns would include equal dropout rates for both of the two groups in the generated dataset. The other two dropout patterns would have unequal dropout rates between the two groups in the dataset. It was further decided that the control group would have a larger dropout than the treatment group, implying a favourable treatment. Using the SO_2S trial as a guide, it was decided that at the second time point 10% of individuals would have dropped out. This increased to 20% dropout at the third time point. For the dropout patterns with unequal dropout, the dropout in the treatment group was halved so that there was a 5% dropout at the second time point, and a 10% dropout at the third time point.

5.3.2.2 Missing completely at random (MCAR)

Using a missing completely at random (MCAR) dropout mechanism, means we assume that there is no relationship between the missing values and any unobserved or observed values in the dataset. Let us denote complete data as Y_{com}, which is made up of $Y\mathit{obs}$, Y_{mis} as the observed and missing parts such that Y_{com}= (Y_{obs}, Y_{mis}). Rubin defined missing data to be MCAR if the missingness does not depend on the observed values (Rubin 1976):

$$\mathrm{P}(\mathrm{R}|\mathrm{Y_{com}}) = \mathrm{P}(\mathrm{R}) \qquad [5.2]$$

This means that the individuals with missing data are just a random subset of the data. When conducting analysis on individuals who have data MCAR, the analyses will be completely unbiased.

5.3.2.3 Missing at random (MAR)

Using a missing at random (MAR) dropout mechanism we assume that there is a relationship between the missing values and one or more observed variables in the dataset. Rubin defined missing data to be MAR if the distribution of missingness does not depend on Y_{mis} (Rubin 1976):

$$\mathrm{P}(\mathrm{R}|\mathrm{Y_{com}}) = \mathrm{P}(\mathrm{R}|\mathrm{Y_{obs}}). \qquad [5.3]$$

This means that the missingness can be fully accounted for by variables with complete information. However, there is still an assumption that the missing values are not dependent on any unobserved variable(s). An assumption of MAR is hard to verify, so reasonable justification may need to be used. MAR data are also called ignorable non-response data, (Schafer & Graham 2002).

5.3.2.4 Missing not at random (MNAR)

It was decided not to include a missing not a random (MNAR) dropout mechanism, which assumes that the missing values are dependent on the variables that are missing. Data are said to be MNAR if the data violates Equation 5.3 and this type of data is known as non-ignorable data, it is hard to identify the difference between ignorable and non-ignorable data in observed data (Schafer & Graham 2002).

5.3.3 Step 3: Analysis

The first step of the analysis was conducted once the first complete dataset was generated. From this dataset, an ordinal logistic regression model was applied to the data using only the first time point of the generated data, and the difference between the two groups was found. This difference was calculated in logit form.

After the replication of the data and the application of the dropout mechanisms, a repeated measures ordinal logistic regression was fitted to the data. The model was fitted, using each of the replications in the dataset, which provided 500 treatment effect estimates at the first time point. This repeated measures model takes into account the value of the ordinal outcome at the other time points in the model and estimates the effect at the first time point from this, taking account of the missing data that we have introduced.

The treatment effect at the first time point that is reported from the repeated measures model after the dropout mechanism has been applied is an average of each of the 500 treatment effects that were calculated with each replication of the data.

5.3.4 Measures of performance

In order to consider the model performance, different assessment criteria tools were used. The main performance assessment used was the difference calculated between the true effect and the treatment effects from the repeated measures model. The standard errors of the model fitted were also recorded. Finally, the percentage change in between the two calculated effects was considered.

5.3.4.1 Difference in estimates

The difference in the estimates is defined as the difference between the average value of the treatment effect and the corresponding parameter – the true treatment effect. The value of the treatment effect is calculated for each of the repetitions and then the overall estimated treatment effect is calculated by taking the average effect size of the 500 repetitions ran. It is the difference between this and the true treatment effect of the model that is calculated:

$$\begin{aligned} Difference &= true\ treatment\ effect \\ &\quad - treatment\ effect\ from\ repeated\ measures\ model \end{aligned}$$

If the value of the difference is negative then the treatment effect has been overestimated, and if the value of the difference is positive there is an underestimation of the treatment effect. Although the above is true, we are specifically interested in the how much the true effect and the repeated measures effect differ, not the direction of the difference and therefore absolute values will be presented.

5.3.4.2 Percentage change

Finally, the percentage change was considered, which looks at the difference as a percentage of the true effect of the model.

$$Percentage\ change = 100 \times \left(\frac{difference\ in\ estimates}{true\ treatment\ effect}\right)$$

5.4 Results

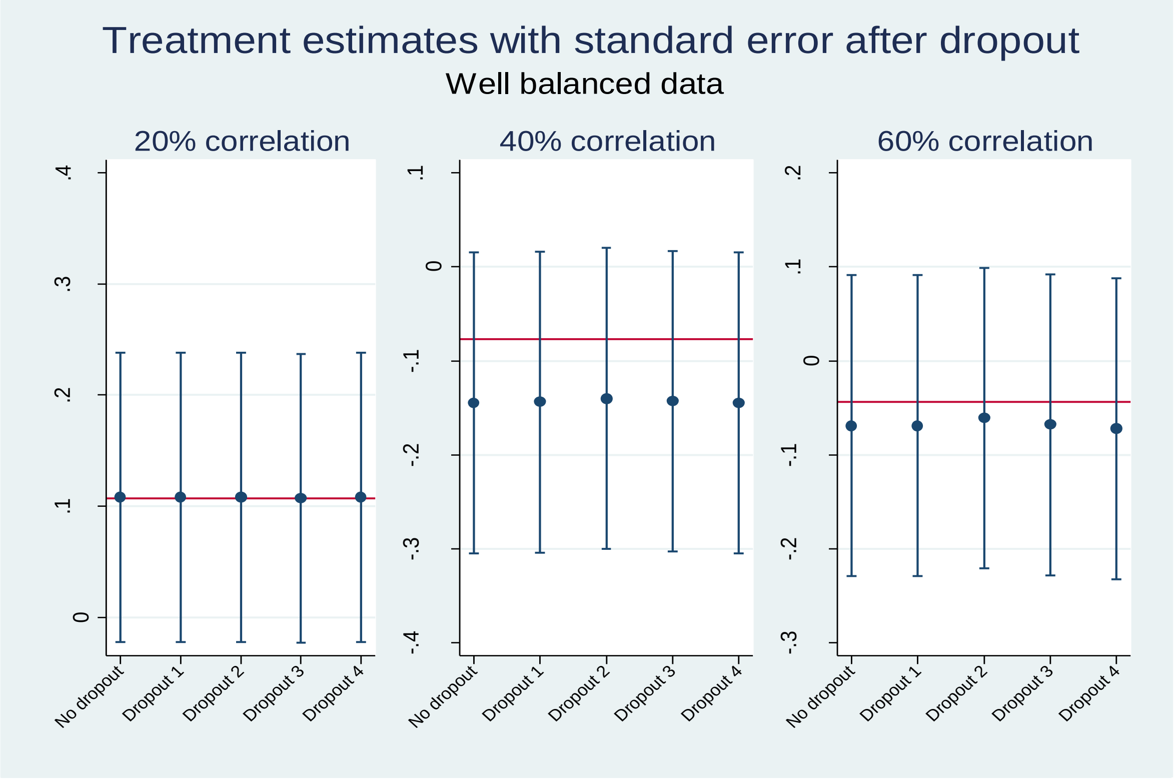
5.4.1 Well-balanced data

The first dataset that was generated was taken from the well-balanced distribution – this is most similar to the mRS distribution within the SO_2S, especially at the first time point, where there is a fairly even spread of individuals across all scores. Each dataset contained 1000 individuals, (500 treatment, and 500 control) and correlation values of 0.2, 0.4 and 0.6 were used to generate the different datasets.

Figure 5.1 shows the results of the simulation for the well-balanced data with 3 time points. The graph shows the true value of the treatment effect at the first time point as a constant line for comparison and estimated value of the treatment effect at the first time point using a repeated measures model before and after the four different dropout mechanisms are applied. A table of results given in Appendix D.

It can be seen that as with the results found in the previous chapter, the use of the repeated measures model causes an overestimation of the treatment effect of the model in some cases. When there is a smaller correlation value, there appears to be almost no difference in the true effect and the effect estimated by the repeated measures model. With increasing correlation size there appears to be more variation seen between the single time point and the repeated measures model. There does not appear to be a pattern with increasing correlation as the 40% correlation looks like it over estimates the treatment effect more than the 60% correlation estimates.

Figure 5.1: Graph of first time point treatment estimates (and standard errors) before and after dropout, for varying correlation clause in well-balanced data with 1000 patients



It is encouraging to see that there appears to be little difference in the effect of the dropout patterns applied, with the most variation across drop out patterns seen in the 60% correlation dataset. The difference in the effect sizes has been calculated and is given in the table below.

Table 5.4: Difference between the average treatment effect at first time point from repeated measures model before and after dropout and the true effect from a single time point model for well-balanced data

		Effect Difference				
Correlation value	**True effect**	**No dropout**	**Dropout 1**	**Dropout 2**	**Dropout 3**	**Dropout 4**
20%	0.107	0.002	0.004	0.004	0.003	0.003
40%	-0.077	0.068	0.067	0.064	0.067	0.069
60%	-0.044	0.025	0.025	0.017	0.024	0.028

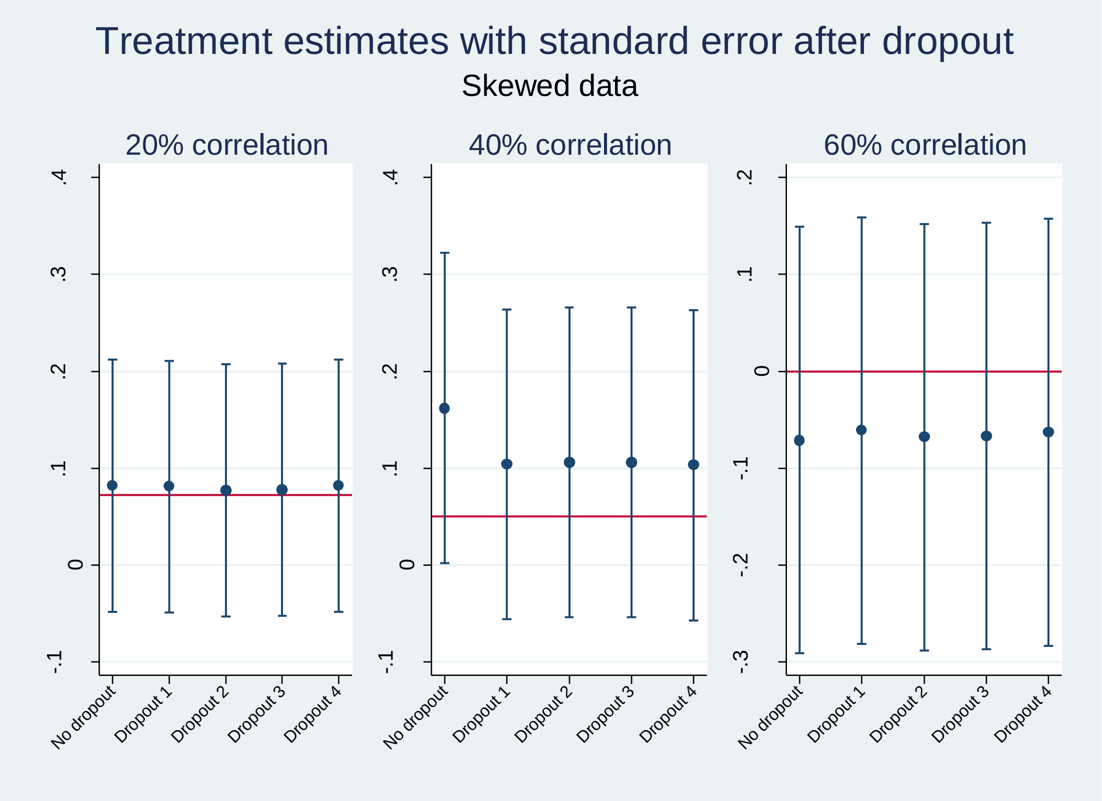
Dropout 1: MCAR – equal dropout; dropout 2: MAR – equal dropout; dropout 3: MAR – unequal dropout; dropout 4: MCAR – unequal dropout

Here we see how small the differences in the effect sizes are, especially for the 20% correlation dataset. The true effect in this data set suggested a positive treatment effect, whereas the larger datasets suggested a negative treatment effect. For both of the larger 2 datasets drop out 2 appear to produce the smallest difference in treatment effect estimates, where the data is MAR with equal dropout rates in both of the groups. We also considered what would happen if the data had been drawn from a skewed distribution, rather than a well-balanced one.

5.4.2 Skewed data

Figure 5.2 shows the results of the simulation for the skewed data with 3 time points for each of the datasets that were generated. As with the well-balanced data, the red line shows the true value of the treatment effect at the first time point along with estimated value of the treatment effect (with standard error) at the first time point using a repeated measures model before and the four different dropout mechanisms are applied. A table of results given in Appendix D.

Figure 5.2: Graph of first time point treatment estimates (and standard errors) before and after dropout, for varying correlation values in skewed data with 1000 patients



As with the well-balanced data, a correlation value of 20% gives the least amount of difference between the true and estimated treatment effects. All 3 of the datasets give an overestimation of the true effect, with a pattern of increasing size with the higher correlation value applied. For the datasets with 20% and 60% correlation there does not appear to be much variation in the different types of dropout applied, or the effect of dropout from the estimates with no dropout applied. However, for the 40% dataset there appears to be a systematic decrease in the estimated treatment effect with the application of the drop out mechanisms. These differences can be quantified and are given in Table 5.5

Table 5.5: Difference between the average treatment effect at first time point from repeated measures model before and after dropout and the true effect from a single time point model for skewed data

		Effect Difference				
Correlation value	**True effect**	**No dropout**	**Dropout 1**	**Dropout 2**	**Dropout 3**	**Dropout 4**
20%	0.072	0.010	0.009	0.005	0.006	0.010
40%	0.05	0.112	0.054	0.057	0.056	0.053
60%	-0.001	0.070	0.060	0.067	0.065	0.062

Dropout 1: MCAR – equal dropout; dropout 2: MAR – equal dropout; dropout 3: MAR – unequal dropout; dropout 4: MCAR – unequal dropout

The table shows that the differences between the estimates are of a similar magnitude for the higher correlations, with small differences seen in the 20% correlated dataset, as shown in Figure 5.2 as well. For the 20% correlated dataset, data MAR with either equal or unequal drop out has a smaller difference in estimates than MCAR data, whereas this is reversed for the larger datasets with data MCAR providing more similar estimates with the repeated measures model.

5.5 Discussion

The aim of this chapter was to use a simulation study in order to identify if there was a scenario where the repeated measures model was able to estimate the value of the true effect at the first time point well. There were four different dropout patterns that were applied to a variety of datasets. The dataset included randomly generating ordinal data with correlations structure of 20%, 40% and 60% between the repeated measures. The randomly generated data were generated using both a well-balanced distribution and a skewed distribution, giving a total of 6 different datasets of varying correlation structures, distributed data and sizes.

After the data had been generated, the four dropout patterns were applied to the data in turn and a repeated measures model generated the treatment effect each time this occurred. It was done 500 times for each scenario. The estimates fitted were summarised and compared in order to look for patterns of how well the data can estimate the true effect estimated at the single time point, without repeated measures. It was found that there was no real pattern to which scenarios were best at predicting the true effect. The dropout mechanism varied between scenarios as to which was best. When there was 20% correlation, the first time point for the repeated measures model had the closest treatment effects to the true effect, which is surprising as you would expect that a larger correlation structure would give more similar results over time and therefore be better at predicting the first time point, as the other values would be similar. A more detailed comparison of the studies is given below.

5.5.1 Comparison of datasets

Having used three different correlation values before applying the various dropout mechanisms, comparisons are able to be drawn with how well the repeated measures models are able to predict the true value of the data before and after dropout mechanisms are able to be applied. We can conclude several things when comparing the well-balanced and skewed datasets. Firstly, there is a large amount of agreement between the true effect of the simulated data and the effects estimates using the repeated measures data, both before and after the dropout mechanisms have been applied. This is most surprising because in the scenario, there is likely to be more variation in the simulated data because there is the lowest correlation structure of the 3 applied to it. There will be more change over time being captured and estimated within the model, as the repeated measures model takes into account all the follow-up points as well as the initial time point. When the data is well-balanced, there does not appear to be a pattern in the different drop out mechanisms applied, as there is some disagreement between the datasets, when the data is skewed there do appear to be differences in effect of the dropout mechanisms.

Secondly, we can look at the percentage change between the repeated measures effects and true effect of the model. These tables are given in Appendix E. The first table, Table E1 looks at the percentage change of the true effect and the estimated effect after dropout. The percentage changes are calculated for the two different types of dropout patterns as well. The average percentage change for the MAR dropout and the MCAR dropout patterns are detailed, to see whether the type of dropout applied has an effect on the size of the bias. The second table looks at the average percentage change for the rates of the dropout to see whether dropout mechanisms with equal dropout rates have a different percentage change from those where the dropout pattern applied with unequal dropout rates.

When the correlation is set at 20%, the smallest percentage change values are produced for all types and sizes of data, approximately 3% difference compared to the size of the true effect. Looking at the two different dropout types when the correlation is 40% or 60%, the MCAR dropout produces more scenarios with smaller percentage change values, with equal percentage change in the 20% correlated data set.

It becomes very obvious from looking at the table, Table E1 that when the treatment effect calculated is very small, the percentage change is very large. This occurs with several of the calculated values when the correlation is at 40% or 60%. Some of the values are as high as a 6600% change when comparing the size of the bias to the size of the treatment effect. However, the treatment effect is so small (- 0.001) that we would expect to see a large value of percentage change even though the bias is no larger than the other bias values calculated in the other datasets.

If we consider the second table, Table E2 where the percentage dropout is compared for equal and unequal dropout rates, it can be found that when the data has 20% correlation, unequal dropout rates produce smaller percentage change than the equal rates; however, this is reversed when the correlation is set at 40% and 60%. Once again there are several percentage change values that are very large for the treatment effects that are very small. These tables, Tables E1 and E2 are used for more of a descriptive value, as a percentage change of 6350% is hard to quantify, although there are justifications for why values of this size are seen.

5.5.2 Further work

Obviously, there are still a variety of different scenarios that could be applied to the data. The first approach would be to look at the size of dropout that was used. Here, when there were equal dropout rates, total dropout of 20% was used. This was a large dropout, but without causing a majority of missing data. When there were unequal dropout rates the rate in the treatment group was halved to 10% missing data. Other dropout rates may provide different results.

Also, it is important to note that the ordinal data generated had seven categories in order to make them comparable to the mRS, our outcome of interest for the analyses throughout the thesis. The mRS has the category of death, therefore an individual who has died will always have a mRS value and will be included throughout the follow-up. The dropout mechanisms applied to the data were unable to capture this and dropped out an individual even if they had a score of 7 (the highest score on the randomly generated data and therefore similar to death on the mRS). This was due to the monotone nature of the dropout mechanism. An extension would be to try to model a missing data mechanism that reflected the mRS more similarly, which produced the missing data, but kept the highest score on the scale complete to see how this affects the treatment estimates from the repeated measures analysis.

5.5.3 Conclusions

The simulation study that has been conducted in this chapter has highlighted that there may not be a pattern to be identified as to when the repeated measures model fits the outcome at the first time point and when there is variation seen in the estimates. The data generated here had 3 different correlation structures applied to it, however there does not seem to be a clear pattern arising as to whether the type of missing data the follow-up has or whether the equal or unequal dropout rates are about to be modelled better.

The following chapter moves away from the ordinal regression models that have been fitted in the previous chapters and moves on to a different approach, which will make difference assumptions about the data and apply multi-state Markov models to model how individuals move over time.

6 Multi-state modelling

Having considered how well the repeated measures regression models fit the data using both the SO_2S trial data and a simulation study, the analysis now moves on to alternative methods. The first of these methods considered for the analysis of longitudinal ordinal data is a multi-state Markov model. This method allows for the ordinality of the scale to be preserved while treating each point on the scale as an individual category. Patients transition between states, which represent each of the categories in the ordinal score.

The first model fitted requires individuals to have at least 2 complete follow-up time points, so individuals with missing data are removed from the analysis. The second model fitted uses the idea of censored states in order to incorporate those individuals who have missing data, allowing them to be in the model without a state assigned at each follow-up point.

6.1 Introduction

Multi-state models based on Markov processes are a well-established method of estimating rates of transition between stages of a disease, and this is one of the proposed methods for analysing the longitudinal mRS scores. There are some considerations that must be made when fitting the multi-state Markov models, due to the nature of the data that is being modelled.

Many previous publications that have used multi-state Markov models to analyse clinical data have considered fewer states in the model than the number of mRS disability states (7 in total including death) (O'Keeffe et al. 2011, Titman 2011, Saint-Pierre et al. 2003). Markov models have been used to relate progression in multiple sclerosis to baseline covariates (Palace et al. 2014). However, such models assume that further progression essentially depends only on the previous measurement and may be less able to cope with issues such as missing data and the need for imputation. Arguments for having an adjusted analysis in stroke trials would introduce multiple parameter estimates, and the aim is to find a parsimonious model, therefore, only the treatment covariate was included in the model, this is in line with the regression models that have previously been fitted.

Previously, there has been research being conducted by Cassarly in the use of multi-state models for acute stroke therapy clinical trials (Cassarly 2015). Other published articles have been found that conduct multi-state models within stroke research but their focuses are different to the project. The first used a 3 state model to look at the transitions between healthy, history of stroke and death (Kapetanakis et al. 2013). The second multi-state model found fitted a model the BI as 3 states; a poor functional state (BI between 0 and 40), a moderate functional state (BI between 45 and 80) and good functional state (BI greater than 80) (Pan et al. 2008b).

The most common use of a multi-state model is in event history analysis, looking at when the event of interest has occurred, assuming that the time at which the event occurred is known. Multi-state models have previously been applied in several areas of medicine, including problems arising from lung and heart transplants, HIV infection and AIDS, breast cancer and cervical cancer screening, diabetic complications, hepatic cancer and liver cirrhosis (Cassarly 2015). There is also very little literature on how well the models work when the data being fitted are sparsely populated. The data that is being fitted here is sparsely populated in parts as there are a large number of health states and the movement between them can be considered restrictive. Most subjects transition into an adjacent state, whereas there are a few subjects that transition into a non-adjacent state. As well as this, the mRS is not recorded at baseline, only at time points during the follow-up.

6.2 Methods

A Markov multi-state model can be used in order to describe how an individual transitions through a series of states during a continuous period of time. It is an alternative method that can be used to model an ordinal outcome with repeated measures. This is due to the nature of the ordinal scale, as each individual category on the scale can be treated as a state in its own right, reflecting the natural process of a disease transitioning through the different stages of severity in the scale. These models allow for all the information about disease severity to be included, with no need for a dichotomisation of the scale, which would lead to a loss of information and reduced statistical power. This type of model also allows those who have died to be included in the model as death can be included as a separate state, although once in that state it is impossible to move out of it, whereas individuals are otherwise free to move between any other states as appropriate.

Usually, a multi-state model is fitted to data that represent a continuously observed process, so the state of the individual is known at all times during the follow-up period. The simplest example of this is a survival model where the patient is either alive or dead, where it is assumed the patient remains alive until the date of death is observed or the patient is censored.

6.2.1 Illness death model

A commonly used multi-state model is the illness-death model, where there are three states in the model representing health, illness and death. In this basic form, an individual is able to make a transition from health to illness, illness to death and also from health to death, as seen in Figure 6.1. The model can also be extended to allow individuals to recover from an illness where it is possible for the individual to transition from illness to health, shown in Figure 6.2. The state of death is known as an absorbing state and so although it is possible for an individual to transition into the state it is not possible for an individual to transition back out of this state.

Figure 6.1: Illness-death multi-state model

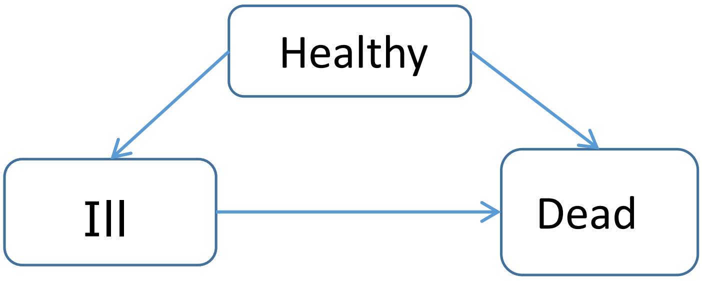
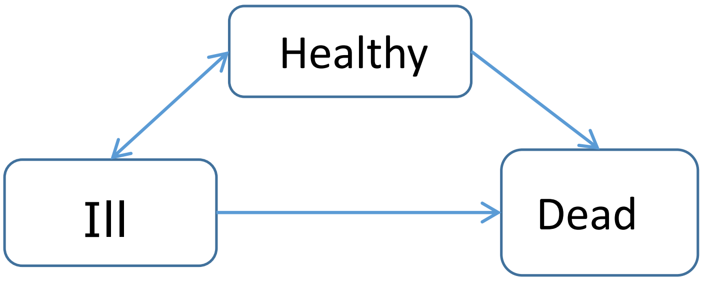


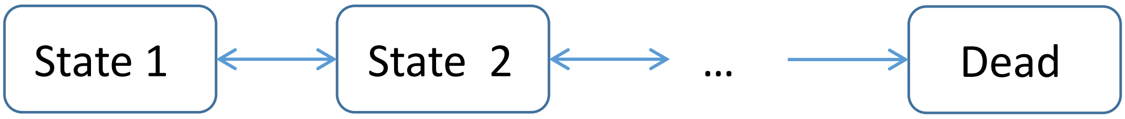
Figure 6.2: Illness-death multi-state model with recovery



6.2.2 Disease progression model

This basic illness-death model with recovery can further be extended to include a series of successively more severe disease stages with an absorbing death state. The individual may move though the disease states by advancing or recovering into the neighbouring state from which they are currently in, shown in Figure 6.3. The final state of the model is the most severe in terms of disease progression and is usually the absorbing state of the model, commonly death. The stages of the disease are able to be modelled as a both homogeneous continuous-time and discrete time Markov process, described in 6.2.4.

Figure 6.3: Disease progression multi-state model



6.2.3 Multi-state model for panel data

In longitudinal data where there are repeated measures for the outcome of interest, although the current state of the individual is recorded at set times during the follow-up, the actual time at which the transition into the state is usually unknown. This is because the individuals are only assessed at specific time points – usually prearranged, such as follow-up visits at which monitoring information is collected. Although a fixed schedule is usually specified, this is usually hard to comply with due to the health of the patient. Date of death is usually known as an exact date within the arranged time points as this is recorded separately and so the transition to death is usually known in more detail than the rest of the outcomes being measured. It is also possible to have censored observations within a multi-state model; this is where it is known that the patient is alive (as they have not been recorded as dead) but the actual current state of the individual is unknown. These censored observations could be individuals that are lost to follow-up during the follow-up period.

Data that take the form of a continuous process but are recorded at discrete time points are known as panel data. Panel data can be modelled using both continuous and discrete multi-state models, although in order for the discrete model to be used, the time in between each time the outcome is recorded during the follow-up period must be equally spaced.

6.2.4 Markov property

In order to fit a multi-state model to panel data, it must be assumed that the Markov property is valid for the data. The Markov property refers to the memoryless property of a stochastic process and requires future states of the model to depend only on the current state and not those that have come before it. This means that the transition intensity (see definition below) that is being estimated is independent of the observation history of all other transitions made up to the present transition.

6.2.5 Transition intensities

The event of interest in a multi-state model is the transition of an individual between two different states. The state that an individual moves into is governed by a set of transition intensities that are calculated for each pair of states within the model. The intensities may also depend on the time at which the event occurred, or even more generally a set of individual-specific or time-dependent explanatory variables, denoted here as *z(t)*. The transition intensity represents the instantaneous risk of an individual moving from one state to the other.

Suppose an individual moves from state *r* to state *s*, and *s≠r*, then the transition intensity *q_{rs}(t, z(t))* is:

$$\mathrm{q}_{rs}(\mathrm{t}, \mathrm{z(t)}) = \lim_{\delta t \to 0} \frac{P(S(t+\delta t) = s | S(t) = r)}{\delta t}$$

The q_{rs} values form an R X R matrix, Q, whose rows sum to 0, such that the diagonal entries are defined by:

$$-\sum_{s \neq r} q_{rs}$$

This matrix is known as the transition intensity matrix, and it is constant throughout the follow-up period in a time-homogeneous model. In time-homogeneous models each q_{rs} is independent of time. The amount of time that an individual is in state *r* is exponentially distributed with a mean of -1/q_{rr}. The probability of an individual moving into state *s* from state *r* is $–q_{rs}/q_{rr}$.

The ‘msm’ package in the software package R can be used to fit multi-state Markov models to the data. This package has been designed to allow a general multi-state model to be fitted to the data and uses methods that allow models for continuous time Markov models for panel data.

6.3 Application to SO₂S trial data

The model for disease progression, in this case, is described by the following transition intensity matrix. There are eleven non-zero, non-diagonal entries which represent the instantaneous risk of moving from state *r* to state *s*. Each non-zero element indicates a transition that is able to occur within the model. The seventh row of the matrix has entries that are all set to 0 as this is the absorbing state in the model and it is impossible to transition out of this state. It was also decided to assume that in order to move between states more than 1 state apart a patient would have moved through the intermediate states. Because of this, patients can only transition into death from the previous state and not every state in the model. A model with a transition straight into death from any other state may be considered as an extension in further work.

$$Q = \begin{pmatrix} q11 & q12 & 0 & 0 & 0 & 0 & 0 \\ q21 & q22 & q23 & 0 & 0 & 0 & 0 \\ 0 & q32 & q33 & q34 & 0 & 0 & 0 \\ 0 & 0 & q43 & q44 & q45 & 0 & 0 \\ 0 & 0 & 0 & q54 & q55 & q56 & 0 \\ 0 & 0 & 0 & 0 & q65 & q66 & q67 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The inclusion of control variables in multi-state models proves tricky, as each time a variable (continuous or dichotomised) is introduced into the model, the number of parameters to be estimated increases by 11, as there are 11 possible transitions that are being modelled. This can lead to complex models when a large number of control variables are included, which can be computationally intensive and the model may fail to converge on a maximum likelihood estimation. Therefore, it is important that the specific control variables to be included are carefully pre-defined using clinical knowledge. The main and only predictor variable included in the current model is the treatment that each individual received: either standard care or oxygen treatment. By including treatment, the model created is interpretable and can be compared to other models that have been fitted to the SO_2S data. The models converged to the maximum likelihood and the Hessian matrices were estimated using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. A Hessian matrix is a square matrix of second-order partial derivatives, which can be approximated using an iterative methods such as the BFGS algorithm, which has been shown to perform well (Liu & Nocedal 1989).

In order to determine whether the Markov property is valid for the model, the observed and expected numbers of individuals in each state are plotted over the follow-up time. An assumption of the model could be considered to be violated if there is a significant deviation of the expected numbers from those that were actually observed. This method is also used to assess the goodness of fit of the model. As there is a covariate included in the model, a likelihood ratio test will be conducted in order to compare nested models. This will allow us to determine whether the included covariate is statistically significant within the model.

6.4 Results

6.4.1 Disability progression model with no censoring

The first model that was applied to the data considered all those patients who had 2 or more recorded mRS values over the 12-month follow-up time, whether the values recorded were at consecutive time points or with a break in between the recorded values.

A total of 8,003 individuals were recruited to the SO_2S trial and followed up over 12-months. In total there were 278 (3.5%) individuals who were completely lost to follow-up, having no mRS value recorded at any of the time points. These individuals were excluded from the analyses along with individuals who had two missing mRS follow-up values. There were 27 individuals missing a mRS score at both the 3- and 6-month follow-up points, 189 missing a mRS score at both the 6 and 12-month follow-up points and 5 individuals missing a mRS score at both the 3- and 12-month follow-up points. These individuals were all excluded from the analyses as they only had one recorded mRS value and so did not make any transition. Finally, individuals who had died at 3 months were excluded from the analysis as they would make no transition between states as they were in the absorbing state at the start. This excluded 633 individuals from the analysis. This left the dataset with 6,871 individuals with 2 or more follow-up points where the mRS value was recorded for inclusion in the model.

The transitions that occur from one visit in the follow-up to the next are described below. The first table, Table 6.1, shows the transitions that occur between the first time point at 3 months and the second time point at 6 months.

Table 6.1: Observed transitions in SO_2S data between 3 months and 6 months

	6 month mRS								
3 month mRS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Missing**	**Total**
0	661	160	33	30	19	4	3	10	920
1	267	1,307	208	143	61	2	13	11	2,012
2	41	234	476	115	59	5	3	6	939
3	32	145	199	722	147	7	8	13	1,273
4	15	49	84	201	695	57	28	21	1,150
5	2	4	2	8	101	285	33	18	453
Missing	2	1	0	1	4	1	0	0	9
Total	1,020	1,900	1,002	1,220	1,086	361	88	79	6,756

The above table shows how many of the individuals remained in the same mRS category between two time points. Approximately 61% (4,146/6,756) of all the observed transitions remain in the same state. It can then be seen that there are 79 individuals who have a missing value for 6 months, who had a value recorded at 3 months (and 9 who had a missing value at 3 months but had a value at 6 months). There are many transitions within the data where an individual moves to the state next to the one that they are in at the first time point, approximately 25% of the observed transitions. For example, 267 individuals have a score of 1 on the mRS at 3 months and have transitioned to a mRS score of 0 by 12 months. These transitions are known as adjacent-state transitions. It can be seen that although non-adjacent-state transitions occur, they are much less frequent than adjacent-state transitions. The second table, Table 6.2, shows the transitions that occur between the second time point of 6 months and the final time point at 12 months.

Table 6.2: Observed transitions in SO_2S data between 6 months and 12 months

	12 month mRS								
6 month mRS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Missing**	**Total**
0	647	211	33	35	29	4	6	55	1,020
1	222	1,176	169	157	60	4	18	94	1,900
2	42	216	407	160	91	8	10	68	1,002
3	27	134	175	570	185	17	18	94	1,220
4	21	38	71	131	609	77	38	101	1,086
5	3	3	1	4	60	198	39	53	361
6	0	0	0	0	0	0	88	0	88
Missing	2	9	0	4	5	3	56	0	79
Total	964	1,787	856	1,061	1,039	311	273	465	6,756

Once again, it can be seen that the majority of individuals do not transition to another state but remain in the same state between the 6-month and 12-month follow-up. Half of those that do transition mainly move into an adjacent category in the mRS scale (24% of observed transitions). There are transitions that occur where an individual moves more than one state away from the original state they are in, with fewer of these types of transitions occurring as the size of the transition increases, as there are decreasing numbers when the number of states that are being transitioned gets larger. This leads us to have what is considered sparse data in the more extreme transitions (further from the diagonal) that are being attempted to be modelled. Also, the 79 individuals that had missing data at 6 months all have a mRS value recorded at 12 months, with two scoring a 0, nine scoring a 1, four scoring a 3, five scoring a 4 and three scoring a 5 on the mRS scale. Of the individuals that have a missing mRS value at 6 months, 56 have died at the 12-month follow-up point. Once again, non-adjacent transitions are sparse amongst the data and the number of patients moving between each of the states is similar to those that move between 3 and 6 months.

Due to the Markov property, the multi-state model only looks at the immediate transition and not the ones before it; therefore it models each transition on its own regardless of the time. This means that a summary table can be created (Table 6.3) of all the transitions that are made by every patient in the trial, regardless of the time period in which the transitions occurred. The summary table excludes those individuals who have a missing value at the start or the end of the transition period. In the table, an individual with 3 follow-up mRS values recorded will feature in the table twice, as each transition is counted separately, although the model accounts for each individual and the number of transitions they have.

Table 6.3: All observed transitions in SO_2S data between 2 consecutive time points

	Time point 2						
Time point 1	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	1,308	371	66	65	48	8	9
1	489	2,483	377	300	121	6	31
2	83	450	883	275	150	13	13
3	59	279	374	1,292	332	24	26
4	36	87	155	332	1,304	134	66
5	5	7	3	12	161	483	72
6	0	0	0	0	0	0	88

6.4.1.1 Model results

The transition intensities that were calculated for each possible transition defined by the model regardless of the treatment group that the individual is in are given in Table 6.4. The transition intensities represent the instantaneous risk of a patient moving from one state to an adjacent one, if the patient is to make a transition. It is possible for individuals to remain in the same state however this is not fitted in the model, therefore the diagonal entries of Table 6.4 and all other tables of transition intensities are calculated from the non-diagonal elements. The intensities on the diagonal are the negatives of the sum of the non-diagonal elements, as detailed in 6.2.5. The model is defined such that the intensities of each row must sum to 0. It can be seen that there is quite a large difference in the transition intensities depending on where the patient started on the mRS scale. A larger transition intensity can be seen as a greater risk of the individual moving to the adjacent state. Those who have an extreme value, i.e. severely disabled or no disability at all, have a much smaller risk of making any transition than those patients who fall within the middle of the scale.

The transition intensities can be separated into the two different treatment groups in the model, the standard care group and also the oxygen treatment group. The transition intensities and 95% confidence intervals for these are given in Table 6.5. Confidence intervals are calculated from the covariance matrix of the estimates by assuming the distribution is symmetric on the log scale. Overall the transition intensities are very similar in the two groups to those calculated for the whole model before the treatment group was considered. Individuals in the standard care group have a fractionally smaller risk of moving to a lower state in the model, suggesting that they may be less likely to recover. There are also wider confidence intervals for the transition intensities in the standard care group, suggesting more uncertainty in the model. From these transition intensities, it is possible to calculate the hazard ratio comparing the oxygen treatment group to the standard care group, to allow interpretation of the treatment effect. The hazard ratios are calculated by taking the exponential of the transition intensities calculated for each pair of adjacent states.

Table 6.4: Transition intensities and confidence intervals of multi-state model

	Transition intensities (95% confidence interval)						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	-0.10 (-0.11, -0.09)	0.10 (0.09, 0.11)					
1	0.06 (0.05, 0.06)	-0.17 (-0.18, -0.16)	0.11 (0.10, 0.12)				
2		0.21 (0.20, 0.23)	-0.54 (-0.59, 0.50)	0.33 (0.29, 0.37)			
3			0.26 (0.23, 0.30)	-0.38 (-0.41,-0.35)	0.12 (0.11, 0.13)		
4				0.11 (0.10, 0.12)	-0.16 (-0.17,-0.15)	0.05 (0.04, 0.05)	
5					0.08 (0.07, 0.09)	-0.15 (-0.16, -0.14)	0.07 (0.06, 0.08)

Table 6.5: Transition intensities and confidence intervals of multi-state model for oxygen treatment and standard care group separately

	Oxygen treatment transition intensities (95% confidence intervals)						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	-0.10 (-0.11, -0.09)	0.10 (0.09, 0.11)					
1	0.06 (0.05, 0.07)	-0.18 (-0.19, -0.16)	0.12 (0.11, 0.13)				
2		0.22 (0.20, 0.24)	-0.54 (-0.60, -0.49)	0.33 (0.28, 0.38)			
3			0.27 (0.23, 0.32)	-0.39 (-0.43, -0.34)	0.12 (0.10, 0.13)		
4				0.11 (0.10, 0.12)	-0.16 (-0.17, -0.15)	0.05 (0.04, 0.05)	
5					0.08 (0.06, 0.09)	-0.15 (-0.17, -0.13)	0.07 (0.06, 0.08)
	Standard care transition intensities (95% confidence intervals)						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	-0.10 (-0.11, -0.09)	0.10 (0.09, 0.11)					
1	0.05 (0.05, 0.06)	-0.15 (-0.17, -0.14)	0.10(0.09, 0.11)				
2		0.21 (0.18, 0.23)	-0.54 (-0.60, -0.49)	0.33 (0.27, 0.41)			
3			0.24 (0.19, 0.29)	-0.35 (-0.42, -0.31)	0.12 (0.10, 0.14)		
4				0.11 (0.10, 0.13)	-0.16 (-0.19, -0.15)	0.05 (0.04, 0.06)	
5					0.08 (0.06, 0.10)	-0.15 (-0.18, -0.13)	0.08 (0.06, 0.09)

The following table, Table 6.6, details the hazard ratios and the 95% confidence intervals detailing the chance an individual has of moving to an adjacent state in the model in the treatment group compared to the standard care group. The hazard ratio shows the association between the treatment and the rate of transition. A hazard ratio of one indicates that there is no association between treatment and the rate of transition meaning there is no difference between the two groups. The interpretation of the hazard ratio depends on the transition that is occurring. In Table 6.6 the first column represents transitions down the mRS scale, indicating recovery whereas the second column represents transitions to a higher mRS score, which is a worse scenario.

A hazard ratio larger than one indicates a positive association between treatment and rate of transition and is favourable on the downward transitions. A hazard ratio of less than one indicates a negative association, which is favourable on the upwards transitions.

Table 6.6: Hazard ratios for oxygen treatment compared to standard care group

Hazard ratios (95% confidence interval)			
Treatment vs Control		**0 to 1**	0.97 (0.81,1.17)
1 to 0	1.12 (0.94,1.34)	**1 to 2**	1.16 (0.98,1.37)
2 to 1	1.05 (0.88,1.25)	**2 to 3**	0.99 (0.76,1.28)
3 to 2	1.14 (0.88,1.47)	**3 to 4**	0.97 (0.82,1.15)
4 to 3	0.98 (0.81,1.18)	**4 to 5**	0.94 (0.75,1.18)
5 to 4	0.99 (0.73,1.35)	**5 to 6**	0.94 (0.73,1.20)
Favourable outcome	1.03 (0.96, 1.11)	**Unfavourable outcome**	1.09 (1.01, 1.18)

Most of the hazard ratios presented show a positive treatment effect, suggesting that treatment may be beneficial when considering the rate at which patients' transition, if a transition occurs. For example, the oxygen treatment group has a 12% increase in the risk of moving from a mRS of 1 to a mRS of 0 compared to the standard care group, suggesting a favourable outcome. However, the oxygen treatment group also has a much higher risk of moving in the opposite direction from a mRS of 1 to a mRS of 2 that the standard care group, approximately 16%, which is an unfavourable outcome. This is the only unfavourable outcome in the data, with a hazard ratio of less than one, suggesting the treatment group is less likely to make that transition, if the transition occurs.

Otherwise, the treatment group has a very small, non-significant, reduced risk of moving into a higher category on the mRS than the standard care group. None of the associations that are seen in the model are statistically significant, suggesting any possible benefit received from the treatment is not large enough to be considered a real benefit. Even though the results produced are not statistically significant, they still provide an understanding of the treatment effect for an individual receiving oxygen treatment when considering the full mRS ordinal scale, and specifically how it varies depending on where an individual is in the scale. Although it is not possible to calculate a single treatment effect, it is possible to apply constraints to the model that restrict the upwards and downward transitions to be equal at all points on the scale. In Table 6.6 it can be seen that these overall hazard ratios calculated indicate an increase in the risk of moving for the treatment group, regardless of the direction of the movement, (recovery transition OR 1.03 (0.96, 1.11, worsening transition OR 1.09 (1.01, 1.18)). By constraining the effects to be the same (in a similar manner to the proportional odds assumption) the benefit of treatment has been masked, as the unfavourable outcome between 1 and 2 appears to be influential.

As well as considering the hazard ratios and instantaneous risk with the transition intensities, it is also possible to calculate the probability of each state *s* being the next state in the process after each *r* state, calculated as $-q_{rs}/q_{rr}$ for each defined entry in the transition intensity matrix. This method gives a more intuitive parameterisation of a multi-state model than the raw transition intensities. Table 6.7 gives the probability of moving into an adjacent state for both the standard care and the oxygen treatment group. This table looks solely at individuals who move to an adjacent state, and ignores those who stay in the same state. The sum of the rows total 100% as if an individual does not move to one adjacent state then they are assumed to move the other way.

Table 6.7: Probability of moving to an adjacent state

Initial State	**Probability of next state (%) (95% Confidence interval)**						
Standard care	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0		1					
1	0.35 (0.3, 0.39)		0.65 (0.61, 0.70)				
2		0.38 (0.33, 0.45)		0.62 (0.55, 0.68)			
3			0.67 (0.61, 0.72)		0.33 (0.28, 0.39)		
4				0.69 (0.63, 74)		0.31 (0.26, 0.37)	
5					0.51 (0.41, 0.58)		0.49 (0.41, 0.58)
Treatment							
0		1					
1	0.34 (0.30, 0.37)		0.66 (0.64, 0.69)				
2		0.39 (0.35, 0.44)		0.61 (0.56, 0.65)			
3			0.70 (0.66, 0.74)		0.30 (0.26, 0.34)		
4				0.70 (0.66, 0.72)		0.30 (0.27, 0.34)	
5					0.52 (0.46, 0.58)		0.48 (0.42, 0.54)

Obviously for those with a mRS value of 0 at the first time point the only adjacent state that they can transition to is a mRS score of 1, therefore if a transition occurs it will be into mRS 1, which is why the value is 100% for both the treatment and standard care groups. For both groups up until mRS of 3, patients are more likely to have a worse score at the next time point than see an improvement in the mRS score, although these transitions do occur. In the standard care group, an individual with a mRS score of 2 has a 62% chance of moving to a score of 3 and a 38% chance of moving to a score of 1; in the treatment group these probabilities are 61% and 39% respectively. However, patients that have a score of 3, 4 or 5 on the mRS are more likely to move to a score lower on the mRS scale at the following time point, which indicates an improvement. Looking at an individual who has a score of 3 on the mRS scale in the standard care group has a 67% chance of moving to a 2 compared with a 70% chance if they were in the oxygen treatment group. Although the values are quite similar throughout, there are slight differences in the probabilities calculated in the oxygen treatment group compared to the standard care group, where those who have a worse score to start in the treatment group are slightly more likely to see an improvement than those in the control group. As well as this, the estimated transition probability matrix can be estimated, which calculates where an individual will be during a given time period in the model. An individual is able to transition to any state in the model it is just assumed that they will have moved through adjacent states in order to get there.

Table 6.8 details the estimated transition probabilities; they take the score at 3 months and estimate the probability of where an individual will be at 12 months. For example, those with a mRS score of 2 only have 19% chance of staying at the same score, a 9% chance of recovering to a score of 0 and a 30% chance of recovering to a score of 1. They may also worsen, with a 23% chance of scoring 3, 15% chance of scoring 4, a 3% chance of scoring 5 and a 1% chance of being dead at the next time point. Those with a score of 5 have a 30% chance of staying at 5 and a 37% chance of being dead at the next time point. They have a 2% chance of scoring a 1, 3% chance of scoring a 2, 7% chance of scoring a 3 and 20% chance of scoring a 4. The two scores, (mRS=2 and 5) are the only scores where at 12 months the probability of an individual moving into an adjacent state is higher than the probability that they will remain in the same state. Table 6.9 details these estimated transition probabilities when stratified into the two groups in the model.

Table 6.8: Probabilities for an individual in each state to move to any other state in the model

	Probabilities of where an individual will move to at the next time point						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	0.50 (0.47,0.52)	0.35 (0.33,0.37)	0.08 (0.08,0.10)	0.06 (0.05,0.06)	0.02 (0.01,0.02)	0.00 (0.00,0.00)	0.00 (0.00,0.00)
1	0.20 (0.19,0.21)	0.44 (0.42,0.45)	0.16 (0.16,0.17)	0.14 (0.13,0.15)	0.06 (0.05,0.06)	0.01 (0.01,0.02)	0.00 (0.00,0.00)
2	0.09 (0.08,0.10)	0.30 (0.29,0.31)	0.19 (0.19,0.20)	0.23 (0.22,0.24)	0.15 (0.14,0.16)	0.03 (0.02,0.03)	0.01 (0.01,0.02)
3	0.05 (0.04,0.05)	0.21 (0.20,0.23)	0.19 (0.18,0.19)	0.26 (0.25,0.28)	0.22 (0.21,0.24)	0.05 (0.04,0.05)	0.02 (0.01,0.02)
4	0.01 (0.01,0.02)	0.08 (0.08,0.09)	0.11 (0.10,0.12)	0.21 (0.20,0.23)	0.38 (0.36,0.40)	0.13 (0.12,0.14)	0.07 (0.06,0.08)
5	0.00 (0.00,0.00)	0.02 (0.01,0.02)	0.03 (0.02,0.04)	0.07 (0.06,0.08)	0.20 (0.18,0.23)	0.30 (0.27,0.34)	0.37 (0.34,0.41)
6	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	1 (1.00, 1.00)

Table 6.9: Probabilities for an individual in each state to move to any other state in the model stratified by treatment group

	Standard care estimated transition probabilities						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	0.49 (0.44,0.54)	0.36 (0.33,0.40)	0.08 (0.07,0.09)	0.05 (0.05,0.07)	0.02 (0.01,0.02)	0.00 (0.00,0.00)	0.00 (0.00,0.00)
1	0.19 (0.17,0.21)	0.46 (0.43,0.49)	0.15 (0.13,0.16)	0.14 (0.12,0.10)	0.06 (0.05,0.06)	0.00 (0.00,0.01)	0.00 (0.00,0.00)
2	0.08 (0.07,0.10)	0.30 (0.28,0.33)	0.18 (0.16,0.19)	0.24 (0.22,0.26)	0.15 (0.14,0.17)	0.03 (0.02,0.03)	0.01 (0.01,0.01)
3	0.04 (0.04,0.05)	0.20 (0.18,0.23)	0.17 (0.16,0.19)	0.28 (0.25,0.30)	0.23 (0.21,0.26)	0.05 (0.04,0.06)	0.02 (0.01,0.02)
4	0.01 (0.01,0.02)	0.08 (0.07,0.09)	0.10 (0.09,0.12)	0.22 (0.20,0.24)	0.38 (0.35,0.41)	0.13 (0.11,0.16)	0.07 (0.06,0.09)
5	0.00 (0.00,0.00)	0.01 (0.01,0.02)	0.03 (0.02,0.04)	0.07 (0.06,0.09)	0.20 (0.16,0.24)	0.30 (0.24,0.35)	0.39 (0.33,0.45)
6	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	1.00 (1.00,1.00)

	Oxygen treatment estimated transition probabilities						
	Final state						
Initial state	0	**1**	**2**	**3**	**4**	**5**	**6**
0	0.50 (0.47,0.53)	0.34 (0.32,0.36)	0.09 (0.08,0.10)	0.06 (0.05,0.06)	0.02 (0.01,0.02)	0.00 (0.00,0.00)	0.00 (0.00,0.00)
1	0.20 (0.19,0.22)	0.43 (0.41,0.44)	0.16 (0.15,0.18)	0.14 (0.13,0.15)	0.06 (0.05,0.06)	0.01 (0.00,0.01)	0.00 (0.00,0.00)
2	0.10 (0.09,0.11)	0.30 (0.29,0.32)	0.19 (0.18,0.21)	0.23 (0.21,0.24)	0.15 (0.13,0.16)	0.02 (0.02,0.03)	0.01 (0.00,0.01)
3	0.05 (0.05,0.06)	0.22 (0.20,0.23)	0.19 (0.18,0.20)	0.26 (0.24,0.27)	0.22 (0.20,0.24)	0.05 (0.04,0.05)	0.01 (0.01,0.02)
4	0.01 (0.01,0.02)	0.08 (0.08,0.09)	0.12 (0.11,0.13)	0.21 (0.19,0.23)	0.38 (0.36,0.41)	0.13 (0.11,0.15)	0.06 (0.05,0.07)
5	0.00 (0.00,0.00)	0.02 (0.01,0.02)	0.03 (0.03,0.04)	0.07 (0.06,0.08)	0.21 (0.18,0.24)	0.31 (0.27,0.35)	0.37 (0.32,0.41)
6	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	1.00 (1.00,1.00)

The goodness of fit can be tested by looking at the observed and expected values predicted by the model and the prevalence plots comparing the expected and observed values. The observed values are shown in Table 6.10. The observed values at months 4 and 5 are the observed values at month 3 carried forward, similarly the observed values at 6 months are carried forward to 7, 8, 9, 10 and 11 months. There are 6,747 individuals that start in the model, this is reduced to 6,291 throughout the rest of the follow-up. These values are both smaller than the 6,871 individuals included in the dataset as the model has removed those with missing values. It can be seen that over time there is an increase in the number of observed individuals with mRS scores of 0 and 6. There are decreases seen in the number of individuals who are observed with a score of 1, 2, 3, 4, or 5 from 3 months to 12 months. It is known that a lot more individuals die within the 12-month follow-up than are seen here; however, due to those individuals with missing values being excluded, all the transitions to death are not able to be captured, only those of individuals already in the model.

Table 6.10: Observed numbers of individuals in each state of the model throughout follow-up

	Model state (mRS score)							
Observed individuals at each time point (month)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	920	2,012	939	1,273	1,150	453	0	6,747
4	920	2,012	939	1,273	1,150	453	0	6,747
5	920	2,012	939	1,273	1,150	453	0	6,747
6	975	1,817	940	1,139	1,006	326	88	6,291
7	975	1,817	940	1,139	1,006	326	88	6,291
8	975	1,817	940	1,139	1,006	326	88	6,291
9	975	1,817	940	1,139	1,006	326	88	6,291
10	975	1,817	940	1,139	1,006	326	88	6,291
11	975	1,817	940	1,139	1,006	326	88	6,291
12	964	1,787	856	1,061	1,039	311	273	6,291

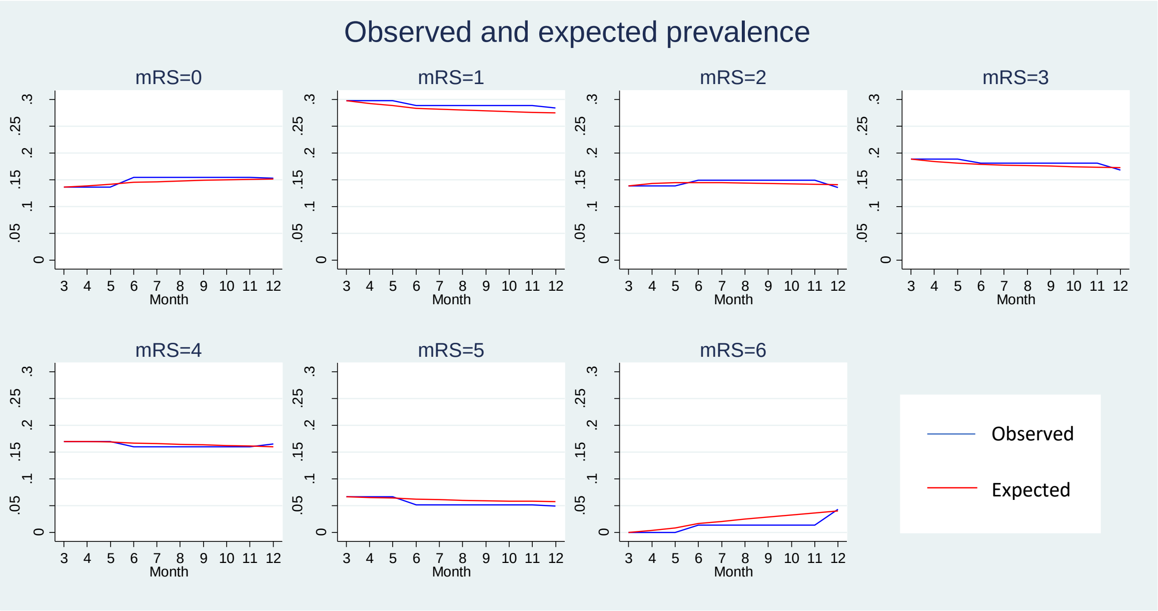
The expected values throughout the model are detailed in Table 6.11. Here there is a steadier change in the number of individuals within each category compared to the large jumps seen in the observed data - where the observed values can only be ‘updated’ at 6 and 12 months, the expected values are updated by the model at each time point. It can be seen that, like the observed data, there are increases seen in the number of individuals who are expected to score a 0 and 6 over time, with decreases in the remaining scores.

Table 6.11: Expected numbers of individuals in each state of the model throughout follow-up

	Model state (mRS score)							
Expected individuals at each time point	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	920	2,012	939	1,273	1,150	453	0	6,747
4	939	1,976	968	1,243	1,147	444	29	6,747
5	956	1,951	980	1,226	1,141	435	58	6,747
6	915	1,786	914	1,126	1,051	392	107	6,291
7	924	1,774	911	1,119	1,044	386	133	6,291
8	932	1,763	907	1,113	1,037	381	158	6,291
9	939	1,754	903	1,106	1,030	376	183	6,291
10	945	1,746	898	1,100	1,023	371	207	6,291
11	950	1,739	894	1,094	1,016	367	232	6,291
12	954	1,732	889	1,088	1,010	363	256	6,291

The prevalence plots in Figure 6.4 plot the observed and expected values for each category in the mRS scale. These can be used to assess how well the model fits the data by comparing how similar the prevalence of the expected and observed values are for each category in the model. Looking at the figure that there is very little variation seen in the expected and observed values from the model. This suggests that the multi-state model with no censoring and the treatment covariate included fit the model well.

Figure 6.4: Observed and expected prevalence plots over follow-up, stratified by mRS score for multi-state model with no censoring



A likelihood ratio test was performed comparing the model with the treatment covariate included to the null model in order to calculate the statistical significance of treatment in the model. According to the likelihood ratio, the model containing treatment was not statistically significant (p=0.55) and therefore the treatment covariate was not necessary for inclusion in the model.

This analysis had anyone with 2 or more missing values removed from the dataset, which is not ideal as it would be beneficial to be able to model everyone in the data set like the regression models. In order to address this problem, a censored state was added to the model. This allows an individual with any missing value to be included in the model as they can be placed in a censored state, meaning that they are known to be alive, but the mRS score is unknown, apart from knowing that it is not a score of 6. Therefore, all those individuals who were removed in the first analysis are now able to be included, as if the value is missing then they will be placed in a censored state, out of which they can transition.

6.4.2 Disability progression with censored states

As the model including treatment was not statistically significant, the treatment variable was removed from the model for the following analyses. Instead, the idea of having censored states within the model was introduced. By introducing a censored state, all individuals who had not died at 3 months could be included in the model, regardless of the number of actual observations recorded for each patient. Table 6.12 gives all the possible transitions that were made by individuals between 3 months and 6 months after the stroke, including a censored category.

Table 6.12: Observed transitions in SO_2S data between 3 months and 6 months including censoring

	6 month mRS								
3 month mRS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Censored**	**Total**
0	661	160	33	30	19	4	3	31	941
1	267	1,307	208	143	61	2	13	70	2,071
2	41	234	476	115	59	5	3	33	966
3	32	145	199	722	147	7	8	37	1,297
4	15	49	84	201	695	57	28	60	1,189
5	2	4	2	8	101	285	33	37	472
Censored	2	2	0	2	5	3	115	27	156
Total	1,020	1,901	1,002	1,221	1,087	363	203	295	7,092

Looking at the table it can be seen that there are no changes in most of the main body of the table from Table 6.1. All the main transitions that were there before are still there; however, the main difference is the additional transition that appears due to allowing censored states into the model. In Table 6.2 for the disability progression model without censoring, there were individuals who had a missing value for 3 months but had a scores recorded at a later time point. These individuals are amongst the censored individuals in Table 6.12. The additional individuals in this model are those individuals who had 2 missing scores, as they were previously removed, but can now be included due to the censored states. There are 27 of these individuals – they will be individuals that have a score recorded at 6 and 12 months, as individuals who had no score recorded at any follow-up time are removed from the model.

Table 6.13: Observed transitions in SO_2S data between 6 months and 12 months including censoring

	12 month mRS								
6 month mRS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Censored**	**Total**
0	647	211	33	35	29	4	6	55	1,020
1	222	1,176	169	157	60	4	18	95	1,901
2	42	216	407	160	91	8	10	68	1,002
3	27	134	175	570	185	17	18	95	1,221
4	21	38	71	131	609	77	38	102	1,087
5	3	3	1	4	60	198	39	55	363
6	0	0	0	0	0	0	203	0	203
Censored	2	9	1	4	5	4	81	189	295
Total	964	1,787	857	1,061	1,039	312	413	659	7,092

Similarly, in Table 6.13, which shows the transitions between 6 months and 12 months post-stroke. There are 189 individuals who have a censored value at 6 and 12 months – these individuals will have all transitioned from a value at 3 months to a censored state at 6 months. Originally, they would have been excluded from the model, but now they can be included in the form of a censored state, as it is known that they have not died, it is just unknown what category of the mRS they fall in. Once again, the Markov property for the multi-state model means that time can be ignored and all possible transitions combined into one state table that the multi-state model will analyse. All the potential transitions that are included in the model are given in Table 6.14.

Table 6.14: All observed transitions in SO$_2$S data between 2 time points including censoring

	Time point 2							
Time point 1	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Censored**
0	1,308	371	66	65	48	8	9	86
1	489	2,483	377	300	121	6	31	165
2	83	450	883	275	150	13	13	101
3	59	279	374	1,292	332	24	26	132
4	36	87	155	332	1,304	134	66	162
5	5	7	3	12	161	483	72	92
Censored	4	11	1	6	10	7	196	216

6.4.2.1 Model results

As before, the transition intensities were calculated for each possible transition defined by the model, regardless of the treatment group that the individual was in, and they are given in Table 6.15. As there is no covariate in the model, the risk of movement cannot be compared between groups as it is assumed that everyone in the model transitions in the same way.

Once again, the risk of moving to an adjacent state is larger in the middle states and those patients who have a score of 3, 4, or 5 on the mRS are more likely to move down a score than up a score, indicating they are more likely to be getting better (there is a greater chance of reduced disability rather than increased disability). These results are similar to those given in Table 6.4, suggesting that this model produces similar results to the disability progression model with the treatment covariate included and no censoring.

Table 6.15: Transition intensities and confidence intervals of multi-state model including censoring

	Transition intensities (95% confidence interval)						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	-0.10 (-0.11,-0.09)	0.10 (0.09, 0.12)					
1	0.06 (0.05, 0.06)	-0.17 (-0.18, -0.16)	0.11 (0.10, 0.12)				
2		0.21 (0.20, 0.23)	-0.54 (-0.59, -0.50)	0.33 (0.29, 0.37)			
3			0.26 (0.23, 0.30)	-0.38 (-0.41, -0.35)	0.12 (0.11, 0.13)		
4				0.11 (0.10, 0.12)	-0.16 (-0.17, -0.15)	0.05 (0.05, 0.06)	
5					0.07 (0.06, 0.08)	-0.16 (-0.17, -0.14)	0.09 (0.08, 0.10)

It is possible to calculate the probability of each state *s* being the next state in the process after each *r* state, calculated as $-q_{rs}/q_{rr}$ for each defined entry in the transition intensity matrix. This method gives a more intuitive parameterisation of a multi-state model than the raw transition intensities. Table 6.16 gives the probability of moving into an adjacent state for all individuals. This table looks solely at individuals who move to an adjacent state, and ignores those who stay in the same state, like table 6.7 from the previous model.

Table 6.16: Probabilities for an individual in each state to move to any other state in the model including censoring

	Probability of next state (%) (95% confidence interval)						
Initial state	**Final state**						
	0	**1**	**2**	**3**	**4**	**5**	**6**
0		1					
1	0.34 (0.31, 0.37)		0.66 (0.64, 0.69)				
2		0.39 (0.36, 0.43)		0.61 (0.57, 0.64)			
3			0.69 (0.66, 0.72)		0.31 (0.28, 0.34)		
4				0.69 (0.66, 0.72)		0.31 (0.28, 0.34)	
5					0.44 (0.40, 0.48)		0.56 (0.51, 0.60)

For individuals with a score of 1 or 2, they are more likely to move to the adjacent score that is higher, indicating that they are more likely to be more disabled after the transition. For those individuals with a mRS score of 3 or 4, they have a 69% chance of moving to a score of 2 and a 31% chance of moving to a score of 4. Finally, individuals who have a mRS score of 5 have a 44% chance of moving to a score of 4 and showing some recovery and a 56% chance of moving to a score of 6 and dying at the next time point. These values are calculated for any instantaneous time point in the model.

The estimated transition probability matrix can also be estimated, which calculates where an individual will be during a given time period in the model (Table 6.17). Here, an individual can transition to any state in the model, rather than just an adjacent state, which was the case for the probabilities before calculated above.

In terms of the estimated transition probabilities, it was once again seen that the highest probabilities were usually assigned to an individual staying in the same state that they were at the start. This was the case for mRS scores of 0 (50%), 1 (44%), 3 (26%) and 4 (38%). For a mRS score of 2 the most likely transition was to a mRS score of 1 (30%) and for a mRS score of 5, the most likely transition was to a score of 6 (44%).

Table 6.17: Probabilities for an individual in each state to move to any other state in the model including censoring

	Probabilities of where an individual will move to at the next time point						
	Final state						
Initial state	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	0.50 (0.48,0.52)	0.35 (0.33,0.37)	0.08 (0.08,0.09)	0.06 (0.05,0.06)	0.02 (0.01,0.02)	0.00 (0.00,0.00)	0.00 (0.00,0.00)
1	0.20 (0.19,0.21)	0.44 (0.42,0.45)	0.16 (0.15,0.17)	0.14 (0.13,0.15)	0.06 (0.05,0.06)	0.01 (0.01,0.01)	0.00 (0.00,0.00)
2	0.09 (0.08,0.10)	0.30 (0.29,0.32)	0.19 (0.18,0.20)	0.23 (0.22,0.24)	0.15 (0.14,0.16)	0.03 (0.02,0.03)	0.01 (0.01,0.01)
3	0.05 (0.04,0.05)	0.21 (0.20,0.23)	0.18 (0.17,0.19)	0.26 (0.25,0.28)	0.22 (0.21,0.24)	0.05 (0.04,0.05)	0.02 (0.02,0.02)
4	0.01 (0.01,0.01)	0.08 (0.07,0.09)	0.11 (0.10,0.12)	0.21 (0.20,0.22)	0.38 (0.36,0.40)	0.13 (0.12,0.14)	0.08 (0.07,0.09)
5	0.00 (0.00,0.00)	0.01 (0.01,0.02)	0.03 (0.02,0.03)	0.06 (0.05,0.07)	0.18 (0.16,0.20)	0.28 (0.25,0.31)	0.44 (0.41,0.47)
6	0	0	0	0	0	0	1.00 (1.00,1.00)

The goodness of fit can be tested by looking at the observed and expected values predicted by the model and from these, the prevalence plots between the expected and observed values can show us how well the model fit with how similar the values are. The observed values in Table 6.18 remain the same throughout intermediate time points until the information at the next time point is given. There were 7,092 individuals starting in the model and this is continued throughout as the individuals that are in a censored state, although no listed in Table 6.17 are included. There is a small increase in the observed numbers of individuals with a score of 0 and 6, with decreases seen in the numbers of individuals who score a mRS of 1, 2, 3, 4, or 5 over time. The data are now more representative of all the individuals who are included in the data set; at 12 months there are now 413 deaths, in total there are 1,046 deaths by 12 months, but the individuals who had died at 3 months still had be removed, all 633 of them as they occurred at the first time point, leaving us with 413 individuals.

Table 6.18: Observed numbers of individuals in each state of the model with censoring throughout follow-up

	Model state (mRS score)							
Observed individuals at each time point	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	943	2,074	966	1,299	1,194	616	0	7,092
4	943	2,074	966	1,299	1,194	616	0	7,092
5	943	2,074	966	1,299	1,194	616	0	7,092
6	1,046	2,000	1,002	1,252	1,147	442	203	7,092
7	1,046	2,000	1,002	1,252	1,147	442	203	7,092
8	1,046	2,000	1,002	1,252	1,147	442	203	7,092
9	1,046	2,000	1,002	1,252	1,147	442	203	7,092
10	1,046	2,000	1,002	1,252	1,147	442	203	7,092
11	1,046	2,000	1,002	1,252	1,147	442	203	7,092
12	1,040	2,036	857	1,180	1,180	386	413	7,092

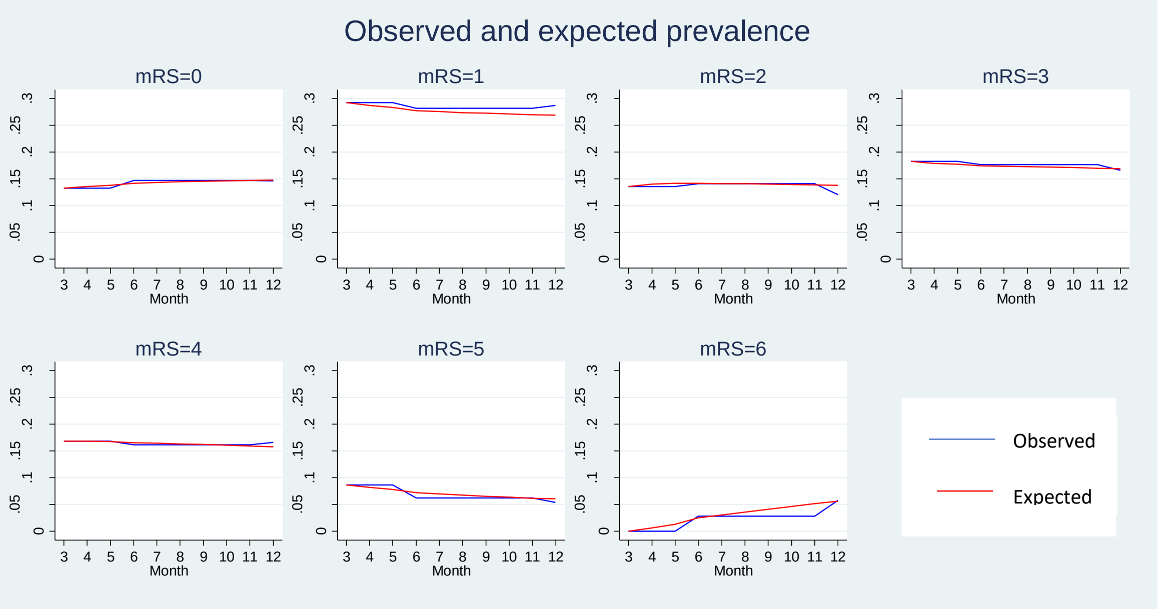
The expected values for each of the categories of the mRS that are estimated by the model are given in Table 6.19. The variation between the expected and the observed values throughout the follow-up as the expected values change at each month rather than at the inputted months like the observed data. Generally, there is a decrease over time of the number of individuals with a mRS score of 0, 1, 3, 4, and 5. For the scores of 2 and 6, there is an increase in the expected numbers of individuals in these categories over the follow-up time. This is slightly different to what happens in the observed data, unlike the disability progression model with no censoring for which the trends in the expected data reflect the trends in the observed data.

Table 6.19: Expected numbers of individuals in each state of the model with censoring throughout follow-up

	Model state (mRS score)							
Expected individuals at each time point	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
3	943	2,074	966	1,299	1,194	616	0	7,092
4	963	2,036	994	1,272	1,194	585	48	7,092
5	981	2,009	1,005	1,257	1,189	558	93	7,092
6	1,007	1,971	1,006	1,239	1,176	514	178	7,092
7	1,018	1,956	1,003	1,233	1,168	496	218	7,092
8	1,027	1,944	999	1,227	1,159	479	257	7,092
9	1,034	1,934	994	1,220	1,150	465	295	7,092
10	1,041	1,924	989	1,214	1,141	452	331	7,092
11	1,046	1,916	984	1,207	1,131	440	367	7,092
12	1,051	1,908	979	1,200	1,122	430	401	7,092

The prevalence plots in Figure 6.5 show the observed and expected values for each category in the mRS scale for the model with censored data. These can be used to assess how well the model fits the data by comparing the observed and expected values of prevalence. Looking at the figure, the observed and expected values of the model appear to be very similar, indicating the fit of the model can be assumed to be reasonable. The observed prevalence values are blue and the expected prevalence values are red.

Figure 6.5: Observed and expected prevalence plots over follow-up, stratified by mRS score for multi-state model with censoring



6.5 Discussion

This chapter aimed to fit a multi-state model to the data from the SO_2S trial, which is a relatively new method in the longitudinal analysis of stroke trials. A multi-state model was chosen because the methods allowed the ordinal nature of the score observed, while also including the category of death in a meaningful way. By including death as an absorbing state, individuals in the model were able to transition to it, but not leave it, meaning that they would not contribute to the analysis anymore, in contrast to the regression models fitted.

Two different models were fitted to the data: one that included censored states and one where no censoring was considered. The models fitted found that although many patients did not undergo any transitions between 3 and 12 months, i.e. remained in the same mRS category, the probability for each individual making a transition into a state that is not their current state can be quantified. As with the regression model, the inclusion of treatment in the model was not statistically significant, and the interpretation of the treatment effect can be seen for each individual possible transition, rather than a treatment effect for the scale as a whole.

A multi-state model can be fitted to data that has regularly spaced or irregularly spaced time points. When the time in between each time the outcome is recorded during the follow-up period is equally spaced then a discrete model can be used, but only the continuous model can be used if the spacing is uneven. The model can also handle an increase in the number of time points included, and 2 time points is the minimum number of points to which a model could be fitted.

Although the models can handle no baseline outcome, the model would benefit from the inclusion of a measurement closer to baseline, which would just be included as the initial values of the model. Missing data is more of an issue with a multi-state model, a missing value would not inform the model of where the individual was transitioning to or from. Therefore, missing data has to be dealt with in a specific way in these models, by the inclusion of censored states.

The models work on modelling the transitions between states and therefore for an outcome with very little change, the transitions estimated would be done so using very little data. This is not dependent on the size of the transitions, the model can deal well with data in which the transitions are to nearby states and also when there are larger transitions, if the data is not sparse. The model can also handle patients who have large numbers of transitions.

6.5.1 Model comparison

It can be seen that the model with no censoring and the model with censored states included produce very similar results. The main difference is the inclusion of the treatment variable, which helped to make the model more interpretable; however, the likelihood ratio test found it to be not statistically significant.

The first model with no censoring and the treatment covariate included shows that when considering the transition intensities calculated from the model there was a larger instantaneous risk of patients moving state when they are towards the middle of the scale (for example with a score of 3 or 4) than those who are at the extremes of the scale. Those who take an extreme value, i.e. severely disabled or no disability, have a much smaller risk of making any transition than those patients who fall within the middle of the scale this may be because these individuals have such a large disability that it is to a large degree irreversible. Whereas those who have a lower score, either had little disability to begin with, or they have already recovered well and therefore they may continue this recovery beyond 3 months. These conclusions are different to the regression model, which fitted common odds ratio regardless of where an individual was on the scale.

When the transition intensities are separated by the treatment that an individual received it can be seen that the transition intensities are very similar between the two groups (and thereby similar to those calculated for the whole model before the treatment group was considered). Individuals in the standard care group have a slightly smaller risk of moving to a lower state in the model, suggesting that they may be less likely to recover. The confidence intervals for the standard care group are slightly larger than the confidence intervals calculated for the oxygen treatment group, suggesting more uncertainty in the estimates, but there are half the number of individuals in the standard care group compared to the treatment group, which may lead to more uncertainty around the transition intensities.

This model with the included covariate produces hazard ratios that are comparisons of the transition intensities that were calculated. For each adjacent state in the model for which a transition is defined in the intensity matrix, the hazard ratio can be calculated. The majority of the hazard ratios calculated showed a positive treatment effect (i.e. favouring oxygen), suggesting that treatment may be beneficial when considering the rate at which patients transition. The transitions where the treatment effect was largest were the transitions between 1 and 2, and 3 and 2 where individuals in the treatment group had a 16% (1 to 2) and 14% (3 to 2) increase in the risk of making the transition between the 2 states respectively. Otherwise, the oxygen treatment group have a very small reduction in the risk of moving into a higher category on the mRS than the standard care group. It is important to note that none of the associations seen were statistically significant and the treatment variable itself was not statistically significant within the model. The model was also constrained in order to calculate 2 single hazard ratios, one for a favourable transition and one for an unfavourable transition. Both of the constrained hazard ratios show that the treatment group was more likely to move in either direction, and may have been influenced by the large effect of the transition between 1 and 2 as it contrasted all the other individual hazard ratios calculated. Although the constrained estimate was calculated, there is merit in the fact that the model produces individual hazard ratios. Clinically, they are most interpretable, as a clinician can be more specific with their patient about their potential to improve. A patient is interested in how they will recover rather than what other people have done or may do.

The second model with the censored states included shows that when considering the transition intensities calculated by the model, these are similar to those seen in the model without the censored states. The risk of moving to an adjacent state is larger in the middle states and those patients who have a score of 3, 4, or 5 on the mRS are more likely to move down a score than up a score, indicating they are more likely to be getting better (there is a greater chance of reduced disability rather than increased disability). Individuals with as score of 1 or 2 are more likely to move to a score higher rather than a score lower and so this indicates that they are at greater risk of increased disability rather than reduced disability. Due to there being no covariate included in the model with censored states, hazard ratios for a covariate in the model cannot be calculated as there is nothing to be compared. The transition to death has increased due to more individuals being included in the model who had a censored state and moved to death, as in the first model this individual would have been excluded. The model fits the data well, as can be seen by the prevalence plots comparing the prevalence of the expected and observed values within the data. There are more individuals included in this analysis than the first model, which is due to the inclusion of the censored states.

All the estimates given in both models are very similar, in term of the transition intensities and the probability of transition assigned from each state in the model. Because the models are so similar it is very hard to actually contrast them. There are benefits to using the model with censored states as it allows more individuals to be included in the model, and will show a slightly more accurate representation of the transition to death as this transition represents most of the individuals who were excluded in the first model. These are the individuals that are of most interest (as an alternative method for treating the score for death was why the multi-state model was fitted in the first place).

6.5.2 Further work

There is already allied work being carried out in this area by (Cassarly 2015). They are looking at furthering the use of multi-state models with the mRS by seeking to include a latent baseline value estimated from the NIHSS recorded for each patient. This would allow the model to estimate transitions from baseline, not only from the 3-month value, allowing the whole year for follow-up to be considered. This work is being carried out using the NINDS dataset, and therefore it was not appropriate to conduct the multi-state modelling on this dataset as the results have already been produced (Cassarly 2015). It is possible to fit multi-state survival models, where the model uses states within a cox model framework, calculating survival probabilities. This project chose to focus less on the survival of the individuals and more on the methods of analysis of the ordinal mRS. However, this could be of interest and could be useful, but would require different software, which has been developed for both R and recently Stata (de Wreede et al. 2011, Crowther 2016) although it requires more detail about transition times. Fitting these models may allow more complex models to be fitted; these are the models that were previously discussed but were not able to converge for the dataset being used here. It may be that because the size of the dataset and the number of states being modelled are both large, the models fitted struggle to converge when models with more complex transitions are being fitted.

An alternative to fitting a multi-state model with censored states included would be to create a complete dataset using multiple imputations, however it was decided that because it was the outcome of interest on which the imputation would be performed, that this was not a good idea. Also the data that has been used throughout was limited to only the treatment group variable and therefore the information used to impute the data would not be included in any of the following analysis. This may be an extension to consider in further work, to see how much variation is seen when the missing data is imputed and assigned a physical value rather than the model using censored states.

6.5.3 Conclusion

Overall, multi-state models are very useful as a novel method for analysing the mRS scale without calling into question the ordinality of the data. The results of the model are easy for a clinician to interpret and the models fitted do not take long to run. The effects seen here are over the whole follow-up period and are calculated for each individual transition rather than one whole treatment effect. By including censored states, the model produces good estimates of the transitions that an individual may make from the current mRS score that they have during the 12-month follow-up period.

The multi-state model has great potential for the longitudinal analysis of stroke trials, it is just unfortunate that the software package used was unable to fit potential models that would have been of interest looking at the recovery of stroke patient, as the absorbing state has the highest value and therefore the model will always be looking at worsening in the disease progression rather than recovery. They provide an alternative approach to analyse repeated measures data with an ordinal outcome. These models describe how a patient moves between states over time, which is desirable in the description of disease progression that naturally moves through increasing stages of severity.

The second alternative to regression modelling is latent class analysis, to be described in the next chapter, which looks at the longitudinal analysis of the stroke trial from a different perspective, that of the individual patients rather than just the whole trial population split into treatment and standard care.

7 Latent class analysis

The previous regression methods discussed have taken a variable-centred approach to modelling the data, whereby the aim of the analysis is to identify the relationships between the dependent and independent variables. There is an alternative way to identifying relationships, which is to take a person-centred approach to the modelling. In a person-centred approach, relationships between patients are explored by classifying them into groups, or classes, based on individual response patterns. The analysis in this chapter will focus on a person-centred approach, specifically a form of latent class analysis (LCA), for two main reasons:

i) It can provide information about individual patient outcomes over time

ii) It can separate the effect of changes observed over time by an individual patient from between-subject differences at baseline.

7.1 Introduction

It is very common to observe heterogeneity in treatment response in clinical trials. It is challenging for researchers to know how best to examine whether there are patterns to this variability. For example, sub populations may exist within larger populations that show differential growth trajectories that may be masked when group means are calculated. In terms of treatment, identifying and categorising differential responders have implications for trial development and clinical management and can potentially highlight information to facilitate the targeting of patients most likely to improve.

The systematic literature review in Chapter 3 identified a latent class growth analysis that had previously been used in a randomised clinical trial in stroke. The study looked at a telephone-based problem-solving intervention for family caregivers of stroke survivors, where the outcome of interest was depressive symptoms and the caregivers' sense of competence (Pfeiffer et al. 2014).

This is obviously a very different outcome being assessed to the mRS. Another looked at a change in handicap over time looking at LCGA but used the modified BI and the London Handicap Scale as the measure of handicap rather than the mRS as well as looking at the changes in depression and quality of life of the individuals (Pan et al. 2008a). There have also been studies that conduct LCA, without the longitudinal data, for example the profiles of the NIHSS at baseline, which was then used to predict patient outcome (Sucharew et al. 2013).

An abstract presented in 2013 by Busija et al at the International Stroke Conference looked at a latent class growth analysis of the mRS in stroke survivors, considering the pre-admission mRS, the mRS at hospital admission and the score 3 months post-stroke. They used data from 202 individuals and were able to identify 6 different clusters in the model. Although this analysis is of a similar form to the analyses in this chapter, it is important to note that the model used pre-stroke Rankin values that are not available for all the data in this project and this was a survivor only model; therefore it does not include a transition to death, which is what is of interest in the models that are fitted in this chapter (Busija et al. 2013).

7.2 Latent class analysis

Traditional approaches to the study of post-stroke recovery assume that recovery follows a homogeneous path; LCA can be used to investigate potential heterogeneity within the recovery pathways. Latent class models aim to measure the change over time in latent variables. LCA attempts to identify unobservable subpopulations with distinct characteristics within the overall population. There is an assumption that each individual can be classified into one of a finite number of unobserved clusters, with the cluster subpopulations based upon patients' responses to multiple unobserved variables. The modelling process uses the probabilistic modelling under finite distributions. This uses probability theory to express all uncertainty in the model. It is assumed that patients in each cluster will express similar characteristics in the observed variables, whilst these variables will differ between the other clusters recorded. The LCA model can be extended in order to look at longitudinal data, where it is possible to conduct latent transition models, longitudinal LCA and latent class growth analysis.

7.2.1 Latent transition analysis

Latent transition analysis (LTA) is a statistical method that uses longitudinal data and attempts to identify movements between different subgroups that have been identified. LTA is an extension of LCA. LCA detects the presence of latent classes within a dataset and looks to create patterns of association. These patterns of association are subpopulations that have distinct characteristics within the overall population. The subjects in the subpopulations have an unmeasured class membership, which is identified using categorical or continuous observed variables. LTA is the longitudinal extension of this, where transitions over time are estimated. The LTA model aims to estimate the prevalence of stage memberships and the incidence of stage transitions. The main difference between LCA and LTA is that in LCA the latent classes are found to represent a stable set of characteristics and individuals are specifically categorised, whereas LTA allows an individual to change cluster membership at different time points in the model. LTA is conducted in a two-step approach. First LCA is performed on a set of variables that are measured at each time point, for all the time points that the model is estimating. This allows subjects with similar characteristics at each time point to be classified into a cluster. Each individual will be assigned to a single cluster at each time point. The second part uses auto-regressive models to calculate the effect of being in a specific cluster at time *t*, dependent on the cluster that the individual was in at the previous time *(t–1).*

LTA could be considered an appropriate technique to be used in stroke research as we have an ordinal scale through which people will change over time, and it is beneficial to see which of those individuals that move there have certain characteristics – which have previously been recorded – that indicate that a patient is more likely to transition. LTA is an ideal method for use when the observed data are categorical in nature (Collins & Lanza 2010) and we are assessing a research question that looks at discrete change over time. However, there are some limits to using LTA: firstly, LTA is limited to small models due to the number of parameters included in the model, and secondly, it is not able to identify changes across multiple time points for each patient, and patients who had similar patterns of recovery would not necessarily be grouped together (Collins & Lanza 2010). This is something that the model chosen for this analysis needs to do, and so LTA is not appropriate as we want to group people with similar trajectories together.

7.2.2 Longitudinal latent class analysis

The second method that was considered for classifying the longitudinal data into recovery trajectories was longitudinal latent class analysis (LLCA). This method is another basic extension of LCA, where multiple variables are recorded as a longitudinal sequence of the same variables on multiple occasions (Collins & Lanza 2010). It is essentially the same method as the basic LCA model; however, there is one variable being considered at multiple time points rather than multiple variables at one time point. The aim of the model is to identify patterns of the state of an individual over time (Vermunt et al. 2008). The model is capable of identifying trajectories over time as people with similar scores at each time point are allocated to the same cluster. The LLCA model is able to capture change across multiple time points, unlike the LTA; however, there is no account taken of the order in which the values for each of the variables are recorded, so time is not taken into account, which is an important consideration when looking at repeated measures data over time. There is definitely an order to the mRS values recorded within the data used within this project and so it is important to take this into account with the method that is chosen.

7.2.3 Latent class growth analysis

The final method that was considered was latent class growth analysis (LCGA). This method is similar to LLCA in that the model aims to classify individuals into different clusters based on the trajectories over the follow-up time, so it was able to identify patterns of change across multiple time points, and unlike the LLCA model, the time order is considered. Each subject in the population is assigned to a specific cluster and, unlike LTA, they are not able to transition between clusters, as the cluster is the pattern over the whole follow-up period (Collins & Lanza 2010). For each cluster fitted to the data, the model estimates the mean growth curve as a function of time with a cluster-specific intercept and slope parameter. Each patient in the cluster is assumed to follow the same mean growth curve, therefore it is assumed that there is no variation between the individuals within the same cluster (Vermunt et al. 2008). This method was the one that was chosen as it was the most appropriate for the modelling this project is undertaking, as it permitted us to model distinct trajectory groups and multiple re-measurement points and accounted for the factor of time. It is a useful method to use if the progress of symptoms is expected to vary across time and among patients. Within both datasets, although there are many individuals who remain the same, there are also many whose disability varies over time. Due to there being no variation between individuals in each cluster, the model is reasonable to interpret, and it is possible to use standard measures of fit, such as log-likelihood and information criterion, to determine optimal cluster membership (Nylund et al. 2007).

7.3 Method

7.3.1 Model description

The method of LCGA fits a semi-parametric model with parameters that describe individual-level trajectories. For continuous variables, these trajectories are assumed to follow a multivariate normal distribution. The heterogeneity seen in the individual sequences of the variables modelled over time is explained using latent variables, and the sequences classified into different clusters. For every subject in the dataset, the posterior cluster membership probability of being in each cluster is calculated. The individual is then assigned to a cluster, using the maximum probability assignment rule, whereby the individual is assigned to the cluster in which their probability of membership is the largest. Each cluster is assumed to have no variation between the individuals in the cluster and so the mean growth curve is the same for all individuals. The mean intercept and linear slopes are calculated using local independence, where the clusters are the only cause of dependence among observed variables and there is no residual covariance among them.

We first start by considering the repeated variable *Y*. Let Y_{it} denote the value *c* of that repeated variable for the i^{th} individual at time *t*, where $t \in \{1, 2 \dots T\}$ and *T* is the maximum number of time periods. The longitudinal sequence of the repeated measures *Y* can be denoted at each time period $Y_i = (Y_{i1}, Y_{i2} \dots Y_{iT})$. Let *X* be a latent nominal variable with *K* clusters. The general form of LCGA is as follows:

$$P(Y_{it}) = \sum_{k=1}^{K} P(X = k) \prod_{t=1}^{T} P(Y_{it} = c|X = k) \qquad [7.1]$$

Where $P(Y_{it})$ is the unconditional probability of the repeated measure *Y*. *P(X=k)* is the probability that a randomly selected patient *i* belongs to cluster *k*, where $k \in \{1, 2 \dots K\}$ and $P(Y_{it} = c \mid X = k)$ is the likelihood that a randomly selected patient has a specific value, *c*, of the repeated measure at time point *t* with the patient in cluster *k*. So the unconditional probability of the repeated measures is calculated by taking the summation of a randomly selected patient belonging in cluster k and the product of the likelihood the patient in cluster k has a specific value of the repeated measure, c, at time point t.

For an ordinal variable, there are adaptations made to the model. There is no adaptation needed to the likelihood form for continuous normal variables, and the time order effect as specified below is described in a standard regression model:

$$Y_{it} = \alpha_i + \beta_{1i}\eta_t + \varepsilon_{it} \qquad [7.2]$$

In the Equation 7.2, α_i is each individual patient *i*'s intercept parameter. This is interpreted as a patient's response for the repeated measure at time *t=1*. β_{1i} is the individual *i*'s linear slope parameter, which is the growth rate of Y_i from one time point to the next time point *(t+1)* for patient *i*. ε_{it} is the residual at time *t* for a patient *i*; this is assumed to be independent across each patient. η_{it} is the time order effect on the repeated measure. It is assumed that at time point *t*, $\eta_{it}= t1$. The correlations between ε_{it}, α_i, and β_{1i} are assumed to be 0 for all *i* and *t*.

There is no variation between individuals within the same cluster, and so all the parameters for individuals within the same cluster are equivalent to the mean across all subjects in cluster *k,* Equation 7.3.

$$\alpha_i^k = \overline{\alpha^k}$$
$$\beta_i^k = \overline{\beta^k} \qquad [7.3]$$

In order to adapt the model for an ordinal variable, an auxiliary threshold model is required in the specification of the time order effect. In that auxiliary model an underlying continuous variable Y_{it}* is defined, which is assumed to be normally distributed. It is the cut-off point of this continuous variable that links Y_{it}* to the ordinal variable Y_{it} with *C* categories. The cut-offs are referred to as the thresholds $\tau_{s-1} - \tau_7$ (where *s* is the order of category *c* of the ordinal variable). The underlying continuous variable is defined in the same way as the regression equation above. So as before, the cluster specific growth factors, $\overline{\alpha^{*k}}\ \overline{\beta^{*k}}$ are established via the underlying continuous variable.

For category, c, of the repeated measures score, a cumulative logit is used to calculate the probability that a randomly selected patient *i* has category *c* or lower on the ordinal variable Y_{it}, in a similar form to ordinal regression.

7.3.2 Cluster allocation

The allocation of an individual to a cluster is based on the posterior membership probability, which is derived from Bayes' theorem. This is the likelihood that an individual should be allocated to a cluster. It is defined for all possible clusters and the sum of the posterior membership probability is equal to one. The probability is estimated from the cluster specific parameters $\overline{\alpha^{*k}}\ \overline{\beta^{*k}}$. The cluster assigned to each individual is the one where the probability is highest.

7.3.3 Goodness of fit

The decision in choosing the optimal latent class model, the best number of classes that the model should have, can be made using several different criteria and tests to inform the decision. There is not common acceptance of the best criteria for deciding the optimal number of classes (Nylund et al. 2007). Considered below are the different criteria and tests that can be used to help inform the decision when choosing the model with the best fit and an optimal number of classes.

7.3.3.1 Information criteria

The Information criterion (IC) is one option that can be used to consider the optimal numbers of classes in a model. These IC indices are based on the maximum log-likelihood of the fitted model, with penalties on the model due to the number of parameters and the sample size of the data. Because of the different penalties across the IC indices, it is possible and likely that each IC may suggest different models have the optimal number of clusters.

Akaike information criterion

The Akaike information criterion (AIC) is a measure of the quality of the model commonly used in model selection (Akaike 1987). Given a range of models fitted on the same data, the AIC value can be compared to judge the model that is most appropriate. The preferred model is the one with the smallest AIC value. Unlike a likelihood ratio test, which can only compare nested models, there are no restrictions on the models that can be compared using the AIC. For a model with *k* parameters the AIC can be calculated by

$$AIC = 2k - 2\ Log\ Likelihood$$

Bayesian information criterion

The Bayesian information criterion (BIC) is a measure closely related to the AIC; however, the BIC usually penalises free parameters more than the AIC, as the penalty term for increased parameters is larger than in the AIC (Schwarz 1978). The BIC can also be compared between models fitted to the same data, and the model with the smallest BIC value would be the preferred model. BIC can be calculated for *k* parameters in a sample size of *n* individuals

$$BIC = 2\ Log\ Likelihood + k\log(n)$$

Adjusted BIC

An adjusted version of the BIC has also been developed where the sample size *n* is replaced by a new variable *n**, which is calculated from the sample size (Sclove 1987). The adjusted BIC should lead to better performance for model selection, suggesting that the optimal number of clusters is more likely to be chosen, when the number of parameters in the model *(k)* is large, or when the sample size is small (Yang 2006).

$$n^{*} = \frac{(n+2)}{24}$$

IC comparison

The AIC, BIC and adjusted BIC can all be used to find the best-fitting model to suggest the optimal number of classes. If these IC indices agree then the optimal number of classes has been found. When these IC indices all suggest different optimal numbers of classes, previous research has suggested that the adjusted BIC is superior to other IC indices (Yang 2006). Also, it has been suggested that the BIC also a more favourable index than the AIC (Vermunt & Magidson 2003). A simulation study conducted to compare the different criteria used to determine the number of classes found that the AIC was not a good indicator of the number of classes needed, as it commonly identifies models with lower cluster sizes, and the BIC was the superior IC for determining the number of classes (Nylund et al. 2007). All 3 IC indices will be considered throughout the analysis; however, care will be taken with interpretation, given the differences seen, and favour given to the BIC.

7.3.3.2 Likelihood ratio tests

The most commonly used likelihood ratio test (LRT), the log-likelihood ratio test, cannot be used to compare nested latent class models. This is because the difference in the log-likelihood of two nested models is assumed to follow a chi-squared distribution. Although LCA models with different classes are considered to be nested models, if you compute the difference between a model with *k*+1 and *k* classes the difference is not distributed as chi-square (McLachlan & Peel

2000). Therefore, standard testing is not applicable. Because of this, adaptations to a LRT have been suggested.

Vuong-Lo-Mendell-Rubin LRT

The Vuong-Lo-Mendell-Rubin LRT (VLMR-LRT) proposes an approximation to the LRT distribution as an alternative method for comparing nested latent class models (Vuong 1898, Lo et al. 2001). The VLMR-LRT compares the improvement of fit between two neighbouring class models and provides a *p*-value that can be used to determine the statistical significance in the improvement of the fit after the inclusion of one more class in the model. A statistically significant *p*-value suggested that the model with one less class should be rejected in favour of the model with more classes. Although there have been criticisms of the model (Jeffries 2003) relating to a flaw in the proof, there is uncertainty in how much this critique affects its use in practice.

Adjusted Lo-Mendell-Rubin LRT

The Vuong-Lo-Mendell-Rubin LRT compares the model fit of a model with *k* classes to a model with *k* -1 classes and assesses the statistical significance. An adjustment by Lo-Mendell-Rubin can also be calculated (Lo et al. 2001). The model with one less class is calculated from the model with *k* classes by removing the first class in the model. Because of this, consideration should be taken when choosing the starting values of the model so that the last class is the largest class in the model where possible. Once again a significant *p*-value suggests that the model with *k–1* classes should be rejected in favour of the model with *k* classes.

Parametric bootstrap LRT

Another alternative method to compare nested latent class models is the parametric bootstrap LRT (BLRT) (McLachlan & Peel 2000). This method uses bootstrap samples to estimate the distribution of the difference in log-likelihood test statistic, as the chi-squared distribution cannot be assumed. So instead of assuming a known distribution, the distribution is empirically estimated by BLRT. In the same way as the VLMR LRT, the BLRT provides a *p*-value comparing the model fit on a *k*–1 and a *k* class model, with a statistically significant *p*-value rejecting the *k–1* model in favour of the *k* class model. The BLRT estimates the distribution for the model comparison with the following method:

1. Estimate both the *k–1* and *k* class models to obtain the likelihoods for calculating the –2 log-likelihood difference.
2. Under the null *k–1* model, generate bootstrap samples and calculate the difference in –2 log-likelihood between the *k–1* and *k* class models
3. Repeat the process independently many times to provide an estimation of the true distribution of the –2 log-likelihood difference.
4. Estimate the *p*-value by comparing the distribution obtained in stage 3 with the –2 log-likelihood difference obtained in stage 1.
5. Use the *p*-value to decide if the null *k–1* model should be rejected in favour of the *k* class model.

The main disadvantage of the BLRT is the amount of computation time needed to fit each model due to the repeated sampling for the bootstrapping. Models fitted in this analysis could take up to up to 30 minutes or more to fit for each cluster size.

Comparison of LRT tests

A simulation study compared the VLMR and BLRT and found that the BLRT performed better in most settings (Nylund et al. 2007). Both versions of the LRT had good power to detect the changes in the likelihood, but BLRT was more accurate in identifying the correct number of classes. It was found that the LMR method overestimated the number of classes. It was also found that the LMR method could estimate *p*-values that changed from being significant to non-significant and back to significant with the inclusion of more classes, whereas once the *p*-value associated with the BLRT was non-significant, it remained that way.

As well as this, the change in log-likelihood between the different models fitted can be used to assess whether the optimal number of clusters has been reached. It has been suggested that when the change in log-likelihood decreases between models and starts to plateau, then the optimal number of clusters may have been reached (Muthén & Muthén 2009).

7.3.3.3 Numbers in groups

In order to be able to draw conclusions from the clusters and to be able to distinguish the characteristics of each cluster, there needs to be a minimum number of individuals in a cluster of a *k* class in order for the *k* class model to be considered an accurate model. In the literature, there does not appear to be a consensus on the minimum cluster size, as it varies depending on the sample size and the study aims. For example, class sizes of 1% (Strauss et al. 2014), 5% (Ploubidis et al. 2007), 7.8% (White et al. 2000) and 8% (Silverwood et al. 2011) have all been reported within previous studies that have conducted either LCA or LCGA, which shows that there is a variety of class sizes that could be chosen and no real consensus amongst previous research. Due to the large numbers of individuals that are included in the analysis conducted, a minimum class size of 5% was chosen (Nylund et al. 2007). This value fits well with the other minimum class sizes that have been used previously. It should allow for more clusters to been seen, without including ones that only have a few individuals in them. It was also decided that no cluster could be larger than 50% of the data; this is to ensure that a cluster does not contain the majority of the data and that there is more of an even split of the data amongst the clusters.

7.3.3.4 Posterior probabilities

The posterior probabilities show the probability of being in each class for each individual. The larger the posterior probability, the greater the chance an individual belongs in that class. It is important that the optimal model should not have posterior probabilities for each class that are similar to another for a given individual. For example, an individual in class one may have 90% chance of being in class 1 and a 10% chance of being in class 2 in a 2 class model. When these posterior probability estimates are close (e.g. a 60% chance of being in class 1 and a 40% chance of being in class 2 in the 2 class model) then there is said to be indistinct classification. It was decided that a model would not be considered if the posterior probability for a cluster in the model was less than 0.7. This threshold was adopted to provide for excluding subjects with equivocal assignment probabilities (Clark & Muthén 2009).

7.3.4 Choosing an optimal model

It is important to consider several factors when determining the number of classes. This includes the goodness of fit of the model, the research question, parsimony, and theoretical justification.

- The fit indices and tests of model fit are useful in the exploratory stages of the analysis.
- A model can be deemed optimal if the posterior probabilities calculated across all clusters are near to 1.
- Although there is no consensus on the minimum cluster size, it is important that they are not too small. It is common to recommend each cluster containing 5% of the data (Nylund et al. 2007). The main importance of the cluster size is that the changes in model stability can be observed.

7.3.5 Model estimation

The LCGA was conducted using the software Mplus (Muthén & Muthén 1998-2015). Mplus uses the maximum likelihood methods based on the expectation maximisation algorithm in order to estimate the model parameters. These include the conditional probabilities and the cluster specific factors.

First, random start values are generated by the software (default in Mplus is 10) and used in the model estimation, with Mplus ranking the resulting log-likelihoods in numerical order. The second stage of the process selects the largest log-likelihoods (the number specified by the user) and examines replication of the log-likelihood. For example, in a model with 500 random start values, and 50 starting value sets, 500 start values will be used to generate log-likelihoods and the 50 largest will be selected. Increasing the number of starting values increases the chance of converging on the largest log-likelihood but also increases the computational time. In addition to this, the user can specify the number of iterations (default is 10), which reflects the number of times Mplus selects the log-likelihood from the start values. Therefore, increasing the number of iterations can increase the probability of converging on the largest optimum log-likelihood. Should the log-likelihood replicate, then this value is used to compute the model estimates.

In the simulations performed in this analysis, 2000 random start values, with the 100 largest log-likelihoods selected, and 100 iterations were used. If the log-likelihood did not replicate then the random starts were increased exponentially from 2000 to 5000, and 10,000 (with the number of values selected and iterations increasing from 100 to 200, and 500 respectively) until model convergence was achieved.

7.4 Results using the SO_2S data

The SO_2S study followed individuals over a 12-month follow-up period. There were 3 follow-up time points – 3 months, 6 months and 12 months post-stroke – and over the 3 follow-up time points the score that each individual received on the mRS is recorded. For each patient in the study, the change on the scale was noted over the three follow-up time points to distinguish the pattern of recovery over time. By looking at each individual, 192 different recovery patterns were found. There were 26 different patterns that occurred in approximately 1% or more patients, and these are listed in Table 7.1. Only patients who had a mRS value recorded at all 3 follow-up points were included in these recovery patterns. This was 6,203 (78%) patients out of the 8,003 patients that were recruited for the SO_2S trial.

Table 7.1: Recovery patterns that occur in more than 1% of individuals in SO_2S dataset

Pattern	N	%	Pattern	N	%
000	452	7.3	**223**	71	1.1
001	110	1.8	**311**	67	1.1
011	81	1.3	**322**	82	1.3
100	143	2.3	**331**	67	1.1
101	89	1.4	**332**	90	1.5
110	150	2.4	**333**	356	5.7
111	869	14.0	**334**	105	1.7
112	94	1.5	**344**	73	1.2
113	81	1.3	**433**	90	1.5
121	65	1.0	**443**	84	1.4
122	87	1.4	**444**	408	6.6
211	134	2.2	**555**	164	2.6
221	99	1.6	**666**	633	10.2
222	203	3.3			

The most frequently observed pattern, 111, was observed in approximately 14% of people. These are patients who were classed as having no significant disability despite the symptoms of a stroke at all time points. The other most frequently observed patterns were 000 (7.3%), 222 (3.3%), 333 (5.7%), 444 (6.6%), 555 (2.6%). As well as these, there were approximately 10% patients who were dead at the 3-month follow-up and these patients obviously had a pattern of 666 as they were unable to change score over time. These frequently observed patterns represent the patients who had no change in their recovery from 3 to 12 months regardless of the disability that they had at the 3-month follow-up period. A patient is described as having had some recovery between 3 and 12 months if they have a lower score at 12 months than 3 months and a patient has deteriorated if the score at 12 months is higher than 3 months There are 10 patterns over the follow-up that show recovery in a patient and 8 patterns that show patients have deteriorated. There were 8 patterns where the 3-month and the 12-month score are the same, showing no change in disability over the follow-up time even if there was a change at 6 months.

7.4.1 All patients at 3 consecutive time points

At each of the 3 consecutive time points, the distribution of the mRS is given in Table 7.2. This is the distribution of the 7,725 individuals that were included in the LCGA, see Appendix F for flow diagram of participants. These were all individuals recruited to the trial, with those individuals who had no mRS value at any time point removed. The proportion of individuals in each category remained quite steady with the main increase seen in the category for death, as obviously individuals do not leave this category.

Table 7.2: Distribution of individuals on the mRS over the 3 follow-up points for SO_2S data

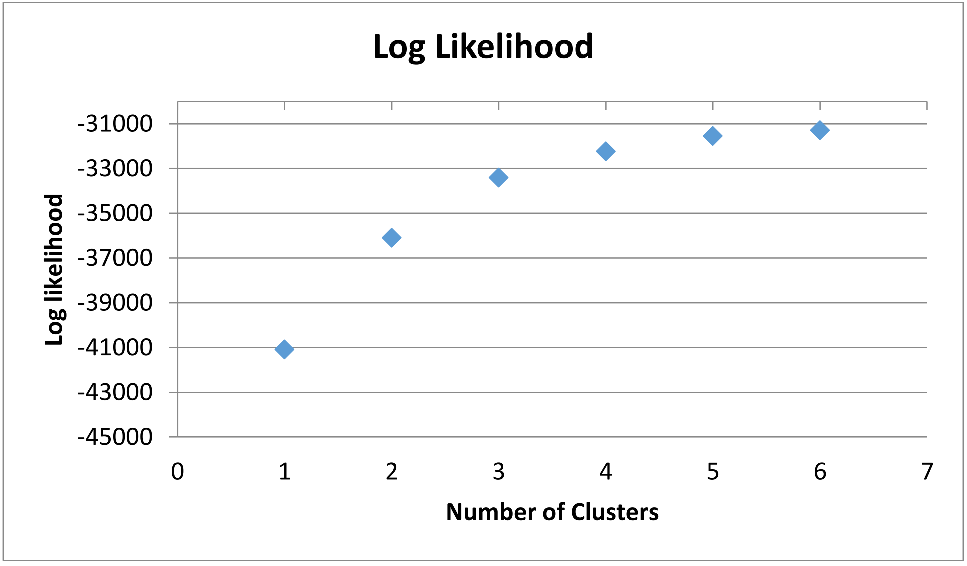
mRS Score	3 months (%)	6 months (%)	12 months (%)
0	12.4	13.7	13.6
1	27.4	25.6	25.3
2	12.8	13.5	12.1
3	17.1	16.4	15
4	15.7	14.6	14.7
5	6.2	4.9	4.4
6	8.4	11.3	14.8

A series of 1, 2 … 6-cluster models were fitted to the data, and the fit indices are defined in Table 7.3. Models with 7-clusters or more were unable to be fitted due to the maximum likelihood not being able to be replicated. We see that the IC indices have decreased with each additional cluster added to the model. This was also replicated in the log-likelihood, with a decrease seen with each additional cluster of the model, shown in Figure 7.1. The BLRT gave *p*-values that were statistically significant for the fit of each model with more classes added, similarly for the VLMR and adjusted LMR until the 5-cluster model was reached. The 5-cluster model did not show a significant improvement on the 4-cluster model, and the 6-cluster model provided an error when trying to estimate the VLMR and adjusted LMR LRT.

Table 7.3: Model fit statistics for LCGA models fitted to SO_2S data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	82198.95	82247.62	82225.37	-	-	-	-41092.5
2	72225.67	72295.19	72263.41	<0.001	<0.001	<0.001	-36102.8
3	66852.38	66942.75	66901.44	<0.001	<0.001	<0.001	-33413.2
4	64499.52	64610.75	64559.91	<0.001	<0.001	<0.001	-32233.8
5	63143.24	63275.33	63214.96	1	1	<0.001	-31552.6
6	62630.27	62783.22	62713.31		Error		-31293.1

Figure 7.1: Graph of the values of the log-likelihood for 6 LCGA models fitted to SO_2S data



With the IC indices and LRTs, the best model that could be fitted to the data was the 4-cluster model. The numbers of individuals in each cluster were considered to check that each cluster contained more individuals than the defined minimum of 5% of the data. The cluster-specific proportions and the average posterior probabilities (AvPP) of the cluster membership of the 4-cluster model are shown in Table 7.4.

Table 7.4: Posterior probabilities and number in each of the assigned clusters for all patients in SO_2S data

	Posterior probabilities					
Assigned cluster	**n**	**%**	**1**	**2**	**3**	**4**
1	877	11	**0.972**	0.028	0	0
2	1,375	18	0.006	**0.901**	0	0.092
3	2,752	36	0	0	**0.931**	0.069
4	2,721	35	0	0.035	0.055	**0.91**

The distribution of the clusters was fairly evenly spread, with two larger clusters, cluster 3 (n= 2,753, 36%) and cluster 4 (n= 2,721, 35%), as well as two smaller clusters, cluster 1 (n= 877, 11%) and cluster 2 (n= 1,375, 18%). There was not one cluster that contained the majority of the data. As well as this, the AvPPs indicated that there was a good agreement with the individuals being assigned to the correct cluster as the smaller AvPP value was 0.9. The 4 derived cluster-specific trajectories in the 4--cluster model were evaluated. Table 7.5 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.5: Conditional probabilities for each of the assigned clusters for all patients in SO_2S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0	0	0	**0**	0	0	0
1	0	0	0	**1**	0.006	0.005	0.003
2	0	0	0	**2**	0.016	0.014	0.01
3	0.001	0	0	**3**	0.106	0.092	0.068
4	0.019	0	0	**4**	0.596	0.579	0.532
5	0.113	0	0	**5**	0.228	0.254	0.309
6	0.867	1	1	**6**	0.048	0.056	0.077
Cluster 3				**Cluster 4**			
0	0.348	0.355	0.371	**0**	0.014	0.013	0.013
1	0.556	0.552	0.542	**1**	0.183	0.181	0.177
2	0.068	0.066	0.063	**2**	0.283	0.282	0.279
3	0.023	0.022	0.021	**3**	0.376	0.378	0.382
4	0.004	0.004	0.004	**4**	0.135	0.136	0.14
5	0	0	0	**5**	0.008	0.008	0.008
6	0	0	0	**6**	0.001	0.001	0.001

In cluster 1 (n= 877, 11%), at 3 months the conditional probabilities suggested that 87% (0.867) of individuals had already died, followed by 11% (0.113) who were classified as being severely disabled and 2% (0.019) who had a moderate-severe disability. By the 6-month follow-up, all individuals who were in this cluster had died.

In cluster 2 (n= 1375, 18%), the conditional probabilities showed that the cluster contained those individuals who were more severely disabled but had not died by 6 months. Within this cluster, 23% were classified as severely disabled at 3 months, and this increased to 25% at 6 months and 31% at 12 months. Most individuals had a mRS score of 4 at each follow-up point, with nearly 60% at 3 months, which decreased to 58% at 6 months and 53% at 12 months. There were 10% of individuals who had a moderate disability (mRS=3) at 3 months, which decreased to 6% at 12 months. The number of individuals who had died increased (5% at 3 months, 6% at 6 months and 8% at 12 months).

The largest cluster in the model was cluster 3 (n=2,752, 36%) in which most individuals had no disability at all and could be classed as independent. There were 35% of individuals who scored a mRS of 0 at 3 months, which increased by 1% for each follow-up time point. The majority of the individuals in this cluster were classified as having a mRS score of 1, indicating no disability despite having symptoms: approximately 56% of individuals at 3 months, 55% at 6 months and 54% at 12 months. The remaining individuals were spread evenly across other categories over time; 6% scored a mRS score of 6 and 2% a mRS score of 3.

Cluster 4 (n= 2,721, 35%) was only marginally smaller than cluster 3 and represented a much broader range of disability than had been seen in the other clusters, which is consistent across the follow-up time period. There were approximately 18% of individuals who had no disability at all (mRS=1). Approximately 28% of the patients in this cluster were defined as having a slight disability (mRS=2). Almost 38% of the individuals in the cluster score had a mRS of 3. Finally, there were still 14% of individuals in the cluster who scored even higher on the mRS scale, with a moderate-to-severe disability.

The cluster specific characteristics for each of the fitted clusters of the model are described in Table 7.6 to give an indication of the profile of each cluster over time. It can be seen that apart from the movement to a score of 6 (death) seen in cluster 1, there was not a lot of movement of individuals within each of the clusters. This may be because many individuals do not have a change in the mRS score over the follow-up period.

Table 7.6: Characteristics of the 4-cluster model fitted to all individuals in SO_2S data

Assigned cluster	n	%	Characteristics
1	877	11	Some severe disability to death
2	1,375	18	Moderate-severe throughout
3	2,752	36	No disability throughout
4	2,721	35	Slight to moderate disability throughout.

Figure 7.2: Graph showing the trajectories the individuals in each cluster, for all patients in SO_2S trial

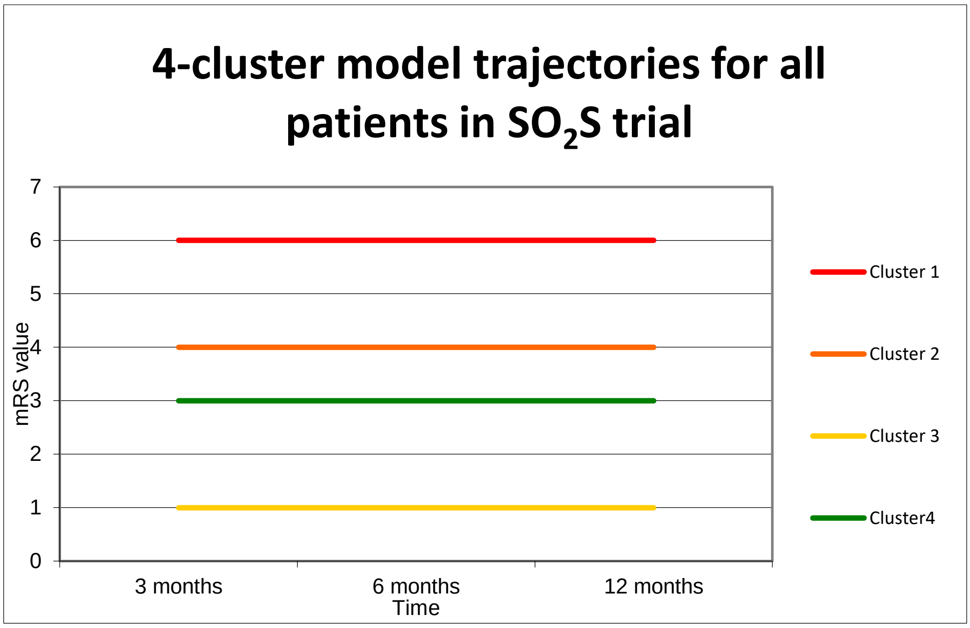


Figure 7.2 shows the model trajectories for the 4 clusters that were fitted in the model. These trajectories were found using the score with the largest percentage of individuals with the score at each time point within each cluster. The trajectories calculated for the 4-cluster model showed no change over time from 3 months to 12 months, as each cluster had a trajectory that was a straight line throughout the follow-up.

To summarise, when considering the full SO_2S data set and the changes in the mRS score over time, these can be described by a 4-cluster model. Patients in each of the clusters defined had very little change over time and any variations seen within the scores of each cluster at each time point were not large enough to change the recovery profiles from straight lines throughout the follow-up period. In order to try and find a model that helped to represent the changes seen in those patients, it was decided to remove individuals who scored the same mRS value at all three of the follow-up time points to see if the model was able to capture the movement of individuals better.

7.4.2 Patients who undergo a transition within the 3 consecutive time points

When we considered the different recovery patterns over time, Table 7.1, patterns where patients who make no transition at all and score the same value at 3, 6 and 12 months were seen in 3,085 of the 6,203 patients, which is just under 50% of the data. Because of this, they were obviously going to dominate the clusters that each LCGA model creates. It was decided to remove these individuals and an LCGA was conducted in only individuals who did not score the same value at all 3 time points. In order to fit the model, the mRS score needed to be complete at all time points. Therefore, patients who had died at 3 months were kept in the analysis so the mRS score included the values between 0 and 6 at each time point. This meant that 2,452 individuals were removed from the 7,725 in the previous analysis, leaving 5,273 individuals; see Appendix F for flow diagram of participants. At each of the 3 consecutive time points, the distribution of the mRS is detailed in Table 7.7. This was the distribution seen in the 5,273 individuals that were included in the analysis looking at patients who did not have repeated scores at each time point:

Table 7.7: Distribution of mRS values at each follow-up point for SO_2S individuals with a transition

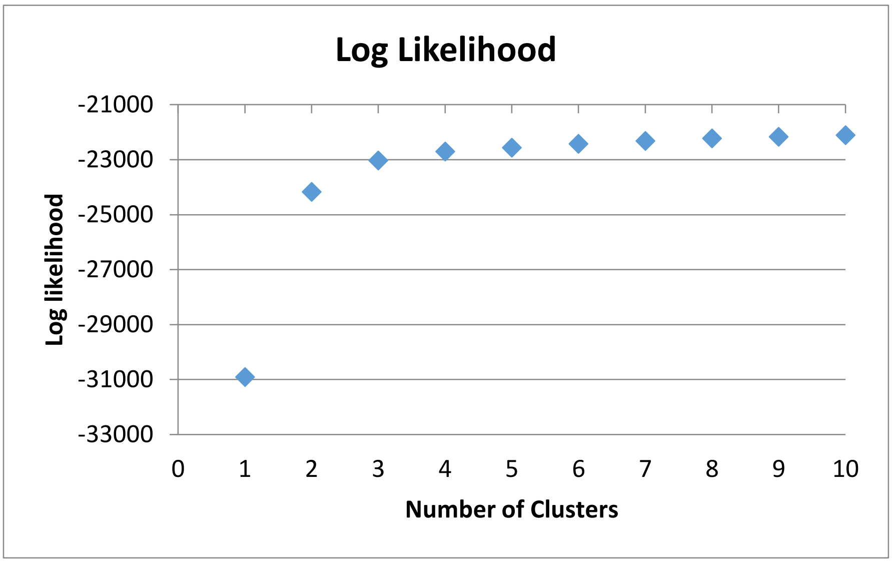
mRS Score	3 months (%)	6 months (%)	12 months (%)
0	9.5	11.4	11.1
1	23.5	20.7	19.9
2	14.9	16.1	14.2
3	18.4	17.4	15.3
4	15.3	13.6	13.7
5	6.0	4.0	3.2
6	12.4	16.8	22.6

A series of 1, 2 … 10-cluster LCGA models were fitted to the data and their fit indices are described in Table 7.8. 10 clusters were chosen to be fitted to the data in order to try and find as many different trajectories as possible in the clusters seen. The 3 IC indices all showed a decrease as one more clusters were added to the model, although the amount of the decrease became smaller as there were more clusters added to the model. This suggested that an optimal model may not be identifiable by using the IC indices alone. When we considered the likelihood ratio tests, we see that the BLRT indicates that each time a cluster was added to the model it provided a significant improvement in the fit of the model. However, the VLMR LRT showed that the 6-cluster model did not show a significant improvement on the model fit of the 5-cluster model, though the 7-cluster model showed an improvement on the 6-cluster model. We also saw no improved fit in the 8-cluster model compared to the 7-cluster model. The log-likelihood decreased when another cluster was added, and we see that similar to the IC indices, the decrease in log-likelihood lessens as more clusters are added. This occurs particularly in the 9 and 10-cluster models, as can be seen in Figure 7.3.

Table 7.8: Model fit statistics for LCGA models fitted to individuals with a transition in SO_2S data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	61814.54	61867.54	61841.68	-	-	-	-30899.3
2	48345.91	48411.62	48379.84	<0.001	<0.001	<0.001	-24163.0
3	46082.85	46168.26	46126.95	<0.001	<0.001	<0.001	-23028.4
4	45419.72	45524.85	45474.00	<0.001	<0.001	<0.001	-22693.9
5	45143.24	45268.07	45207.70	<0.001	<0.001	<0.001	-22552.6
6	44880.88	45025.43	44955.52	0.4268	0.4268	<0.001	-22418.4
7	44672.01	44836.27	44756.82	<0.001	<0.001	<0.001	-22311.0
8	44497.32	44681.29	44592.32	0.3089	0.3089	<0.001	-22220.7
9	44373.27	44576.96	44478.45	0.0049	0.0054	<0.001	-22155.6
10	44265.82	44489.21	44381.17	<0.001	<0.001	<0.001	-22098.9

Figure 7.3: Graph of the values of the log-likelihood for 10 LCGA models fitted to SO_2S data when an individual undergoes at least one transition in the follow-up



The IC indices and LMR tests suggested that a model with a higher number of clusters could be the best model; however, it is important to consider the number of individuals in each cluster of the model. When this was considered this, the models with 7, 8, 9 and 10- clusters did not satisfy this criterion.

The model with the highest number of clusters that had over 5% of the data in each category was the 6-cluster model. However, this model had not improved the fit of the model compared to the 5-cluster model. The cluster-specific proportions and the average probabilities of the cluster membership of the 5-cluster model are shown in Table 7.9.

Table 7.9: Posterior probabilities and number in each of the assigned clusters for all patients with a transition in SO_2S data

			Posterior Probabilities				
Assigned Cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**
1	822	16	**0.972**	0	0.001	0.027	0
2	2,088	40	0	**0.79**	0.116	0.001	0.093
3	1,057	20	0.001	0.147	**0.758**	0.093	0.001
4	419	8	0.06	0.014	0.115	**0.811**	0.001
5	887	17	0	0.12	0.001	0	**0.879**

The distribution of the individuals was fairly well spread between each cluster. We saw that the largest cluster did not include the majority of individuals (n=2,088, 40%), and the smallest cluster was not too small in comparison (n=419, 8%). As well as this, looking at the AvPPs for each cluster we see that the small cluster has a high AvPP assignment (0.811). In addition, all clusters had good AvPPs of assignment ranging from 0.758 to 0.972.

Finally, the 5 derived cluster-specific trajectories in the 5-cluster model were evaluated. Table 7.10 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.10: Conditional probabilities for each of the assigned clusters for patients with transition in SO$_2$S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0	0	0	**0**	0.06	0.058	0.054
1	0	0	0	**1**	0.337	0.33	0.315
2	0	0	0	**2**	0.296	0.297	0.298
3	0.001	0	0	**3**	0.228	0.233	0.244
4	0.015	0	0	**4**	0.073	0.075	0.081
5	0.076	0	0	**5**	0.006	0.006	0.006
6	0.907	1	1	**6**	0.001	0.001	0.001
Cluster 3				**Cluster 4**			
0	0.007	0.007	0.006	**0**	0.001	0	0
1	0.058	0.057	0.054	**1**	0.007	0.004	0.001
2	0.127	0.125	0.12	**2**	0.019	0.01	0.002
3	0.358	0.355	0.35	**3**	0.098	0.054	0.014
4	0.389	0.394	0.403	**4**	0.516	0.408	0.166
5	0.051	0.052	0.054	**5**	0.274	0.37	0.392
6	0.011	0.011	0.012	**6**	0.085	0.155	0.424
Cluster 5							
0	0.418	0.446	0.504				
1	0.462	0.446	0.408				
2	0.081	0.074	0.06				
3	0.03	0.027	0.022				
4	0.007	0.006	0.005				
5	0.001	0	0				
6	0	0	0				

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.11 to give an indication of the profile of each cluster over time. There was some variation seen in the model. However, the clusters still seemed to be fairly rigid in the sense that there still were not many transitions seen throughout the follow-up in each cluster fitted in the model.

Table 7.11: Characteristics of the 5-cluster model fitted to all individuals with transition in SO_2S data

Assigned cluster	n	%	Characteristics
1	822	16	Severe disability & death
2	2088	40	Low to moderate throughout
3	1057	20	Moderate to moderate-severe throughout
4	419	8	Moderate-severe worsening to severe and death
5	887	17	Very little disability with slight recovery to no disability

Figure 7.4: Graph showing the trajectories the individuals in each cluster, for all patients in SO_2S trial who undergo a transition within the follow-up

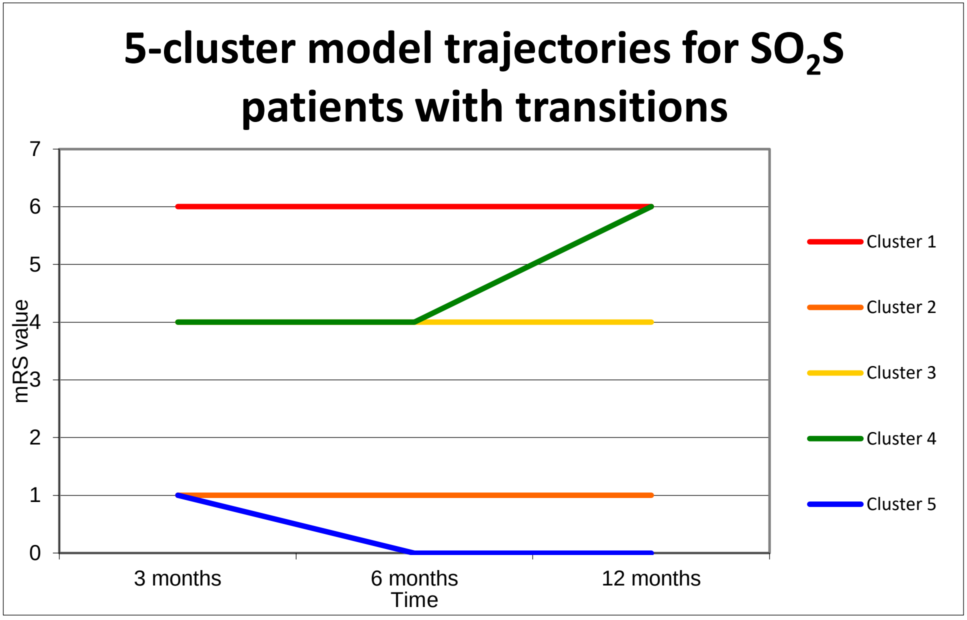


Figure 7.4 shows the model trajectories for the 5 clusters that were fitted in the model. There was more movement in the trajectories now that the patients who had no change at all over the follow-up period had been removed. The trajectories showed 3 clusters that, like before,

showed no change in the largest percentage of individuals over time. There was a profile showing recovery, cluster 5, when individuals moved from slight disability to no disability, and also a profile that showed individuals who worsened over time, cluster 4, where individuals moved two scores over time, so at 3 and 6 months, an mRS score of 4 had the largest percentage of patients in the cluster and by 12 months it had changed to a score of 6.

To summarise, when considering the SO_2S data set using only those patients who had experienced a transition throughout the follow-up period, and the changes in the mRS score over time, these can be described by a 5-cluster model. There was more movement seen within the clusters that were defined in this model, as could be seen by the cluster descriptions and the recovery trajectories predicted from the model.

The aim of this thesis was to look at the effect of treatment on the mRS and to look at different methods and the conclusions that they draw. Next we considered how the distribution of mRS scores varies between the standard care group and the oxygen treatment group, see Table 7.12.

Table 7.12: Distribution of mRS scores for patients with a transition in the SO_2S data at each follow-up point stratified by treatment

	Standard Care			**Oxygen Treatment**		
mRS score	**3 months (%)**	**6 months (%)**	**12 months (%)**	**3 months (%)**	**6 months (%)**	**12 months (%)**
0	9.0	10.7	10.7	9.8	11.8	11.3
1	23.8	21.8	19.6	23.3	20.2	20.1
2	15.1	16.0	13.6	14.8	16.1	14.5
3	17.5	16.8	16.1	18.8	17.7	14.9
4	16.6	14.2	14.2	14.6	13.3	13.4
5	6.5	4.0	3.4	5.8	4.0	3.1
6	11.5	16.5	22.5	12.8	17.0	22.7

Looking at the percentage distribution of the 5,273 individuals that were included in this analysis, it can be seen that there was very little difference to be seen in the distribution of individuals on each score in the oxygen treatment and standard care groups. There were slightly more moderate-severe (mRS=4) and severely disabled (mRS=5) individuals in the standard care group at 3 months, although the distribution was more even at 12 months between the 2 groups. Looking at those who could be classed as independent (mRS = 0, 1 and 2), although they start off evenly distributed across the oxygen treatment and standard care groups, by 12 months the oxygen treatment group seems to have slightly more of these individuals than the standard care group.

7.4.2.1 Standard care group

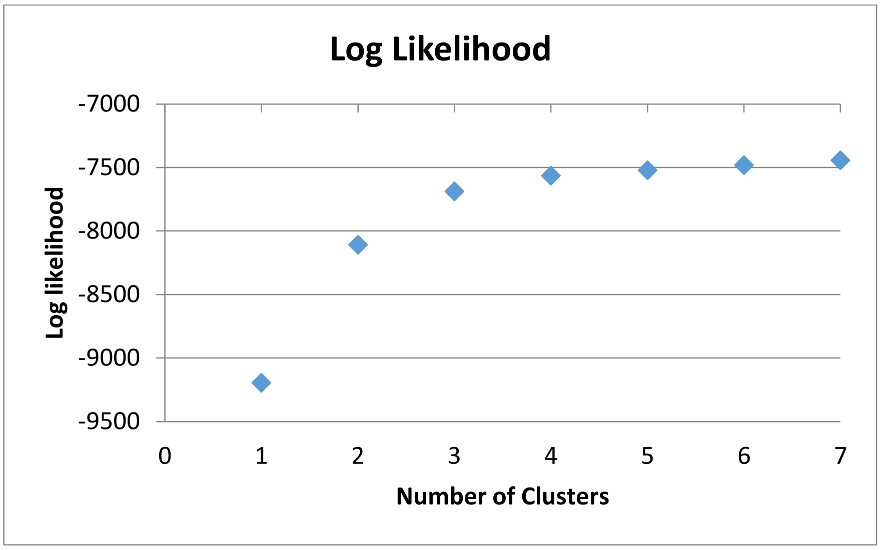
First, we considered the individuals who only received standard care. This was 1,760 of the 5,273 individuals who were fitted in the previous model, approximately 1/3 of the data. A new model was fitted to just the individuals who received standard care.

A series of 1, 2 … 7-cluster models were fitted to the data; after this, the maximum likelihood of the models could not be replicated, Table 7.13. Looking at the IC indices, it can be seen that they indicate that the larger the number of clusters in the model, the better the fit of the model, and additionally the log-likelihood is decreasing each time a new cluster is added. Once again, Figure 7.5 shows that the change in log-likelihood appears to be slowing down toward the 7-cluster model. The BLRT indicates that each additional cluster improves the fit of the model; however, the 5 and 6-cluster models appear not to have any improvement of fit on the 4 and 5-cluster models respectively.

Table 7.13: Model fit statistics for LCGA models fitted to standard care group with transition for SO_2S data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	18404.99	18443.3	18421.06				-9195.49
2	16237.83	16292.56	16260.79	<0.001	<0.001	<0.001	-8108.92
3	15404.35	15475.5	15434.2	<0.001	<0.001	<0.001	-7689.17
4	15162.11	15249.68	15198.85	<0.001	<0.001	<0.001	-7565.05
5	15082.36	15186.34	15125.98	1	1	<0.001	-7522.18
6	15007.78	15128.18	15058.29	0.5429	0.5429	<0.001	-7481.89
7	14941.77	15078.59	14999.17	<0.001	<0.001	<0.001	-7445.88

Figure 7.5: Graph of the values of the log-likelihood for 7 LCGA models fitted to standard care group with transition for SO_2S data



From the above IC indices and tests, it appears that the more clusters in the model the better the fit of the data. However, the numbers of individuals in each cluster needed to be examined as each cluster in the model must meet the minimum cluster size. Unfortunately, the 7-cluster model contained a cluster with only 21 individuals (1%) and therefore does not meet the minimum requirement. Although the 5- and 6-cluster models met the requirement for the number of individuals in each category, the VLMR LRT indicates that they showed no improvement on the fit of the previous models, and the last model that did so was the 4-cluster model. The 4-cluster model satisfied the criteria for the minimum numbers of individuals in each cluster, and each cluster has a distinct trajectory. The cluster-specific proportions and the average probabilities of the cluster membership of the 4-cluster model are shown in Table 7.14.

Table 7.14: Posterior probabilities and number in each of the assigned clusters for standard care group patients with a transition in SO_2S data

Assigned Cluster	n	%	Posterior Probabilities			
			1	2	3	4
1	314	18	**0.962**	0.036	0	0.002
2	418	24	0.015	**0.868**	0	0.116
3	312	18	0	0	**0.888**	0.112
4	716	41	0	0.08	0.065	**0.855**

Looking at Table 7.14, the largest cluster was cluster 4 (n=314, 41%), which contained 41% of the individuals in the analysis. This largest cluster did not contain a majority of the data. The remaining clusters were evenly spread with the rest of the individuals (Cluster 1: n=314, 18%, Cluster 2: n=418, 24%, Cluster 3: n=312, 18%). The AvPPs were all larger than 0.86 and indicated that there is a good probability that individuals that belong in each cluster have been assigned to the correct cluster. The 4 derived cluster-specific trajectories in the 4-cluster model were evaluated. Table 7.15 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.15: Conditional probabilities for each of the assigned clusters for standard care group who transition in SO_2S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0	0	0	**0**	0.003	0.003	0.002
1	0	0	0	**1**	0.034	0.029	0.021
2	0.001	0	0	**2**	0.081	0.07	0.052
3	0.007	0.001	0	**3**	0.283	0.26	0.212
4	0.057	0.008	0	**4**	0.448	0.464	0.484
5	0.183	0.033	0.001	**5**	0.115	0.131	0.169
6	0.751	0.958	0.999	**6**	0.037	0.043	0.059
Cluster 3				**Cluster 4**			
0	0.424	0.445	0.49	**0**	0.045	0.042	0.038
1	0.478	0.464	0.433	**1**	0.325	0.315	0.294
2	0.067	0.062	0.053	**2**	0.3	0.3	0.3
3	0.024	0.022	0.019	**3**	0.241	0.248	0.264
4	0.006	0.005	0.004	**4**	0.078	0.082	0.091
5	0.001	0.001	0	**5**	0.009	0.01	0.011
6	0	0	0	**6**	0.003	0.003	0.003

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.16 to give an indication of the profile of each cluster over time. These clusters appeared to be slightly less rigid than those seen in previous models fitted, as the clusters represent individuals at different stages of disability, with some change seen in the conditional probabilities of the scores during the follow-up time period.

Table 7.16: Characteristics of the 4-cluster model fitted to individuals in the standard care group with transition in SO_2S data

Assigned Cluster	n	%	Characteristics
1	314	18	Severe disability to death
2	418	24	Moderate – severe disability with some worsening but little change
3	312	18	No disability with improvement
4	716	41	Slight to moderate disability with some worsening

Figure 7.6: Graph showing the trajectories the individuals in each cluster, for all patients in standard care group with transition for SO_2S trial

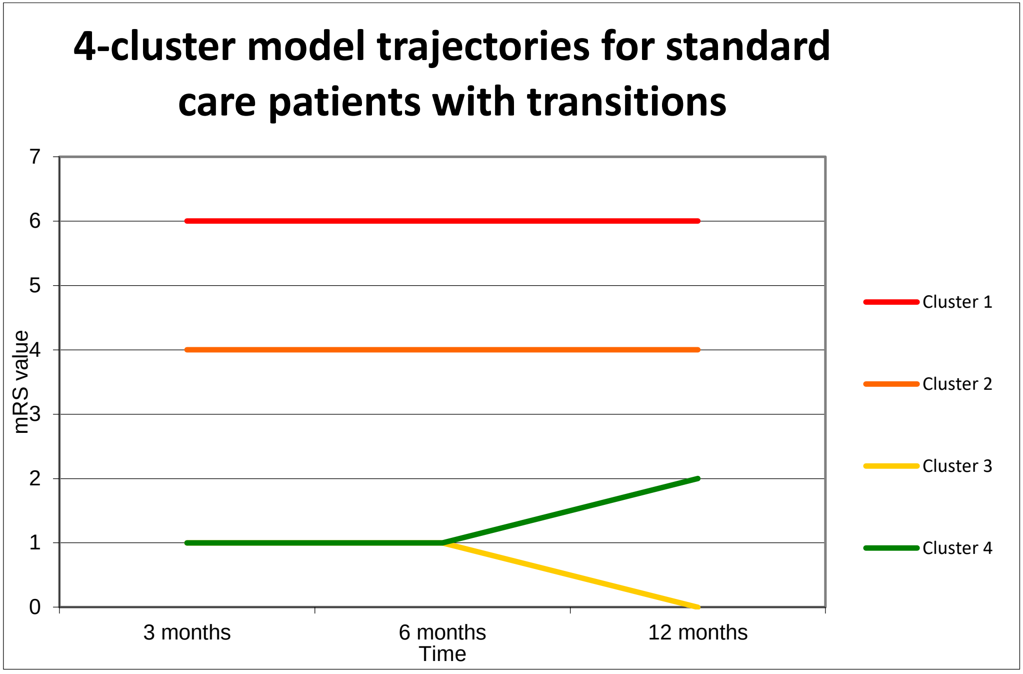


Figure 7.6 shows the model trajectories for the 4 clusters that were fitted in the model. For those individuals who received only standard care, it could be seen that the recovery trajectories calculated from the model were quite different from the trajectories in those models including all patients with a transition. There was still a trajectory with recovery from slight disability to no disability but the movement occurs later, between 6 and 12 months rather than 3 and 6 months. There were also still 2 clusters where a large percentage of individuals in that cluster remain at the same mRS score throughout follow-up. There was also a cluster with a worsening trajectory, cluster 4. These individuals had slightly more disability by the end of follow-up than at the start of follow-up, with the majority of people scoring 1 at 3 and 6 months and a mRS score of 2 at 12 months.

To summarise, when considering the SO_2S data set using individuals who received only standard care, and only those who experienced a transition throughout the follow-up period, the changes in the mRS score over time can be described by a 4-cluster model. There was some movement of individuals over the follow-up period but these changes occurred in individuals who were less disabled at the start of the follow-up period, with clusters that contained individuals who were more disabled showing less movement with constant mRS values over time.

7.4.2.2 Oxygen treatment group

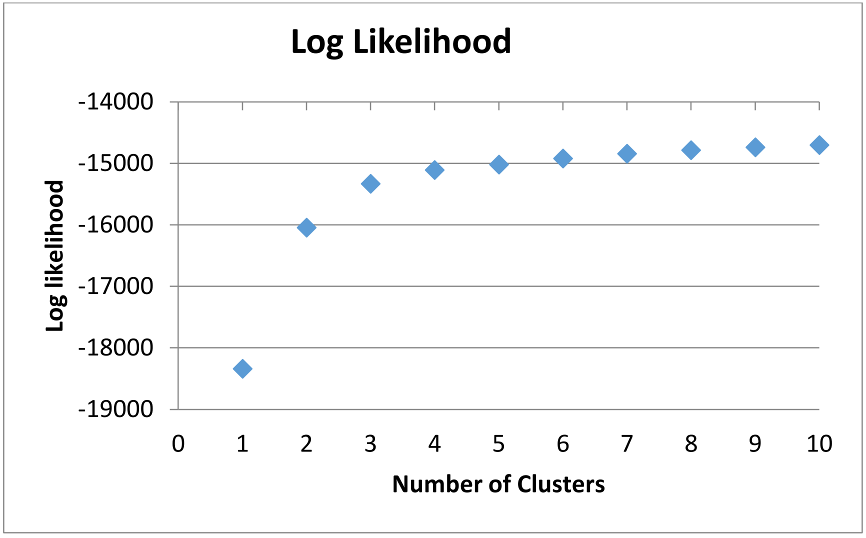
After considering the standard care group, attention was turned to those individuals who received the oxygen treatment as part of the study. Out of the 5,273 individuals included in first analysis, 3,513 received the oxygen treatment and were included in the LCGA.

A series of 1, 2 … 10-cluster models were fitted to the data as there were no issues with the maximum likelihood being replicated as in the standard care group, Table 7.17. The 10-cluster model had the lowest values for each of the IC indices, suggesting that it had the best goodness of fit out of all the models fitted to the data. The log-likelihood was also lowest for the 10-cluster model, and Figure 7.7 shows the plateauing of the log-likelihood. Once again the BLRT showed that every model with an additional cluster had an improvement in fit on the previous model, as does the VLMR LRT and adjusted LMR LRT.

Table 7.17: Model fit statistics for LCGA models fitted to treatment group with transition in SO_2S data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	36697.1	36740.25	36718.01				-18341.6
2	32117.89	32179.53	32147.76	<0.001	<0.001	<0.001	-16048.9
3	30691.29	30771.43	30730.12	<0.001	<0.001	<0.001	-15332.6
4	30255.76	30354.39	30303.55	<0.001	<0.001	<0.001	-15111.9
5	30078.86	30195.98	30135.61	<0.001	<0.001	<0.001	-15020.4
6	29892.12	30027.73	29957.83	<0.001	<0.001	<0.001	-14924.1
7	29743.96	29898.06	29818.63	<0.001	<0.001	<0.001	-14847.0
8	29625.63	29798.23	29709.26	<0.001	<0.001	<0.001	-14784.8
9	29544.7	29735.79	29637.28	0.0384	0.0408	<0.001	-14741.3
10	29473.67	29683.25	29575.22	<0.001	<0.001	<0.001	-14702.8

Figure 7.7: Graph of the values of the log-likelihood for 10 LCGA models fitted to oxygen treatment group for SO_2S data



Looking at the numbers of individuals in each cluster, models with 7 clusters or more had clusters with less than 5% of individuals assigned to them. The 6-cluster model was the largest model where the minimum number of individuals in a cluster was met. The cluster-specific proportions and the average probabilities of the cluster membership of the 6-cluster model are shown in Table 7.18.

Table 7.18: Posterior probabilities and number in each of the assigned clusters for treatment group patients with transitions in SO_2S data

			Posterior Probabilities					
Assigned Cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**	**6**
1	543	15	**0.981**	0.018	0	0	0	0
2	275	8	0.061	**0.819**	0	0	0.108	0.011
3	234	7	0	0.001	**0.841**	0.078	0.002	0.077
4	376	11	0	0	0.069	**0.836**	0.001	0.093
5	695	20	0	0.088	0	0.001	**0.757**	0.153
6	1389	40	0	0.001	0.017	0.048	0.116	**0.818**

The largest cluster in the model was cluster 6 (n=1389, 40%). This cluster was twice as big as the next largest cluster and almost 6 times larger than the smallest cluster; however, this was not a cause for concern as it still did not contain the majority of individuals in the analysis. The smallest two clusters had only 234 and 275 patients in each of them, approximately 7% and 8% respectively. The AvPPs indicated that generally, an individual who belonged to a specific cluster was likely to have been assigned to it. Most of the AvPPs were about 0.8, which was larger than predefined value of 0.7. The 6 derived cluster-specific trajectories in the 6-cluster model were evaluated. Table 7.19 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.19: Conditional probabilities for each of the assigned clusters for treatment group with transition in SO_2S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0	0	0	**0**	0	0	0
1	0	0	0	**1**	0.007	0.004	0.001
2	0	0	0	**2**	0.018	0.009	0.002
3	0.001	0	0	**3**	0.094	0.051	0.013
4	0.011	0	0	**4**	0.507	0.395	0.156
5	0.058	0	0	**5**	0.285	0.381	0.39
6	0.93	1	1	**6**	0.088	0.16	0.437
Cluster 3				**Cluster 4**			
0	0.879	0.624	0.073	**0**	0.201	0.421	0.866
1	0.112	0.338	0.475	**1**	0.593	0.496	0.124
2	0.006	0.026	0.257	**2**	0.135	0.057	0.007
3	0.002	0.009	0.151	**3**	0.056	0.021	0.002
4	0	0.002	0.041	**4**	0.013	0.005	0.001
5	0	0	0.003	**5**	0.001	0	0
6	0	0	0.001	**6**	0	0	0
Cluster 5				**Cluster 6**			
0	0.005	0.004	0.004	**0**	0.039	0.04	0.041
1	0.061	0.06	0.058	**1**	0.347	0.35	0.357
2	0.127	0.125	0.121	**2**	0.295	0.295	0.294
3	0.36	0.359	0.355	**3**	0.236	0.234	0.229
4	0.386	0.39	0.397	**4**	0.075	0.074	0.072
5	0.051	0.052	0.054	**5**	0.006	0.006	0.006
6	0.01	0.011	0.011	**6**	0.001	0.001	0.001

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.20 to give an indication of the profile of each cluster over time. These clusters were split slightly to include some clusters that were rigid with no movement with them, and others in which individuals definitely moved amongst scores over the follow-up of the study.

Table 7.20: Characteristics of the 6-cluster model fitted to individuals in the treatment group with transition in SO_2S data

Assigned Cluster	**N**	**%**	**Characteristics**
1	543	15	Severe disability & death
2	275	8	Moderate to severe disability worsening to severe death
3	234	7	No to slight disability worsening to slight to moderate disability
4	376	11	Slight disability with recovery to no disability
5	695	20	Moderate disability with no change throughout
6	1389	40	Slight disability with no change throughout

Figure 7.8: Graph showing the trajectories the individuals in each cluster, for all patients in to oxygen treatment group for SO_2S data

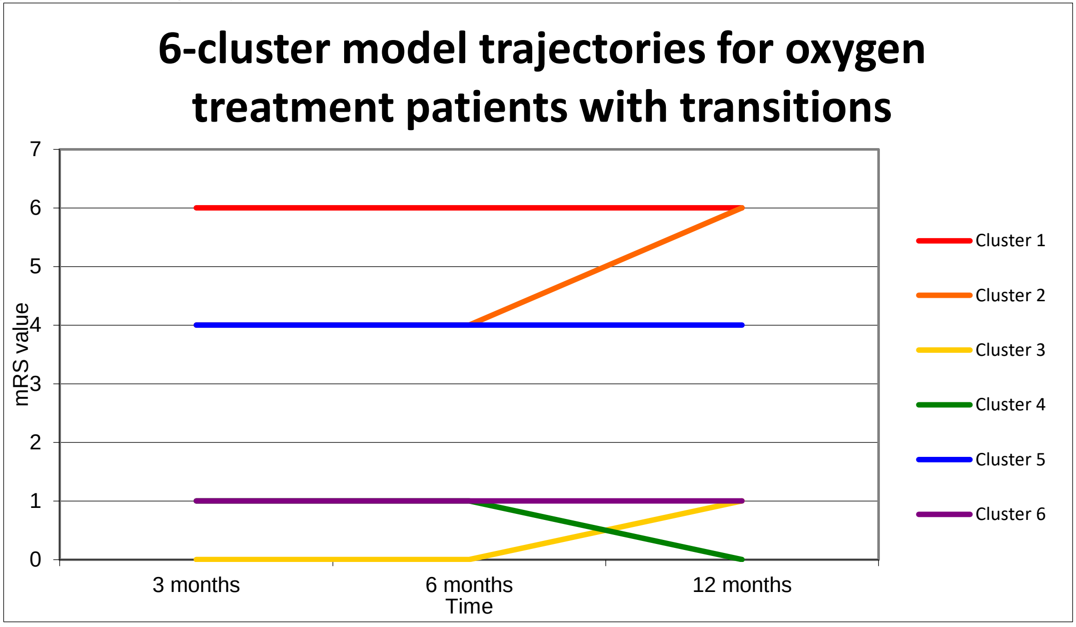


Figure 7.8 shows the model trajectories for the 5 clusters that were fitted in the model. With more clusters fitted in the model it can be seen that there was more movement in the scores

that had the largest percentage of individuals within each cluster. There were 3 clusters that contained no movement on the score with the largest percentage of individuals over time. There were 2 worsening profiles and 1 recovering one. The recovering one was the same as for the standard care group where patients moved from slight disability to no disability over time. One of the worsening profiles was seen before the individuals were split into treatment groups, where the largest percentage of patients scored a mRS of 4 at 3 and 6 months and had died at 12 months. The second profile were patients who had no disability at 3 and 6 months, but by 12 months they were experiencing some form of disability. There were most likely other underlying causes for this disability shown.

To summarise the model fitted to individuals who received oxygen treatment when considering the SO_2S data set using only those patients who experienced a transition throughout the follow-up period, the changes in the mRS score over time can be described by a 6-cluster model. Having 6 clusters allowed more movement of the cluster, as previously fitted clusters with no movement over time could be separated into clusters that may have a change in the trajectory rather than a constant line.

7.4.2.3 Comparison of standard care and oxygen treatment groups

The model fitted to the standard care group showed that the best model that was fitted was the 4-cluster model. There were 1,760 individuals split into 4 clusters and each cluster had a distinct trajectory over the 12-month follow-up period. The model fitted to the treatment group shows that the best model that was fitted was the 6-cluster model. There were 3,513 individuals split into 6 clusters in the model all took distinct trajectories over the 12 months follow-up period.

Both models contained the cluster in which individuals were severely disabled and died during the follow-up period, the majority making that transition between 3 and 6 months. These clusters were of similar sizes in both models with 18% of individuals in the model fitted to the standard care group and 15% of individuals in the oxygen treatment group. There were 2 other clusters that feature in both models – the first was the cluster where individuals have little or no disability and recovered and the second was the cluster who had a slight or moderate disability that worsened to have a moderate disability by the 12-month assessment.

The 4-cluster model had a cluster that individuals changed very little over time in, like two of the clusters in the treatment model, however the initial severity in the 4-cluster model was higher than the 2 clusters seen in the 6-cluster model. There were definitely differences seen between the clusters fitted to each of the stratified treatment groups, suggesting that there may be different recovery trajectories depending on which treatment group you are in.

By looking at the model fitted to all the data, it can be seen whether these distinct clusters in each model were still seen, or whether when the data are combined these distinct trajectories were lost. The model fitted to all the data with no distinction between treatment groups showed that the best model that was fitted was the 5-cluster model. There were distinct trajectories for the 12-month follow-up with the 5,273 individuals assigned to one of the 5 distinct clusters in the model.

The model containing all individuals had two clusters that matched up with ones that had previously been fitted. The clusters where individuals were severely disabled at 3 months and died by the end of the follow-up and the cluster where individuals had no or little disability at 3 months and recovered to show no disability at all at 12 months. These clusters represent individuals at the opposite ends of the mRS scale.

Surprisingly, there was no cluster that had individuals with a slight disability moving to moderate disability over time, which had been seen in both the 4 and 5-cluster models previously fitted. The remaining clusters were a mixture of those seen in the other models fitted. There was a cluster that had moderate-severe disability throughout the follow-up like in the 4-cluster model and one where individuals have low to moderate disability throughout the follow-up like in the 5-cluster model.

7.4.3 Survivor only model

One of the issues with fitting the models that had individuals removed if they made no transition over the follow-up period was that individuals who had died at 3 months had to be included as otherwise the models would have outcomes of different sizes at different follow-up time points that the software could not compute, as the outcome needed to have the same number of categories at each follow-up point. One way to overcome this was to look at the recovery in individuals who had survived the whole follow-up time period. Next, models were fitted that removed any individual who had died at any stage in the follow-up. The first included all patients in the study, so long as they did not die during follow-up, and this included 6,679 out of the original 8,003 individuals. The second looked at individuals who did not die in the follow-up time who had some transition between two categories on the scale at some point, this model included 4,227 out of the original 8,003 individuals, see Appendix F for flow diagram of participants.

7.4.3.1 Survivor only – all patients

The first survivor model fitted included at all patients within the study. As before, the distribution of the percentages of patients with each score on the mRS was considered. The table below, Table 7.21, details the distribution at the 3 consecutive time points. We see that the number of individuals in each category on the mRS remained fairly stable across the time points. Obviously, because those individuals who had died have been completely removed there is no 7th category on the scale.

Table 7.21: Distribution of survivors on the mRS over the 3 follow-up points in SO_2S data

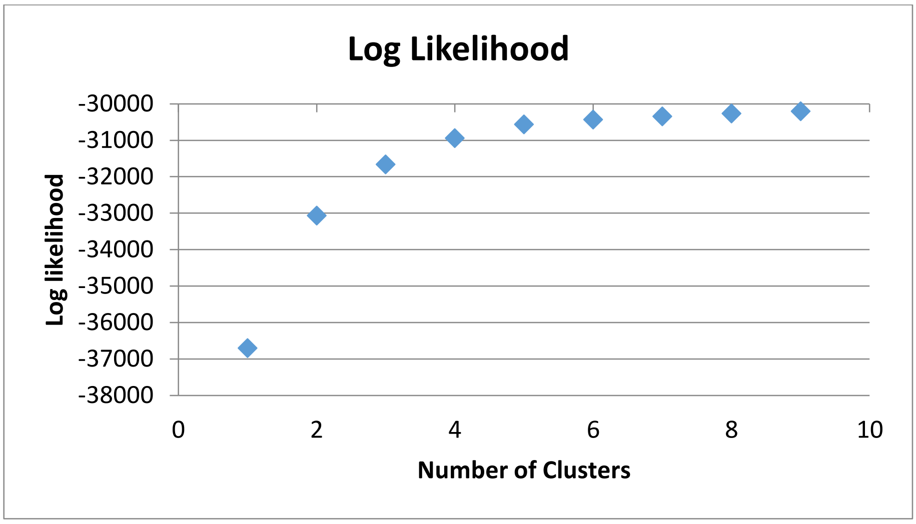
mRS score	3 months (%)	6 months (%)	12 months (%)
0	13.9	15.7	16
1	30.5	29.1	29.7
2	14.2	15.3	14.2
3	18.8	18.6	17.6
4	16.8	16.2	17.3
5	5.8	5	5.2

1, 2 … 9-cluster LCGA models were fitted to the data and their fit indices are described in table 7.22. 10 clusters were chosen to be fitted to the data in order to try and find as many different trajectories in the clusters seen, however, the model with 10 clusters failed to converge on a likelihood. When we considered the likelihood ratio tests, we see that the BLRT indicates that each time a cluster is added to the model it provided a significant improvement in the fit of the model, as do the VLMR and the adjusted LMR. The log-likelihood also decreased for each model when another cluster was added, and we saw that similar to the IC indices the decrease in log-likelihood lessens as more clusters are added. This occurs particularly in the 8 and 10-cluster models, as can be seen in Figure 7.9.

Table 7.22: Model fit statistics for LCGA models fitted to survivors in SO_2S data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	73421.83	73469.48	73447.23				-36703.9
2	66162.78	66237.66	66202.7	<0.001	<0.001	<0.001	-33070.4
3	63347.36	63449.47	63401.8	<0.001	<0.001	<0.001	-31658.7
4	61952.53	62054.86	61994.48	<0.001	<0.001	<0.001	-30943.8
5	61173.88	61330.44	61257.35	<0.001	<0.001	<0.001	-30563.9
6	60919.74	61103.52	61014.72	<0.001	<0.001	<0.001	-30432.9
7	60759.99	60970.91	60872.4	<0.001	<0.001	<0.001	-30348.9
8	60596.59	60834.82	60723.6	<0.001	<0.001	<0.001	-30263.3
9	60498.7	60756.46	60632.22	<0.001	<0.001	<0.001	-30206.3

Figure 7.9: Graph of the values of the log-likelihood for 9 LCGA models fitted to survivors in SO_2S data



The IC indices and LMR tests suggested that models with a higher number of clusters could be the best model; however, it was important to consider the number of individuals in each cluster of the model. When we considered this, the models with 6, 7, 8 and 9 clusters did not satisfy this criterion. The model with the highest number of clusters that had over 5% of the data in each category was the 5-cluster model. This model had improved fit compared to the 4-cluster model. The cluster-specific proportions and the average probabilities of the cluster membership of the 5-cluster model are shown in Table 7.23.

Table 7.23: Posterior probabilities and number in each of the assigned clusters for survivors in SO_2S data

	Posterior Probabilities						
Assigned Cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**
1	2,168	32	**0.873**	0	0	0.069	0.058
2	811	12	0	**0.897**	0	0	0.103
3	322	5	0.001	0	**0.943**	0.056	0
4	933	14	0.082	0	0.029	**0.888**	0.001
5	2,446	37	0.099	0.037	0	0	**0.864**

There were two very large clusters, cluster 1 has 2,168 individuals (32%) and cluster 5 has 2,446 individuals (37%). The next 2 largest clusters were very similar in size to each other, cluster 2 has 811 individuals (12%) and cluster 4 has 933 individuals (14%). The final cluster was a lot smaller than the other clusters; cluster 3 has only 322 individuals, which was only 5% of the data. The AvPPs are all higher than 0.864. The 5 derived cluster-specific trajectories in the 5-cluster model were evaluated. Table 7.24 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.24: Conditional probabilities for each of the assigned clusters for all survivors in SO_2S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0.006	0.006	0.006	**0**	0.838	0.843	0.853
1	0.135	0.136	0.138	**1**	0.154	0.15	0.141
2	0.276	0.277	0.279	**2**	0.006	0.005	0.005
3	0.434	0.433	0.431	**3**	0.001	0.001	0.001
4	0.145	0.144	0.142	**4**	0	0	0
5	0.003	0.003	0.003	**5**	0	0	0
Cluster 3				**Cluster 4**			
0	0	0	0	**0**	0	0	0
1	0	0	0	**1**	0.006	0.005	0.005
2	0	0	0	**2**	0.019	0.019	0.018
3	0.002	0.002	0.004	**3**	0.145	0.142	0.136
4	0.099	0.123	0.189	**4**	0.754	0.755	0.758
5	0.899	0.874	0.807	**5**	0.076	0.078	0.082
Cluster 5							
0	0.124	0.129	0.139				
1	0.665	0.668	0.671				
2	0.153	0.148	0.139				
3	0.05	0.048	0.045				
4	0.007	0.007	0.007				
5	0	0	0				

The cluster specific characteristics for each cluster of the model are described in Table 7.25 to give an indication of the profile of each cluster over time. It can be seen that with the exception for cluster 3, where those individuals who were severely disabled have a small recovery, the clusters had individuals who made no transition and remained in the same state of disability throughout.

Table 7.25: Characteristics of the 5-cluster model fitted to all survivors in SO_2S data

Assigned Cluster	n	%	Characteristics
1	2,168	32	Moderate disability throughout
2	811	12	No disability throughout
3	322	5	Severe disability with recovery
4	933	14	Moderate to moderate-severe disability throughout
5	2,446	37	Slight to no disability throughout

Figure 7.10: Graph showing the trajectories the individuals in each cluster, for all survivors in SO_2S trial

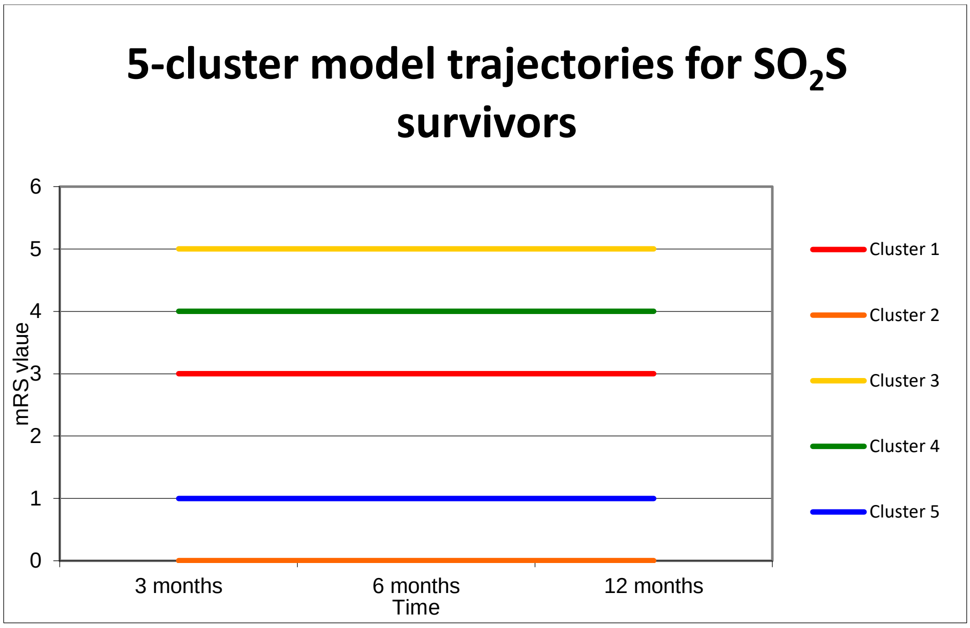


Figure 7.10 shows the model trajectories for the 5 clusters that were fitted in the model. It can be seen that for all clusters in the model the score with the largest percentage of individuals did not change at any time point throughout the follow-up, leading to constant recovery trajectories.

To summarise, when considering the SO_2S data set using all patients who survive the follow-up period, and the changes in the mRS score over time, then these can be described by a 5-cluster model. It can be seen from the model descriptions that with the exception of cluster 3, where individuals undergo some recovery over time, each cluster was fairly constant in the severity of disability throughout the follow-up period. Cluster 3, although there was movement present, still had a score of 5 as the score with the largest percentage of individuals at each time point. This was consistent with what has been seen before in previous models that we have fitted to the SO_2S using all individuals with no stratification for treatment. Once again individuals who did not make a transition were removed to see how this affected the clusters fitted to the survivors of the study.

7.4.3.2 Survivors who make a transition somewhere at one of the 3 consecutive time points

The second model, which looked only at patients who survived the whole 12 months follow-up, was fitted to individuals who had at least 1 change in the mRS score over the 12-month follow-up period. This was 4,227 individuals out of the original dataset. The percentage of individuals with each score on the mRS is given in the table below, Table 7.26, in order to look at the distribution of mRS scores over time. It could be seen that there was slightly more change in the percentages of individuals with each score over time, especially for the lower categories on the mRS.

Table 7.26: Distribution of survivors with transition on the mRS over the 3 follow-up points in SO_2S data

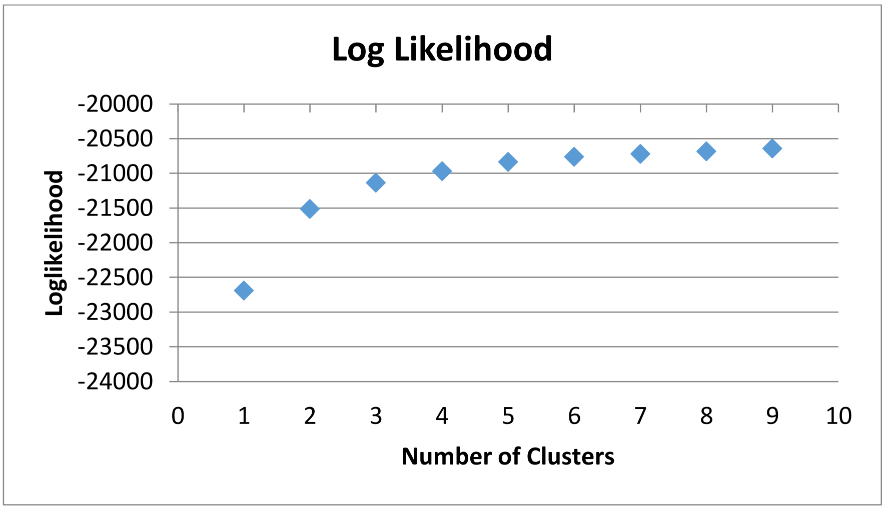
mRS score	3 months (%)	6 months (%)	12 months (%)
0	11.3	14.0	14.3
1	27.7	25.3	25.7
2	17.6	19.7	18.3
3	21.3	21.1	19.8
4	16.9	16.0	17.7
5	5.3	4.0	4.1

1, 2 … 9-cluster LCGA models were fitted to the data and their fit indices are described in Table 7.27. 10 clusters were chosen to be fitted to the data in order to try and find as many different trajectories in the clusters seen. As with the survivor only model that included all patients the model with 10 clusters failed to converge on the likelihood. The 3 IC indices all showed a decrease with an increase in the number of clusters being fitted in each model, although the amount of the decrease was getting smaller as there were more clusters added to the model, Figure 7.11. When we considered the likelihood ratio tests, we see that the BLRT indicates that each time a cluster is added to the model it provides a significant improvement in the fit of the model, as do the VLMR and the adjusted LMR.

Table 7.27: Model fit statistics for LCGA models fitted to survivors with transition in SO_2S data

Cluster	AIC	BIC	ad BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	45392.92	45437.36	45415.11				-22689.5
2	43051.75	43121.59	43086.63	<0.001	<0.001	<0.001	-21514.9
3	42302.37	42397.61	42349.94	<0.001	<0.001	<0.001	-21136.2
4	41976.43	42097.07	42036.7	<0.001	<0.001	<0.001	-20969.2
5	41724.62	41870.66	41797.57	<0.001	<0.001	<0.001	-20839.3
6	41578.93	41750.35	41664.56	<0.001	<0.001	<0.001	-20762.5
7	41501.52	41698.34	41599.84	<0.001	<0.001	<0.001	-20719.8
8	41438.09	411660.3	41549.1	<0.001	<0.001	<0.001	-20684
9	41360.93	41608.55	41484.62	<0.001	<0.001	<0.001	-20641.5

Figure 7.11: Graph of the values of the log-likelihood for 9 LCGA models fitted to survivors with a transition during follow-up SO_2S data



After considering the IC indices and the likelihood ratio tests we next considered the proportion of individuals in each cluster of the model, in order to find the best fitting model that has at least 5% of the data within each cluster of the fitted model. The 7, 8 and 9-cluster models all had at least one cluster where there was less than 5% of the individuals in the dataset within the cluster. The 6-cluster model was the largest cluster model that satisfied the criteria for the proportion of individuals in each cluster. This model also had an improved fit compared to the 5-cluster model when considering the IC indices and LRT results. The cluster-specific proportions and the average probabilities of the cluster membership of the 6-cluster model are shown in Table 7.28.

Table 7.28: Posterior probabilities and number in each of the assigned clusters for survivors who make a transition during follow-up in SO_2S data

	Posterior Probabilities							
Assigned Cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**	**6**
1	486	11	**0.771**	0.075	0.001	0.049	0.104	0
2	453	11	0.131	**0.756**	0	0.014	0.098	0
3	466	11	0.001	0	**0.893**	0	0.058	0.049
4	310	7	0.041	0.036	0	**0.922**	0.001	0
5	2073	49	0.046	0.048	0.042	0	**0.844**	0.02
6	439	10	0	0.001	0.11	0	0.094	**0.794**

The smallest cluster in the model was cluster 4, which only had 310 individuals in it, approximately 7% of the data. This was closely followed by cluster 6, which had 439 individuals, which was 10% of the data. Clusters 1, 2 and 3 all had 11% of the data in them. The largest cluster, cluster 5, contained 2,073 individuals, which was 49% of the data. Although there were no clusters that had the majority of the data in them, this was very close to containing 50% of the individuals in the dataset. All of the calculated AvPPs were greater than 0.7 and suggested the model was good at predicting cluster membership, although some of these values were much lower than have been seen in previous models.

Finally, the 6 derived cluster-specific trajectories from the 6-cluster model were evaluated. Table 7.29 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.29: Conditional probabilities for each of the assigned clusters for survivors with transition in SO_2S data

mRS scale	3 months	6 months	12 months	mRS scale	3 months	6 months	12 months
Cluster 1				**Cluster 2**			
0	0.001	0.002	0.011	**0**	0.009	0.003	0
1	0.013	0.029	0.148	**1**	0.126	0.045	0.005
2	0.033	0.073	0.247	**2**	0.226	0.107	0.013
3	0.226	0.366	0.433	**3**	0.451	0.428	0.104
4	0.681	0.51	0.157	**4**	0.184	0.404	0.761
5	0.047	0.02	0.004	**5**	0.004	0.013	0.117
6	0.001	0.002	0.011	**6**	0.009	0.003	0
Cluster 3				**Cluster 4**			
0	0.179	0.421	0.897	**0**	0	0	0
1	0.612	0.505	0.097	**1**	0	0.001	0.001
2	0.141	0.052	0.005	**2**	0.001	0.002	0.003
3	0.059	0.019	0.002	**3**	0.01	0.014	0.023
4	0.009	0.003	0	**4**	0.388	0.449	0.572
5	0	0	0	**5**	0.6	0.535	0.401
6	0.179	0.421	0.897	**6**	0	0	0
Cluster 5				**Cluster 6**			
0	0.034	0.034	0.035	**0**	0.83	0.586	0.1
1	0.344	0.347	0.353	**1**	0.159	0.375	0.559
2	0.309	0.309	0.308	**2**	0.008	0.028	0.216
3	0.257	0.255	0.25	**3**	0.003	0.01	0.107
4	0.055	0.055	0.053	**4**	0	0.001	0.018
5	0.001	0.001	0.001	**5**	0	0	0
6	0.034	0.034	0.035	**6**	0.83	0.586	0.1

The cluster specific characteristics for each cluster of the model are described in Table 7.30 to give an indication of the profile of each cluster over time. It could be seen from the table that there was a lot more movement in the model fitted to the survivors who have some change in the mRS score over the follow-up period, than any of the other models previously fitted to the data, with 4 of the 6 clusters showing individuals recovering or deteriorating over time.

Table 7.30: Characteristics of the 4-cluster model fitted to all survivors with transition in SO_2S data

Assigned Cluster	n	%	Characteristics
1	486	11	Moderate-severe and severe disability with recovery
2	453	11	Moderate disability with worsening over time
3	466	11	No to slight disability with recovery
4	310	7	Severe disability throughout
5	2073	49	Slight disability throughout follow-up
6	439	10	No disability with deterioration over time

Figure 7.12: Graph showing the trajectories the individuals in each cluster, for all survivors who have a transition during follow-up in SO_2S trial

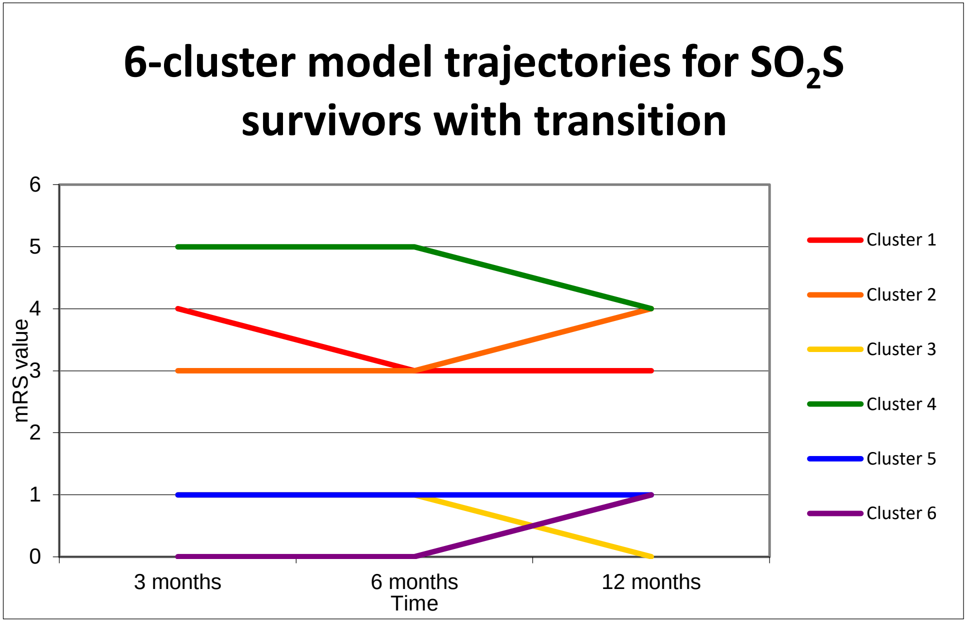


Figure 7.12 shows the model trajectories for the 6 clusters that were fitted in the model Only 1 cluster showed no change over time, cluster 5, those with no-slight disability. Recovery could be seen in 3 of the trajectories, cluster 1 showed individuals with score a 4 at 3 months were the largest group within that cluster, which decreased to a mRS score of 3 at 6 months. The other recovery clusters showed slightly later recovery between 6 and 12 months, with cluster 4 having the largest percentage of individuals with a score of 5 then a score of 4, and cluster 3 having the largest percentage of individuals scoring a 1 and then 0 with no disability by 12 months. Cluster 2 had most individuals with moderate disability at 3 months, which increased to a moderate-severe

disability at 12 months. The final cluster showed the individuals who have no disability at 3 months but slight disability at 12 months.

To summarise, when considering the changes in the SO_2S data set using those patients who survived the follow-up and experience a transition throughout the 12-month period, these can be described by a 6-cluster model. This model had the most transitions that have been seen in all the models fitted to the SO_2S data, with 5 of the 6 clusters fitted showing some form of movement over time whether that be recovery or a worsening in the score.

7.4.3.3 Comparison of models fitted to SO_2S survivor data

There were 2 different models fitted to the SO_2S data that only included individuals who survived the whole of the 12-month follow-up of the trial. The first was fitted to all patients in the trial and the second was fitted to only those individuals in the dataset who had a change in their mRS score over the 12-month follow-up.

The best model fitted to all the patients in the trial was a 5-cluster model and was fitted to 6,679 individuals. The best model fitted to the individuals who had a transition was a 6-cluster model and was fitted to 4,227 individuals. There were only 2 clusters that were the same between the two models fitted. The first was those individuals who had a moderate-severe or severe disability with recovery over time and the second was those who had slight disability throughout the follow-up. The other clusters in the model with all patients were clusters where individuals remained on the same scores throughout the follow-up. For the model fitted to patients who survived and made a transition, there was a lot more variation seen in the mRS scores, with people worsening or recovering over time rather than staying within the same mRS categories. The posterior probabilities for these clusters were, however, much lower than the posterior probabilities that have been seen in other fitted models, suggesting there was more chance of the model not classifying individuals into the correct clusters, although the values are not below the threshold that was deemed to be a good posterior probability.

A challenge of using the SO_2S data was the limited number of data points, therefore the NINDS dataset was used, which increased the number of follow-up points to 4.

7.5 Results using the NINDS data

Over the 4 follow-up time points the score that each individual received on the mRS was recorded. Each patient in the study had his or her own distinct recovery pattern. The different recovery patterns seen in the data were recorded in order to identify the change in mRS score over the 4 consecutive time points. There were 163 different patterns of mRS score over the 4 time points. Of these 163 patterns, there were 22 recovery trajectories that contained over 1% of the data, Table 7.31. Patterns were only considered if they had full mRS data at each time point, which meant 619 out of the 624 patients in the study were considered. The remaining 5 individuals had missing values during the follow-up.

Table 7.31: Recovery patterns that occur in more than 1% of individuals in NINDS dataset

Pattern	N	%	Pattern	N	%
0000	41	6.6	**4222**	6	1.0
0011	6	1.0	**4333**	25	4.0
0111	6	1.0	**4444**	21	3.4
1000	15	2.4	**4666**	14	2.2
1110	7	1.1	**5333**	7	1.1
1111	19	3.0	**5444**	16	2.6
2111	20	3.2	**5466**	6	1.0
3111	7	1.1	**5555**	14	2.2
3222	8	1.3	**5566**	9	1.4
3333	9	1.4	**5666**	60	9.6
4111	12	1.9	**6666**	38	6.1

The most frequently observed pattern in the data was an individual who scored 5666. These patients were severely disabled at 7 days and by the 3 month follow-up had died; approximately 10% of the individuals in the study fell into this trajectory over time. Other frequently observed trajectories included 0000 (6.6%), 4333 (4%), 4444 (3.4%), and 6666 (6.1%). Over the 4 time points individuals did tend to make a transition at one time point during the follow-up. The change was mainly seen between the 7-day and 3-month follow-ups, when a patient appears to be experiencing most of his or her recovering. Patients who made no transition at all were also frequently observed, a trajectory of 1111 occurring in 3% of patients, 3333 in 1.4%, and 5555 in 2.2% of patients.

A patient is described to have had some recovery between 7 days and 12 months if they have a lower score at 12 months than at 7 days, and a patient has deteriorated if the score at 12 months was higher than at 7 days then there are 6 trajectories show deterioration and 10 trajectories that show recovery in a patient.

7.5.1 All patients at 4 consecutive time points

At each of the 4 consecutive time points, the percentage distribution of the mRS score can be detailed, as seen in Table 7.32. These values are the percentage distributions of the individuals included in the LCGA, excluding those individuals who have missing values. Between 3 and 12 months the distribution of individuals was stable, however, there were many transitions between 7 days and 3 months as the distribution seen was very different between these 2 time points. These values included all the individuals in the data set, as the LCGA models were fitted to all 624 individuals, regardless of whether there were missing values.

Table 7.32: Distribution of individuals on the mRS over the 4 follow-up points in the NINDS data

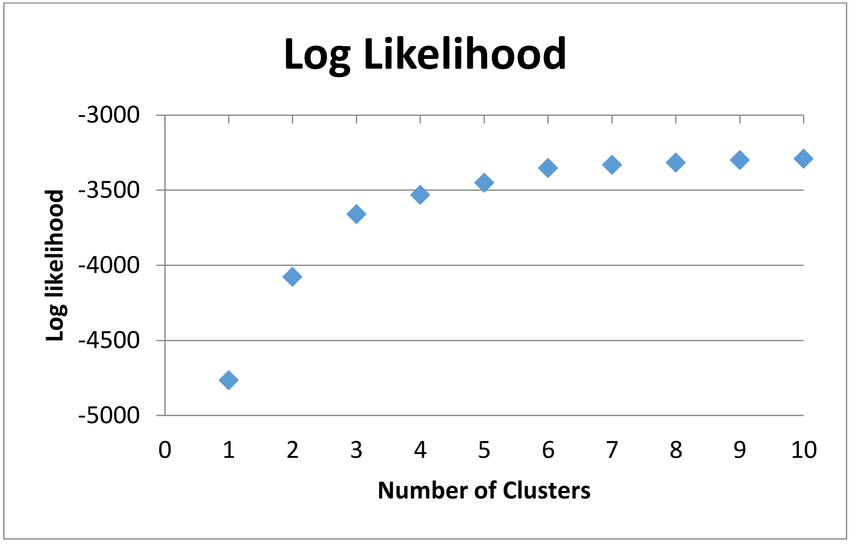
mRS score	7 days (%)	3 months (%)	6 months (%)	12 months (%)
0	11.6	14.4	14.7	14.6
1	13.4	20.2	20.0	18.6
2	9.0	9.8	9.3	9.3
3	9.4	13.6	14.7	12.3
4	25.5	16.7	11.5	8.5
5	24.9	6.4	5.1	5.1
6	6.1	18.9	24.5	31.6

A series of 1, 2 … 10-cluster models were fitted the data, see Table 7.33 for the results. The IC indices were decreasing as each extra class was added. The AIC and adjusted BIC were lowest in the 10-cluster model and the BIC was lowest in the 9-cluster model. We also see that the log-likelihood was reduced with each added cluster of the model. From Figure 7.13, the decreasing log-likelihood can be seen, and the amount that the likelihood was decreasing flattens off from the 8-cluster model onwards. When considering the LRTs, the BLRT suggested that there was an improvement in the fit of the model for every model as new clusters were added. The VLMR and adjusted LMR LRTs agree with this; however, the 10-cluster model had non-significant *p*-values suggesting that there was no increase in the fit of the 10-cluster model compared to the 9-cluster model.

Table 7.33: Model fit statistics for LCGA models fitted to NINDS data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	9541.653	9572.706	9550.482				-4763.83
2	8172.521	8216.883	8185.134	<0.001	<0.001	<0.001	-4076.26
3	7344.751	7402.421	7361.147	<0.001	<0.001	<0.001	-3659.38
4	7095.035	7116.013	7115.216	<0.001	<0.001	<0.001	-3531.52
5	6938.504	7022.79	6962.468	<0.001	<0.001	<0.001	-3450.25
6	6749.413	6847.008	6777.161	<0.001	<0.001	<0.001	-3352.71
7	6713.192	6824.096	6744.725	0.0254	0.0301	<0.001	-3331.6
8	6686.745	6810.957	6722.061	<0.001	<0.001	<0.001	-3315.37
9	6660.617	6798.138	6699.718	<0.001	<0.001	<0.001	-3299.31
10	6648.485	6799.314	6691.369	0.199	0.2096	<0.001	-3290.24

Figure 7.13: Graph of the values of the log-likelihood for 10 LCGA models fitted to NINDS data



The IC indices and the LRTs give an indication that having more clusters gave models that fitted the data better; however, each cluster must contain at least 5%. Looking at the numbers in each class, both the 6- and 7-cluster models contained at least 5% of the data in each cluster. However, the 5-cluster and the 8-, 9- and 10-cluster models had less than 5% of the data in at least one cluster of the model. The 7-cluster model was chosen as it was the model with the largest number of clusters that still satisfied the minimum number of individuals in each cluster. The cluster-specific proportions and the average probabilities of the cluster membership of the 7-cluster model are shown in Table 7.34.

Table 7.34: Posterior probabilities and number in each of the assigned clusters for all patients in NINDS data

			Posterior Probabilities						
Assigned Cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	34	5	**0.839**	0.055	0.029	0.056	0.009	0	0.011
2	136	22	0.006	**0.917**	0	0.038	0.038	0	0
3	140	22	0	0	**0.984**	0.016	0	0	0
4	105	17	0.002	0.072	0.017	**0.909**	0	0	0
5	106	17	0	0.033	0	0	**0.865**	0.002	0.1
6	70	11	0	0	0	0	0.015	**0.952**	0.033
7	33	5	0	0	0	0	0.094	0.041	**0.865**

The smallest cluster, cluster 7, had only 33 people in it, closely followed by cluster 1 with 34, and these clusters were both borderline as they contain just 5% of the data in each. However, there was not one cluster that contained a large proportion and the others are fairly similar, with clusters 2, and 3 containing 136 and 140 individuals respectively, which were approximately 22% of the data in each cluster. The AvPP assignments were all high, ranging from 0.84 to 0.98, suggesting that individuals in each cluster were highly likely to be assigned that cluster.

Finally, the 7 derived cluster-specific trajectories in the 7-cluster model were evaluated. Table 7.35 displays the item conditional probability of each level of the mRS ordinal scale.

Table 7.35: Conditional probabilities for each of the assigned clusters for all patients in NINDS data

mRS scale	7-10 days	3 months	6 months	12 Months	mRS scale	7-10 days	3 months	6 months	12 months
Cluster 1					**Cluster 2**				
0	0.369	0.033	0.002	0	**0**	0.001	0.001	0.002	0.006
1	0.588	0.531	0.055	0	**1**	0.033	0.05	0.076	0.175
2	0.035	0.3	0.171	0	**2**	0.113	0.158	0.216	0.338
3	0.008	0.111	0.413	0.003	**3**	0.363	0.405	0.421	0.348
4	0.001	0.023	0.306	0.028	**4**	0.402	0.327	0.246	0.118
5	0	0.002	0.045	0.159	**5**	0.075	0.05	0.033	0.013
6	0	0	0.008	0.809	**6**	0.013	0.009	0.005	0.002
Cluster 3					**Cluster 4**				
0	0	0	0	0	**0**	0	0	0	0
1	0.001	0	0	0	**1**	0.005	0.005	0.004	0.003
2	0.004	0	0	0	**2**	0.021	0.018	0.016	0.013
3	0.026	0.002	0	0	**3**	0.112	0.103	0.093	0.075
4	0.212	0.022	0.001	0	**4**	0.478	0.464	0.448	0.409
5	0.455	0.129	0.01	0	**5**	0.304	0.322	0.341	0.378
6	0.301	0.846	0.989	1	**6**	0.08	0.088	0.098	0.122
Cluster 5					**Cluster 6**				
0	0.019	0.037	0.074	0.262	**0**	0.641	0.896	0.979	0.999
1	0.406	0.554	0.676	0.668	**1**	0.344	0.101	0.02	0.001
2	0.358	0.285	0.187	0.055	**2**	0.012	0.002	0	0
3	0.172	0.101	0.053	0.013	**3**	0.003	0.001	0	0
4	0.039	0.02	0.01	0.002	**4**	0	0	0	0
5	0.004	0.002	0.001	0	**5**	0	0	0	0
6	0.001	0	0	0	**6**	0	0	0	0
Cluster 7									
0	0.475	0.393	0.31	0.175					
1	0.496	0.568	0.634	0.713					
2	0.023	0.031	0.044	0.087					
3	0.005	0.007	0.01	0.021					
4	0.001	0.001	0.002	0.004					
5	0	0	0	0					
6	0	0	0	0					

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.36 to give an indication of the profile of each cluster over time. Having 4 time points in the follow-up gave individuals more room to transition than only having 3 follow-up time points.

Table 7.36: Characteristics of the 7-cluster model fitted to all individuals in NINDS data

Assigned cluster	n	%	Characteristics
1	34	5	No-slight disability with deterioration to severe disability or death
2	136	22	Slight to moderate-severe disability with recovery
3	140	22	Moderate-severe disability moving to death
4	105	17	Moderate to severe disability with little movement over time
5	106	17	Slight to moderate disability with recovery
6	70	11	Slight or no disability to no disability
7	33	5	Slight disability with deterioration over time

Figure 7.14: Graph showing the trajectories the individuals in each cluster, for all patients in NINDS trial

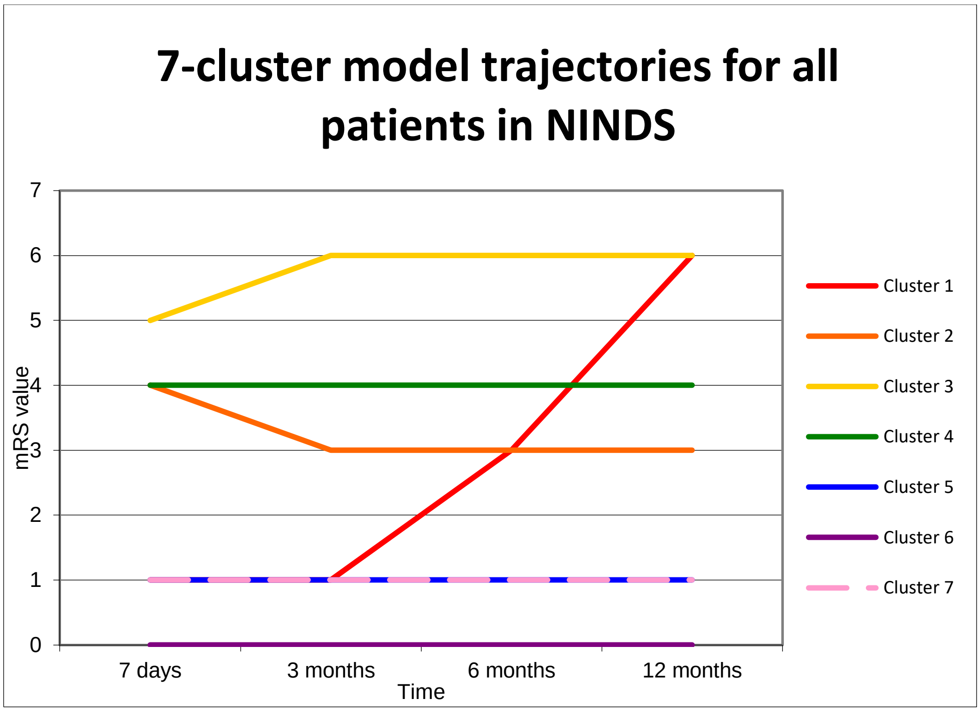


Figure 7.14 shows the model trajectories for the 7 clusters that were fitted in the model. With the exception of cluster 1, which showed a recovery trajectory that worsened at each time point in the model, the trajectories show the most change in the mRS score occurred between 7-10 days and 3 months, if a change occurred. Clusters 5 and 7 had the same mRS score for the largest percentage of individuals at each cluster. So although the movement was slightly different within each cluster, cluster 5 had some recovery and cluster 7 had some deterioration, overall the score

with the largest percentage of individuals was the same. Individuals in cluster 3 worsened between 7-10 days and 3 months, and the majority had died at 3 months. The majority of individuals in cluster 2 had a moderate-severe disability at 7-10 days, which decreased to moderate disability at 3 months.

To summarise, when considering the NINDS dataset and the changes in the mRS score over time, these can be described by a 7-cluster model. These clusters showed some movement, with most movement occurring early in the follow-up period. Figure 7.14 shows us that we had 2 clusters that follow the same trajectory for the majority of individuals and so this model was not really appropriate, even though it was chosen as the best fitting model and appears to had distinct clusters within it

A comparison to look at whether there was a difference between those individuals who received the placebo and those individuals who received the rt-PA treatment during the trial was conducted by fitting the treatment and control groups in separate models. The distribution of the mRS scores has been stratified between the treatment and the placebo group and is detailed in the table below, Table 7.37.

Table 7.37: Distribution of mRS values for each follow-up point stratified by treatment for NINDS trial

	Placebo group				**rt-PA treatment group**			
mRS score	**7 days (%)**	**3 months (%)**	**6 months (%)**	**12 months (%)**	**7 days (%)**	**3 months (%)**	**6 months (%)**	**12 months (%)**
0	7.4	10.6	10.6	10.3	15.8	18.3	18.9	18.9
1	10.0	16.0	17.9	16	16.8	24.4	22.1	21.2
2	9.7	11.9	10.3	11.5	8.4	7.7	8.3	7.1
3	9.4	14.4	16.0	12.2	9.4	12.8	13.5	12.5
4	29.1	19.9	13.5	10.9	21.9	13.5	9.6	6.1
5	26.5	6.7	5.8	5.8	23.2	6.1	4.5	4.5
6	7.8	20.5	26.0	33.3	4.5	17.3	23.1	29.8

Looking at Table 7.37, individuals in the placebo group seemed to have worse mRS scores at 7 days than those in the treatment group. This difference did not appear to have evened itself out over the follow-up period, with more individuals in the treatment group having no-to-slight disability than in the placebo group. Apart from the increasing number of individuals who had died by 12 months, both groups showed that after 12 months follow-up patients do appear to exhibit some recovery, as there are fewer individuals distributed in the higher mRS categories.

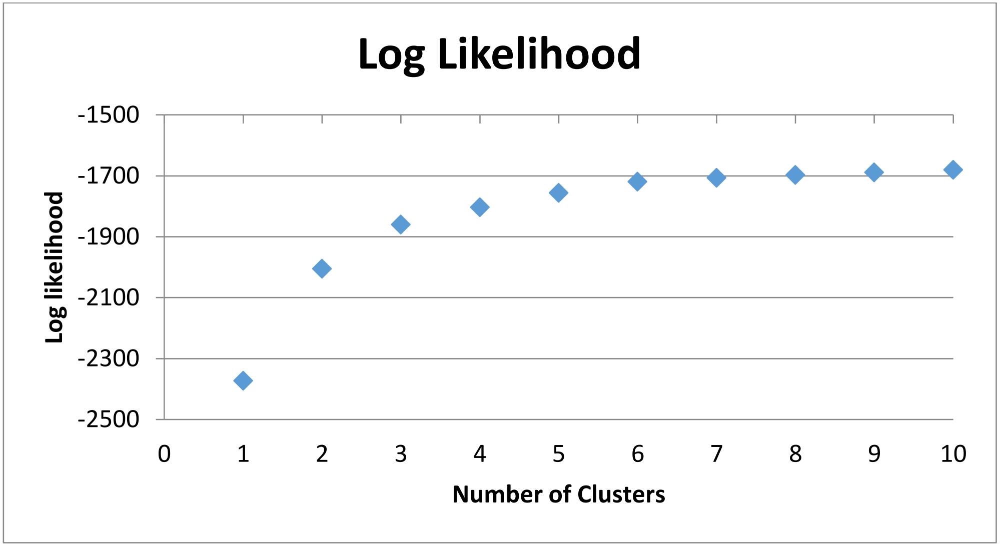
7.5.2 Placebo treatment group

First, we considered those individuals who received the placebo treatment. In the trial, individuals were randomised 1:1 to each treatment so there are 312 individuals in the placebo group. A series of 1, 2 … 10-cluster models were fitted to the data containing the only those who received the placebo treatment, as was done with all individuals, Table 7.38. It was found that the AIC and adjusted BIC were lowest for the 10-cluster model, but the BIC was lowest for the 9-cluster model. The log-likelihood also reduced with every additional cluster added to the model, and the reduction in log-likelihood can be seen in Figure 7.15. Once again, the BLRT indicated that every additional cluster in the model improved the fit of the model. It was seen from the VLMR LRT and the adjusted LMR LRT that most models had an improved fit on the one before, except the 7- and 8-cluster models, which had a statistically non-significant *p*-value.

Table 7.38: Model fit statistics for LCGA models fitted to placebo group in the NINDS data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	4759.415	4785.616	4763.414	-	-	-	-2372.71
2	4109.934	4147.364	4115.648	<0.001	<0.001	<0.001	-2004.97
3	3746.686	3795.345	3754.114	<0.001	<0.001	<0.001	-1860.34
4	3638.001	3697.889	3647.143	<0.001	0.0012	<0.001	-1803
5	3549.641	3620.758	3560.496	<0.001	<0.001	<0.001	-1755.82
6	3482.453	3564.799	3495.022	0.0174	0.021	<0.001	-1719.23
7	3462.878	3556.453	3447.161	0.75	0.7545	<0.001	-1706.44
8	3450.265	3555.069	3466.262	0.6509	0.6539	<0.001	-1697.13
9	3438.951	3554.984	3456.663	0.0134	0.0098	<0.001	-1688.48
10	3429.705	3556.967	3449.13	<0.001	<0.001	<0.001	-1680.85

Figure 7.15: Graph of the values of the log-likelihood for 10 LCGA models fitted to placebo treatment group of NINDS data



Having considered the IC statistics and the LRTs, we next looked at which model met the minimum requirement for numbers of individuals within each category. The 6-cluster model was the model with the largest number of clusters that had the minimum number of individuals in each category. The IC indices and LRTs suggested that this model had the best fit, as it was an improvement on the 5-cluster model, and so this was the chosen model. The cluster-specific proportions and the average probabilities of the cluster membership of the 6-cluster model are shown in Table 7.39.

Table 7.39: Posterior probabilities and number in each of the assigned clusters for placebo group patients in NINDS data

			Posterior probabilities					
Assigned cluster	**n**	**%**	**1**	**2**	**3**	**4**	**5**	**6**
1	75	24	**0.985**	0	0	0	0.015	0
2	58	19	0	**0.937**	0	0.007	0	0.056
3	16	5	0.024	0.012	**0.888**	0	0.047	0.029
4	23	7	0	0.048	0	**0.952**	0	0
5	63	20	0.006	0	0.006	0	**0.941**	0.047
6	77	25	0	0.05	0.005	0	0.036	**0.91**

Compared to previous models that had been fitted, there was not one very large cluster in this model, and although there were 2 smaller clusters with 5% and 7% of individuals in them respectively, they were not much smaller than the rest. The largest cluster, cluster 6, contained 25% of the data, which was closely followed by cluster 1 with 24% of the individuals in the model. This cluster had the best AvPP at 0.985. In fact, all the AvPPs were large with the smallest being 0.888 in cluster 3, suggesting that the model was very good at assigning individuals to the correct cluster of the model.

Finally, the 6 derived cluster-specific trajectories over time in the 6-cluster model were evaluated. Table 7.40 displays the item conditional probability of each level of the mRS ordinal scale at each of the 4 follow-up points.

Table 7.40: Conditional probabilities for each of the assigned clusters for placebo group patients in NINDS data

mRS scale	7-10 days	3 months	6 months	12 months	mRS scale	7-10 days	3 months	6 months	12 months
Cluster 1					**Cluster 2**				
0	0	0	0	0	**0**	0.083	0.113	0.155	0.28
1	0.001	0	0	0	**1**	0.498	0.547	0.582	0.576
2	0.004	0	0	0	**2**	0.292	0.246	0.196	0.111
3	0.023	0.002	0	0	**3**	0.102	0.076	0.054	0.027
4	0.218	0.02	0.001	0	**4**	0.023	0.016	0.011	0.005
5	0.446	0.111	0.007	0	**5**	0.002	0.001	0.001	0
6	0.307	0.867	0.992	1	**6**	0	0	0	0
Cluster 3					**Cluster 4**				
0	0.464	0.052	0.003	0	**0**	0.456	0.913	0.994	1
1	0.466	0.406	0.038	0	**1**	0.472	0.081	0.006	0
2	0.055	0.35	0.135	0	**2**	0.057	0.005	0	0
3	0.012	0.151	0.37	0.002	**3**	0.013	0.001	0	0
4	0.002	0.037	0.386	0.026	**4**	0.003	0	0	0
5	0	0.003	0.058	0.141	**5**	0	0	0	0
6	0	0.001	0.011	0.831	**6**	0	0	0	0
Cluster 5					**Cluster 6**				
0	0	0	0	0	**0**	0.002	0.004	0.007	0.019
1	0.005	0.004	0.004	0.004	**1**	0.033	0.053	0.085	0.209
2	0.019	0.018	0.017	0.015	**2**	0.12	0.174	0.243	0.367
3	0.098	0.094	0.089	0.08	**3**	0.354	0.397	0.404	0.297
4	0.489	0.482	0.474	0.456	**4**	0.412	0.323	0.23	0.097
5	0.304	0.313	0.322	0.341	**5**	0.066	0.042	0.026	0.009
6	0.084	0.089	0.093	0.104	**6**	0.012	0.008	0.005	0.002

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.41 to give an indication of the profile of each cluster over time. Comparing these cluster specific characteristics to those of all the individuals in the trial there are lots of similarities in the cluster characteristics.

Table 7.41: Characteristics of the 4-cluster model fitted to individuals in the placebo group in NINDS data

Assigned Cluster	n	%	Characteristics
1	75	24	Moderate-severe & severe to death
2	58	19	No to moderate disability with some recovery over time
3	16	5	No to slight disability deteriorating greatly over time to death
4	23	7	No to slight disability with full recovery
5	63	20	Moderate-severe & severe throughout
6	77	25	Slight to moderate-severe disability with recovery

Figure 7.16: Graph showing the trajectories the individuals in each cluster, for patients in placebo treatment group of NINDS trial

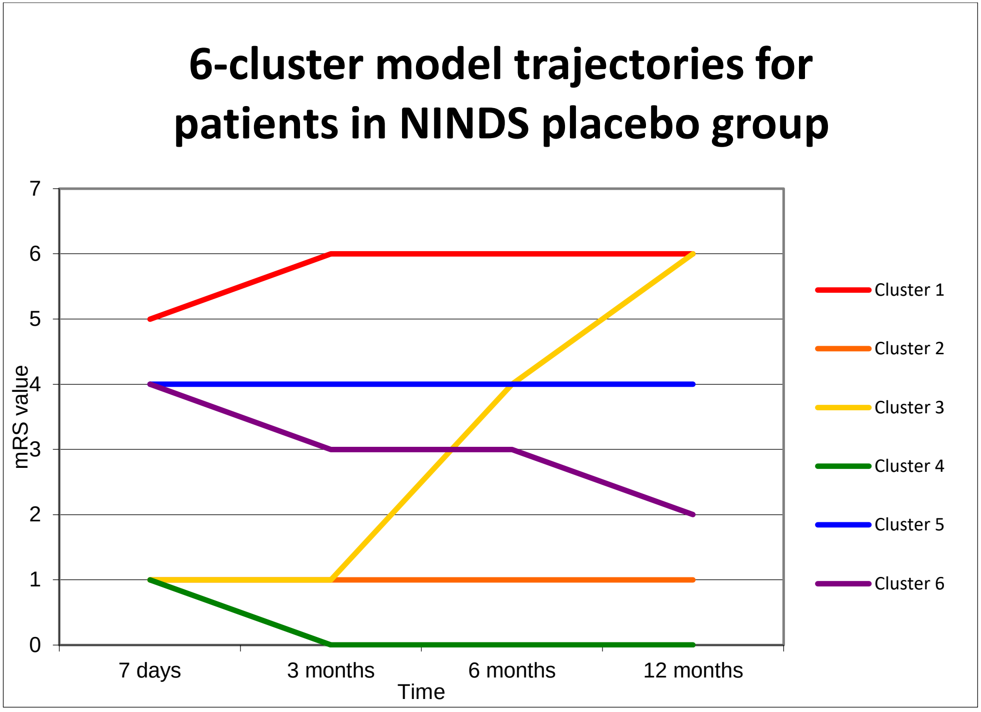


Figure 7.16 shows the model trajectories for the 6 clusters that were fitted in the model. The trajectories that showed deterioration over time were clusters 1 and 3 and these were the same trajectories as two of the clusters in the model that included all individuals in the trial. Two clusters showed no movement in the score of the majority of individuals over time. Finally, two clusters showed recovery over time; both clusters had a one-score decrease on the mRS between 7-10 days and 3 months, with cluster 6 showing a second recovery step and a further one-score decrease between 6 and 12 months.

To summarise, when considering the individuals in the NINDS dataset who received the placebo treatment, the changes in the mRS score over time can be described by a 6-cluster model. There were a lot of similarities seen in the clusters fitted in just those individuals who had placebo and all individuals, with similar deterioration clusters. The clusters that showed recovery were specific to those who recovered the placebo treatment.

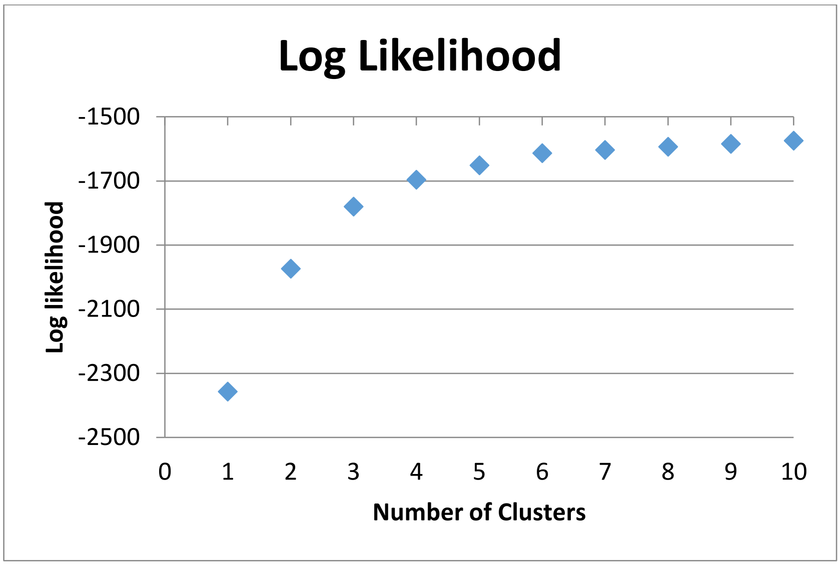
7.5.3 rt-PA treatment group

Next, we considered the other 312 individuals who received the rt-PA treatment during the trial. As with the other 2 models using the NINDS data, 1, 2… 10-cluster models were fitted to the data. There were no problems with the estimation of the maximum likelihood. As has been seen before, the lowest values of the IC indices were for the 10-cluster model, and they agreed with each other, unlike the model fitted to the placebo group, Table 7.42. As each cluster was added to the model, the log-likelihood decreased; this can be seen in Figure 7.17. At the start, the change in log-likelihood was large between models, but it began to decrease for higher-cluster models. The BLRT indicated that every model improves the model fit on the one with one less cluster. However, the VLMR LRT suggested that the models with 7 clusters or more show no better fit than the previous model as all the *p*-values are not statistically significant.

Table 7.42: Model fit statistics for LCGA models fitted to the treatment group in the NINDS data

Cluster	AIC	BIC	adj BIC	VLMR LRT	adj LMR LRT	BLRT	LL
1	4728.729	4754.93	4732.728	-	-	-	-2357.36
2	3967.758	4005.188	3973.472	<0.001	<0.001	<0.001	-1973.88
3	3586.351	3635.01	3593.778	<0.001	<0.001	<0.001	-1780.18
4	3424.637	3484.525	3433.778	<0.001	<0.001	<0.001	-1696.32
5	3340.759	3411.873	3351.612	<0.001	<0.001	<0.001	-1651.38
6	3271.679	3354.025	3284.248	0.0058	0.0075	<0.001	-1613.84
7	3257.032	3350.607	3271.316	0.1281	0.1422	<0.001	-1603.52
8	3241.991	3346.95	3257.989	0.6756	0.6933	<0.001	-1593.0
9	3231.47	3347.503	3249.181	0.1444	0.1497	<0.001	-1584.74
10	3216.392	3343.654	3235.818	0.2175	0.228	<0.001	-1574.2

Figure 7.17: Graph of the values of the log-likelihood for 10 LCGA models fitted to treatment group if NINDS trial



The minimum number of individuals for each cluster were considered and it was found the 6-cluster model had enough individuals to satisfy the minimum criterion. This was the largest model to do so, as the 7-cluster model and models with a higher number of clusters contained clusters with only 1% of individuals in them, as did the 5-cluster model. The cluster-specific proportions and the average probabilities of the cluster membership of the 6-cluster model are shown in Table 7.43.

Table 7.43: Posterior probabilities and number in each of the assigned clusters for treatment group patients in NINDS data

			Posterior probabilities					
Assigned cluster	N	%	1	2	3	4	5	6
1	67	21	**0.969**	0	0	0	0.03	0
2	44	14	0	**0.982**	0	0	0	0.018
3	61	20	0	0	**0.938**	0.002	0.045	0.015
4	20	6	0.018	0	0.085	**0.825**	0.045	0.027
5	37	12	0.013	0	0.089	0.003	**0.894**	0
6	83	27	0	0.026	0.022	0.001	0	**0.95**

The largest cluster in the model was cluster 6 (n=83, 27%) and the smaller was cluster 4 (n=20, 6%). The largest cluster had 4 times as many individuals in the cluster than the smallest, but still contained nowhere near enough individuals to have the majority of those in the dataset. All of the AvPPs showed that the model fitted the data well; the smallest was 0.825 and every individual in the model had a very good chance of being assigned to the correct cluster. The 6 derived cluster-specific trajectories in the 6-cluster model were evaluated. Table 7.44 displays the item conditional probability of each level of the mRS ordinal scale over the 4 follow-up time points.

Table 7.44: Conditional probabilities for each of the assigned clusters for treatment group in SO_2S data

mRS scale	7-10 days	3 months	6 months	12 months	mRS scale	7-10 days	3 months	6 months	12 months
Cluster 1					**Cluster 2**				
0	0	0	0	0	**0**	0.763	0.904	0.968	0.997
1	0.001	0	0	0	**1**	0.225	0.092	0.031	0.003
2	0.004	0	0	0	**2**	0.009	0.003	0.001	0
3	0.027	0.002	0	0	**3**	0.003	0.001	0	0
4	0.2	0.023	0.002	0	**4**	0	0	0	0
5	0.477	0.151	0.013	0	**5**	0	0	0	0
6	0.29	0.823	0.985	1	**6**	0	0	0	0
Cluster 3					**Cluster 4**				
0	0.002	0.002	0.003	0.008	**0**	0.347	0.036	0.002	0
1	0.035	0.052	0.077	0.168	**1**	0.584	0.452	0.049	0
2	0.094	0.13	0.177	0.282	**2**	0.051	0.302	0.125	0
3	0.368	0.414	0.439	0.39	**3**	0.016	0.171	0.409	0.003
4	0.402	0.334	0.258	0.133	**4**	0.003	0.034	0.342	0.031
5	0.086	0.06	0.04	0.017	**5**	0	0.004	0.062	0.189
6	0.013	0.009	0.006	0.002	**6**	0	0.001	0.01	0.776
Cluster 5					**Cluster 6**				
0	0	0	0	0	**0**	0.095	0.123	0.16	0.266
1	0.006	0.005	0.004	0.003	**1**	0.631	0.657	0.668	0.635
2	0.018	0.016	0.013	0.009	**2**	0.187	0.154	0.122	0.071
3	0.12	0.103	0.088	0.062	**3**	0.073	0.056	0.042	0.023
4	0.462	0.44	0.412	0.347	**4**	0.013	0.01	0.007	0.004
5	0.319	0.348	0.379	0.434	**5**	0.001	0.001	0.001	0
6	0.074	0.087	0.104	0.146	**6**	0	0	0	0

The cluster specific characteristics for each fitted cluster of the model are described in Table 7.45 to give an indication of the profile of each cluster over time. There were a lot of similarities seen in the clusters fitted in individuals who received the rt-PA treatment and those fitted in individuals who received the placebo.

Table 7.45: Characteristics of the 6-cluster model fitted to the treatment group in NINDS data

Assigned Cluster	N	%	Characteristics
1	67	21	Moderate-severe & severe to death
2	44	14	No to slight disability with full recovery
3	61	20	No to moderate disability with some recovery over time
4	20	6	No to slight disability deteriorating greatly over time to death
5	37	12	Moderate severe & severe with some deterioration
6	83	27	Slight to moderate-severe disability with recovery

Figure 7.18: Graph showing the trajectories the individuals in each cluster, for patients in treatment group of NINDS trial

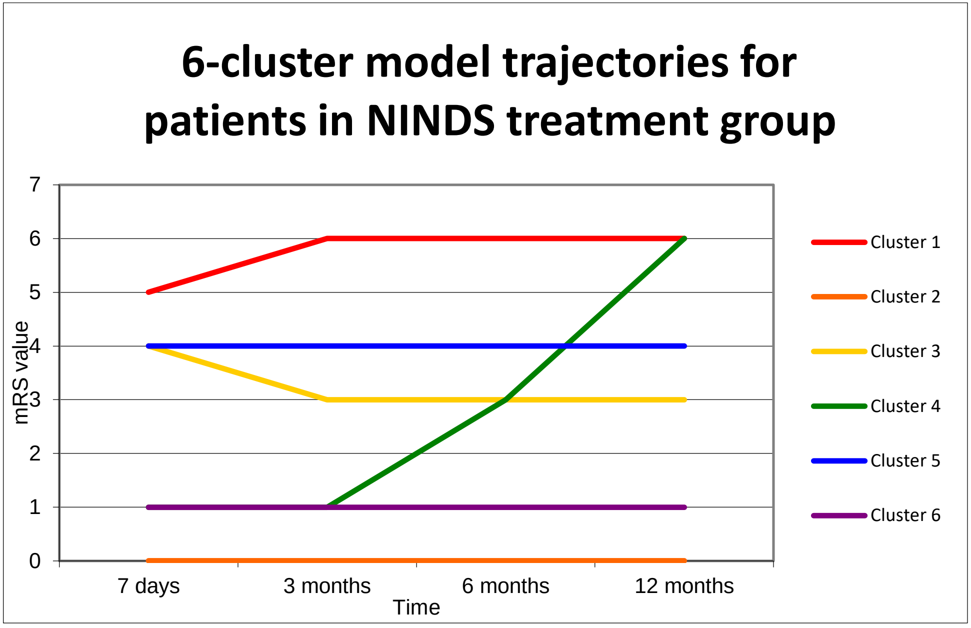


Figure 7.18 shows the model trajectories for the 6 clusters that were fitted in the model. Although there were similarities in the cluster descriptions and the small movements within a cluster between the treatment and the placebo, the trajectories for the treatment group showed less movement over time. Apart from cluster 4 showing deterioration at each time point of the model, all movement occurs early in the follow-up. There were only 2 clusters that had movement, one with those who were severely disabled and died at 3 months and one where the majority of

individuals had a moderate-severe disability and showed some recovery in the first few months to have only a moderate disability.

To summarise, when considering the individuals in the NINDS dataset who received the rt-PA treatment and the changes in the mRS score over time, these can also be described by a 6-cluster model. Although there were similarities in the cluster descriptions and the small movements within a cluster between the treatment and the placebo, there was less movement over time in the treatment group. With the exception of the cluster that transitioned from no disability to death, all recovery or deterioration was done early.

7.5.4 Comparison of placebo and rt-PA groups

The models fitted to the placebo group showed that the best model was the 6-cluster model. There were 312 individuals split into 6 clusters and each cluster had a distinct trajectory over the 12-month follow-up period. The models fitted to the rt-PA treatment group showed that the best model was the 6 cluster model. There were distinct trajectories for the 12-month follow-up with the 312 individuals assigned to one of the 6 distinct clusters in the model.

The models fitted to the placebo group and the rt-PA treatment group are very similar in the clusters that are fitted to the data. Out of the 6 clusters in each model, 5 of the clusters are very similar in the characteristics of the trajectories that they show. Both models had clusters where individuals move from moderate and severe disability at 7 days to death at 12 months, and no or slight disability at 7 days with deterioration to death at 12 months. There are also clusters where individuals had no or slight disability with no change. There are two clusters in which recovery is seen, where individuals show recovery from between slight and moderate-severe disability and from no and moderate disability. The only cluster where the models differ slightly is the cluster where the placebo group individuals had moderate-severe disability throughout, whereas in the rt-PA treatment group there is some deterioration seen to slightly more severely disabled. The treatments were also considered when they were combined together, which seems sensible as the clusters fitted to the placebo and rt-PA treatments are very similar.

The model fitted to the data combined across treatment groups shows that the best model that was fitted was the 7-cluster model. There were 624 individuals split into 7 clusters and each cluster had a distinct trajectory over the 12-month follow-up period.

As expected, the model fitted to all the data provided very similar clusters to those seen in the 6-cluster model for the placebo and the 6-cluster model for the rt-PA. The main difference was that there was a 7th cluster in the model with all the data, which was a cluster that shows individuals with a slight disability at 7 days who then go on to deteriorate over time. The only other difference between the models was that the cluster in the placebo and rt-PA fitted models had a cluster with no to moderate disability with recovery, which in the model with all the data has individuals who are slightly more disabled and range from slight to moderate disability at 7 days before recovery. Otherwise, the clusters produced by the models are consistent across all the models fitted.

7.6 Discussion

This chapter aimed to use a person-centred approach to the longitudinal analysis of the mRS, by conducting LCGA. The analysis was conducted in order to look at the clustering of individuals due to similar trajectories of their recovery profiles over time. LCGA models were fitted firstly to the SO_2S data using all patients at 3 consecutive time points, followed by individuals who experienced some change in the mRS score during the follow-up time period, and finally individuals who did not die during the 12-month follow-up period. Models were also fitted using 4 consecutive time points using the NINDS dataset for all individuals that were followed up over the 12 months. The models fitted found that similar results in these individuals were mainly clustered into trajectories that remained the same over time, even after those individuals were removed. More variable profiles were seen with 4 consecutive follow-up time points, although still very similar.

These models as with all those models in the thesis can be extended to various other scenarios. As with multi-state models, data that change very little over time are less suitable with this type of modelling, as there would be little movement within the trajectories over time, meaning little variation to distinguish between the clusters. The models need a minimum of 3 time points in order to be able to draw a useful trajectory, but with a larger number of time points the model can handle both regular and irregularly spaced follow-up points.

The trajectories are modelled from the first time point, therefore it there was a baseline measurement of the outcome then this would be the starting point of the trajectories rather than the first follow-up time point. LCGA models are able to handle missing data well, patients with missing data are able to be included in the analysis as the model is able to cope with small amounts of missing data.

7.6.1 Model comparisons

The first model fitted used all the individuals that had been followed up for 12 months in the SO$_2$S dataset. The model was fitted to 7,725 individuals who were split into 4 distinct clusters within the model. Patients in each of the clusters defined had very little change over time and any variations seen within the scores of each cluster at each time point were not large enough to change the recovery profiles from straight lines throughout the follow-up period. The 4 clusters had individuals who were severely disabled, moderate-severely disabled, had no disability at all and with slight to moderate disability throughout the follow-up.

The second set of models fitted to the data removed those individuals who did not make any transition to another mRS value over the follow-up period. The first model was fitted to all 5,273 individuals who showed some movement throughout the follow-up, and had 5 distinct clusters. For the model fitted to the standard care group there were 1,760 individuals split into 4 clusters and in the treatment group there were 3,513 individuals split into 6 clusters in the model all took distinct trajectories over the 12 months follow-up period.

There were 2 clusters that featured consistently about throughout the three models, one where the individuals were severely disabled and had died by 12 months and one where individuals had no or slight disability and showed signs of recovery other the 12 months. Other clusters were unique to each of the models but included a cluster of individuals with slight/moderate disability who worsened over time and clusters with low/moderate or moderate sever disability who remained the same throughout the follow-up.

The third set of models were fitted to only those individuals who were alive at the 12-month follow-up assessment. The first of these models was fitted to 6,679 individuals who were all individuals alive at 12 months and there were 5 distinct clusters found. The second model fitted to those individuals who were alive at 12 months only considered individuals who had a change in the mRS score at some point over the 12-month follow-up. This applied to 4,227 individuals and a 6-cluster model was fitted to the data. The only common cluster in the models was the cluster in which there are individuals with slight disability throughout follow-up. Otherwise the clusters were distinct in the two models, with no movement overtime seen in the model with all individuals and changes in trajectories seen in the second model, where individuals need to have at least one transition.

The fourth set of models were fitted to the NINDS dataset, looking at 4 consecutive follow-up time points rather than the 3 consecutive time points in the SO_2S dataset. The first of these models was fitted to the full 624 individuals and there were 7 distinct clusters found. The second model fitted used just the placebo group, which was 312 individuals, and there were 6 distinct clusters fitted to the data. The final model was fitted the rt-PA group, which was 312 individuals, and there were 6 distinct clusters fitted to the data. There are 5 of the clusters are very similar in the characteristics of the trajectories that they show. All the models had clusters where individuals move from moderate and severe disability at 7 days to death at 12 months, and no or slight disability at 7 days with deterioration to death at 12 months. There are also clusters where individuals had no or slight disability with no change. There are two clusters in which recovery is seen, where individuals show recovery from between slight and moderate-severe disability and from no and moderate disability.

It can be seen across the entire collection of models fitted that there was not a great deal of change in the conditional probabilities calculated over time, giving clusters an overall impression that there was not much movement of individuals between scores on the mRS scale. This was consistent with other results found in the thesis. The biggest surprise was that even when individuals who do not make a transition are removed, the movement between scores is very small and that there appears to be no overall change within the cluster. However, this may have been expected, given that a change of 1 was seen to be of clinical importance in the mRS, this transition may be hard to capture on such a large scale. When we remove individuals who have died at 12 months and restrict the model to individuals who have a change in the mRS during the follow-up, we begin to see more movement of the individuals within the clusters; however, this is probably the worst fitting model of all the LCGA models fitted to the data. We also see more movement when there are 4 consecutive time points rather than 3. This could also be due to the time at which the mRS was recorded, as the first time point is at 7 days and the second at 3 months, this time interval is where most of the recovery of a stroke patient occurs, and so we would expect to see more changes here than we would expect to see from 3 months onwards.

7.6.2 Further work

There are other models that could be fitted the data that use a person centred approach The main one that is most comparable is growth mixture modelling (GMM). It can be seen as an extension to LCGA where the assumption of no variation between individuals within the clusters has been relaxed. This means that individuals who are in the same cluster are allowed to have varying growth curves, but the growth curves are assumed to be similar to the mean curve for each cluster. The GMM adds a random effect to allow individuals to vary the growth curve that they follow. There are different types of GMM that can be fitted, including quadratic or cubic trajectory models, that model linear growth; however, they require 4 and 5 time points respectively, which is obviously more than is recorded in the SO_2S dataset. Alternatively, it is possible to fit a freely estimated time factor model, which requires only 3 time points and models non-linear growth, which may be a model that is possible to fit to extend the person-centred analysis. This technique requires a large amount of time to compute, and often provides no additional benefit to LCGA (Strauss et al. 2014).

As well as this, it would be of interest to investigate the baseline covariates that were recorded to see if the individuals in each of the clusters in the model have specific characteristics that could be used to predict the membership of an individual to each cluster. For the purpose of this analysis the main interest lay in the effect of treatment on the mRS, with each individual assigned to a cluster depending in the trajectory of the mRS over time. Once assigned to a cluster it is possible to look at characteristics of the individuals in each cluster of the final model, for example age, gender and assessed baseline severity, or other things that could be used as covariate adjustment in a regression model. Questions that this could answer may include are there different ages ranges in the clusters found, are the older individuals in the clusters that have more disability or are there clusters that a made up of considerably more women or men? This extension may be of interest to help identify the potential trajectories of individuals who have had a stroke with specific mRS scores at a 3 month time point. Although this was not investigated in full, due to there being so many different models fitted to the data, as an example the baseline characteristics for the SO_2S model with all patients who make a transition are detailed in the table below.

Table 7.46: Baseline characteristics of each of the identified groups in the 5 cluster model for all patients who make a transition in the follow-up time

	Age (years)	% Female	week 1 NIHSS	% Treatment Group
Cluster 1	80.1	56%	13 (0-38)	67%
Cluster 2	70.1	44%	3 (0-30)	67%
Cluster 3	72.7	48%	6 (0-34)	66%
Cluster 4	79.4	52%	9 (0-34)	64%
Cluster 5	69.6	39%	2 (0-16)	68%

It can be seen that there do appear to be some differences in the baseline characteristics detailed here. For example, cluster 1, in which patients were severely disabled and moved to death, were on average 10 years older than the other clusters, with more females and a high NIHSS score at week 1, indicating a more severe stroke. In contrast, cluster 5, in which patients had little to no disability, had the lowest average age and the smallest NIHSS score at week 1, suggesting that the effects of the stroke were not so severe. This type of investigation could be conducted for any of the LCGA models fitted to the data, and for any range of baseline indicators. These could then be used to inform clinicians about the potential recovery that an individual has, depending on their characteristics and the cluster into which they best fit.

7.6.3 Conclusions

Looking at the LCGA models fitted to both datasets, there does appear to be a difference in the recovery trajectories when the treatment and control groups are fitted separately. The model is unable to quantify this difference as a treatment effect cannot be calculated. This method can provide a visual representation of how individuals change over time after they have had a stroke. These methods have already been applied in the field of stroke, but rarely in a longitudinal setting, and these results suggest that using LCGA may provide additional information on how the treatment effects recovery by comparing the profiles of the trajectories.

The graphs produced are appealing to clinicians as the visual trajectories are easy to interpret. The trajectories found for most of the models show little change over time, but speaking to clinicians this could be clinically accurate, as although recovery can occur over a 12 month follow-up, the most common time for recovery is early after the stroke, which is not recorded within this data, as the first follow-up point is 3 months.

Having fitted several different models to the data, the next chapter looks at how each different model fitted estimates the treatment effect of the data, provides a commentary on what the different models mean, and reflects on key decisions that have been made throughout the thesis. It also looks at how the work conducted tries to address the aims of the thesis, as well as future work that could be conducted.

8 Discussion

In this PhD project, I have investigated several methods that can be used to conduct a longitudinal analysis of stroke data. The focus of the research has been the mRS, a common primary end point for stroke trials and a scale rarely utilised for a longitudinal analysis. This chapter reviews the reasons for undertaking the analysis and the principal findings from the different longitudinal methods that have been fitted. This is followed by a discussion of the key decisions that were made throughout the thesis and their impact on the project. Finally, the chapter concludes with future research and the implications of this work.

8.1 Current gaps in knowledge and thesis objectives

From a review of the literature, it was found that very little research was conducted using longitudinal methods in stroke trials. The systematic review identified that nearly a third of trials that recorded measurements at multiple time points failed to conduct a longitudinal analysis. One reason for this may be that there is currently no consensus on the best method to use for serial measurements, which highlights the importance of the aim of this project. It was found that when a longitudinal analysis was conducted in a stroke trial, the methods that were chosen may not have given the best interpretation of the data, and usually failed to fit models that used all available data. It is important to remember that when collecting data on several scales, these scales should be used to their full potential and not dichotomised just because the interpretation is considered to be easier, as there is a loss of statistical power when dichotomisation is applied, which is not offset by the ease of interpretation.

Recently there has been much research conducted on how best to analyse the outcomes of stroke trials, with particular emphasis on the mRS (Bath et al. 2012). The Optimising Analysis of Stroke Trials (OAST) Collaboration conducted work on this question in 2007, and recommended methods that keep the functional outcome data in its ordinal form is more statistically efficient than those methods that dichotomise the data (Bath et al. 2007). The results were consistent across the different outcome measures considered. Even after these recommendations, there has not been a dramatic shift in the types of analysis performed when analysing acute stroke trials (Nunn et al. 2016).

Functional recovery over time is not well researched in stroke. This may be due to the fact that general recovery is hard to specify and that a lot of recovery is done in the first few weeks, which is often before follow-up interviews and questionnaires have been conducted, and so they fail to capture the moment at which patients' recovery begins. The aim of this thesis was to apply different statistical models to the mRS in order to look at how the different models affect the estimate of treatment effect during the follow-up time.

8.2 Summary of main findings

8.2.1 Using regression models to analyse SO$_2$S dataset

Exploring the use of repeated measures regression models in order to analyse longitudinal data (Chapter 4) revealed that using ordinal regression models is preferable in the analysis of the mRS to dichotomising the data and conducting binary logistic regression. The chapter looked at how well a repeated measures model was able to estimate the treatment effect at 3 months (usually the primary endpoint of a trial), compared to a regression model with a single 3-month time point. It was found that both the dichotomised mRS and the mRS kept in its full ordinal form produced an overestimate of the treatment effect at 3 months. However, the magnitude of the overestimation was smaller when the mRS was kept in an ordinal form (3% overestimation vs 7% overestimation in the effect of treatment compared to control). The repeated measures analysis fitted to the longitudinal data uses all data that are recorded in the follow-up and not just a single isolated point.

8.2.2 Using a simulation study to investigate the association between effect estimates of the repeated measures model and the single ordinal regression model.

There was a high correlation between the mRS at each of the time points in the follow-up, which prompts the question: is this always the case, or are there certain situations for which the difference in the single time point model and the repeated measures model treatment estimates at the same time point are very similar, or even more different than has been observed, suggesting the 2 could not be used interchangeably. If a difference is always seen, this would suggest that it would be reasonable to use a repeated measures model to find the treatment effect as a main analysis for a trial, when longitudinal data have been collected. This simulation study fitted ordinal data with exchangeable correlation structures of 0.2, 0.4 and 0.6. Surprisingly, the smallest correlation values appears to give the best similarities between the results of the single time point and the repeated measures model, even though the outcomes vary more over time than with a higher correlation structure. There did not appear to be any distinguishable pattern between the effect estimates and the different dropout patterns that were applied to the data, with MAR or MCAR, suggesting that the agreement between the repeated measures model and the single ordinal regression model cannot be predicted.

It was thought that using only 3 time points for the follow-up may be too short in order to be able to draw meaningful conclusions about how the repeated measures model performs.

8.2.3 Using multi-state models to analyse SO_2S dataset

Multi-state models were applied to the SO_2S dataset in order to keep the mRS as an ordinal variable, whilst allowing each category in the scale to be defined as a state in the model, which meant that the transitions of individuals between each state could be identified.

It was found that fitting models that had no censored states and those that included censored states within the model produced very similar results, and that the model was able to calculate a treatment effect by producing a hazard ratio for each potential transition in the model. There are benefits to using the model with censored states as it allows more individuals to be included in the model, and will show a slightly more accurate representation of the transition to death, as these are most of the individuals who were excluded in the first model.

As had been found in the regression models that were fitted to the data, the treatment effect was calculated to be not statistically significant, and although a treatment effect was unable to be calculated as a single effect, it is useful to be able to distinguish between different transitions in the model.

Although the model was restricted to allow individuals to transition to the adjacent state from the current state, an individual was assumed to have moved to further states in the model by assuming that they had moved through consecutive states to reach the final state. By using a multi-state model, having fewer time points in the model is less of a limitation as the Markov property of the multi-state model suggests that previous state transitions are to be ignored and the model works on calculating only the instantaneous risk of an individual in the model. Restrictions were placed in upwards and downwards transitions to give single estimates.

8.2.4 Using LCGA models to analyse both SO_2S and NINDS datasets

There have been applications of LCGA within stroke trials (Busija et al. 2013, Pan et al. 2008a), though not necessarily using the mRS, and an extension to fit a LCGA model to the data was therefore a natural extension. The LCGA analysis took a different approach again to analysing the data and took a person-centred approach, whereby groups of individuals were clustered together based upon the trajectories of the mRS score over time. Identifying potential recovery patterns over time showed that many individuals had no change over the follow-up period, and so after an initial model had been fitted to the data, these individuals were removed in order to try to classify individuals that had a change of score between 3 and 12 months. These individuals were also modelled as treatment and standard care groups separately, so that, although an effect estimate for the difference in treatment and standard care groups was not estimable, visual differences in the trajectories could be observed. The LCGA analysis was the only model fitted in the study that excluded individuals who had died, as the inclusion of death on the mRS was one of the main interests.

Having 3 follow-up points means that there are only 2 potential points at which an individual can move states, so by using the NINDS dataset the LCGA could be fitted to 4 points, which showed that there was more movement within the clusters of the model. The majority of this movement was seen within the earliest 2 time points, before the time at which the SO_2S dataset recorded the mRS, where individuals are expected to have more recovery.

8.3 Comparison of techniques applied

There are several different statistical techniques that have been used to conduct the analyses of longitudinal data in this project. The different models fitted use different numbers of individuals and so it is hard to use conventional techniques to compare the different models fitted – for example, AIC and BIC – as these may not have been calculated like in the multi-state models that have been fitted. Therefore, this section aims to compare the different models through discussion, looking at the effect estimates that are produced and the assumptions that the models require in order to be fitted to the data.

Without considering the results of the models, the first thing that can be compared between them is how missing data are dealt with. Both the regression models (ordinal and binary) as well as the LCGA models are able to accommodate missing data within the analysis that is being conducted because the models are able to deal with the missing outcome values for 1 or 2 of the time points. For the multi-state models, this is not that case, and those individuals with a missing outcome values are not immediately included within the analysis, as a missing value would not inform the model of where the individual was transitioning to or from. In this scenario a censored state can be included in the model, which allows the model to know that the individual has not died, but could be in any of the other states in the model.

All three models that were fitted to the data were able to model the irregularly spaced time-points of the outcome included. The regression model did this by including dummy variables and treating time as a categorical value, with a sensitivity analysis conducted to consider the effects of continuous time. The use of categorical time meant that the treatment effect could only be estimated at the times for which the outcome was actually recorded. The multi-state model fits a transition between 2 time points, and assumes a Markov property such that each time point does not depend on the previous time points; therefore the 2 potential transitions between the 3 time points could be estimated, using methods for panel data, which were able to able to deal with the unequal time spacing. Finally, the LCGA model looks at transitions between each consecutive time point in the model; therefore, this method also had no issues with modelling unequally spaced data.

The regression models were able to estimate specific overall treatment effects, for the odds of being dependent compared to independent for the binary model, and the odds of being in a higher category than the cut off in the ordinal model. These overall effects are useful because they are easy to interpret; however, there is the potential for some information to be lost, due to the assumption of proportional odds producing the same estimates for each cut-off rather than estimates the vary dependent of the comparison that is being made. The proportional odds assumption was assessed in the regression models fitted and, although there was no improvement in the fit if the model, it could be seen that by relaxing the assumption different estimates were produced for treatment at the different cut off points of the model. These differing estimates are lost when proportional odds are assumed.

The multi-state model fitted to the data calculated a treatment effect for each of the potential adjacent state transition, with small differences seen in the treatment effect depending on where an individual started on the scale. The individual estimates for the treatment effect on the worsening or recovery of individuals on the mRS are potentially useful because it may be that there is a subgroup within the scale that is more likely to benefit from treatment, for example those with an mRS between 2 and 4, and therefore the treatment may have the potential to target these individuals, for whom a benefit may not have been seen with an overall treatment effect. Unlike the other models in the analysis, the differences between the treatment and control groups could not be quantified by the LCGA models that were fitted to the data. Visual comparisons can be made, and indeed presenting the results as a visual trajectory of movement over time is a novel way of presenting the data and may be useful as an aid for explaining potential recovery to patients, who may not understand the odds ratios and hazard ratios calculated in the other models.

There are different restrictions placed on the individuals that were able to be included in the models that were fitted. All individuals were able to be included in the regression models that were fitted, due to the fact that that the model could deal with missing data and could include individuals if they had died at 3 months using the last observation carried forward. The multi-state models needed censored states to deal with individuals with missing state values, but the model was not able to model individuals who had died at 3 months, as this would require modelling a transition into the same state (the absorbing one) and although the model allows individuals to remain in the same state, it is defined for an individual's moving to an adjacent state, which is not possible for the absorbing state. Because of this, individuals who had died at 3 months were dropped and those who died at 6 or 12 months were included with a missing value at 12 months if an individual had died at 6 months. The opposite happened for the LCGA, when considering individuals who had a transition in the model, those who had died at 3 months had to be include, even though they made no transition throughout the follow-up. This is because the outcome included in the LCGA needed to have the same number of categories at each time point (0 – 6), without those who had died at 3 months included the first time point had an outcome that only ranged from 0 – 5.

8.4 Strengths and limitations

A major strength for all the different analyses used within the project is the size of the dataset that has been included; due to the large numbers in the trial, over 7,700 of the 8,003 individuals could be included in the analysis. The large sample size helps to reduce the amount of uncertainty seen in the model because of the larger numbers it removes the idea that the differences seen in the analysis are seen by chance. As well as this the estimates produced are more reflective of what would happen in the general population, as more individuals are included in deriving them.

8.4.1 Regression models

Strengths of the regression models that were fitted were that these models could be fitted in readily available statistical packages, were easy to fit, and also provided single treatment effect estimates that were easy to interpret. The findings here supported the work that has previously been conducted in stroke research (Bath et al. 2007) in suggesting that an ordinal regression model was the best model to be applied to the mRS at both a single time point, which was previously recommended, and when using a repeated measures model to consider the effect over time. If researchers want to investigate the effects of a predictor variable on an ordinal scale, it is important that these effects can be calculated at all stages of the ordered categorical response and not just at a single point on the scale; ordinal regression should therefore be used in preference to a binary model (Javali & Pandit 2010).

A further strength of the work conducted here was that the assumptions of the proportional odds model were checked and the models extended in a systematic way, rather than using a package with an auto fit option, for example, gologit2 in Stata. By doing this it meant that the random effect for the individuals could still be incorporated in the repeated measures model, which is important to account for the correlation between the outcomes at each time point due to their being observed repeatedly for each individual in the model. This could not be done in available software and required the use of a written SAS program, given in Appendix B.

A limitation of the work completed is the inability to compare how well the single time point and repeated measures models fit the data, due to the inability to compare goodness of fit statistics. This is because there are different numbers of data points within each model. Because of this, each model could be assessed on how good the fit of the model was within each setting and the treatment effects could be calculated and compared; however, the goodness of fit statistics of the two types of model fitted could not be compared. Roozenbeck et al. (2011) calculated the ratio of the Wald statistics, which can be interpreted as the gain in information density, and therefore, could be considered a suitable measure for the efficiency of the different approaches and allow two different types of model to be compared,

There was also a limitation with the number of follow-up points recorded in the SO_2S study. With only 3 follow-up points there are few opportunities for changes to be observed, and once an individual had died, the series of repeated measures they have was reduced further, with a last observation carried forward used to impute a value at the following time points, as they were known to have died. A longer follow-up period, or just more regularly recorded repeated measures outcomes within the year-long follow-up, would increase the number of serial observations and allow more complex models like a joint ordinal regression and survival model to be fitted to the data, but with the limited series that was recorded in this data, a joint model was unable to produce accurate results.

8.4.2 Simulation study

A strength of the simulation study lies in the variety of scenarios that were tested to see how well the repeated measures model can estimate the true effect. The number of time points was chosen based on the SO_2S dataset that has been used in the thesis for different types of analyses.

There was a decision made that because there were 500 repetitions of the dropout pattern application and it was computationally intensive to fit the regression models, 1000 individuals would be sufficient, although not as many as were included in the SO_2S trial data. Having completed the study, this decision seems appropriate as there is no real pattern seen, therefore including more individuals would not produce better estimates than the ones that were seen, and so 1000 individuals were sufficient due to a lack of consistency in the results seen.

A limitation of the study was that each of the generated datasets produced a different true effect. This meant that sometimes the results were influences by the size of that initial effect (i.e. when it was very small). In hindsight, it would have perhaps been better to simulate data with the same common treatment effect across all the different dataset generated in order to be able to make better comparisons of the effects.

8.4.3 Multi-state model

There are several advantages to fitting multi-state models to longitudinal stroke data. The first is that it provides a way to use the whole of the mRS scale without having to worry about violating the ordinality of the scale. Although each score on the scale is considered as a different category, ordinality is still preserved in the model as patients are able to transition to categories (albeit only to adjacent categories) in the model and it is also assumed that patients will transition throughout the model on the way to death.

Secondly, by only using adjacent states as defined transitions, the transitions being used are using information in the dataset where the data are not sparse, which is the case with transitions that move further off the diagonal. As well as this, the inclusion of censored states allows a number of individuals to be included in the analysis where they would normally be excluded for having missing values. Furthermore, a change of one point on the mRS is deemed to be clinically relevant (Harrison et al. 2013) and this is what the multi-state model fitted to the data is modelling. The adjacent-state modelling does not restrict individuals just to transitioning to the next state; it just means that they are assumed to have transitioned through each adjacent state in order to get to one further away.

Having only a few time points in the model is less of a limitation, as the Markov property of the multi-state model suggests that previous state transitions are to be ignored and the model works on calculating only the instantaneous risk of an individual in the model. As well as this, it is possible that using these types of model will bring out important insights that may be missed by ordinal regression models. With the assumption that the odds proportional, the estimates are the same for each cut-off point on the scale and the overall estimate represents transitions in a single direction, whereas each individual transition and direction of transition is able to be modelled using multi-state models.

Lastly, the computation time needed to fit the model is small, with the multi-state models converging faster than some of the more complex regression models fitted in Chapter 4, and definitely in a shorter time than the latent class analysis models fitted in Chapter 7.

There are some limitations to using multi-state models, including fitted models that would not converge for the data at hand. Models that had censored states and included the treatment variable would not converge to the Hessian matrix and the maximum likelihood values could not be estimated even after several increases in the number of iterations. Secondly, convergence was also an issue for models that had a transition to death at any stage in the model and not just from the adjacent state. Once again, the maximum likelihood values could not be estimated and there was no convergence after numerous increases in the number of iterations and adjustments to starting values.

The use of censored states allowed more individuals to be included in the model; however, the added complexity made the models hard to fit and converge. An alternative to the censored state would have been to impute the missing data.

Another limitation is the lack of an overall treatment effect from the model. Each individual transition has his or her own treatment effect in the first model; however, it is hard to generalise this into one effect for the standard care group and one effect for the oxygen treatment group, although having an effect for each individual transition means the model has good interpretation for the movement of an individual with a specific mRS score.

There is the potential for these methods to be used more widely by practitioners, but at the moment there is a lack of software that can be used for these models, which could be responsible for their limited use (Meira-Machado et al. 2009).

8.4.4 LCGA

A strength of this type of analysis was that it uses the whole mRS scale without loss of information and it is taken from a person-centred approach, unlike other models fitted to the data, which take a variable-centred approach to modelling the data. Using a person-centred approach makes no assumption that the individuals come from a population with a single growth trajectory that can effectively approximate the entire population, but assumes that there are several underlying trajectories into which individuals can be grouped.

There are other strengths of using LCGA, which include the fact that there can be unequal spacing in the observations measured over time, and the model works well even when there is a limited number of time points for the transitions that are being modelled to be observed in. As well as this there are goodness of fit indices and tests, even if there is still some disagreement about which are the best ones to use to decide the number of clusters that should be included in a model.

This type of analysis can have quite a long computation time, especially when there are lots of clusters being attempted to be fitted and more being added to the model. It is important to balance the fact that we would like to fit as many clusters as possible, with the computation time that is needed to fit those clusters. The computational burden and computation time is also increased when there are larger numbers of start values included in the model; this happens when there is non-convergence in the model. It was found in the models fitted here that models with large numbers of clusters that failed to converge would not be considered due to the small numbers of individuals chosen for each cluster. This additional complexity also may not be warranted, as the extra computational time and burden may not benefit the aims of the research (Strauss et al. 2014).

A limitation of the LCGA models fitted to the data is the need for the outcome variable being analysed to have the same number of categories at each of the time points included in the model, so for example the outcome has 7 categories at all time points, rather than 6 at one time point and 7 at the others. Because of this, for the analyses fitted to patients with at least one transition, individuals who had died at 3 months needed to be included in the models, even though they did not undergo a transition themselves, and indeed dominated one of the clusters in each of the different models fitted to the data.

There are also potential convergence issues within the LCGA models fitted to the data, sometimes the maximum likelihood was not able to be replicated, even after increasing the number of random starts. It is known that convergence can be an issue, especially when the numbers of clusters are increasing, with local solutions being found and also non-convergence (Hipp & Bauer 2006).

Finally, although the difference in the recovery trajectories of the two different treatment groups has been compared by fitting models to each group separately, there is no overall treatment effect produced by the model. Therefore, it is hard to quantify the difference in the treatment effect. It is possible to add a treatment effect into the model to estimate a model with all individuals, which then compares the numbers of individuals in the treatment and control groups within each cluster. However, this is still unable to quantify the difference specifically between the treatment and control groups. Obviously, fitting the two groups separately allows you to visually compare the trajectories of the model, and we saw that there were common clusters between these models but also some different ones, and both were reflected in models with all the potential individuals.

8.5 Key decisions

8.5.1 Use of only the mRS

The decision was made to conduct the analysis in this thesis using the mRS as the outcome of interest, and to ignore the other stroke scales that were recorded in the datasets. There were several reasons for this. First, the mRS is the most commonly used primary outcome of a stroke trial, which would mean that by using the mRS for the analysis in this thesis, it has the best chance of being able to be used and make recommendations for stroke trials in the future. Secondly, there was a great interest in the category on the scale for death. This is not a common occurrence, as was seen in a review of the literature of outcome scores that include death within them. This inclusion of death posed statistical challenges, especially when the ordinality of the scale had the potential to be questioned due to its inclusion. Usually in a follow-up study, once patients have died they are dropped out of the study, with the date of death recorded along with other information such as the cause of death. They will then make no further contribution to the study follow-up. However, with the inclusion of a category for death this will not be the case as the individual will have a score to contribute to the model.

Research conducted by Diehr et al (2015) proposed several approaches for incorporating death when considering health-related variables in longitudinal studies (Diehr et al. 1995). These included adding a category for death to the variable of interest as an extreme value on the scale or dichotomising the scale into healthy and not healthy and including those individuals who have died as not healthy. Death in our case is the most extreme end of the scale and the research conducted shows that binary regression loses information when the scale is ordinal and, according to Bath et al (2007), is not the recommended way to analyse the mRS. It has been advocated including the score for death so long as care was taken, as strategies that gave death less influence tended to show more favourable changes in health over time (Diehr et al. 1995).

It has been seen throughout this thesis that the mRS has remained largely constant across the follow-up period of the trial. This suggests that the mRS may be insensitive to change, when it comes to conducting a longitudinal analysis. This may be one of the reasons why there has been little research done using the mRS in a longitudinal setting. The lack of movement in the mRS raises the question of whether the mRS should be considered at all for the longitudinal analysis of stroke trials. Would it be better if the BI was used, although this scale would bring its own problems? For example, any analysis would have to try and account for the floor and ceiling effects of the BI.

8.5.2 Use of SO$_2$S and NINDS data

The data that were used in this thesis came from two randomised controlled trials in stroke. Both completed follow-ups on the individuals recruited in the study for 12 months. Both of these datasets were chosen for practical reasons. The SO$_2$S trial was a large trial with over 8,000 individuals recruited to receive oxygen therapy. The recruitment for the trial had just finished when the idea for this project was developed. The size of the trial was very promising with such a large number of individuals recruited. There was one major problem with using the SO_2S trial data, which was that the follow-up period of the trial was relatively short for a longitudinal study and this left us with a limited series of data with which to work. Therefore, the analysis that has been conducted has tried to make the best of the available data and to present the most efficient analyses that are possible using data with such a short follow-up. This short series of repeated measures meant that there were some methods that would have been appropriate for the data, for example a joint model of ordinal regression and the time to event data, which it would have been advantageous to apply; however, it was not possible to fit reasonable models with the limited data available. Hence, a second dataset with more data points was also used, despite the length of the follow-up for the NINDS trial being the same as in the SO_2S trial. This is due to the fact that the trial recorded the mRS at a much earlier time point to the SO_2S trial, between 7 and 10 days. At this time point in the SO_2S study, the only outcome recorded was the NIHSS.

This particular dataset was chosen for practical reasons, as it was open access and similar analyses to those applied in this thesis to the SO_2S data had previously been applied to the NINDS data (ordinal regression in a joint model and multi-state model (Li et al. 2010, Cassarly 2015)). This did mean however, that we were not able to apply all of the methods of the thesis to both datasets as it would be conducting the same analysis twice. A limitation of the NINDS dataset is that it is very small in comparison to the SO_2S trial, with the size of the trials less than 10% of the SO_2S trial. Both trials recruited from multiple centres, the SO_2S trial based in the UK and the NINDS from acute stroke centres in USA.

8.5.3 Decision to focus solely on treatment effect

A decision was made to focus on the different models as applied solely to the treatment variable. This was because when the project was developed, the main results SO_2S trial analysis had not been published. The trial analysis plan included a longitudinal analysis of the data and so it was decided that in order to avoid duplication by conducting analyses that could be included in the trial analysis, taking one specific focus would be best. This is why only the treatment variable was considered. It was considered whether to look at a completely different covariate, such as the effect of a thrombolysis, as 17% of individuals in the data had this treatment, but this would then cause issues with randomisation (which was on the basis of oxygen treatment, not thrombolysis), and so it was decided that using just a treatment effect would be suitable and the analysis undertaken would be appropriately different from the trial analysis.

The results of the SO_2S trial found that the treatment had no effect on the primary outcome, which was the mRS at 3 months. There were discussions about whether, since it was known that treatment was ineffective, the trial could be used as one whole natural cohort for a study. However, it was decided that the effect of different models in the treatment variable was to be the interest of this project. The project was not interested in what the effect of treatment was, only how the different models that were applied to the data changed the effect of treatment that was calculated and presented. The fact that treatment was always going to be non-significant in the model could be accepted, the inclusion of treatment, time and the interaction of treatment and time pre-specified as the only predictor variables included in the model.

8.5.4 Choice of software

The thesis used several different software packages in order to fit the different models that were required. The ordinal and binary logistic regression models were completed in Stata. However, in order to fit the proportional odds model in a stepwise form, as was shown in section 4.7.3, the models needed to be fitted using a SAS program, Appendix B. Additionally, as a sensitivity analysis, the ordinal regression model with proportional odds fitted in Stata was also fitted using the SAS program and it was found that there was very little difference in the estimates that both models produced. This meant that it could be sure that although fitted using different software the models were comparable. It is also why the -2 log-likelihood was compared as this was the value produced by the SAS package. It was important to check that both statistical packages produced the same estimates for the proportional odds model, in order to be able to make comparison with the new models fitted in SAS and the previous model fitted in Stata. If these models had produced different results, this would have been concerning as it was the same model being fitted to the same dataset just in different packages.

SAS was also used in order to generate the discrete ordinal data that were used in the simulation study. The reason that this package was chosen over Stata was that a paper was found that generated longitudinal ordinal data (Ibrahim & Suliadi 2011), in the exact form that was required for the simulation study. This study also went on the add missingness to the data, but then went on to look at multiple imputation methods in order to deal with it.

R was used in order to fit the multi-state models to the data. There was a software package that had previously been written, the msm package, that allowed multi-state Markov models to be fitted to the data, with or without censored states and it also allowed for the inclusion of predictor variables. The computation time in this package was small for these models; however, with the attempted inclusion of several predictor variables in the more complex models that were considered, convergence was an issue. The number of iterations was increased, as well as reducing the size of the data set by taking a random sample in order to see if converge would occur, but this was not the case.

Initially, R was also considered in order to estimate the LCGA models. However, the computation time for fitting models that contained only 4 clusters was upwards of 24 hours and so it was decided that a better software package needed to be used. Mplus was chosen because it is known for its flexibility in fitting different forms of LCA and was chosen over LatentGOLD, which is another common package used for LCA and its extensions. Mplus was chosen to be used over

LatentGOLD because of online support forums that are available to discuss problems that arose with the analysis that was conducted.

8.6 Future research

There are many challenges that have been faced with having such a limited number of follow-up points. Having only considered the mRS in isolation, it was found that using ordinal regression models to model the repeated measures data was better than dichotomising the scale and fitting logistic regression models. There are several studies that advocated the idea of modelling more than one outcome variable simultaneously, even if they consist of data of mixed types.

Teixeira-Pinto and Mauri (2011) described how clinical trials and observational studies involving multiple outcomes are common in medical research (Teixeira-Pinto & Mauri 2011). It is considered that disease complexity may often not be adequately characterised by using a single outcome, and several aspects of the patients' response must be considered. Furthermore, one can examine how the association between outcomes evolves over time, how outcome-specific evolutions are related to each other (Fieuws & Verbeke 2004), or one can even consider joint testing of a treatment effect on a set of outcomes. These situations require a joint model for all outcomes.

Using a joint analysis avoids the need for multiple testing and leads to a natural global test, which results in increased power and better control of type 1 error rates. Significant efficiency gains over separate univariate analyses have also been reported, especially in settings where there are missing data. By modelling all the outcomes jointly, we may be able to correct for this bias through the correlation between the outcomes obtained from the complete cases, in which all the outcomes are observed. To achieve this, Verbeke et al (2012) suggest that the final modelling choice will depend on the research questions, the data structure (balanced/unbalanced), the desire to model observed outcomes rather than latent constructs, the dimension of the problem, and the nature of the outcomes (Verbeke et al. 2012).

So should an investigator just pick one outcome of choice, the mRS? An alternative would be to conduct an analysis with more than one stroke scale included. If the treatment has a beneficial effect on all scales, then combining them will increase the power to demonstrate the advantage of the treatment. Tilley et al. (1996) combined four scales in the trial of rT-PA as a treatment in acute stroke conducted by the National Institute of Neurological Disorders and Stroke (Tilley et al. 1996). The results of the trial were positive and the approach was well received. The approach used by Tilley et al. combined three analyses using generalised estimating equations (based on an independence covariance structure). That is, one would analyse as if the three scores were independent, but adjust the standard error of the treatment effect estimate using the sandwich estimator. But within this study all outcomes were dichotomised and we prefer to retain the ordinal nature of the mRS.

We propose to consider the approach of Ivanova (2016) who describes approaches for the joint modelling of different types of responses. For a bivariate model involving a continuous and ordinal outcome we need to define the core components to developing a joint model (Ivanova et al. 2016). Obviously these methods proposed would all benefit from a longer series of follow-up points to be included in the model, in order to provide more information of where individuals change mRS scores within the follow-up. Using this approach could provide more information as more other stroke scales may not be as insensitive to change as the mRS scale.

8.7 Conclusions

The conclusions of the research conducted in this project are in agreement with previous research that concludes that because the mRS is an ordinal scale, the best way of including it within any analysis is to keep it in its ordinal form. This project has shown that there are methods that can be used in order to conduct a longitudinal analysis of the mRS.

These methods have application not only within future stroke research but also with other clinical areas that have assessment scales that are ordinal in nature, especially if there is a category of death in the model that may call the ordinality of the scale into question. All of the methods included within the thesis can also be applied within cohort studies, as all the methods are suitable for longitudinal analyses, and are able to be extended out of a trial setting. Specifically, LCGA may be more applicable in a cohort study, rather than a trial setting, due to what is usually a lengthier follow-up and there are therefore more potential time points to use in order to construct the trajectories, although the method is still valid within clinical trials.

These methods have highlighted that the use of longitudinal analysis applied to repeated measures data needs to be encouraged, and that the results obtained, especially when considering the regression models do not change the conclusions drawn about the treatment effect when using repeated measures models compared to regression models at a single fixed time point. This means that statisticians should be encouraged to plan to conduct a longitudinal analysis on all the follow-up time points that they record rather than taking a single fixed time point to conduct the analysis on, whilst using the ordinal scale in its full form in order to draw the best conclusions from the data and use as much information from the outcomes recorded as possible.

Out of the proposed methods used in the thesis, the method that I would select for use in a longitudinal stroke trial would be the Markov multi-state model. This method uses the ordinal scale in such a way that is meaningful and is easily interpretable for clinicians. One of the main benefits of fitting this type of model is that it is able to estimate each transition separately allowing for a more accurate estimation of treatment depending on what your starting value of the mRS is. The use of the scale as states within the model allows for various transitions to occur and it works well with no baseline value because the memoryless property means that a transition is not dependent on where the patient has moved previously, only where the patient is at the time of the transition. The estimates of risk calculated are easily interpretable and the treatment effect can also be obtained easily.

8.8 Summary

The research conducted highlights the importance of trying to find a suitable model to try and capture the change in the mRS over time, due to the limited movement of patients within the scale. As the recommended scale to assess disability in stroke patients after they have left hospital, it is an important scale for which to determine optimum analyses. The research in this thesis has shown that the application of different models will provide different values of the treatment effect when considered in the longitudinal analysis of the mRS, and therefore consideration should be taken in advance of the clinical question that is trying to be answered in order to analysis a scale such as the mRS, which is ordinal in nature and has a category included within the scale for death.

Following on from the research by Bath et al (2007), it is also recommended that the mRS is analysed in its full ordinal scale for longitudinal regression models as well as at the single time point previously considered. Although, this research is not widely followed give the research conducted by Nunn et al. (2016), however hopefully individuals conducting stroke research will take heed and will start to use the mRS in an ordinal form more often.

The use of the ordinal scale with death calls into question the ordinality of categories and therefore some individuals may feel like the scale is no longer ordinal. In this case, a multi-state model is a good modelling tool to use as each category is considered individually, but not compared to a single reference category, like multinomial modelling and therefore the comparisons made are more clinically meaningful.

Finally, researchers may wish to analyse the data from a patient-level perspective, which is what the LCGA models do, allowing for individuals who a have similar trajectories to be clustered together. This method is good because it produces trajectories of individuals through time. The only downside to using these models is that changes do not occur often on the mRS scale, therefore the LCGA is not very sensitive to change, however it is a useful tool for stroke research because it is able to be applied to other stroke scales that include more variation in the scores seen over time.

To conclude, this thesis has investigated potential methods that could be used for the longitudinal analysis of future stroke trials, using novel techniques that have not often been applied in stroke trials, but have successfully been applied elsewhere that are applicable in the analysis of the mRS and potential other stroke scales that are used to assess the recovery of an individual after they have had a stroke.

References

Akaike, H. 1987, “Factor analysis and AI”, *Psychometrika,* no. 52, pp. 317.

Ali, K., Warusevitane, A., Lally, F., Sim, J., Sills, S., Pountain, S., Nevatte, T., Allen, M. & Roffe, C. 2013, “The Stroke Oxygen Pilot Study: A randomized controlled trial of the effects of routine oxygen supplementation early after acute stroke—effect on key outcomes at six months”, *PLOS ONE,* pp. e59274.

Ali, M., Bath, P.M.W., Curram, J., Davis, S.M., Diener, H.C., Donnan, G.A., Fisher, M., Gregson, B.A., Grotta, J., Hacke, W., Hennerici, M.G., Hommel, M., Kaste, M., Marler, J.R., Sacco, R.L., Teal, P., Wahlgren, N.G., Warach, S., Weir, C.J. & Lees, K.R. 2007, “The Virtual International Stroke Trials Archive”, *Stroke,* vol. 38, no. 6, pp. 1905-1910.

Altman, D.G. & Royston, P. 2006, “The cost of dichotomising continuous variables”, *BMJ: British Medical Journal,* vol. 332, no. 7549, pp. 1080.

Anderson, J.A. 1984, “Regression and ordered categorical variables”, *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 46, no. 1, pp. 1-30.

Armstrong, B.G. & Sloan, M. 1989, “Ordinal regression models for epidemiologic data”, *American Journal of Epidemiology,* vol. 129, no. 1, pp. 191-204.

Bamford, J.M., Sandercock, P.A., Warlow, C.P. & Slattery, J. 1989, “Inter-observer agreement for the assessment of handicap in stroke patients” *Stroke,* vol. 20, no. 6, pp. 828-828.

Banks, J.L. & Marotta, C.A. 2007, “Outcomes validity and reliability of the modified Rankin scale: Implications for stroke clinical trials: A literature review and synthesis” *Stroke*, vol. 38, no.3, pp. 1091-1096

Bath, P.M., Gray, L.J., Collier, T., Pocock, S. & Carpenter, J. 2007, “Can we improve the statistical analysis of stroke trial?”, *Stroke*, vol. 38, no. 6, pp.1911-1915

Bath, P.M., Lees, K.L., Schellinger, P.D., Altman, H., Bland, M., Hogg, C., Howard, G. & Saver, J. 2012, “Statistical analysis of the primary outcome in acute stroke trials”, *Stroke,* vol. 43, no. 4, pp. 171.

Bowling, A. 2005, *Measuring Health,* 3rd Edition, Open University Press, Buckingham.

Bowling, A. 2001, *Measuring Disease,* 2nd Edition, Open University Press, Buckingham.

Brant, R. 1990, “Assessing proportionality in the proportional odds model for ordinal logistic regression”, *Biometrics,* vol. 46, no. 4, pp. 1171-1178.

Briere F.N., Rohde P., Stice E. & Morizot J. 2016, “Group-based symptom trajectories in indicated prevention of adolescent depression” *Depression and Anxiety,* vol. 33, no. 5, pp. 444-451.

Brott, T., Adams, H.P., Olinger, C.P., Marler, J.R., Barsan, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R. & Hertzberg, V. 1989, "Measurements of acute cerebral infarction: a clinical examination scale", *Stroke,* vol. 20, no. 7, pp. 864-870.

Bulpitt, C.J. 1982, "Quality of life in hypertensive patients" in *Hypertensive cardiovascular disease: pathophysiology and treatment*, pp. 929.

Busija, L., Liew, D., Yan, B., Weir, L., Hand, P. & Davis, S. 2013, "Recovery trajectories in acute non-fatal stroke", *Stroke*, vol.44, Supplement 1, ATP384

Carriere, I. & Bouyer, J. 2006, "Random-effect models for ordinal responses: application to self-reported disability amoung older persons", *Epidemiology and Public Health,* vol. 54, pp. 61-72.

Cassarly, C. 2015, *Multistate Markov models for analysis of the modified Rankin scale in phase III clinical trials of acute stroke therapy*, Medical University of South Carolina.

Clark, S.L. & Muthén, B.O. 2009, "Relating latent class analysis results to variables not included in the analysis" *http://www.statmodel.com.*

Coggon, D., Rose, G. & Barker, D. 2007, "Longitudinal studies" in *Epidemiology for the uninitiated*, 4th Edition, BMJ.

Collins, L.M. & Lanza, S.T. 2010, *Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences,* 1st Edition, Wiley, United States.

Cozby, P.C. & Bates, S.C. 2012, *Methods in behavioral research,* 11th Edition, McGraw-Hill Education.

Crowther, M.J. 2016, *MULTISTATE: Stata module to perform multi-state survival analysis*.

Cui, L.Y., Zhu, Y.C., Gao, S., Wang, J.M., Peng, B., Ni, J., Zhou, L.X., He, J. & Ma, X.Q. 2013, "Ninety-day administration of dl-3-n-butylphthalide for acute ischemic stroke: a randomized, double-blind trial" *Chinese Medical Journal,* vol. 126, no. 18, pp. 3405.

das Nair, R., Moreton, B.J. & Lincoln, N.B. 2011, "Rasch analysis of the Nottingham extended activities of daily living scale", *Journal of Rehabilitation Medicine,* vol. 43, no. 10, pp. 944-950.

de Vos, B.,C., Runhaar, J., Verkleij, S.P.J., van Middelkoop, M. & Bierma-Zeinstra, S. 2014, "Latent class growth analysis successfully identified subgroups of participants during a weight loss intervention trial", *Journal of Clinical Epidemiology,* vol. 67, no. 8, pp. 947-51.

de Wreede, L.C., Fiocco, M. & Putter, H. 2011, "mstate: An R package for the analysis of competing risks and multi-state models", *Journal of Statistical Software,* vol. 38, no. 7, pp. 1.

Department of Health 2007, *National Stroke Strategy*.

Department of Health 2005, *Reducing brain damage: Faster access to better stroke care,* National Audit Office, London.

Diehr, P., Johnson, L.L., Patrick, D.L. & Psaty, B. 2005, “Methods for incorporating death into health-related variables in longitudinal studies”, *Journal of Clinical Epidemiology,* vol. 58, no. 11, pp. 1115-1124.

Diehr, P., Patrick, D., Hedrick, S., Rothman, M., Grembowski, D., Raghunathan, T.E. & Beresford, S. 1995, “Including deaths when measuring health status over time”, *Medical Care,* vol. 33, no. 4, pp. AS164-AS172.

Dona, G., Preatoni, E., Cobelli, C., Rodano, R. & Harrison, A.J. 2009, “Application of functional principal component analysis in race walking: an emerging methodology”, *Sports Biomechanics,* vol. 8, no. 4, pp. 284-301.

Donneau, A.F., Mauer, M., Molenberghs, G. & Albert, A. 2015, “A simulation study comparing multiple imputation methods for incomplete longitudinal ordinal data”, *Communications in Statistics - Simulation and Computation,* vol. 44, no. 5, pp. 1311-1338.

Fearon, P., McArthur, K.S., Garrity, K., Graham, L.J., McGroarty, G., Vincent, S. & Quinn, T.J. 2012, “Pre-stroke modified Rankin stroke scale has moderate inter-observer reliability and validity in an acute stroke setting”, *Stroke*, vol.43, no.12, pp.3184-3188

Feigin, V.L., Forouzanfar, M.H. & Krishnamurthi, R. 2014, “Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010”, *Lancet*, vol.383, no.9913, pp. 245-254

Fieuws, S. & Verbeke, G. 2004, “Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach”, *Statistics in Medicine,* vol. 23, no. 20, pp. 3093-3104.

Ford N.D., Martorell R., Mehta N.K., Ramirez-Zea M. & Stein A.D. 2016, “Life-course body mass index trajectories are predicted by childhood socioeconomic status but not exposure to improved nutrition during the first 1000 days after conception in Guatemalan adults”, *Journal of Nutrition,* vol. 146, no. 11, pp. 2368-2374.

Ghandehari, K. 2013, “Challenging comparison of stroke scales”, *Journal of Research in Medical Sciences,* vol. 18, no. 10.

Gill T.M., Guralnik J.M., Pahor M., Church T., Fielding R.A., King A.C., Marsh A.P., Newman A.B., Pellegrini C.A., Chen S.-H., Allore H.G. & Miller M.E. 2016, “Effect of structured physical activity on overall burden and transitions between states of major mobility disability in older persons: Secondary analysis of a randomized trial”, *Annals of Internal Medicine,* vol. 165, no. 12, pp. 833-840.

Hanger, H.C., Fogarty, B., Wilkinson, T.J. & Sainsbury, R. 2000, “Stroke patients' views on stroke outcomes: death versus disability”, *Clinical Rehabilitation,* vol. 14, no. 4, pp. 417-424.

Harbison, J., Hossain, O., Jenkinson, D., Davis, J., Louw, S.J. & Ford, G.A. 2003, “Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the face arm speech test”, *Stroke,* vol. 34, no. 1, pp. 71-76.

Harrison, J.K., McArthur, K.S. & Quinn, T.J. 2013, “Assessment scales in stroke: clinimetric and clinical considerations”, *Clinical Interventions in Aging,* no. 8, pp. 201-211.

Hedeker, D. & Mermelstein, R.J. 2000, “Analysis of longitudinal substance use outcomes using ordinal random-effects regression models”, *Addiction,* vol. 95, no. Supplement 3, pp. S381-S394.

Hipp, J.R. & Bauer, D.J. 2006, “Local solutions in the estimation of growth mixture models”, *Psychological Methods,* vol. 11, no. 1, pp. 36-53.

Huybrechts, K.F. & Caro, J.J. 2007, “The Barthel index and modified Rankin scale as prognostic tools for long-term outcomes after stroke: a qualitative review of the literature”, *Current Medical Research and Opinion,* vol. 23, no. 7, pp. 1627-1636.

Ibrahim, N.A. & Suliadi, S. 2011, “Generating correlated discrete ordinal data using R and SAS IML”, *Computer Methods and Programs in Biomedicine,* vol. 104, no. 3, pp. e122-e132.

Intercollegiate Stroke Working Party 2012, *National Clinical Guideline for Stroke,* 4th Edition, London: Royal College of Physicians.

Ivanova, A., Molenberghs, G. & Verbeke, G. 2016, “Mixed models approaches for joint modelling of different types of responses”, *Journal of Biopharmaceutical Statistics,* vol. 26, no. 4, pp. 601-618.

Javali, S.B. & Pandit, P.V. 2010, “A comparison of ordinal regression models in an analysis of factors associated with periodontal disease”, *Journal of Indian Society of Periodontology,* vol. 14, no. 3, pp. 155-159.

Jeffries, N. 2003, “A note on “testing the number of components in a normal mixture”, *Biometrika,* no. 90, pp. 991.

Jennett, B. & Bond, M. 1975, “Assessment of outcome after severe brain damage” *Lancet,* vol. 1, no. 7905, pp. 480-484.

Johnston, K. & Wagner, D. 2006, “Relationship between 3-Month National Institutes of Health Stroke Scale Score and dependence in ischemic stroke patients”, *Neuroepidemiology,* vol. 27, no. 2, pp. 96-100.

Johnston, S., Gress, D.R., Browner, W.S. & Sidney, S. 2000, “Short-term prognosis after emergency department diagnosis of TIA”, *JAMA,* vol. 284, no. 22, pp. 2901-2906.

Joseph, R. 2015, *Statistical analysis of longitudinal randomized clinical trials with missing data: a comparison of approaches*, Keele University.

Kapetanakis, V., Matthews, F.E. & Hout, A. 2013, “ A semi-Markov model for stroke with piecewise-constant hazards in the presence of left, right and interval censoring” *Statistics in Medicine,* vol. 32, no. 4, pp. 697.

Karnofsky, D.A., Abelmann, W.H., Craver, L.F. & Burchenal, J.H. 1948, “The use of the nitrogen mustards in the palliative treatment of carcinoma. With particular reference to bronchogenic carcinoma”, *Cancer,* vol. 1, no. 4, pp. 634-656.

Knapp, T.R. 1990, “Treating ordinal scales as interval scales: an attempt to resolve the controversy”, *Nursing Research,* vol. 39, no. 2, pp. 121-123.

Kurtzke, J.F. 1983, “Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)”, *Neurology,* vol. 33, no. 11, pp. 1444.

Kwiatkowski, T.G., Libman, R.B., Frankel, M., Tilley, B.C., Morgenstern, L.B., Lu, M., Broderick, J.P., Lewandowski, C.A., Marler, J.R., Levine, S.R. & Brott, T. 1999, “Effects of tissue plasminogen activator for acute ischemic stroke at one year”, *New England Journal of Medicine,* vol. 340, no. 23, pp. 1781-1787.

Lansberg, M.G., Schrooten, M., Bluhmki, E., Thijs, V.N. & Saver, J.L. 2009, “Treatment time-specific number needed to treat estimates for tissue plasminogen activator therapy in acute stroke based on shifts over the entire range of the modified Rankin scale”, *Stroke,* vol. 40, no. 6, pp. 2079-2084.

Lansky, S.B., List, M.A., Lansky, L.L., Ritter-Sterr, C. & Miller, D.R. 1987, “The measurement of performance in childhood cancer patients”, *Cancer,* vol. 60, no. 7, pp. 1651-1656.

Lees, K.R., Bath, P.W., Schellinger, P.D., Kerr, D.M., Fulton, R., Hacke, W., Matchar, D., Sehra, R. & Toni, D. 2012, “Contemporary outcome measures in acute stroke research: Choice of primary outcome measure”, *Stroke,* vol. 43, no. 4, pp. 1163-1170.

Levine, S.R. & Hill, M.D. 2014, “NeuroThera effectiveness and safety trial 3”, *Stroke,* vol. 45, no. 11, pp. 3175-3177.

Li, N., Elashoff, R.M., Li, G. & Saver, J. 2010, “Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial”, *Statistics in Medicine,* vol. 29, no. 5, pp. 546-557.

Lindley, R.I., Waddell, F., Livingstone, M., Sandercock, P., Dennis, M.S., Slattery, J., Smith, B. & Warlow, C. 1994, “Can simple questions assess outcome after stroke?”, *Cerebrovascular Diseases,* vol. 4, no. 4, pp. 314-324.

Liu, D.C. & Nocedal, J. 1989, “On the limited memory BFGS method for large scale optimization”, *Mathematical Programming,* vol. 45, no. 1, pp. 503.

Lo, Y., Mendell, N.R. & Rubin, D.B. 2001, “Testing the number of components in a normal mixture”, *Biometrika,* vol. 88, no. 3, pp. 767.

Long, J.S. & Freese, J. 2006, *Regression models for categorical dependent variables using Stata,* 2nd Edition, Stata Press, Texas.

Lovett, J.K., Dennis, M.S., Sandercock, P.A.G., Bamford, J., Warlow, C.P. & Rothwell, P.M. 2003, “Very early risk of stroke after a first transient ischemic attack”, *Stroke,* vol. 34, no. 8, pp. 138-140.

Lunt, M. 2001, “Stereotype ordinal regression”, *Stata Technical Bulletin,* no. 61, pp. 12.

Lyden, P., Lu, M., Jackson, C., Marler, J., Kothari, R., Brott, T. & Zivin, J. 1999, “Underlying structure of the National Institutes of Health Stroke Scale: Results of a factor analysis”, *Stroke,* vol. 30, no. 11, pp. 2347-2354.

Lyden, P.D., Lu, M., Levine, S.R., Brott, T.G., Broderick, J. & the NINDS rtPA Stroke Study Group 2001, “A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: Preliminary reliability and validity”, *Stroke,* vol. 32, no. 6, pp. 1310-1317.

Mackay, J. & Mensah, G. 2004, *WHO :The Atlas of heart disease and stroke section 15: Global burden of stroke*. Available: http://www.who.int/cardiovascular_diseases/en/cvd_atlas_15_burden_stroke.pdf?ua=1 [2014, 17,03].

Mahoney, F.I. & Barthel, D.W. 1965, “Functional evaluation: the Barthel index”, *Maryland State Medical Journal,* no. 14, pp. 61-65.

McArthur, K., Beagan, M.L.C., Degnan, A., Howarth, R.C., Mitchell, K.A., McQuaige, F.B., Shannon, M.A.C., Stott, D.J. & Quinn, T.J. 2013, “Properties of proxy-derived modified Rankin scale assessment”, *International Journal of Stroke,* vol. 8, no. 6, pp. 403-407.

McHugh, G.S., Butcher, I., Steyerberg, E.W., Marmarou, A., Lu, J., Lingsma, H.F., Weir, J., Maas, A.I.R. & Murray, G.D. 2010, “A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project”, *Clinical Trials,* vol. 7, no. 1, pp. 44-57.

McLachlan, G.J. & Peel, D. 2000, *Finite mixture models,* 1st Edition, Wiley, New York.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. & Andersen, P.K. 2009, “Multi-state models for the analysis of time-to-event data” *Statistical Methods in Medical Research.* vol. 18, no.2, pp195-222

Muir, K.W., Weir, C.J., Murray, G.D., Povey, C. & Lees, K.R. 1996, “Comparison of neurological scales and scoring systems for acute stroke prognosis”, *Stroke,* vol. 27, no. 10, pp. 1817-1820.

Muthén, L.K. & Muthén, B.O. 2009, *Mplus Users guide: statistical analysis with latent variables*.

Muthén, L.K. & Muthén, B.O. 1998-2015, *Mplus User’s Guide,* 7th edition, Muthén & Muthén, Los Angeles, CA.

Newman, G.C., Bang, H., Hussain, S.I. & Toole, J.F. 2007, “Association of diabetes, homocysteine, and HDL with cognition and disability after stroke”, *Neurology,* vol. 69, no. 22, pp. 2054-2062.

NHS Choices 2012a*, Stroke diagnosis*. Available: http://www.nhs.uk/Conditions/Stroke/Pages/Diagnosis.aspx [2014, 17, 03].

NHS Choices 2012b, *Treating stroke*. Available: http://www.nhs.uk/Conditions/Stroke/Pages/treatment.aspx [2014, 18, 03].

NHS Choices 2012c*, Stroke recovery*. Available: http://www.nhs.uk/Conditions/Stroke/Pages/recovery.aspx [2014, 17, 03].

Nouri, F.M. & Lincoln, N.B. 1987, “An extended activities of daily living scale for stroke patients”, *Clinical Rehabilitation,* vol. 1, no. 4, pp. 301-305.

Nunn, A., Bath, P.M. & Gray, L.J. 2016, "Analysis of the modified Rankin scale in randomised controlled trials of acute ischaemic stroke: a systematic review", *Stroke Research and Treatment,* vol. 2016.

Nylund, K.L., Asparouhov, T. & Muthén, B.O. 2007, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study" *Structural Equation Modeling,* no. 14, pp. 535.

O'Keeffe, A.G., Tom, B.D.M. & Farewell, V.T. 2011, "A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments", *Journal of the Royal Statistical Society C Applied Statistics*, vol. 60, no. 5, pp. 675.

O'Connell, A. 2006, *Logistic regression models for ordinal response variables,* 1st Edition, SAGE Publications, California.

Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T. & Carbone, P.P. 1982, "Toxicity and response criteria of the Eastern Cooperative Oncology Group", *American Journal of Clinical Oncology,* vol. 5, no. 6.

Olsson, G., Lubsen, J., van Es, G.S. & Rehnqvist, N. 1986, "Quality of life after myocardial infarction: effect of long term metoprolol on mortality and morbidity" *BMJ (Clinical research ed.),* vol. 263, no. 6534, pp. 1491.

Palace, J., Bregenzer, T., Tremlett, H., Oger, J., Zhu, F., Boggild, M., Duddy, M. & Dobson, C. 2014, "UK multiple sclerosis risk-sharing scheme: a new natural history dataset and an improved Markov model", *BMJ Open,* vol. 4, no. 1.

Pan, J.H., Song, X.Y., Lee, S.Y. & Kwok, T. 2008a, "Longitudinal analysis of quality of life for stroke survivors using latent curve models", *Stroke,* vol. 39, no. 10, pp. 2795-2802.

Pan, S.L., Lien, I.N., Yen, M.F., Lee, T.K. & Chen, T.H.H. 2008b, "Dynamic aspect of functional recovery after stroke using a multistate model", *Archives of Physical Medicine and Rehabilitation,* vol. 89, no. 6, pp. 1054-1060.

Pfeiffer, K., Beische, D., Hautzinger, M., Berry, J.W., Wengert, J., Hoffrichter, R., Becker, C., van Schayck, R. & Elliott, T.R. 2014, "Telephone-based problem-solving intervention for family caregivers of stroke survivors: a randomized controlled trial", *Journal of Consulting and Clinical Psychology.* vol.82, no.4, pp.628-643

Ploubidis, G.B., Abbott, R.A., Huppert, F.A., Kuh, D., Wadsworth, M.E.J. & Croudace, T.J. 2007, "Improvements in social functioning reported by a birth cohort in mid-adult life: A person-centred analysis of GHQ-28 social dysfunction items using latent class analysis", *Personality and Individual Differences,* vol. 42, no. 2, pp. 305.

Poisson, S.N., Johnston, S.C. & Josephson, S.A. 2010, "Urinary tract infections complicating stroke", *Stroke,* vol. 41, no. 4, pp. 180-184.

Potter, J., Mistri, A., Brodie, F., Chernova, J., Wilson, E., Jagger, C., James, M., Ford, G. & Robinson, T. 2009, "Controlling Hypertension and Hypotension Immediately Post-Stroke (CHHIPS) - a randomised controlled trial", *Health Technology Assessment,* vol. 13, no. 9, pp. 96.

Quinn, T.J., Langhorne, P. & Stott, D.J. 2011, “Barthel index for stroke trials: Development, properties, and application”, *Stroke,* vol. 42, no. 4, pp. 1146-1151.

Rankin, J. 1957, “Cerebral vascular accidents in patients over the age of 60. II. Prognosis”, *Scottish Medical Journal,* no. 5, pp. 200-215.

Roffe, C., Ali, K., Warusevitane, A., Sills, S., Pountain, S., Allen, M., Hodsoll, J., Lally, F., Jones, P. & Crome, P. 2011, “The SOS pilot study: a RCT of routine oxygen supplementation early after acute stroke—effect on recovery of neurological function at one week”, *PLOS ONE,* vol. 6, no. 5, pp. e19113.

Roffe, C., Nevatte, T., Crome, P., Gray, R., Sim, J., Pountain, S., Handy, L. & Handy, P. 2014, “The Stroke Oxygen Study (SO2S) - a multi-center study to assess whether routine oxygen treatment in the first 72 hours after a stroke improves long-term outcome: study protocol for a randomized controlled trial”, *Trials,* no. 15, pp. 99.

Roozenbeek, B., Lingsma, H.F., Perel, P., Edwards, P., Roberts, I., Murray, G.D., Maas, A.I.R. & Steyerberg, E.W. 2011, “The added value of ordinal analysis in clinical trials: an example in traumatic brain injury*”*, *Critical Care,* vol. 15, no. 3, pp. 127.

Rotolo F., Dunant A., Le, C.T., Pignon J.-P. & Arriagada R. 2014, "Adjuvant cisplatin-based chemotherapy in nonsmall-cell lung cancer: New insights into the effect on failure type via a multistate approach”, *Annals of Oncology,* vol. 25, no. 11, pp. 2162-2166.

Royal College of Physicians, Clinical Effectiveness and Evaluation Unit 2014, *Sentinel stroke national audit programme (SSNAP) clinical audit July - September 2014 public report.*

Rubin, D.B. 1976, “Inference and missing data”, *Biometrika,* vol. 63, no. 3, pp. 581-592.

Saint-Pierre, P., Combescure, C., Daurès, J.P. & Godard, P. 2003, “The analysis of asthma control under a Markov assumption with use of covariates”, *Statistics in Medicine,* vol. 22, no. 24, pp. 3755-3770.

Sajobi, T.T., Zhang, Y., Menon, B.K., Goyal, M., Demchuk, A.M., Broderick, J.P. & Hill, M.D. 2015, “Effect size estimates for the ESCAPE trial: Proportional odds regression versus other statistical methods” *Stroke,* vol. 46, no. 7, pp. 1800-1805.

Sarker, S.J., Rudd, A.G., Douiri, A. & Wolfe, C.D.A. 2012, “Comparison of 2 extended activities of daily living scales with the Barthel index and predictors of their outcomes: Cohort study within the south London stroke register (SLSR)”, *Stroke,* vol. 43, no. 5, pp. 1362-1369.

Saver, J.L. 2011, “Optimal endpoints for acute stroke therapy trials: best ways to measure treatment effects of drugs and devices”, *Stroke*, vol.42, no.8, pp.2356-2362.

Schafer, J.L. & Graham, J.W. 2002, “Missing data: Our view of the state of the art”, *Psychological Methods,* vol. 7, no. 2, pp. 147-177.

Schepers, V.P.M., Ketelaar, M., Visser-Meily, J., Dekker, J. & Lindeman, E. 2006, “Responsiveness of functional health status measures frequently used in stroke research”, *Disability Rehabilitation,* vol. 28, no. 17, pp. 1035-1040.

Scherphof C.S., Van Den, E.R., Lugtig P., Engels R.C.M.E. & Vollebergh W.A.M. 2014, "Adolescents' use of nicotine replacement therapy for smoking cessation: Predictors of compliance trajectories", *Psychopharmacology,* vol. 231, no. 8, pp. 1743-1752.

Schwarz, G. 1978, "Estimating the dimension of a model", *Annals of Statistics,* no. 6, pp. 461.

Sclove, S.L. 1987, "Application of model-selection criteria to some problems in multivariate analysis", *Psychometrika,* no. 52, pp. 333.

Silverwood, R.J., Nitsch, D., Pierce, M., Kuh, D. & Mishra, G.D. 2011, "Characterizing longitudinal patterns of physical activity in mid-adulthood using latent class analysis: Results from a prospective cohort study", *American Journal of Epidemiology,* vol. 174, no. 12, pp. 1406-1415.

Spence, J.D., Howard, V.J., Chambless, L.E., Malinow, M.R., Pettigrew, L.C., Stampfer, M. & Toole, J.F. 2001, "Vitamin Intervention for Stroke Prevention (VISP) trial: rationale and design", *Neuroepidemiology,* vol. 20, no. 1, pp. 16-25.

Strauss, V.Y., Jones, P.W., Kadam, U.T. & Jordan, K.P. 2014, "Distinct trajectories of multimorbidity in primary care were identified using latent class growth analysis", *Journal of Clinical Epidemiology,* , no. 67, pp. 1163.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. & Carpenter, J.R. 2009, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls", *BMJ,* vol. 338.

Stroke Association 2016, *State of the nation - stroke statistics*. Available: https://www.stroke.org.uk/sites/default/files/stroke_statistics_2015.pdf [2016, 11, 21].

Stroke Association 2013, *Stroke statistics*. Available: http://www.stroke.org.uk/resource-sheet/stroke-statistics [2014, 17, 03].

Stroke Association 2012, *Ischaemic stroke*. Available: http://www.stroke.org.uk/factsheet/ischaemic-stroke [2014, 17, 03].

Sucharew, H., Khoury, J., Moomaw, C.J., Alwell, K., Kissela, B.M., Belagaje, S., Adeoye, O.K., Khatri, P., Woo, D., Flaherty, M.L., Ferioli, S., Heitsch, L., Broderick, J.P. & Kleindorfer, D. 2013, "Profiles of the National Institutes of Health Stroke Scale items as a predictor of patient outcome", *Stroke,* vol. 44, no. 8, pp. 2182-2187.

Teasdale, G. & Jennett, B. 1974, "Assessment of coma and impaired consciousness", *Lancet,* vol. 304, no. 7872, pp. 81-84.

Teixeira-Pinto, A. & Mauri, L. 2011, "Statistical analysis of noncommensurate multiple outcomes", *Circulation: Cardiovascular Quality and Outcomes,* vol. 4, no. 6, pp. 650-656.

The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group 1995, "Tissue plasminogen activator for acute ischemic stroke", *New England Journal of Medicine,* vol. 333, no. 24, pp. 1581-1588.

Tilley, B.C., Marler, J., Geller, N.L., Lu, M., Legler, J., Brott, T., Lyden, P. & Grotta, J. 1996, “Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA stroke trial”, *Stroke,* vol. 27, no. 11, pp. 2136-2142.

Titman, A.C. 2011, “Flexible nonhomogeneous Markov models for panel observed data”, *Biometrics,* vol. 67, no. 3, pp. 780-787.

Townsend, N., Wickramasinghe, K., Bhatnagar, P., Smolina, K., Nichols, M., Leal, J., Luengo-Fernandez, R. & Rayner, M. 2012, *Coronary Heart Disease Statistics,* 2012 Edition, British Heart Foundation, London.

Upshaw J.N., Konstam M.A., Van, K.D., Noubary F., Huggins G.S. & Kent D.M. 2016, "Multistate model to predict heart failure hospitalizations and all-cause mortality in outpatients with heart failure with reduced ejection fraction", *Circulation: Heart Failure,* vol. 9, no. 8, pp. e003146.

Uyttenboogaart, M., Stewart, R.E., Vroomen, P.C.A.J., De Keyser, J. & Luijckx, G.J. 2005, “Optimizing cutoff scores for the Barthel index and the modified Rankin scale for defining outcome in acute stroke trials”, *Stroke,* vol. 36, no. 9, pp. 1984-1987.

van Swieten, J.C., Koudstaal, P.J., Visser, M.C., Schouten, H.J. & van Gijn, J. 1988, “Interobserver agreement for the assessment of handicap in stroke patients”, *Stroke,* vol. 19, no. 5, pp. 604-607.

Verbeke, G., Fieuws, S., Molenberghs, G. & Davidian, M. 2012, “The analysis of multivariate longitudinal data: A review”, *Statistical Methods in Medical Research,* vol. 23, no. 1, pp. 42 - 59

Vermunt, J.K. & Magidson, J. 2003, “Latent class models for classification “, *Computational Statistics & Data Analysis,* no. 41, pp. 531.

Vermunt, J.K., Tran, B. & Magidson, J. 2008, “Latent class models in longitudinal research”, *Handbook of longitudinal research: Design, Measurement, and Analysis,* pp. 373-385.

Vuong, Q.H. 1898, “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses”, *Econometrica,* vol. 57, no. 2, pp. 307.

Wade, D.T. & Collin, C. 1988, “The Barthel ADL Index: a standard measure of physical disability?”, *International Disability Studies,* vol. 10, no. 2, pp. 64-67.

White, H.R., Johnson, V. & Buyske, S. 2000, “Parental modeling and parenting behavior effects on offspring alcohol and cigarette use: a growth curve analysis”, *Journal of Substance Abuse,* vol. 12, no. 3, pp. 287-310.

Whiteley, W., Jackson, C., Lewis, S., Lowe, G., Rumley, A., Sandercock, P., Wardlaw, J., Dennis, M. & Sudlow, C. 2009, “Inflammatory markers and poor outcome after stroke: A prospective cohort study and systematic review of Interleukin-6”, *PLOS Medicine,* vol. 6, no. 9, pp. e1000145.

Wilson, J.T., Pettigrew, L. & Teasdale, G. 1998, “Structured interviews for the Glasgow outcome scale and the extended Glasgow outcome scale: Guidelines for their use”, *Journal of Neurotrauma,* vol. 15, no. 8, pp. 573-585.

Wilson, J.T.L., Hareendran, A., Grant, M., Baird, T., Schulz, U.G.R., Muir, K.W. & Bone, I. 2002, "Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin scale", *Stroke*, vol.33, no. 9, pp.2243-2246

Winship, C. & Mare, R.D. 1984, "Regression models with ordinal variables", *American Sociological Review,* vol. 49, no. 4, pp. 512-525.

World Health Organisation 2013*, The top 10 causes of death, Fact Sheet N°310*. Available: http://www.who.int/mediacentre/factsheets/fs310/en/ [2014, 17, 03].

World Health Organisation 1978, *Cerebrovascular disorders: a clinical and research classification,* World Health Organization.

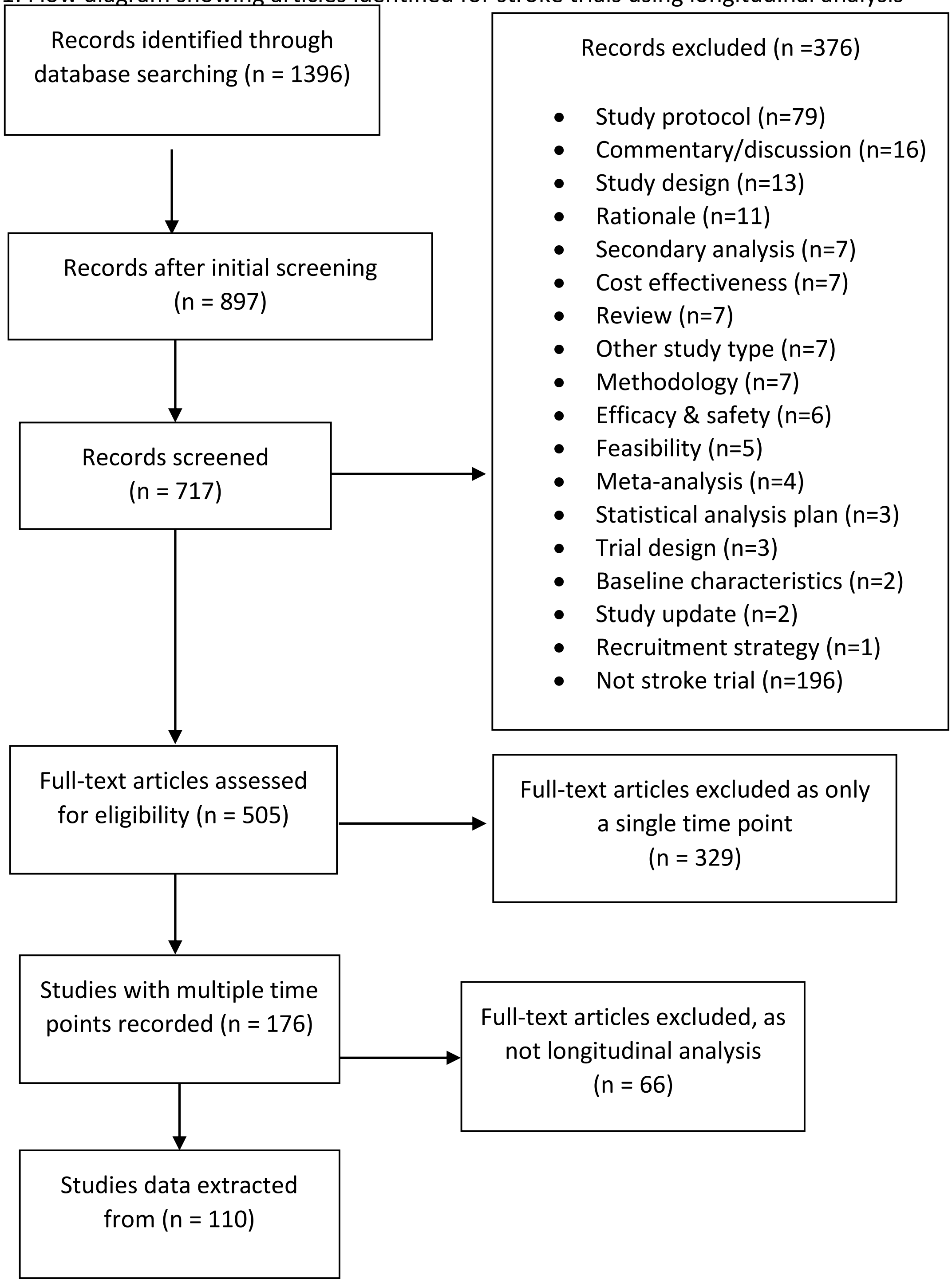
Yang, C.C. 2006, "Evaluating latent class analysis models in qualitative phenotype identification" *Computational Statistics & Data Analysis,* no. 50, pp. 1090.

Yen A.M.-F., Auvinen A., Schleutker J., Wu Y.-Y., Fann J.C.-Y., Tammela T., Chen S.L.-S., Chiu S.Y.-H. & Chen H.-H. 2015, "Prostate cancer screening using risk stratification based on a multi-state model of genetic variants", *Prostate,* vol. 75, no. 8, pp. 825-835.

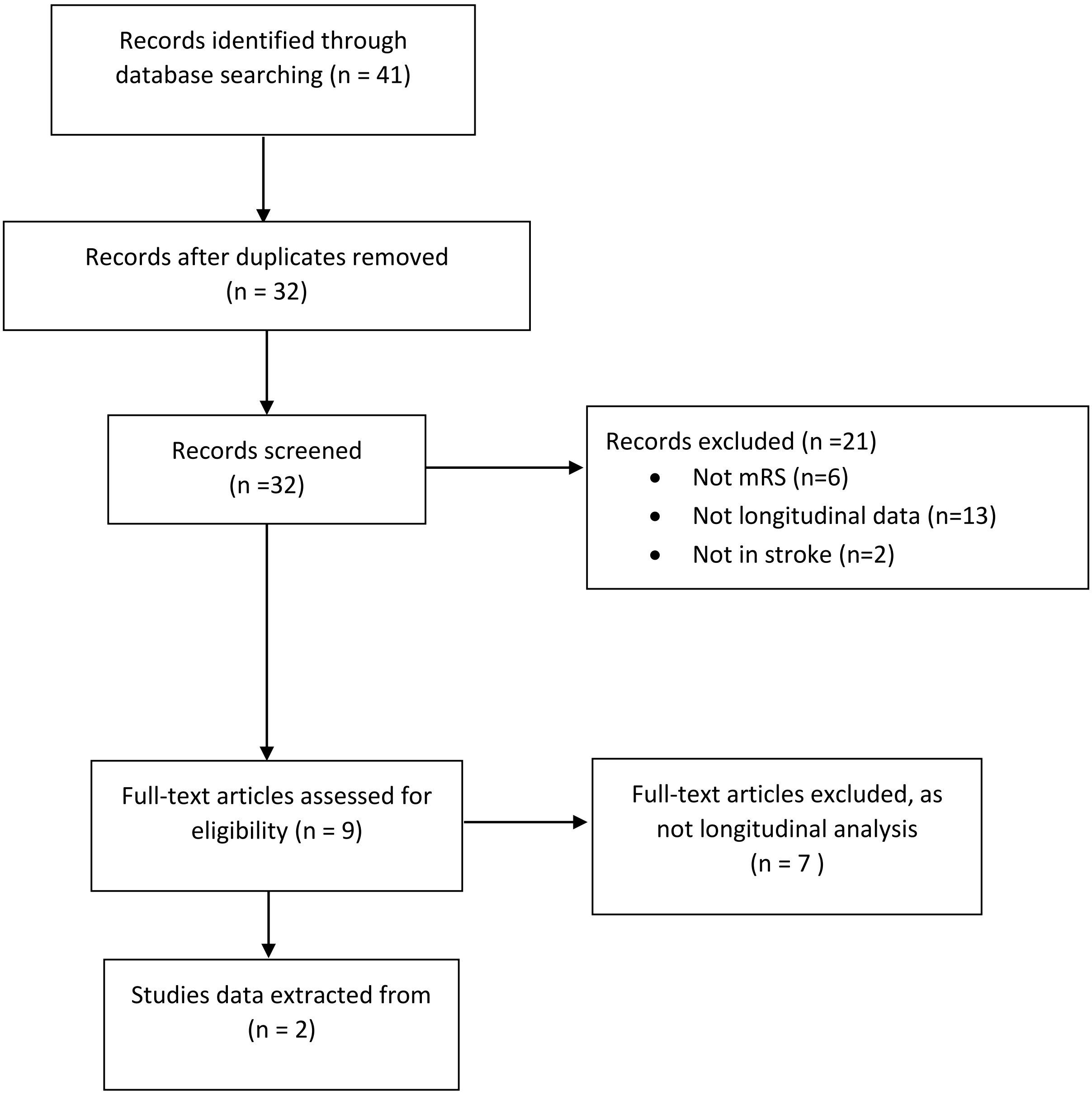
Zhang, X., Li, Q., Rogatko, A., Tighiouart, M., Hardison, R.M., Brooks, M.M., Kelsey, S.F., Kaul, S. & Bairey Merz, C.N. 2015, "Analysis of the bypass angioplasty revascularization investigation trial using a multistate model of clinical outcomes", *American Journal of Cardiology,* vol. 115, no. 8, pp. 1073-9.

Appendix A

Appendix A.1: Flow diagram showing articles identified for stroke trials using longitudinal analysis


Records identified through database searching (n = 1396)
Records after initial screening (n = 897)
Records screened (n = 717)
Records excluded (n =376)
Study protocol (n=79)
Commentary/discussion (n=16)
Study design (n=13)
Rationale (n=11)
Secondary analysis (n=7)
Cost effectiveness (n=7)
Review (n=7)
Other study type (n=7)
Methodology (n=7)
Efficacy & safety (n=6)
Feasibility (n=5)
Meta-analysis (n=4)
Statistical analysis plan (n=3)
Trial design (n=3)
Baseline characteristics (n=2)
Study update (n=2)
Recruitment strategy (n=1)
Not stroke trial (n=196)
Full-text articles assessed for eligibility (n = 505)
Full-text articles excluded as only a single time point (n = 329)
Studies with multiple time points recorded (n = 176)
Full-text articles excluded, as not longitudinal analysis (n = 66)
Studies data extracted from (n = 110)

Appendix A.2: Flow diagram showing articles identified for stroke trials conducting a longitudinal analysis with the mRS


Records identified through database searching (n = 41)
Records after duplicates removed (n = 32)
Records screened (n =32)
Records excluded (n =21)
• Not mRS (n=6)
• Not longitudinal data (n=13)
• Not in stroke (n=2)
Full-text articles assessed for eligibility (n = 9)
Full-text articles excluded, as not longitudinal analysis (n = 7)
Studies data extracted from (n = 2)

Appendix B

Appendix B1: Repeated measures proportional odds models SAS code

```
title1 ' Generalized ordered logit model' ;
title2 ' TIME, treatment' ;
proc nlmixed data=reg.sos_data qmax=5000 ;
bounds i1>0, i2>0, i3>0, i4>0, i5>0;
parms beta11=0.56 beta12=0.61 beta13=0.64 beta14=0.64 beta15=0.64
beta16=0.64 beta31=0.56 beta32=0.61 beta33=0.64 beta34=0.64
beta35=0.64 beta36=0.64 beta41=0.56 beta42=0.61 beta43=0.64
beta44=0.64 beta45=0.64 beta46=0.64 beta51=0.56 beta52=0.61
beta53=0.64 beta54=0.64 beta55=0.64 beta56=0.64 i1=0.2 i2=0.2
i3=0.3 i4=0.4 i5=0.3 su=0.4 beta0=-10 beta26=0.3 beta25=0.3
beta24=0.3 beta23=0.3 beta22=0.3 beta21=0.3;

eta6 = beta0 + beta16*treat + beta26*dummy6 + beta36*dummy12 +
beta46*inter6 + beta56*inter12 + u ;
eta5 = beta0 + i1 + beta15*treat + beta25*dummy6 + beta35*dummy12
+ beta45*inter6 + beta55*inter12 + u ;
eta4 = beta0 + i1 + i2+ beta14*treat + beta24*dummy6 +
beta34*dummy12 + beta44*inter6 + beta54*inter12 + u ;
eta3 = beta0 + i1 + i2 + i3+ beta13*treat + beta23*dummy6 +
beta33*dummy12 + beta43*inter6 + beta53*inter12 + u ;
eta2 = beta0 + i1 + i2 + i3 + i4+ beta12*treat + beta22*dummy6 +
beta32*dummy12 + beta42*inter6 + beta52*inter12 + u ;
eta1 = beta0 + i1 + i2 + i3 + i4 + i5+ beta11*treat +
beta21*dummy6 + beta31*dummy12 + beta41*inter6 + beta51*inter12 +
u ;

if      (imrs=6) then z= 1/(1+exp(-eta6)) ;
else if (imrs=5) then z= 1/(1+exp(-eta5)) - 1/(1+exp(-eta6)) ;
else if (imrs=4) then z= 1/(1+exp(-eta4)) - 1/(1+exp(-eta5)) ;
else if (imrs=3) then z= 1/(1+exp(-eta3)) - 1/(1+exp(-eta4)) ;
else if (imrs=2) then z= 1/(1+exp(-eta2)) - 1/(1+exp(-eta3)) ;
else if (imrs=1) then z= 1/(1+exp(-eta1)) - 1/(1+exp(-eta2)) ;
else if (imrs=0) then z= 1 - 1/(1+exp(-eta1));
ll=log(z) ;
model imrs ~ general(ll);
random u ~ normal(0,su**2) subject=id;
estimate 'int6' beta0;
estimate 'int5' beta0+i1;
estimate 'int4' beta0+i1+i2;
estimate 'int3' beta0+i1+i2+i3;
estimate 'int2' beta0+i1+i2+i3+i4;
estimate 'int1' beta0+i1+i2+i3+i4+i5;
predict eta1 out=eta1;
predict eta2 out=eta2;
predict eta3 out=eta3;
predict eta4 out=eta4;
predict eta5 out=eta5;
predict eta6 out=eta6;
run;
```

Beta values were adjusted for the partial proportional odds models

Appendix C

Appendix C1: Example Code for generation of ordinal data for simulation study

```
/**********************************/
/**********************************/
/*** MODALITY = 7  et TIME = 3 ***/
/**********************************/
/**********************************/
/*Marginal probabilities for X=0*/
data simul.marginal0;input C1 C2 C3;
cards;
0.216 0.233 0.252
0.148 0.154 0.159
0.121 0.123 0.124
0.080 0.079 0.078
0.117 0.114 0.110
0.135 0.128 0.121
0.184 0.169 0.156
;run;
/*Correlation matrix*/
data simul.good;input C1 C2 C3;
cards;
1 0.2 0.2
0.2 1 0.2
0.2 0.2 1

;run;
/*Marginal probabilities for X=1*/
data simul.marginal1;input C1 C2 C3;
cards;
0.208 0.199 0.192
0.145 0.141 0.138
0.120 0.119 0.118
0.080 0.080 0.080
0.118 0.119 0.120
0.138 0.141 0.145
0.192 0.199 0.208
;run;

/****************************************/
/**** Generation COMPLETE DATA N= 100****/
/****************************************/
/*X=0*/
%simul_ordi(marginal=simul.marginal0,corr=simul.good,N=100);
data simul.simul0;set simul;X=0;run;
/*X=1*/
%simul_ordi(marginal=simul.marginal1,corr=simul.good,N=100);
data simul.simul1;set simul;X=1;run;
data simul.final_100;set simul.simul0 simul.simul1;run;
/****************************************/
/**** Generation COMPLETE DATA N= 300****/
/****************************************/

/*X=0*/
```

```
%simul_ordi(marginal=simul.marginal0,corr=simul.good,N=300);
data simul.simul0;set simul;X=0;run;
/*X=1*/
%simul_ordi(marginal=simul.marginal1,corr=simul.good,N=300);
data simul.simul1;set simul;X=1;run;
data simul.final_300;set simul.simul0 simul.simul1;run;


/***************************************/
/**** Generation COMPLETE DATA N= 500****/
/***************************************/

/*X=0*/
%simul_ordi(marginal=simul.marginal0,corr=simul.good,N=500);
data simul.simul0;set simul;X=0;run;
/*X=1*/
%simul_ordi(marginal=simul.marginal1,corr=simul.good,N=500);
data simul.simul1;set simul;X=1;run;
data simul.final_500;set simul.simul0 simul.simul1;run;
```

Appendix D

Legend: Dropout 1: MCAR – equal dropout; dropout 2: MAR – equal dropout; dropout 3: MAR – unequal dropout; dropout 4: MCAR – unequal dropout

Appendix D.1: First time point estimates before and after dropout – well-balanced data

		No dropout	**Dropout 1**	**Dropout 2**	**Dropout 3**	**Dropout 4**
Dataset size	**True effect**	**RM treatment effect at time point 1**				
20%	0.107	0.108	0.108	0.108	0.107	0.108
40%	-0.077	-0.145	-0.144	-0.140	-0.143	-0.145
60%	-0.044	-0.069	-0.069	-0.061	-0.068	-0.072

Appendix D.2: First time point estimates before and after dropout – skewed data

		No dropout	**Dropout 1**	**Dropout 2**	**Dropout 3**	**Dropout 4**
Dataset size	**True effect**	**RM treatment effect at time point 1**				
20%	0.072	0.082	0.081	0.077	0.078	0.082
40%	0.050	0.162	0.104	0.106	0.106	0.103
60%	-0.044	-0.069	-0.069	-0.061	-0.068	-0.072

Appendix E.1: The average percentage change between the simulated effect and the true effect for each correlation value, and dataset split by type of missingness mechanism used.

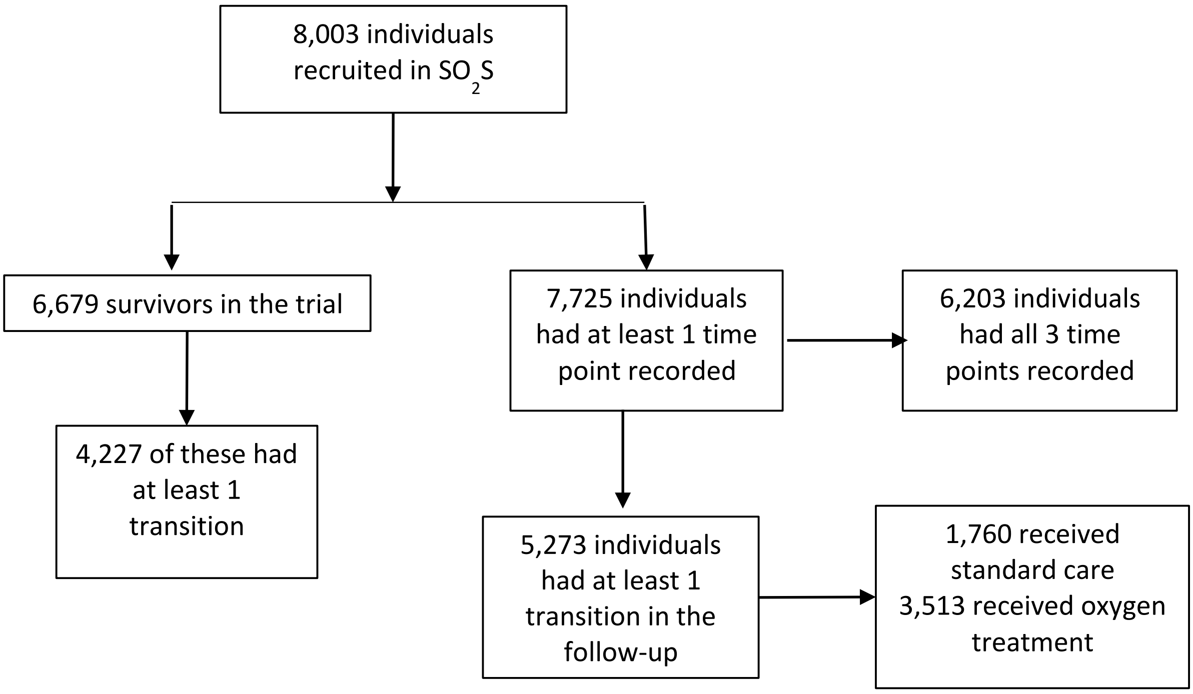
	20% Correlation			**40% Correlation**			**60% Correlation**		
		Average percentage change			**Average percentage change**			**Average percentage change**	
	True Effect	**MCAR**	**MAR**	**True Effect**	**MCAR**	**MAR**	**True Effect**	**MCAR**	**MAR**
Well balanced	0.107	3	3	-0.077	88	96	-0.044	60	82
Skewed	0.072	13	8	0.05	107	113	-0.001	6100	6600

Appendix E.2: The average percentage change between the simulated effect and the true effect for each correlation value, and dataset split by whether the dropout rate was equal or unequal between groups.

	20% Correlation			**40% Correlation**			**60% Correlation**		
		Average percentage change			**Average percentage change**			**Average percentage change**	
	True effect	**Equal dropout**	**Unequal dropout**	**True Effect**	**Equal dropout**	**Unequal dropout**	**True Effect**	**Equal dropout**	**Unequal dropout**
Well-balanced	0.107	4	3	-0.077	85	88	-0.044	48	59
Skewed	0.072	10	11	0.050	111	109	-0.001	6350	6350

Appendix F

Appendix F.1: Flow diagram indicating flow of individuals in SO_2S study, used in the different LCGA model



8,003 individuals recruited in SO_2S
6,679 survivors in the trial
4,227 of these had at least 1 transition
7,725 individuals had at least 1 time point recorded
6,203 individuals had all 3 time points recorded
5,273 individuals had at least 1 transition in the follow-up
1,760 received standard care
3,513 received oxygen treatment