Development of a minimum stroke dataset for electronic collection in routine stroke care

Dr Elizabeth Ann Teale

MBChB MRCP MPH

Submitted in accordance with the requirements for the degree of
Doctor of Medicine

The University of Leeds
School of Medicine

November 2011

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 form the basis for a paper "A systematic review of case-mix adjustment models for stroke" E A Teale, J B Young, A Forster, T Munyombwe. The final definitive version of this paper has been published ahead of print in Clinical Rehabilitation, January 2012 doi: 10.1177/0269215511433068 by SAGE Publications Ltd, all rights reserved ©. This paper can be accessed at http://cre.sagepub.com/content/early/2012/01/17/0269215511433068. Assistance with developing the searches for this systematic review was provided by Deirdre Andre (University of Leeds Medical School Library). Initial screening of titles was performed by Anita Ranjendram (University of Leeds Medical School) and Anne Forster (University of Leeds). Review of abstracts and full text citations was performed by myself and Anne Forster. Double data extraction was performed by myself and Ruth Lambley (Academic Unit of Elderly Care and Rehabilitation, Leeds Institute of Health Sciences) and statistical appraisal of identified models by myself and Theresa Munyombwe (Department of Biostatistics, University of Leeds). I co-ordinated the review and drafted the manuscript. Professor John Young provided support in writing the manuscript.

The outcomes dataset as outlined in section 4.2.1 of the thesis was developed through a systematic review submitted and examined for a Master of Public Health degree (E A Teale) and subsequently published: "A review of stroke outcome indicators valid and reliable for administration by postal survey" E A Teale, JB Young Reviews in Clinical Gerontology 2010 20; 338-353. I performed the review and drafted the manuscript. John Young offered support in writing the manuscript.

Acknowledgements

I am pleased to acknowledge funding from the National Institute for Health Research Collaborations for Leadership in Applied Health Research and Care funding scheme. The views and opinions expressed in this thesis are those of the author and not necessarily those of the NHS, the NIHR, or the Department of Health.

I have received help and support from many people through the inception, execution and writing of this thesis. In particular I thank the patients and research staff who contributed their time and expertise in the data collection phase of the project. I thank especially Ian Sleigh for his Information Technology expertise, Rachel Jones for co-ordinating the data collection and Theresa Munyombwe, Brian Cattle and Robert West at the department of Biostatistics at the University of Leeds for their comprehensive support in the statistical analysis of the data.

Supervision of the thesis was provided by Professor John Young and Dr Nigel Dudley and I thank them for their patience, time, support and encouragement. I also thank the team that assisted with the systematic review of case-mix adjustment models: Deirdre Andre at the University of Leeds Medical School Library for assistance with development of the search strategy, Anita Ranjendram at the University of Leeds Medical School, Anne Forster at the University of Leeds and Ruth Lambley from the Academic Unit of Elderly Care and Rehabilitation in Bradford for assistance with data extraction and Theresa Munyombwe at the University of Leeds Department of Biostatistics for assistance with statistical appraisal of the models.

Finally, I thank my friends, family and colleagues for their support, encouragement and understanding during the writing of this thesis.

Abstract

Improving stroke care is a national priority and adherence to national policy and guidelines is closely monitored by numerous organisations using a considerable number of overlapping indicators of stroke care processes. Demonstration that the processes of stroke care are linked to patient outcomes in empirical post-stroke populations is confounded by the complexities of patient case-mix.

Electronic, real-time, point of care data capture of care processes that are demonstrably linked to appropriately case-mix adjusted patient reported outcomes would increase confidence that the important aspects of patients' care are measured, monitored, and improved. This thesis aims to determine the best available case-mix adjuster, process measures and preferred patient reported outcome instruments and, through exploration of the relationships between these factors, to develop a dataset for use within an electronic data system. The best available case-mix adjuster was identified through a systematic literature review as the Six Simple Variable (SSV) model. Through group decision making workshops, and informed by a previous systematic review, the Subjective Index of physical and Social Outcome (SIPSO) was identified as the preferred postal outcome measure. I demonstrate how existing process markers for stroke lack variability, such that when recorded in their current format, their relative impact on patient outcome is difficult to discern.

Process measures which feature as important predictors of patient outcome are shown to act as proxy measures of stroke severity. The SSV case-mix adjustment model is overshadowed by a simple univariable predictor (length of stay) which is also likely to be acting as a proxy for stroke severity. In this context, length of stay may offer a pragmatic alternative to more complex case-mix adjustment models to examine the relationships between processes of care and outcome in populations of stroke survivors.

Table of Contents

List of Tables

List of Figures

List of Models

Part I Rationale for study and development of preliminary datasets for further testing

Chapter 1 Introduction

1.1 A brief history of stroke

Until the early part of the 20th Century, stroke was referred to in the medical literature as apoplexy. The term originated with the ancient Greeks and, etymologically, derives from the Greek 'to disable by means of a stroke', with stroke in this context taken to mean "as if struck by lightning" or "the stroke of God" (Pound P et al 1997 p 337). In 1802, Heberden offered a description of apoplexy:

> ***"…a sudden, or rapid weakness in some of the muscles of voluntary motion, constitutes a palsy, and in this manner it most usually begins; and a total loss of motion in every part of the body except the heart and organs of respiration, together with insensibility, is called an apoplexy; the cause of which is sometimes strong enough to put a stop to the motion even of the heart and lungs, and to occasion instant death." (Heberden W 1892 p338)***

This accurate description of the onset of stroke is remarkably similar to the current World Health Organisation definition of "rapidly developing clinical signs of focal disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin" (Hatano S et al 1976).

Interruption to cerebral blood flow as the pathological cause of stroke was recognised as long ago as the ancient Greeks (Galen AD 131) (Pound P et al 1997). Blood-letting was commonly employed in an attempt to relieve the symptoms of stroke and, remarkably, it was not until the early 20th Century that venesection was deemed to be of no benefit (Pound P et al 1997). Treatment options for acute stroke remained dishearteningly limited with the Hippocratic aphorism that: "It is impossible to remove a strong attack of apoplexy, and not easy to remove a weak attack" remaining a remarkably insightful observation. As recently as the 1980s, treatment of stroke remained largely supportive, with emphasis being placed on prevention of further events (Petersdorf RG et al 1983 p 2041). In the middle of the 20th Century, pioneers of Geriatric Medicine such as Marjorie Warren demonstrated the benefits of rehabilitation in longer term conditions including stroke (Barton et al 2003); the importance of organised multidisciplinary therapy in stroke has been increasingly recognised and encouraged since the early 1960s (Pound P et al 1997).

However, it was the 1997 Stroke Unit Trialists' systematic review of randomised trials of organised stroke unit care versus general ward care that catalysed a paradigm shift towards organised inpatient stroke care delivered in dedicated stroke units (Stroke Unit Trialists' Collaboration 2007). This review demonstrated a clear benefit in terms of likelihood of survival, return to independence and living at home following a stroke; a benefit that was seen in all patients regardless of stroke type or severity.

Specific interventions have now been shown in randomised controlled trials to be effective in improving patient outcomes following stroke. Early Supported Discharge schemes promote the use of community based specialist stroke rehabilitation teams and have been shown to be cost effective, improve patient outcomes and reduce length of stay in those with less severe strokes (Early Supported Discharge Trialists 2005). Thrombolytic therapy has been shown in meta-analysis of large multicentre randomised trials to be of benefit in terms of increasing the likelihood of independent survival in specific subgroups of patients with acute ischaemic stroke (Wardlaw JM et al 2009).

These complex interventions require significant organisational infrastructure to enable them to be routinely available to all who may benefit from them - for example the timely availability of brain imaging to allow administration of thrombolytic agents in those for whom it is indicated, or sufficient capacity on the specialist acute stroke unit to allow direct admission from the Emergency Department or from the community. Over the last decade, there has been considerable work to define best practice in stroke care based on the emerging evidence base and to identify areas of deficiency in stroke care provision. This has prompted significant investment in the development of stroke services in an attempt to improve access to these interventions. In the next section I discuss the evolution of the definition and monitoring of high quality stroke care in England, Wales and Northern Ireland.

1.2 The evolution of stroke care monitoring

Since their inception in 1998, sequential biennial Royal College of Physicians Clinical and Organisational National Sentinel Stroke Audits (RCP NSSA) have allowed local services to assess changes over time and in relation to the national situation as regards stroke care provision in England, Wales and Northern Ireland (Intercollegiate Stroke Working Party 2010; Intercollegiate Stroke Working Party 2011). Stroke process data are extracted retrospectively from consecutive patients admitted during the audit period (the first 60 patients in the last audit) and submitted to the RCP via a web based form. The audits continue to provide useful information regarding clinical and organisational aspects of stroke care provision and, over time, have become central to a number of stroke metrics as indicators of the quality of stroke care.

The National Service Frameworks were introduced in the 1997 White Paper 'The New NHS; modern, dependable' with the aim of consolidating clinical best practice and cost-effectiveness to improve service provision in several key areas of healthcare (Department of Health 2007c sect. 3.5). In contrast to coronary heart disease for which a dedicated NSF was published, and despite 25% of strokes occurring in those under the age of 65 (National Audit Office 2005), stroke featured as one of the eight standards in the 2001 NSF for older people (Department of Health 2007c).

The first National Clinical Guideline for Stroke (NCGS), devised by the Intercollegiate Stroke Working Party at the Royal College of Physicians, was published in 2000 (Intercollegiate Stroke Working Party 2008). The guideline has undergone two subsequent revisions (2004 and 2008), with the latest version of the guideline incorporating the National Institute for Health and Clinical Excellence (NICE) guidelines for the management of acute stroke and TIA (National Institute for Health and Clinical Excellence 2008). The NCGS defines best practice in stroke management through consolidation of trial evidence and expert consensus opinion, describing the components of a quality stroke service and offering recommendations as to how these should be achieved (Intercollegiate Stroke Working Party 2008).

In 2005, the National Audit Office produced a critical report for the Committee of Public Accounts that highlighted deficiencies in the provision of care against the evidence base and guidance documents in several key areas of the stroke pathway (National Audit Office 2005). The National Stroke Strategy (NSS) (Department of Health 2007b) was the policy response from the Department of Health (DH) to address these deficiencies.

The NSS outlines a ten year strategic framework to drive stroke service reconfiguration and deliver improvements in the quality of stroke care along the entire stroke care pathway. However, the delivery of quality, personalised stroke care in a timely and cost-effective manner requires definition of the components of quality care, demonstration that delivered care has a positive effect on patient and carer outcome, and reliable metrics with which to quantify these effects along the whole stroke care pathway. Drawing heavily on previous consensus documents such as the NICE clinical guidelines for acute stroke and TIA (National Institute for Health and Clinical Excellence 2008) and the NCGS (Intercollegiate Stroke Working Party 2008), the NSS offers 20 ‘Quality Markers’ of a quality stroke service and a series of ‘measuring success’ metrics to facilitate quantitative analyses (both within and between services) (Department of Health 2007b).

The NSS has become a major driver of stroke service improvement. Implementation of the strategy features in the NHS Operational Framework and, as a result, a number of markers and metrics have been developed in an attempt to measure and monitor its delivery. These data are requested by a variety of disparate bodies for the purposes of performance monitoring, remuneration or service improvement (see section 1.2.1 below). The quality of these data is imperative to ensure that robust, consistent and comparable conclusions regarding the delivery and quality of stroke care may be drawn.

Quality data are “accurate, up-to date, free from duplication and free from confusion” (NHS Connecting for Health 2011). In short, data items and how these are used to derive indicators should be explicitly defined, captured once and in a timely manner (ideally at the point of patient care). The existing ‘cacophony’ of stroke requirements and datasets obfuscates the collection of ‘quality data’ in many, if not all, of these areas. A brief discussion of the current data that are requested from provider trusts is outlined here to highlight its complexity. A timeline of the key policy documents and data collections is given in Figure 1 (page 9).

1.2.1 Existing datasets for monitoring the delivery of stroke care

1.2.1.1 Integrated Performance Measures

The annual NHS Operating Framework defines national priorities in health care, the direction of health reform, and financial objectives of the NHS (Department of Health 2009b). Every quarter, since the 2008/09 review, service providers have been required to provide mandatory performance indicators to the Department of Health (DH) via Primary Care Trusts (PCTs) (Department of Health 2008b). In stroke and TIA, these 'Vital Sign' (VS) indicators (Department of Health 2008d) are designed to demonstrate implementation of the NSS. In June 2010, following the formation of the new Government, the Operating Framework underwent a series of revisions (Department of Health 2010c). Implementation of the NSS has remained in 'Tier 1' of the Operating Framework retaining stroke and transient ischaemic attack (TIA) as high national priorities (Department of Health 2009b). However, in the 2010/11 Operating Framework, the Vital Signs will be renamed 'Integrated Performance Measures' (IPM), although their content will remain identical (Department of Health 2008d; Department of Health 2010e).

1.2.1.2 CQUINs

In 2008, 'High Quality Care for All' (Darzi A 2008) introduced the Commissioning for Quality in Innovation (CQUIN) framework. Goals are locally agreed between commissioners and providers to encourage the provision of quality services at a contractual level. A proportion of a provider's income is reliant on meeting these goals (Department of Health 2008c; Institute for Innovation and Improvement 2010). Dependent on local priorities, stroke data may be required to fulfil CQUIN requirements.

1.2.1.3 Payment by Results (PbR)

In England, Payment by Results (PbR) is the mechanism through which, providers are remunerated by commissioners for delivery of services. Payments are made according to national tariffs calculated from adjusted 'average' service costs across similarly grouped activities (Healthcare Resource Groups (HRGs)) (Department of Health 2007a). The HRG for a particular hospital spell is calculated from ICD-10 and OPCS-4 coding data following the patient's discharge (ibid).

National Best Practice Tariffs were introduced in England in 2010/11 for selected conditions (including stroke) where 'best practice' is well defined, but variations in delivered care occur (Department of Health 2010b). The BPT framework offers incentives for the delivery of high quality care through additional payments above the base tariff for aspects of an individual patient's care that meet specific quality standards (Department of Health 2010b). The base tariff for stroke care is set below the average cost of a stroke hospital spell (and below the 'conventional tariff' for stroke that had been used within the PbR framework), such that the remuneration of care not meeting the BPT quality standards is, on average, less than the cost of the admission (Department of Health 2010b). Moreover, specific aspects of care (e.g. CT brain imaging) are not included within the base tariff, such that patients scanned outwith the BPT criteria will not be remunerated (Department of Health 2010b). Adjustments may be paid outside the BPT framework to cover the cost of drugs and the additional resource associated with thrombolytic agent treatment (Department of Health 2010b).

1.2.1.4 National Sentinel Stroke Audits

The Royal College of Physicians (RCP) Clinical and Organisational National Sentinel Stroke Audits (RCP NSSA) (see section 1.2) have had significant and sustained effects on national improvements in stroke services (Intercollegiate Stroke Working Party 2010; Intercollegiate Stroke Working Party 2011). However, there are a number of deficiencies with the audits. They comprise multiple process markers resulting in a large dataset which is complex, unwieldy and neither designed nor feasible for prospective, real-time collection. Moreover many of the indicators within the dataset are of unproven association with patient outcome (e.g. being weighed during the course of the hospital admission) and it is possible that such process markers are acting as proxy markers for more complex factors (e.g. stroke severity).

1.2.1.5 Stroke Improvement National Audit Programme (SINAP)

Funded by the DH, the Stroke Improvement National Audit Programme (SINAP) dataset has been developed by the RCP Stroke Programme to capture real-time prospective data describing acute stroke care (Royal College of Physicians Stroke Programme 2010). Although the dataset is large, only care delivered within the first 72 hours following acute stroke is considered, and many of the questions relate to the provision of thrombolysis. The audit started in 2010 with data being entered into a web-based form and submitted electronically to the Royal College of Physicians. The first report of data for England was published in July 2011 covering the reporting periods June 2010 to June 2011 (Royal College of Physicians 2011). Although not currently a mandatory requirement, the SINAP audit features as one of the National Clinical Audits within the Quality Accounts for 2011 and as such, participation is required implicitly (Healthcare Quality Improvement Partnership (HQIP) 2011).

1.2.1.6 NICE quality standards

The National Quality Board (NQB) was created in March 2009 as a recommendation of 'High Quality Care for All' (Darzi A 2008). The board has a remit to "oversee improvement of quality indicators" and to "ensure overall alignment of the quality system" (Department of Health 2010d). The NICE Quality Standards Programme, under the direction of the NQB, aims to extract or devise quality standards from available evidence, existing guidance documents and expert consensus of health and social care professionals. In early 2010, a 'Topic Expert Group' developed and refined a set of eleven quality standards and metrics to describe the stroke pathway (National Institute for Health and Clinical Excellence 2010b). It is anticipated that these standards will, in part, inform service commissioning and act as a stimulus for high quality stroke care (National Institute for Health and Clinical Excellence 2010a).

1.2.1.7 Stroke Improvement Programme, Accelerating Stroke Improvement.

The Stroke Improvement Programme (SIP), developed in 2007 as part of NHS Improvement (NHS Improvement 2010b), oversees 28 regional Stroke Care Networks tasked with the local implementation of the NSS (NHS Improvement 2010b). In early 2010, as a follow up to the 2005 report (National Audit Office 2005), the NAO re-examined the national situation regarding the provision of stroke care (National Audit Office 2010). Improvements in the acute end of the stroke care pathway (with notable exceptions such as direct admission to a stroke unit) were identified, but deficiencies in the longer-term management of stroke remained (National Audit Office 2010). The DH responded by committing the NHS to a year of accelerated improvement in stroke care. The Accelerating Stroke Improvement Programme, as part of the SIP, was launched in April 2010 (NHS Improvement 2010a). The programme has been extended and continues into 2011/12. Nine aspects of the NSS have been targeted for ambitious accelerated improvement with particular emphasis on the longer-term care of stroke patients (NHS Improvement 2010a).

1.2.1.8 Emerging datasets 'SSNAP' (Sentinel Stroke National Audit Programme)

It has been proposed that from spring 2012, SINAP and the Sentinel audits will be combined into one prospective audit to cover aspects of the whole stroke pathway. This will involve the development of a further, new dataset in place of the existing data collections. It is proposed that this new dataset will be funded by HQIP and, at the time of writing, the selection of provider for this new data collection is out to tender (Health Quality Improvement Partnership (HQIP) 2011).

1.2.1.9 Data definitions and accuracy

Accurate between institution or within institution comparisons of performance and quality rely on like being compared with like. Effective case-mix adjustment constitutes one aspect of this (see Chapter 3), but explicit data definitions form another important factor. Consistency in reporting of metrics requires every step in the derivation of indicators to be unequivocally defined. Application of these data definitions should occur at the point of data capture (or data extraction if these are different). Although derivation of some indicators (such as the IPM and BPT) is well defined (Department of Health 2008d; Department of Health 2010b), other datasets are more open to interpretation (e.g. the Accelerating Stroke Improvement metrics (NHS Improvement 2010a; Stroke Improvement Programme 2011)).

1.2.1.10 Overlapping datasets and duplication of data capture

The existing stroke datasets as outlined in section 1.2.1 are complex and replicative often containing similar yet subtly different indicators. In the absence of robust IT systems, the greater the data burden, the greater the data collection resource required to extract it. Capture of data is expensive and resource intensive. Every data item that is requested comes at a cost. Duplication of data extraction for different bodies therefore reflects wasted resource. The benefit that every data item confers should be weighed up against the cost of collecting it. Thus the capture of large and unwieldy datasets comprising process markers of little or unproven link to patient outcome are unlikely to be cost or resource effective (See section 2.1.5). Intuitively, the more data that are requested, the less likely it is that these data will be extracted and reported accurately. Moreover, frequent changes to data requirements are likely to have an impact on the accuracy and consistency of reporting.

1.3 Background to CIMSS

This MD thesis forms part of the preliminary work for the CIMSS project (Clinical Information and Management System for Stroke), the stroke theme of the National Institute for Health Research (NIHR) funded Leeds, York, Bradford (LYBRA) Collaboration for Leadership in Applied Health Research and Care (CLAHRC).

The CIMSS project has the overall aim of defining, iteratively refining and implementing, a novel core stroke dataset that is clinically relevant and feasible for electronic collection at the point of care by members of the multidisciplinary team responsible for delivering that care. The CIMSS dataset has an emphasis on patient reported outcomes, with the anticipation that routine collection and feedback of relevant CIMSS data to healthcare professionals and commissioners will result in measurable improvements in the effectiveness of stroke care in the stroke services within the Yorkshire and the Humber region that are participating in the implementation phase of the CIMSS CLAHRC project. This thesis describes the research led process to define and test the preliminary CIMSS dataset and refine the preliminary fields to a dataset suitable for wider implementation.

This CIMSS dataset differs from existing and previous data collections as it includes routinely collected patient reported functional outcomes data within an infrastructure that allows these outcomes data to be linked directly to information relating to patients' inpatient stay (process data) and their individual characteristics (case-mix). Moreover, through consideration of the current 'data environment' in stroke, an alternative approach to stroke data collection is suggested, offering standardisation, reproducibility and consistency of data collection and reporting across provider trusts. This solution ensures that existing data requirements are met (e.g. participation in SINAP - see section 1.2.1.5), but that stroke data collection is simplified, rather than compounded by the collection of CIMSS data.

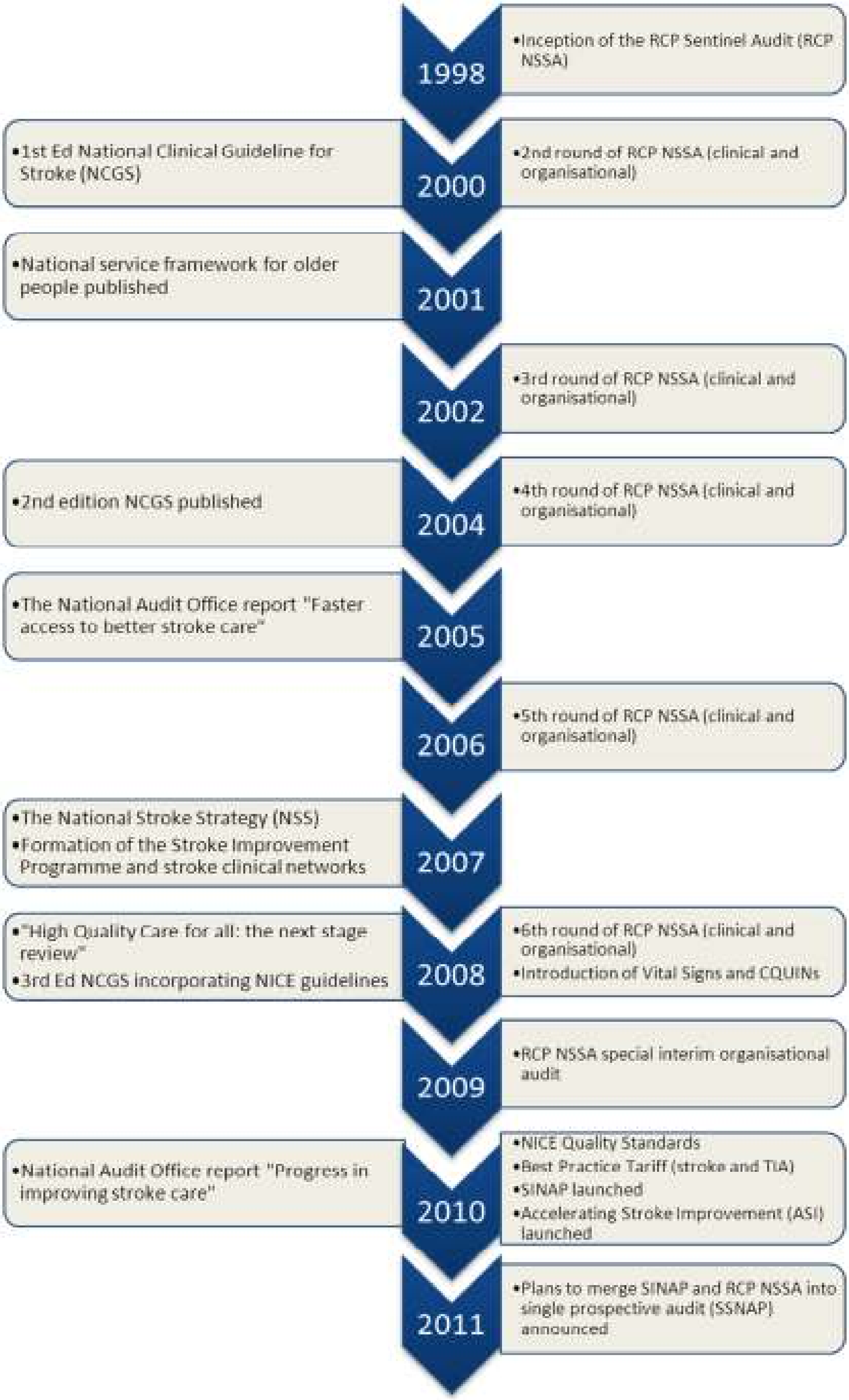
1.4 Aims of thesis

Through examination of the literature and a prospective observational cohort study I aim to address the following research questions:

- Which combination of postal outcomes instruments best captures the physical and social functioning of patients following stroke?
- Which is the best available case-mix adjuster in stroke?
- How does care process relate to patient outcome after stroke?
- Which process, case-mix and outcomes markers should be included in a routinely collected stroke dataset'?

I begin by critically appraising the literature regarding both process and outcomes driven approaches to the monitoring of healthcare, and will consider the implications of the application of these arguments to stroke care. I then describe a systematic review to identify the best available case-mix adjuster in stroke. Subsequent chapters describe the design, execution and results of a prospective cohort study to test the utility of the preliminary dataset to capture care process, case-mix and physical and social outcomes following stroke. Finally I discuss the refinement of this dataset to a set of core fields for wider implementation where CIMSS fields are combined with existing datasets to provide a flexible core minimum dataset from which all existing stroke metrics may be derived.

Figure 1 Timeline of best practice guidelines, reports and data collections


1998
•Inception of the RCP Sentinel Audit (RCP NSSA)
2000
•1st Ed National Clinical Guideline for Stroke (NCGS)
•2nd round of RCP NSSA (clinical and organisational)
2001
•National service framework for older people published
2002
•3rd round of RCP NSSA (clinical and organisational)
2004
•2nd edition NCGS published
•4th round of RCP NSSA (clinical and organisational)
2005
•The National Audit Office report "Faster access to better stroke care"
2006
•5th round of RCP NSSA (clinical and organisational)
2007
•The National Stroke Strategy (NSS)
•Formation of the Stroke Improvement Programme and stroke clinical networks
2008
•"High Quality Care for all: the next stage review"
•3rd Ed NCGS incorporating NICE guidelines
•6th round of RCP NSSA (clinical and organisational)
•Introduction of Vital Signs and CQUINs
2009
•RCP NSSA special interim organisational audit
2010
•National Audit Office report "Progress in improving stroke care"
•NICE Quality Standards
•Best Practice Tariff (stroke and TIA)
•SINAP launched
•Accelerating Stroke Improvement (ASI) launched
2011
•Plans to merge SINAP and RCP NSSA into single prospective audit (SSNAP) announced

Chapter 2 Literature review

2.1 Measuring quality in stroke care

There is much debate in the medical literature as to how the quality of delivered health care should be measured, monitored and improved. Two discrete approaches have attracted significant debate: the reflection of quality through measurement of care processes, or through patient outcome. In this chapter, through discussion of the benefits and pitfalls of process and outcomes driven approaches, I will discuss how assessment of healthcare delivery depends on the definition of quality, the purpose of quality measurement and the perspective from which these assessments are made. Moreover, in order to achieve a broad quality perspective that captures the entire stroke pathway, I will argue that measures of both process and outcomes are required.

2.1.1 What is quality in healthcare?

2.1.1.1 The political background in England, Northern Ireland and Wales

The National Health Service (NHS) is underpinned by the principles of Clinical Governance, the system "through which NHS organisations are accountable for continuously improving the quality of their services and safeguarding high standards of care" (Department of Health 1999). The system was introduced in 1998, to ascribe formal accountability to the requirements of clinical audit and quality assessment, assurance and improvement that had previously been the informal responsibility of healthcare professionals, commissioners and health services management (Buetow SA et al 1999).

A decade of health reforms following the 1997 General Election saw a series of government policy initiatives aimed at increasing capacity in the NHS (Darzi A 2008 preface). Key policies included the introduction of Performance Assessment Frameworks, disease specific National Service Frameworks (with compliance markers) and the Quality Outcomes Framework in Primary Care. In addition, the establishment of the National Institute for Clinical Excellence (NICE) (latterly the National Institute for Health and Clinical Excellence) introduced a series of frameworks of evidence based, cost effective best practice guidance for a range of health technologies, pharmaceuticals and interventions across a range of disease areas (McLaughlin V et al 2001).

In 2008, the political focus shifted from capacity building within the NHS towards the delivery of care based on quality, productivity and value (Department of Health 2009a). "High Quality Care for All" (Darzi A 2008), a report commissioned by the DH and led by a senior clinician, was a vision of a 21st Century NHS with specific focus on achieving improvements in patient centred, quality care through the provision of safe and effective treatments. Notable outputs from the Darzi review include the National Quality Board (NQB), developed in March 2009 (Department of Health 2010d).The NICE Quality

Standards Programme, under the direction of the NQB, aims to extract or devise quality standards from available evidence, existing guidance documents and expert consensus of health and social care professionals (see also section 1.2.1.6).

Since the formation of the Conservative/Liberal Democrat coalition government in May 2010, there has been a further shift of focus in the NHS with emphasis being placed firmly with the measurement of patient outcome indicators as markers of the quality of care (Department of Health 2010e). These outcomes, however, tend to focus on hard objective endpoints such as mortality and length of hospital stay and any patient reported outcomes are limited to quality of life and satisfaction surveys (Department of Health 2010e). Although at the time that the Outcomes Framework was written it was proposed that an indicator for stroke recovery should be included, it was yet to be developed.

2.1.2 Defining quality in healthcare

Quality of care is a complex, multi-dimensional concept that has been variously defined and described. Campbell et al (2000) have suggested that care quality centres on two constructs: efficiency (in the use of resources including needs-based access to care) and effectiveness (in both the delivery of personalised care and in technical aspects of clinical care) (Campbell SM et al 2000). Anavedis Donabedian described three interrelated aspects of quality in healthcare in his influential 1966 paper "Evaluating the Quality of Medical Care" (Donabedian A 1966; Frenk J 2000). *Processes* of care describe technical aspects of care delivery – whether particular aspects of care occurred e.g. patients undergoing timely imaging or the most appropriate operation. These hard aspects of healthcare delivery are often measured through process metrics which may be used for clinical audit and benchmarking to encourage the delivery of care according to evidence-based practice. Whilst some of these care processes are directly related to an individual clinician (e.g. operative skill, choice of drug), some will require adequate staffing or care pathways and are the responsibility of the organisation. These latter, organisational aspects reflect the *structure* of care – the infrastructure necessary to deliver high quality care. Finally, there is patient *outcome* - this is defined in the NHS Outcomes Framework (Department of Health 2010e) as "… a change in the health status of an individual, group or population, which is attributable to an intervention from a healthcare provider". The premise that the way in which care is organised (care structure) affects the care that is delivered (care processes) which in turn affect patient outcome remains fundamental to the considerable and ongoing debate in the medical literature as to whether care process or patient outcomes should be monitored in order to reflect the quality of patient care. There is a dynamic relationship between these three dimensions of quality, the full understanding of which requires a fourth factor: *case-mix,* to account for the severity of an index condition in an individual patient (Figure 2). Case-mix is an important, but often ignored factor if between-organisation quality comparisons are to be attempted.

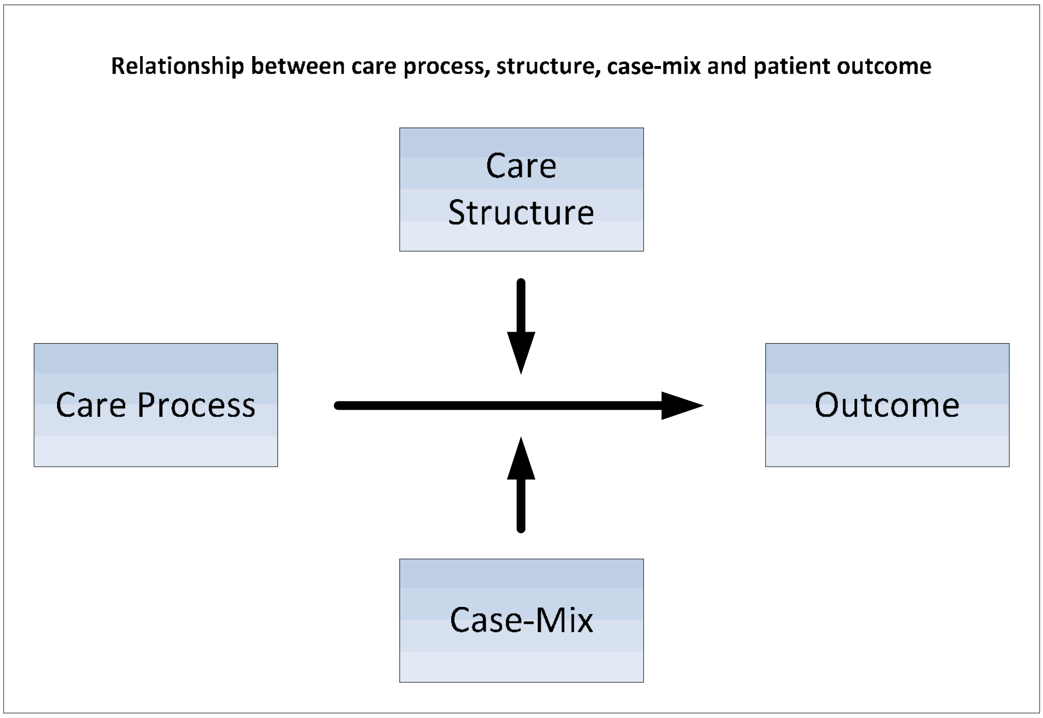

Relationship between care process, structure, case-mix and patient outcome
Care Structure
Care Process
Outcome
Case-Mix

Figure 2 The interrelationship between care process, structure, case-mix and patient outcome

The Care Quality Commission (CQC) (formerly the Healthcare Commission), is the independent regulator of health and adult social care in England. The CQC broadly define quality care as that which is "safe; has the right outcomes (including clinical outcomes); is a good experience for the people who use it, their carers and their families; helps to prevent illness and promotes healthy, independent living; is available to those who need it when they need it; and provides good value for money" (The Care Quality Commission 2009).This definition encompasses three main perspectives of quality care: Patients and their families (safety, experience and clinical outcome), commissioners and service providers (resource availability and value for money), and society as a whole (prevention and health promotion). The CQC definition of quality has formalised the need to account for and quantify the experience of a health care encounter from the perspective of the patient, family and carer. The patient experience of care adds an additional layer of complexity to the measurement of quality. In order to ensure that patients are receiving care that meets emotional as well as physical needs, the Department of Health defines good patient experience as that which ensures that patients are "getting good treatment in a comfortable, caring and safe environment, delivered in a calm and reassuring way; having information to make choices, to feel confident and to feel in control; being talked to and listened to as an equal and being treated with honesty, respect and dignity" (Department of Health 2007d).

It can be seen therefore, that although much discussed, quality in healthcare remains a concept complicated by its multidimensional nature; measurement needs to encompass the perspectives of patient and healthcare professional as well as logistical, organisational, financial and procedural aspects of care. Markers of quality are not, therefore, synonymous with markers of performance.

2.1.3 How can quality be measured?

> ***"Not everything that counts can be counted and not everything that can be counted counts."***
>
> ***(Albert Einstein 1879-1955)***

In order to improve the quality of delivered care, it must be measurable. There are no 'units of quality' and therefore proxy markers must be used. The nature of these markers will depend on the purpose of the measurement. Regardless of which markers are used, they must be valid (in their reflection of quality), explicitly defined to allow measurement against agreed standards or against similar services, and their measurement reliable (stable) over time and between raters. Indicators should also be sensitive to change (i.e. be able to detect and discriminate between small changes), relevant and acceptable to clinicians and patients, and provide relevant and useful information to wider stakeholders (Davies HTO 2005).

The populations in which quality markers are used should be standardised for baseline characteristics and case-mix (variation in e.g. stroke severity between individual patients and populations of patients) to allow legitimate and meaningful comparative measurements. The important issue of case-mix is discussed further in Chapter 3.

2.1.4 Why measure healthcare quality?

2.1.4.1 Performance monitoring and remuneration

Commissioning bodies require reassurance that services are being planned and delivered in line with commissioning contracts and national guidance. As such, specific (usually process) markers may be used to examine performance often with associated financial incentives (Department of Health 2010b). Remuneration of individual service providers has been based on volume and activity through the Payment by Results framework. However, increasingly, remuneration is only provided for care provided in line with explicit 'Best Practice' guidelines within the Best Practice Tariff structure (Department of Health 2010b). The origins and current data requirements for performance monitoring and remuneration of stroke service delivery have been described in detail in sections 1.2.1. Open competition for the commissioning of health services from NHS and non-NHS (public and private) organisations was introduced by the last Labour administration (Department of Health 2009a sect 4.19 p 54). Planned NHS reform under the coalition government aims to extend this competition and proposes that commissioning responsibilities are moved from Primary Care Trusts to consortia of General Practitioners (GPs) (Department of Health 2010a). This is likely to heighten the emphasis on the requirements for, and demonstration of, value for money through performance monitoring.

The requirements for data regarding quality of care from the perspective of patients and commissioners highlight two different aspects of quality care – patient satisfaction and value for money. Indeed, the data obtained through ‘quantification of quality’ through the measurement of patient satisfaction surveys as compared with performance metrics provide very different types of information. The first could be considered an outcome (a broad patient-centric opinion on the healthcare experience), whist the second reflects measurement of process based around volume (e.g. PbR) or delivery of care against pre-specified standards (e.g. BPT). Collecting data to describe aspects of the patient experience may be achieved through patient surveys, although focus groups or patient interviews may provide richer information in specific areas. There are, however, issues with the representativeness and feasibility of collecting data in this way. Therefore, the aspect of quality that is measured should reflect the purpose of the ‘quality assessment’.

2.1.4.2 Patient centred care

There has been a gradual, yet sustained evolution of the concept of personalised care within the NHS since the introduction of the Patients’ Charter in 1991 (The King's Fund 2011). Since 2009, as a consequence of the Darzi Review (Darzi A 2008), NHS trusts have been required to collect and report routinely Patient Reported Outcome Measures (PROMs) following specific operative procedures (Department of Health 2008a). Although the scope of this framework is currently limited, it is anticipated that the scheme will “extend … across the NHS wherever practicable” (Department of Health 2010a).

The PROMs framework is designed to capture patient reported disease specific and subjective outcomes data in an attempt to reassure patients, commissioners, healthcare providers and the tax-payer that delivered care has a positive effect on the types of outcomes that are relevant to patients. Moreover, it is intended that these may be used to differentiate good from poor quality care. Indeed, the Information Centre (IC) website (currently responsible for the PROMs data) states: “The health status information collected from patients by way of PROMs questionnaires before and after an intervention provides an indication of the outcomes or quality of care delivered to NHS Patients” (The Information Centre 2011b). Raw and case-mix adjusted PROMs data collected since 2009 are available in the public domain from the Information Centre via the Hospital Episode Statistics (HES) website (Hospital Episode Statistice (HESonline) 2011). In this context PROMs – subjective measures of an individual’s disease specific outcome and quality of life– are likely to be used to make assumptions about the relative quality of delivered care (process) between institutions, regardless of whether or not this is the intent of capturing the data.

2.1.5 Linking process and outcome

The reflection of quality through markers of process is dependent on the demonstration of robust linkages between the process marker and patient outcomes in unselected populations - in order for the processes of care to reflect quality they must be known to explain some variability in patient outcome. The measurement of process becomes an abstract concept “of little intrinsic interest” (Mant J 2001), a marker of the quality of process and not the quality of care, unless it has been demonstrated to have some impact on outcome. Similarly, the use of outcomes of care as markers of quality is of little benefit in terms of improving the quality of care unless it is known which specific aspects of care process are responsible for the variation in patient outcome and whether optimisation of specific aspects of process could indeed improve these outcomes (Lilford RJ et al 2007). Therefore, regardless of whether processes or outcomes of care are to be used to monitor care quality, it should be clear that variation in outcome is explained through variation in care processes rather than, for example, unexplained differences in case-mix or chance (Lilford RJ et al 2007; Mant J 2001).

Evidence based healthcare relies on the translation of processes and interventions shown to be beneficial in the clinical trial setting into routine care. However, an important caveat to the development and legitimacy of process markers based on trial interventions for the purposes of monitoring quality of patient care is that this depends on the demonstration of linkages between process and outcome in unselected populations. Clinical trials often involve the measurement of both processes of care (or specific interventions) and patient outcomes. However, there are important differences between measurement of process and outcome in the research setting and for quality assessment.

Randomised controlled trials (RCTs) and meta analyses of RCTs are generally considered the ‘gold standards’ in terms of the hierarchy of research evidence (Scottish Intercollegiate Guidelines Network (SIGN) 2008 p 51), as the effect of confounding variables can be minimised through randomisation. Indeed, RCTs have demonstrated that many processes of care to be effective in reducing the hard endpoints of death or dependency in stroke (e.g. thrombolysis, stroke unit care, early supported discharge, aspirin). However, direct translation of RCT results into routine care (generalizability to unselected populations) may be limited by trial inclusion and exclusion criteria (Black 1996). For example, the proportion of patients recruited over the age of 80 in the Cochrane review of thrombolysis for acute ischaemic stroke, was just 0.5% (Wardlaw JM et al 2009). It is possible that the benefits (and risks) of some interventions may be attenuated or accentuated in certain subgroups of patients that were excluded from the original trials. Extrapolation of research findings to these subgroups is not necessarily valid and, in the absence of studies to demonstrate generalizability into unselected populations, subgroup treatment effects remain untested. Randomised rehabilitation trials pose particular challenges. For example, blinding and sustained adherence to rigid treatment protocols may be difficult due to the complexities of the interventions (Horn SD et al 2005; Black N 1996). Basing generic performance markers on randomised trial evidence should therefore be with the caveat that this may be in the absence of empirical evidence of generalizability into unselected populations.

Evans et al (2001) aimed to identify processes of care within the complex intervention of stroke unit care that may predict dichotomised patient modified Rankin Score (mRS) at three months (Evans A et al 2001). Logistic regression analysis was performed using data from 304 patients collected for a previous randomised controlled trial of stroke unit vs. general ward care. Limited case-mix variables (age and baseline Barthel Index) were also entered into the regression models. Of the factors that were identified as being associated with outcome in this study (prevention of aspiration pneumonia, early feeding, stroke progression, chest infection, dehydration and management on a stroke unit) (Evans A et al 2001), many are likely to reflect the severity of stroke rather than a discrete process of care; chest infections, stroke progression and dehydration are more likely to be markers of stroke severity or comorbidity rather than deficiencies in care process. Although these data were taken from a randomised trial, case-mix adjustment is still important as patients were randomised to receive stroke unit care and not individual care processes. Additionally, the data for intervention and control arms were pooled to form the study dataset such that the randomisation is no longer effective.

Observational studies offer an alternative way to examine populations that may produce more generalizable (externally valid) results than RCTs (Black N 1996). Capture of data describing processes of care that actually occur rather than through RCT treatment protocols allows a pragmatic exploration of delivered care. The broad external validity that is conferred through examination of non-randomised, unselected populations is, however, attenuated through the uncertainty that is introduced through the heterogeneity of these populations (i.e. a loss of internal validity) (Horn SD et al 2005). Observational studies can therefore facilitate exploration of correlation (as opposed to causation) between processes of care and outcomes. However, these relationships are complicated by the effects of additional and potentially unmeasured confounding factors (Lilford RJ et al 2007).

Several groups have tried to correlate process markers with patient outcomes (at an institutional level) across a variety of conditions in empirical (unselected) populations with limited success (Lilford RJ et al 2004). In a review of 36 studies examining 51 such relationships between process and outcomes, Pitches et al found a positive correlation in 51%, no correlation in 31% and a paradoxical relationship (where ‘better’ care process was associated with higher mortality) in 18% (Pitches DW et al 2007). Of the four studies included in this review that specifically examined stroke (Dubois et al 1987; McNaughton H et al 2003; Mohammed MA et al 2005; Weir N et al 2001), one study found no correlation between individual processes of care and patient outcomes (Dubois RW et al

1987), and one identified a paradoxical association where the hospital with the highest summed (unweighted) process scores (RCP NSSA process markers) (Intercollegiate Stroke Working Party 2011) reported poorer patient outcomes (McNaughton H et al 2003). The two remaining (multicentre) stroke studies each found higher mortality rates persisted at one of their study sites following adjustment for case-mix (using the variables of the Six Simple Variables case-mix adjustment model (Counsell et al 2002)), and that these sites were also deficient in aspects of care process (Mohammed MA et al 2005; Weir N et al 2001). No differences in between site mortality (Weir N et al 2001), or relationships between processes of care and standardised mortality rate (SMR) (Mohammed MA et al 2005) remained at other sites following case-mix adjustment, despite significant differences in process delivery across sites.

A systematic review performed in 2007, pooling data from 16 observational stroke studies where adjustment for case-mix or baseline variables had been performed (N=42,236), demonstrated a clear survival benefit at one year for patients receiving organised stroke unit care vs. general ward care (OR 0.79 [0.73,0.86]) (Seenan P et al 2007). These figures are comparable to those demonstrated in the SUT systematic review of randomised trials of stroke unit vs. general ward care (OR 0.86 [0.76-0.98]) (Stroke Unit Trialists' Collaboration 2007).

However, commentators have argued that variations in observed outcome between institutions are more likely to reflect systematic differences in between institution populations (e.g. differences in case-mix, data quality or the role of chance) than true differences in the quality of care.

2.1.6 Case-mix

Case-mix, discussed in detail in Chapter 3, represents the range of disease severity and baseline characteristics that may be the cause of variation in outcomes between individuals and populations (Lilford et al 2004; Mant J 2001). Case-mix has been argued to be a major barrier to the demonstration of process-outcome linkages in empirical studies. Differences in observed outcome between groups have been shown to be wholly or partly attributable to case-mix in a number of experimental stroke studies (Davenport RJ et al 1996; Lingsma et al 2008; Mohammed MA et al 2005; Weir N et al 2001). In other studies, process markers identified as potentially important in determining patient outcome in unadjusted analyses were no longer significant following case-mix adjustment whilst other variables, became statistically significant predictors of outcome following adjustment (Bravata DM. et al 2010).

Where differences in outcome remain after case-mix adjustment in observational studies or empirical populations, it has been argued that this is more likely to reflect unmeasured case-mix variables or confounders than true differences in delivered care (Lilford RJ et al

2004; Mant J 2001). It is unlikely that any case-mix adjustment model will ever account for all potential confounders and as such, many prognostic or specific case-mix variables may remain unmeasured and their effect unaccounted for (Mant J 2001).

2.1.7 Process saturation

A further barrier to the linkage of processes of care with outcome is a consequence of the heightened delivery of specific aspects of care according to existing stroke care monitoring indicators. Many existing markers of process for stroke care have evolved from a systematic evaluation of the key aspects of stroke care process consistently delivered in the effective stroke units identified in the SUT systematic review of organised stroke unit care (Langhorne et al 2002; Stroke Unit Trialists' Collaboration 2007). However, robust RCT evidence to link many of these and other individual processes of care that occur on a stroke unit with improved patient outcomes is often lacking. Where there is expert consensus on specific aspects of care delivery (for example early mobilisation), processes are often adopted into clinical guidance in the absence of RCT trial evidence (Intercollegiate Stroke Working Party 2008). Inclusion of these processes as an accepted part of standard quality care precludes, on ethical grounds, the randomisation that would be required to allow formal clinical trial evaluation of the potential benefit of receiving the process (Black N 1996). Comprehensive adoption of these processes into routine care reduces the variability in care process delivery such that detecting the effect of omitting the process through observational studies becomes more difficult due to a lack of statistical power (see also section 6.1.2.1). As saturation of the process reaches 100%, demonstrating a process is effective becomes impossible. For processes of care where there is a logical rationale for clinical benefit this is unlikely to be problematic. However, for process indicators where the potential benefit to individual patients is not clear cut (for example being weighed at least once during the course of the admission (Intercollegiate Stroke Working Party 2011) it remains unclear whether the process has any impact on patient outcome.

2.1.8 The role of chance

Statistical analyses can offer confidence limits and levels of statistical significance, however, it should be remembered that these are simply reflections of the likelihood of an event being due to chance. Differences in outcome between centres or within centres over time may therefore be due to random variation rather than delivered care. The risk of associations being due to chance is higher when either the numerator (outcome) or denominator (total number of cases) is small (Mant J 2001) – i.e. for rare events or small sample sizes. For example, in an attempt to demonstrate the effect of the introduction of a stroke unit on 30 day and 1 year stroke mortality, Davenport et al collected data on 216 patients pre and 252 patients post introduction of the new service (Davenport RJ et al

1996). Having adjusted for case-mix adjustment, no difference in mortality was observed following introduction of the stroke unit. In response, Mant et al identified that in order to be adequately powered to detect the differences in mortality comparable to those seen in the SUT RCT of organised stroke unit care with 5% confidence and at 80% power, the before and after study would have needed to recruit 2066 patients (Mant J et al 1996).

Where multivariable models have been constructed to explore relationships between process and outcome whilst adjusting for confounding factors (e.g. linear or logistic regression analyses), ensuring that the number of variables entered into the model is appropriate for the sample size is critical in reducing the identification of spurious or chance associations. This is discussed in detail in 3.5.8.1**.**

2.1.9 Data quality

In order for the routine measurement of process markers to be reliable in longitudinal or cross-sectional (between institution) comparisons, there needs to be standardisation in the data that are recorded (Lilford RJ et al 2004; Mant J 2001). This relies on strict data definitions such that it is clear that precisely the same aspects of care are being measured between individuals and between institutions (Lilford RJ et al 2007). In order for this to occur, there should be minimal subjectivity in measurement. Ideally, process markers should be observations (or derivations) of whether and when an explicitly defined event occurred. Variations in measurement of processes (such as could occur if data definitions do not exist or if there are no validation checks to ensure that they have occurred) can lead to spurious data and unfair comparisons.

Some processes appear to have been particularly poorly completed within the SINAP dataset in the 2010/11 data collection period (where data collection and submission were not mandatory) (Royal College of Physicians Stroke Programme 2010). For example, bundle 12 reflects the proportion of eligible patients receiving antiplatelet therapy within 72 hours, *and* “adequate” fluids and nutrition within each 24 hours of the 72 hour audit period. The SINAP data reveal that 25% of trusts achieve this process marker in just 43% of their patients (Royal College of Physicians 2011). The RCP national sentinel audit report (2010) states that 99% of patients (nationally) receive fluids within 24 hours, 95% receive nutrition within 72 hours and 93% are commenced on antiplatelet therapy within 72 hours of admission. It therefore would appear somewhat incongruous that the SINAP data reflect such low achievement of a similar (but not directly comparable) marker, perhaps reflecting differences or difficulties in data reporting rather than true deficiencies in patient care. It should, however, be considered that SINAP as a prospective audit, includes all patients admitted to a particular trust with a stroke diagnosis, whereas the RCP NSSA is performed on the first 60 consecutive stroke admissions in the reporting period (Intercollegiate Stroke Working Party 2011; Royal College of Physicians Stroke Programme 2010).

Missing data is a further important consideration in data quality. Routine documentation of the delivery of specific processes of care may be deficient, thus complicating retrospective data extraction (Walsh et al 2002). In terms of measurement of process, data may be missing because a process was not performed or because data were either not extracted or recorded incorrectly (i.e. non-sense data). If possible, differentiation between these types of missing data adds additional and useful information. For example, actively identifying a process that is consistently not performed (i.e. recording that the data are not available) may indicate a problem with staff training or resource, whilst data that are consistently not recorded may indicate that these data are difficult to extract, highlighting a problem with the indicator itself. Examination of missing data can help to identify missingness patterns (i.e. subgroups of patients that tend to have missing data) as these may lead to bias. Metadata ('data about data') can help to explore these patterns – for example examination of the missing data in relation to disease severity, or by institution.

2.1.10 Data sources

Studies that have attempted to link care process with patient outcome have utilised data from a variety of sources: retrospective routine data (e.g. stroke registers or routine hospital data), retrospective data obtained from case-note review, secondary use of trial data (e.g. data from control arms of RCTs) and prospective observational data defined a priori and obtained expressly for the purposes of the study. The data source affects both data quality, and the conclusions that may be drawn. Often, more than one of these approaches is employed to obtain the necessary data.

There are many sources of routine healthcare data such as locally held stroke databases, hospital records systems (such as Patient Administration Systems (PAS)) and anonymised data held in large central databases (e.g. Hospital Episode Statistics (HES) held by the Information Centre) (The Information Centre 2011a). However, the specific data fields that are recorded and available in these routine databases are likely to limit their use, i.e. the information to answer specific questions may not have been routinely captured within existing systems. If a proxy marker is available this could be used to reflect data that are unavailable (e.g. marital status could be used as a proxy marker of living alone). However, the validity of the proxy marker will depend on how well it reflects the underlying construct (there are many reasons why patients who are married may live alone and many reasons why people who are not married may not). Definitions for data that are collected routinely may not be standardised, and this will be reflected in the quality of the data. In addition, data available from routine data sources for case-mix adjustment is often limited to variables such as age, sex and comorbidity based on hospital coding data and these may not be sufficiently detailed to support complex case-mix adjustment (The Information Centre 2011a).

A further caveat to the use of retrospectively collected data is that the outcomes data that are routinely recorded (or that may be extracted retrospectively) are limited and often restricted to mortality and length of stay. Prevention of post-stroke mortality is not the only goal of therapy or stroke unit care. Many of the complex therapy interventions that occur on a stroke unit are aimed not at preventing death, but at achieving improvements in function and independence. Using death as an outcome fails to capture a 'middle band' of patients who survive but with disability. There are crude measures which would allow this middle band to be quantified (i.e. independent survival, discharge home or modified Rankin Score), although these measures fail to capture the nuances of an individual's post-stroke recovery, and are not currently recorded routinely in England, Wales and Northern Ireland.

Patient case-notes are a rich source of patient specific data. However, data within patient case-notes are not usually recorded in a standard format and as such data extraction from the case-note narrative requires specific expertise. The data that are required may not always be available and extrapolation or 'best guesses' may occur especially if data extraction is not performed by stroke experts. The data extraction process is therefore time consuming and as a result expensive and resource intensive. Case-note data captured retrospectively are often not timely; the 2010 RCP sentinel stroke clinical audit report was published 11 months after the end of the data extraction period (Intercollegiate Stroke Working Party 2011).

An alternative approach to obtaining process and outcomes data in empirical populations is through prospective data collection as part of usual care. The development of electronic systems has facilitated data capture and submission such that routine, prospective and cumulative data collection is now feasible. As a consequence many countries now host electronic stroke databases (Australian Stroke Clinical Registry 2011; Dennis M et al 2011; Royal College of Physicians Stroke Programme 2010; Asplund K et al 2011)**.** Routine collection of functional outcomes is currently in operation in the Australian and Swedish registries (Australian Stroke Clinical Registry 2011; Asplund K et al 2011). Co-ordinating large scale data collection requires robust electronic infrastructure and methods for ensuring data quality, cleaning and obtaining missing data.

2.2 Measuring quality through Process

2.2.1 Benefits of process driven care

It has been argued that identification of deviations from care processes can detect discrepancies in quality of patient care with more sensitivity than can be achieved through the measurement of outcomes (Mant J 2001).

In a simulation study, Mant and Hicks (1995) demonstrated the relative sensitivity of process and outcomes measurement to detect discrepancies in quality of care using the specific example of mortality following myocardial infarction (MI) (Mant J et al 1995). The calculated combined effect of proven acute pharmacological interventions was used to model the effect of different rates of uptake of therapies on mortality between two theoretical 'hospitals' identical in all other respects. Calculations of sample size revealed that the deviations from care process (defined as a failure to administer treatment to patients in whom it is indicated) that would result in a difference in mortality of 9% between institutions (0% vs. 55% process compliance) could be detected within 2 weeks (12 patients in each institution). Detecting this difference through direct measurement of mortality (with power of 80%, 2p=0.05) would take just over ten months and 389 patients (Mant J et al 1995). This demonstration does, however, assume a linear relationship between mortality rates and the use of effective interventions, and a linear cumulative effect of interventions (Mant J et al 1995). This simulation study elegantly demonstrates that detection of deviation from process is likely to be a more efficient way of detecting poor compliance with care processes than waiting for the effect of defective processes to be borne out through mortality rates. However, the measurement of process for the purposes of quality monitoring, simply informs whether a particular process occurred or not. No inference can be drawn about the effect of missing a particular process on an individual patient's outcome.

Detection of deviation from care processes identifies deficiencies in procedural aspects of care. Through this direct detection of deficiencies in care process delivery, improvements to the average quality of care can be achieved in all institutions regardless of baseline quality of care or observed outcomes (Lilford RJ et al 2007; Lilford RJ et al 2004). In other words, there is a paradigm shift towards improved care (Lilford RJ et al 2004; Lilford RJ et al 2010).

A major argument in support of the measurement of process as a marker of quality care is the immediacy with which data are available. Lilford et al have consistently argued that the delay between the delivery of process and detection of the effects of deviations from process through outcome measurement makes it hard to ascribe differences in outcome to deficiencies in care process (Lilford RJ et al 2007). Moreover, it is easier and more timely to record and rectify deficiencies in process delivery through audit, than waiting to estimate the effect of an event not occurring through observing patient outcomes (Lilford RJ et al 2007; Lilford RJ et al 2010; Lilford RJ et al 2004).

2.2.2 Drawbacks to process driven care

The validity of datasets made up of process metrics for the monitoring of quality of care in clinical situations rely on some key assumptions:

a. Patients benefit equally from interventions.

The use of process metrics as measures of quality of care could increase the likelihood that, providing specific interventions are not contraindicated in individuals, all patients receive similar care. An assumption that all patients require the same ‘bundles’ of care processes, overlooks the possibility that particular interventions may be of more, or less relevance to particular subgroups of patients and could “result in standardised care of little relevance to individuals” (Walsh K et al 2002). Some existing process measures for stroke (e.g. the RCP NSSA markers) circumvent this problem through the inclusion of explicit criteria to allow patients to be allocated a ‘no but’ code. These codes are allocated to patients in whom an intervention is either not indicated (e.g. a patient with no speech or language deficit that does not require a SLT assessment) or contraindicated (e.g. patients receiving palliative care). Patients allocated ‘no but’ codes are removed from the denominator in RCP NSSA process scores.

Other datasets, however, calculate percentage compliance with interventions using the whole population as the denominator, but build in ‘tolerances’ to account for patients in whom interventions are not indicated, or contraindicated. For example, the ASI metric that 60% of patients admitted with stroke and in atrial fibrillation should be on anticoagulation, or have a plan for anticoagulation by discharge from hospital (Stroke Improvement Programme 2011). Here, patients in whom anticoagulation is contraindicated, or that refuse, are included in the 40% tolerance limit. This approach requires careful planning to ensure that the tolerances are reasonable and to reduce the risk of gaming, or inappropriate prescribing.

The Stroke Unit Trialists’ (SUT) concluded that all patients benefit from stroke unit care (regardless of the severity of stroke) (Stroke Unit Trialists' Collaboration 2007) and that the improved outcomes observed in patients treated on a stroke unit vs. general wards is likely to be due to the prevention of post-stroke complications (Langhorne P et al 2002). However, there is likely to have been significant heterogeneity in the processes of care delivered on the units within the included trials as participants were randomised to organised stroke unit or other “conventional care” rather than to specific treatment protocols (Stroke Unit Trialists' Collaboration 2007)**.** The SUT review could therefore be interpreted to suggest that *the average* care delivered on the stroke units included in the systematic review is beneficial to a population of heterogeneous post-stroke patients as compared with the *average* care that is delivered on conventional wards.

b. The measurement of individual processes is equally important

There are likely to be some care processes that are more important than others in terms of the effect that their omission may have on patient outcomes (Sudlow C et al 2009). If care process is to be used to assess the quality of delivered care, especially if institutions are to be compared on the basis of the delivery of these processes, weighting to account for this relative importance may be appropriate. For example, a unit that achieves delivery of complex care processes of benefit a few patients may do so to the detriment of delivery of care that is of potential benefit to all. A failure to apply weighting to metrics designed to reflect capture of care processes that are only indicated in small subgroups of a population may result in disproportionate emphasis on particular aspects of care. For this reason, the denominator (case-volume or the number of patients in whom a process is indicated) should form an important consideration of between or within institution comparisons of process delivery (see section 2.4). However, it has been argued that the denominator for the measurement of process should be the number of opportunities for a process to have occurred, rather than the number of patients in whom it is indicated (Lilford RJ et al 2004). This approach incorporates a form of case-mix adjustment for process measures as there is more scope for omission of processes of care that occur repeatedly (e.g. administration of aspirin), or in patients who require more interventions (e.g. those with severe strokes) (Lilford RJ et al 2004).

Some stroke care processes (e.g. stroke unit care or antiplatelet therapy) have been shown to be of benefit to patients regardless of stroke severity (Chen Z-M et al 1997; Stroke Unit Trialists' Collaboration 2007), whilst the use of other interventions (for example thrombolysis) is restricted to those fulfilling specific criteria (Boehringer Ingelheim 2009; Intercollegiate Stroke Working Party 2008).

The numbers needed to treat (NNT) to prevent an adverse outcome is greater for some processes of care than others. For example the NNT with aspirin to prevent one death or dependent outcome following an ischaemic stroke is 67, as compared with a NNT of 10 to avert the same outcomes following treatment with alteplase (rtPA) (Sudlow C et al 2009). However, the relatively small *treatment effect* of aspirin is offset by the treatment *achievability* and *eligibility* i.e. the administration of aspirin is a relatively simple task and should be achieved consistently and completely in all patients in whom it is not contraindicated (the proportion of the acute post stroke population in whom antiplatelet therapy is indicated is estimated at 80% (Langhorne P et al 2009). In contrast, an estimated 10% of patients admitted to hospital with acute stroke are eligible for thrombolytic therapy (Langhorne P et al 2009). Moreover, achievability of thrombolysis is currently limited by the considerable infrastructure required to deliver it safely (Sudlow C et al 2009) (5% of patients admitted to hospital with acute stroke in England, Wales and Northern Ireland received thrombolysis in the 2010 RCP NSSA (Intercollegiate Stroke Working Party 2011)) . Despite a relatively small NNT to prevent an adverse outcome following thrombolysis (large treatment effect), the number of adverse outcomes actually averted is attenuated by limited achievability and eligibility as compared with other treatments (Sudlow C et al 2009).

c. The process is important and missing it has a detrimental effect to individuals rather than the institution

Some processes may be correlated with patient outcome, although the nature of this relationship may be complex (Rubin HR et al 2001). For example, it is difficult to see how the RCP NSSA process measure "Is there evidence that the patient was weighed at least once during their admission?" (Intercollegiate Stroke Working Party 2011) would relate directly to patient outcome. However, patients who are weighed may be more likely to also receive other aspects of care relating to nutrition that may have a more causal relationship with outcome; such processes have been termed 'tracers' (Walsh et al 2002). Weighing a patient may therefore be acting as a proxy marker for other aspects of care that may be more difficult to capture. Close examination of markers where there is correlation with outcome may reveal what these additional markers could be, although it may be difficult to estimate the effect of failing to achieve proxy measures of care on patient outcome if the underlying constructs are unmeasured and unknown.

2.2.3 Quality and interpretability of data

2.2.3.1 Gaming

The use of process data for remuneration, performance ratings or for between institution comparisons runs the risk that the care is focused on the meeting of targets rather than reflecting the broader context of care that the process markers are designed to represent (Davies HTO 2005). There is a possibility that situations will be manipulated to allow such targets to be met (Mears et al 2010), indeed examples of gaming in the health service are well documented (Bevan G et al 2006). Pejorative comparison of institutions based on unadjusted process (or outcome) measures can therefore be potentially damaging both to those institutions and to the patients they treat:

> ***"Reward and punishment strategies do not produce knowledge;***
> ***they produce fear and anxiety often leading to distortion of the data***
> ***or the process" (Lilford RJ et al 2004)***

2.2.3.2 Representation of data

The importance of data interpretation and presentation is highlighted here using the example of the first Royal College of Physicians SINAP report (2011). During the reporting period, participation in data collection for SINAP was not mandatory, indeed nationally only 82 out of 157 trusts submitted data, with nine of these trusts submitting insufficient data for analysis (Royal College of Physicians 2011). Seventy-three trusts were therefore included in the data analysis. The SINAP report presented the number and percentages of (eligible) patients receiving specific processes, 'bundles' of care (patients receiving combinations of processes) and an average process score as the unweighted mean of the percentages of eligible patients receiving each of 12 key processes.

There are a number of problems with this approach. Firstly, presentation of an unweighted score fails to account for the relative importance or difficulty of delivering individual processes of care – i.e. processes that are simple to achieve are given the same scoring weight as more complex processes. The use of total scores as a summary measure also has implications in terms of scaling properties (i.e. an assumption that a summary score may be treated as an interval scale may not be valid).

Calculating a mean from a series of percentages, as has occurred with the SINAP data, is a flawed approach, as the denominator (case volume) for each individual process has not been accounted for. Centres with small volumes of cases are more likely to demonstrate extreme values as variations in process delivery would have a disproportionately influential effect on their overall score (O'Brien S et al 2008). The denominator for each process marker can be calculated from the numerator and proportion of eligible patients as detailed in the SINAP data spreadsheet (available from the RCP SINAP website (Royal College of Physicians 2011)). Notwithstanding the fundamental problems with calculating summed process scores (see above), the analyses performed by SINAP have been repeated here simply to demonstrate the effect of consideration for case volume on the relative position of patients in 'league tables'. If the sum of patients at each site receiving all the processes of care for which they are eligible are presented as a proportion of the sum of patients eligible for each of the processes, 13 of the 73 trusts move up a quartile, 10 move down a quartile and one trust moves from the top (1st) to the third quartile of all the trust scores. These marked movements in the relative 'position' of trusts based purely on the volume of patients treated, without consideration of the problems encountered with summed total scores or more complex factors such as case-mix, are an indication of the potential difficulties in publishing data in the public domain. This is discussed further in section 2.4.

2.3 Outcomes driven stroke care

The distinction between outcome indicators and patient reported outcomes (as outlined in section 2.1.4.2) is an important one. Hard, objective endpoints (e.g. mortality following stroke) offer no information as regards the complexities of stroke recovery from the perspective of the individual, indeed prevention of mortality is not the only goal of therapy. Patient reported outcomes can offer information regarding aspects of patient outcome that cannot be measured through other means (Mant J 2001), but the question remains: how can or should this information be used at a population level?

Comparison of institutions based on patient reported outcomes as subjective measures runs the risk that any between institution differences are attributed to the quality of delivered care. However, aside from arguments regarding the availability of more appropriate methods to detect deficiencies in the quality of care processes (2.1.5), differences in patient reported outcome are dependent on the way in which individuals perceive their healthstate (Lilford RJ et al 2007). Moreover, good outcome may occur despite failures in the delivery of care process and vice versa (Mant J 2001; Walsh K et al 2002).

2.3.1 What is outcomes measurement good for?

Subjective outcomes assessment by patients offers some benefits that cannot be achieved through the isolated measurement of care processes:

a. Useful information at individual level

Outcomes measurement is arguably of more importance and relevance to individuals than individual aspects of care process. When considered at an individual level, patient reported outcomes offer a valuable resource. Interventions may be tailored to specific needs to inform care at an individual patient level i.e. to facilitate discussions regarding targeted longer-term treatments or therapy goals for individuals. Cumulatively, this information could help to inform the identification of gaps in local service delivery. Outcomes measurement may therefore be useful as part of a feedback loop to plan ongoing care, rather than as a method of evaluating the care that has already been delivered.

A broad overview of delivered healthcare at a population level

> ***"…every hospital should follow every patient it treats long enough to determine whether or not the treatment has been successful, and then should inquire, "If not, why not?" with a view to preventing similar failure in the future"***
>
> ***Ernest Codman (1869-1940)***

The measurement of outcomes measurement can provide a 'broad barometer' of delivered care (Lilford RJ et al 2007; Mant J 2001) – i.e. the cumulative effects of complex interventions, service structure and individual characteristics. Identification of substantial outliers, i.e. where there is deviation beyond that expected through 'normal variation' following case-mix adjustment could allow identification of institutions where outcomes are particularly good, or less good than expected (Lilford RJ et al 2004; Mohammed MA 2001). Examination of these institutions may reveal areas for further exploration, or systematic differences may lead to generation of hypotheses regarding 'what might work' (Lilford RJ et al 2007). Identification of such outliers is likely to be best achieved through funnel plots and this approach is discussed in detail in 2.4.1.

2.3.2 Drawbacks of outcomes measurement

Aside the problems in linking care process with outcome (2.1.5), there are a number of technical and logistical difficulties in the routine collection of patient reported outcomes.

One of the major considerations in terms of outcomes measurement is which instrument should be used. Outcomes such as mortality or length of stay data are more readily available and may already be routinely recorded in existing systems. However, these offer no information regarding post stroke function in stroke survivors, and are not useful to inform individual patient care. The alternative is the collection of patient reported outcomes. There are a number of stroke specific and generic questionnaires to assess various domains of patient functioning following stroke. Any measurement scale for this purpose should be valid and reliable in stroke populations and should ideally have had these psychometric properties tested in a number of different datasets. Previous reviews (Jenkinson C et al 2009; Teale EA et al 2010) and online resources (Salter K et al 2010) aim to identify the optimal outcomes instrument for stroke, but consensus is lacking.

Routine collection of patient reported outcomes requires considerable infrastructure. Collection of outcomes data face to face is unlikely to be feasible in routine care due to resource costs and time restraints. Existing infrastructure could be exploited (e.g. clinic attendance, or the community stroke team) in order to collect outcomes information, although these assessments are unlikely to be sufficiently standardised to ensure robust data collection. Postal questionnaires are an alternative, but introduce problems with return and completion rates, proxy completion and stroke specific problems such as the impact of dysphasia on the questionnaire completion. Postal questionnaires also need to be triggered at an appropriate time, following checks of residency and survival. All patient-completed questionnaires (unless completed and submitted electronically) will require some form of data entry resource.

2.4 Problems with presentation of data in the public domain

Process and outcomes data are readily available in the public domain, including patient reported outcome data (The Information Centre 2011c), satisfaction ratings (Ipsos MORI 2011) and league tables (Dr Foster 2010) for a variety of conditions including stroke (Dr Foster 2010; Intercollegiate Stroke Working Party 2011; Royal College of Physicians 2011). Often the data presented are standardised or adjusted for case-mix variables, but these are often those that are available from existing resources such as the Hospital Episode Statistics database (HES) (e.g. Dr Foster 2010). Data may therefore have been adjusted by the case-mix variables that are available, rather than those that have been validated through research or that make the most clinical sense.

Appropriate interpretation of complex activity or outcomes data by the public, clinicians and commissioners is dependent on the way in which these data are presented (Davies HTO 2005). As patient choice becomes more prevalent and increasingly likely to drive commissioning decisions (Department of Health 2010a), the interpretation of data that may be used to inform decision making is key. It is unclear how well the public are able to interpret these complex data (Scott IA et al 2006). However, the presentation of data that are flawed or difficult to interpret is likely to be damaging to individual provider trusts both financially and in terms of reputation (Davies HTO et al 1997; Lilford RJ et al 2004).

2.4.1 Common cause versus special cause variation

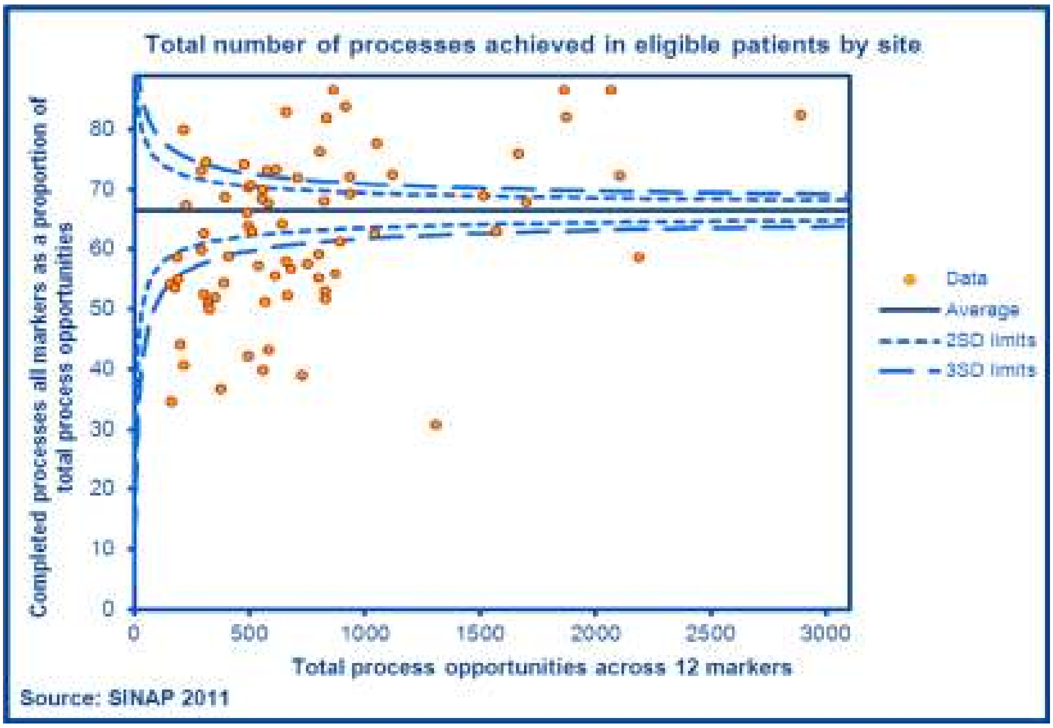
Mohammed et al explored the potential use of Shewhart charts (or statistical process control charts) to identify common cause (expected random) variation in outcome from special cause variation (due to an external influence on an otherwise stable process) (Mohammed MA 2001). A funnel plot presents similar information, but takes account of case volume. The funnel plot represents interval level data, and can be used for normally distributed data, proportions (based on a binomial distribution) or count data (based on a Poisson distribution). For example, observed cases as a proportion of potentially eligible cases could be plotted against the potentially eligible cases to account for differences in sample size (Speigelhalter DJ 2005; Speigelhalter D 2002).

Providing that the markers approximate the appropriate distribution, funnel plots may be used to present either deviations from process or special cause variation in objective or patient reported outcomes either between institutions (Gale CP et al 2006) or within institutions over time (Henderson GR et al 2008). Special cause variations (greater than 3 standard deviations from the mean or a chance probability of 1 in 500 (0.2%)) are highly unlikely to be due to chance and could be examined further for external causes or influences (Mohammed MA 2001; Speigelhalter DJ 2005). Special cause variation could indicate an important case-mix variable or confounder that has not been accounted for but that is particularly important at an individual site or at a particular time (e.g. temporary loss of scanning resource due to a broken scanner). Examination of the data in this way allows rational interpretation and may lead further exploration or investigation as required.

Figure 3 uses data extracted from the 2011 RCP SINAP data spreadsheet (Royal College of Physicians 2011) to highlight that similarly, funnel plots may be a more useful way to present complex stroke data than the summed averages of percentage process scores that have been presented on the SINAP website (Royal College of Physicians 2011) (see also section 2.2.3.2). The proportion of patients receiving the processes that they require across trusts and the sum of the patients in whom these processes are indicated were calculated as discussed in section 2.2.3.2. The sample sizes from which these proportions have been calculated are large enough for the binomial distribution of the proportions to approximate a normal distribution (according to the central limit theorem).These data were therefore used to create a funnel plot (Figure 3). A number of hospitals are outwith the 3SD limits and therefore show ‘special cause variation’ from the average proportion (across all trusts) of patients in whom required processes are achieved. Trusts within the outer limits are within "common cause" variation – i.e. their deviation from the mean is within the bounds of chance variation. There are benefits to examining both high and low outliers – to learn lessons from those that perform well and to explore further those that appear to perform less well. There are a number of possible reasons for the apparent differences between sites e.g. differences in measurement or data recording or case-mix that should be explored before the disparity is attributed to a true failure to achieve the process markers (Lilford RJ et al 2004).

In 2011, The Information Centre moved towards the presentation of PROMs data in the form of funnel plots as they are "…relatively easy to produce, readily interpretable and allow for additional variability in institutions with small volumes." (Department of Health 2011b)

Figure 3 Funnel plot to present total process scores from SINAP 2011 audit as a function of case volume



Source: SINAP 2011

2.5 Conclusions

Both process and outcomes driven approaches to the measurement of the complex and multifaceted construct of quality in healthcare have been advocated. In order for such measurements to be meaningful it is important to be mindful of the purposes of the data collection. It is unlikely that exclusive measurement of process, objective outcomes indicators or patient reported outcome measures will capture all aspects of patient care.

Process monitoring is useful for detecting deviations from agreed protocols and best practice, although the effect of these deviations may not be detectable through the measurement of outcome. Monitoring of processes of care may also lead to a data driven approach to care provision that may fail to meet the needs of individual patients. Measurement of patient reported outcomes offers a unique insight into broader aspects of care, and may identify areas of service deficiency and need at an individual and population level.

The complexity of the relationships between process, outcome and case-mix make interpretation of routine data problematic. Monitoring of quality through process measurement requires knowledge that the process (or omission of the process) has an effect on patient outcome, whilst routine measurement of patient outcomes would be enhanced through knowledge of whether processes of care are affecting outcome such that they may be monitored and improved. Both approaches are therefore dependent on the demonstration of robust adjusted process outcome linkages. Case-mix adjustment is key to the exploration of these linkages and this will be discussed in more detail in Chapter 3**.**

It is likely that a combined approach to the measurement of stroke care that encompasses aspects of care process and patient reported outcome will give the most useful perspective of the stroke care pathway of interest to commissioners, service providers, clinicians, patients, and researchers. Routine capture of a dataset that includes these key aspects of care could also allow further exploration of the complexities of case-mix adjustment in stroke care.

Chapter 3 Systematic review case-mix adjustment model(s) in stroke

3.1 Introduction

Stroke is a heterogeneous and complex clinical syndrome. The clinical course and outcomes for individual patients following stroke are dependent not only on the site and/or size of the pathological lesion, but on the context of the injury in relation to combinations of mediating factors that are unique to individuals. For example, pre-stroke function, co-morbidities, social environment and rehabilitation potential are all likely to affect an individual's functional, cognitive and social outcomes. It is the combination of these complex factors that contribute to case-mix and make direct comparisons between individuals or empirical populations following stroke problematic and unreliable. It is, therefore, over simplistic and potentially misleading to consider the effect of treatments and therapies on patient outcome following stroke without considering the mediating effects of these other factors (see Chapter 2, section 2.1.6).

Randomisation in clinical trials aims to balance the unmeasured confounders and biases between intervention and control groups. As such, in adequately randomised stroke trials, outcomes in two (or more) populations may be legitimately compared. Nevertheless, through the role of chance in random allocation, differences in important prognostic factors may remain between groups (Altman D 1985) and differences in outcome may reflect these imbalances rather than the true effect of the intervention. For recognised and measurable risk factors, effects may be tempered by minimisation procedures (active balancing of patients with certain characteristics between intervention and control arms). Biases introduced by more immeasurable confounders may be attenuated by stratification according to predicted outcome prior to randomisation.

In routine care and in unadjusted observational studies, the possible contributions of mediating factors are not accounted for and their effect on outcomes remains unmeasured and unknown. Inadequate (or absent) case-mix adjustment may therefore preclude meaningful examination of the relationships between care process and outcomes in observational studies (Mant J 2001) (see also section 2.1.6)**.** Through prognostic modelling, case-mix adjustment of empirical post-stroke populations allows stratification into more homogenous groups according to predicted outcomes. Within such strata, observed outcomes between groups of individuals may be directly compared more legitimately. The influence of specific prognostic factors on patient outcome may be non-uniform across the spectrum of stroke severity (i.e. some factors are more or less important in certain subgroups of patients) (Lilford RJ et al 2007). In observational studies, adjustment for these factors can result in the 'constant risk fallacy' – a paradoxical increase in bias between groups (Nicholl 2007).

Stratification of populations according to predicted outcomes therefore has uses in both clinical care (e.g. targeting of appropriate therapies) and research (e.g. in examination of non-randomised populations or for stratified randomisation) (Counsell C et al 2001). However, such analyses must still be interpreted with caution as important differences and a degree of heterogeneity are likely to remain within strata even following case-mix adjustment (Mant J 2001).

3.2 Assessment of prognostic models

There are no universally accepted criteria to assess the quality of prognostic studies (Altman D 2001; Hemingway H et al 2010; Mallett S et al 2010b). However, there is both generic and disease specific literature that has identified key clinical and statistical criteria that should be considered in model development or assessment (Altman D 2001; Counsell C et al 2001; Harrell FE et al 1996; Hayden JA. et al 2006; Kwakkel et al 1996; Laupacis A et al 1997; Mallett S et al 2010b; Mallett S et al 2010a; Perel P et al 2006; Wyatt JC 1995). This broadly concerns consideration of model internal, external and statistical validities.

A "systematic review of reviews" published by Hayden et al in 2006 considered the quality of reviews to identify clinical prediction models across a range conditions (Hayden JA. et al 2006). This review suggests a "framework of potential biases" that should be considered in the assessment of the quality of studies to develop prognostic models. This framework largely considers internal validity of models across the broad categories of representativeness of the study population, attrition (and consideration of whether loss-to follow up could be systematic), inclusion and accurate measurement of appropriate prognostic information, accurate definition and measurement of valid and reliable outcomes, consideration of confounding and the use of appropriate modelling techniques. Within each of these categories a number of criteria are specified to aid consideration of whether or not the study to develop the model is adequate (Hayden JA. et al 2006). However, this framework does not address key issues relating to the models that have been developed, for example external validity or feasibility of prognostic models in terms of their clinical utility and ease of data collection.

Several authors have provided detailed discussion regarding the development of robust models (Harrell FE et al 1996; Mallett S et al 2010a; Wyatt JC 1995). There are several factors that should be taken into account during model development and failure to do so may affect the stability and utility of models. This includes consideration of sample size and the number of variables that may be entered into the model, prospective data collection, representativeness of population samples, coding of variables (and proper classification of continuous variables) and variable selection.

In a 2001 review, Counsell et al tabulate 25 separate criteria for assessing the quality of studies to develop prognostic models in stroke. The broad categories of internal, external and statistical validities, model evaluation, and feasibility are considered. These criteria incorporate many of the methodological quality markers identified in other studies (Counsell C et al 2001).

Previous reviews have been undertaken to identify prognostic models specifically in stroke (Counsell C et al 2001; Hier HB et al 1991; Jongbloed L 1986; Kwakkel G et al 1996; Meijer et al 2003a; Meijer et al 2003b; Meijer et al 2004; Segal M et al 1997) and show these models to be generally poor. One of these reviews identified studies describing models to predict stroke survival, survival in an independent state or alive and at home (Counsell C et al 2001). The vast majority of the 83 discrete prognostic models identified demonstrated significant flaws in statistical or internal validities. Only four met 8 simple quality criteria of internal and statistical validity defined by the authors, and none was fit for purpose to case-mix adjust in routine clinical care (Counsell C et al 2001). Other authors have attempted to identify case-mix adjustment models which were developed to predict functional outcomes following stroke (Jongbloed L 1986; Kwakkel G et al 1996). However, these have tended to be limited to prediction of activities of daily living; most commonly the Barthel Index which has limitations due to its marked, and well documented, ceiling effects (Salter K et al 2010).

Since these reviews were performed, clear evidence demonstrating the benefits of organised specialist multidisciplinary stroke care over general ward care (Stroke Unit Trialists' Collaboration 2007) has led to the widespread adoption of this model and fundamental changes to the delivery and monitoring of stroke care across health care systems (American Stroke Association's Task Force on the Development of Stroke Systems 2005; Lindsay MP et al 2010; Thomassen L et al 2006). It is possible that prognostic factors previously unknown or overlooked are important in determining patient outcomes and these should be modelled explicitly. In addition, increasing scrutiny of the quality of prognostic research (Altman D 2001; Hayden JA. et al 2006; Hemingway H et al 2010) and more sophisticated statistical modelling techniques (e.g. multilevel modelling, latent variable analysis and structural equation modelling) are likely to have altered the type and quality of models to predict outcomes following stroke.

A systematic review of the literature was therefore undertaken in order to update previous reviews (in light of the above factors) and identify any externally validated prognostic model to predict outcome in unselected populations following acute stroke comprising simple clinical variables feasible for collection in routine care.

3.3 Methods

3.3.1 Review team

To minimise bias and in accordance with suggested guidelines (Altman D 2001; Centre for Reviews and Dissemination 2009), this review was designed and conducted by a team with a variety of skills led and co-ordinated by Elizabeth Teale (ET). The search strategy was developed in collaboration with a colleague at Leeds University Library (Deidre Andre (DA), Research Support Officer). Development and definition of inclusion criteria was undertaken by ET. Screening of titles was performed by Anita Ranjendran (AR, Medical Student University of Leeds Medical School) and Anne Forster (AF, Professor of Stroke Rehabilitation, University of Leeds). Subsequent screening of titles for which consensus had not been met was performed by ET. Review of abstracts and selection of studies for inclusion was performed by AF and ET. Double data extraction was performed by Ruth Lambley (RL, Research Assistant, Academic Unit of Elderly Care and Rehabilitation) and ET. Statistical appraisal of identified models was performed by Theresa Munyombwe (TM, Medical Statistician, University of Leeds) and ET. Consolidation and synthesis of findings was performed by ET.

3.3.2 Information sources

A comprehensive search strategy was developed with a colleague at Leeds University Medical Library (DA) combining terms to identify stroke studies (as developed by the Cochrane Stroke Group (Cochrane Database of Systematic Reviews Stroke Review Group 2009) with terms to describe prognostic modelling. The full search strategy is included in Appendix A-1 Searches were run in MEDLINE, EMBASE, CINAHL, PsycInfo, AMED and ISI Web of Science with no date or language limits until 30th May 2009. Results were downloaded into EndNote™ (version X2.0.1) and duplicates removed.

3.3.3 Study selection

Titles were examined by two independent reviewers (AR and AF) and obviously irrelevant titles excluded. A third reviewer (ET) examined titles where there was no agreement to ensure all relevant titles were included. Abstracts of potentially relevant papers where there was agreement between at least two of the three reviewers were then further examined by two reviewers (AF and ET). Papers fulfilling inclusion criteria were retrieved in full text.

Studies were included if they described development or external validation of a discernible prognostic model at a fixed time point following ischaemic or haemorrhagic stroke. Studies referring to ‘adjustment for baseline variables’ were excluded unless the method of adjustment was further qualified.

Only studies describing models with variables considered to be feasible for collection in a routine care setting, by ward staff and within two weeks of stroke were included. Prognostic models that required specific radiological or laboratory test results were excluded as the aim is to identify a case-mix adjustment model comprising variables that may be collected at the bedside.

For the purposes of this review, an assumption was made that not all services are currently set up to facilitate data collection requiring specialist assessment on, or within a few hours of, admission to hospital. Models that require collection or measurement of case-mix variables requiring a level of expertise above that expected on a typical medical assessment unit by non-specialist stroke clinicians (for example the National Institute of Health Stroke Score (NIHSS)) were therefore excluded.

Similarly, models requiring the collection of case-mix variables within six hours of presentation were excluded as patients not admitted directly to a stroke unit (or those transferred to a stroke unit from the Emergency Department within four hours) are unlikely to have case-mix variables collected reliably within this time frame. The rates of direct admission to specialist stroke units are improving in England, Wales and Northern Ireland. The proportion of patients admitted directly to an acute stroke unit increased from 29% to 56% between the 2008 and 2010 RCP clinical audits (Intercollegiate Stroke Working Party 2011; Royal College of Physicians 2009b). A continuation in this trend is likely to occur with the expansion of thrombolysis services with initial assessments increasingly likely to be made by more senior and specialist stroke clinicians. This may, in the future, facilitate the collection of more complex baseline data at the point of admission (e.g. complex clinical prognostic scoring systems) and shorten the time frames within which collection is feasible.

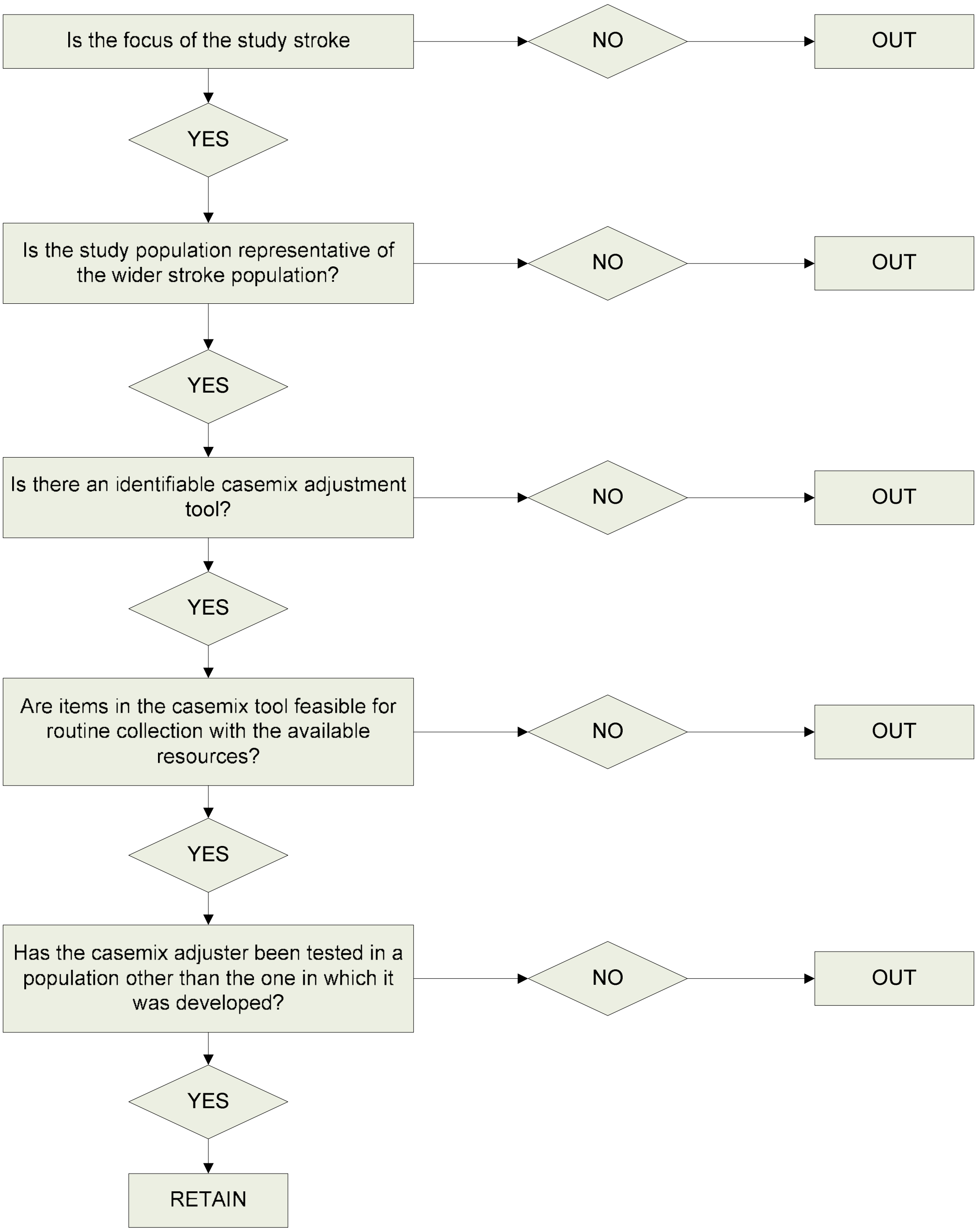
Previous reviews have either limited their searches to identify models predicting functional outcomes (Kwakkel G et al 1996) or have limited the scope of their review to consider the outcomes of death and dependency (as defined by a dichotomised modified Rankin Score)(Counsell C et al 2001). We aimed to identify all available prognostic models to predict any post stroke outcome (including mortality, disability and functional outcomes).

Models developed in populations unlikely to be representative of the wider stroke population (e.g. exclusion of the oldest old, or patients at the extremes of stroke severity) were excluded as models developed in such populations are unlikely to be generalisable to unselected stroke populations. Similarly, prognostic factors for stroke are likely to differ from those of transient ischaemic attack (TIA) and subarachnoid haemorrhage (SAH). Models developed to predict outcome following TIA or SAH were therefore excluded.

External validation refers to the testing of models in populations independent to those in which the model was developed. Ideally, this should occur in an unselected population in a different institution from that in which the model was developed. Models without evidence of external validation were excluded from this review. Where it was not clear whether a model had been externally validated, the paper was retained for further scrutiny.

A flow chart of inclusion and exclusion criteria is presented in Figure 4.

Figure 4 Citation screening inclusion and exclusion criteria


Is the focus of the study stroke
NO
OUT
YES
Is the study population representative of the wider stroke population?
NO
OUT
YES
Is there an identifiable casemix adjustment tool?
NO
OUT
YES
Are items in the casemix tool feasible for routine collection with the available resources?
NO
OUT
YES
Has the casemix adjuster been tested in a population other than the one in which it was developed?
NO
OUT
YES
RETAIN

3.3.4 Data extraction

Data extraction was performed in duplicate by ET and a fourth independent reviewer (RL).

Details regarding the name of the case-mix model, author, model variables, reference population (inception cohort and study exclusion criteria), prospective or retrospective data collection, losses to follow up, outcome measures (and time point of measurement), sample size, external validation of model and feasibility of collection of independent variables were extracted if available. Studies describing the development of models and subsequent validation studies were then grouped together.

3.3.5 Data items

Initially, criteria to assess the quality of each model were applied. These were extracted from a framework of criteria used to assess internal validity of models (Hayden JA. et al 2006) and a broader set of criteria to examine quality of models as used by Counsell et al in their 2001 review (Counsell C et al 2001). The criteria used at this stage in the review were selected to cover the aspects of model quality considered to be essential: adequate inception cohort, prospective data collection, a description of patients lost to follow up (and no systematic exclusion or drop out of particular patient subgroups), clinical relevance of prognostic factors, assessment of valid and reliable outcomes at a fixed time point, no inclusion or exclusion criteria that might limit generalisability, and variables that are feasible to collect in routine care.

Independent statistical appraisal of the model development was then performed by TM to assess aspects of statistical quality as regards model fitting. This included consideration of sample size, variable selection techniques, consideration of collinearity and interaction terms (Harrell FE et al 1996). Regression models require a linear relationship between the independent and dependent variables, and normally distributed model residuals with constant variance (the difference in observed and predicted outcome for each case) (Fox J 1997p 113). These assumptions should have been checked explicitly to demonstrate that that the models are statistically robust. Information regarding any testing of modelling assumptions was therefore also extracted.

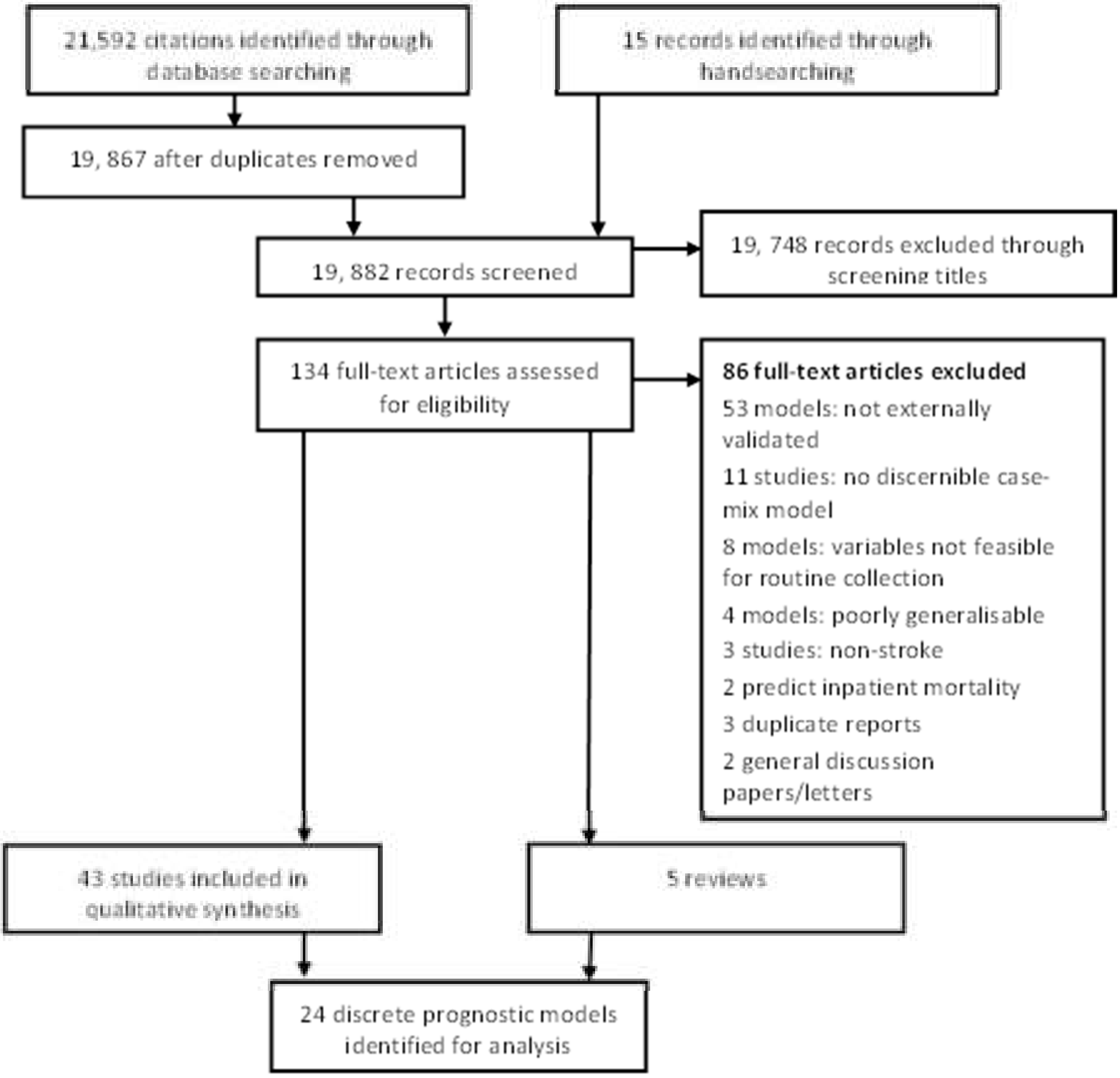
3.3.6 Model performance

Measures of model performance were extracted from external validation studies for each model. These comprise measures of discriminatory function (e.g. the c statistic) sensitivity/specificity analysis and calibration of models in independent populations (Altman D et al 2000; Altman D et al 2009).

3.4 Results

After the removal of duplicates, the initial search identified 19,867 titles. Screening of titles (to exclude obviously irrelevant citations) and abstracts (based on inclusion and exclusion criteria (Figure 4) led to two independent reviewers agreeing to the retention of 176 citations. In 487 further citations where consensus between these two reviewers was not met, the opinion of a further independent reviewer (ET) resulted in the inclusion of an additional 183 potentially relevant citations. A discussion (based on abstracts) between AF and ET regarding relevance for inclusion in the review of the 359 identified papers resulted in the retention of 119 papers for examination in full text. Handsearching of the reference lists of these papers (ET) identified a further 15 potentially relevant citations. A total of 43 of the papers were retained for data extraction. In addition, five previous reviews were examined to identify any models that may otherwise have been overlooked. A flow-diagram of the selection process can be found in Figure 5.

Figure 5 Identification of citations for inclusion in the review (Teale et al 2012)


21,592 citations identified through database searching
15 records identified through handsearching
19, 867 after duplicates removed
19, 882 records screened
19, 748 records excluded through screening titles
134 full-text articles assessed for eligibility
86 full-text articles excluded
53 models: not externally validated
11 studies: no discernible case-mix model
8 models: variables not feasible for routine collection
4 models: poorly generalisable
3 studies: non-stroke
2 predict inpatient mortality
3 duplicate reports
2 general discussion papers/letters
43 studies included in qualitative synthesis
5 reviews
24 discrete prognostic models identified for analysis

Twenty-one case-mix or prognostic models predicting a range of outcomes were described in 43 papers (Table 1). In addition, two studies described the use of 3 existing impairment scales to predict patient outcome post-stroke (Lai et al 1998; Muir KW. et al 1996). Of these 3 models, only one was used in isolation to predict outcomes whilst the other two were incorporated into existing models, such that their independent performance was not discernible. Following data extraction, therefore, these two models, the Canadian Neurological Score and the Middle Cerebral Artery Neurological Score (MCANS or Orgogozo score) were not considered further. Of the remaining 22 models, one was developed to predict outcome following intracerebral haemorrhage and was retained (Weimar et al 2006). Examination of previous reviews did not identify any additional externally validated models comprising variables that were feasible for collection in routine care.

Table 1 Prognostic models identified through review (Teale et al 2012)

Model	**Citation**
Anderson	(Anderson et al 1994)
Belfast	(Fullerton KJ et al 1988)
Bristol	(Wade DT et al 1983)
Edinburgh	(Prescott et al 1982)
G score	(Gompertz et al 1994)
Guys	(Allen CMC 1984)
Johnston	(Johnston et al 2000)
Lincoln	(Lincoln et al 1990)
Masiero	(Masiero et al 2007)
Modified National Institute of Health Stroke Scale (mNIHSS)	(Lyden et al 2001)
Shortened National Institute of Health Stroke Scale (NIHSS_8)	(Tirschwell et al 2002)
National Institute of Health Stroke Scale + age (NIHSS+age)	(Weimar et al 2004)
Orpington	(Kalra et al 1993)
Six Simple Variables (SSV)	(Counsell C et al 2002)
Tilling	(Tilling et al 2001a)
Uppsala	(Frithz G et al 1976)
Wang	(Wang et al 2003)
Weimar	(Weimar et al 2002)
Weimar intracerebral haemorrhage model (Weimar_ICH)	(Weimar C et al 2006)
Williams	(Williams et al 2000)
Young	(Young et al 2001)
Existing prognostic models	Citation
Canadian Neurological Scale (CNS)	(Muir KW. et al 1996)
Middle Cerebral Artery Neurological Score (Orgogozo score)	(Muir KW. et al 1996)
National Institute of Health Stroke Scale (NIHSS)	(Lai SM et al 1998; Muir KW. et al 1996)

Four models (Anderson CS et al 1994; Masiero S et al 2007; Wang Y et al 2003; Williams G et al 2000) were validated using a ‘split-sample’ technique. Here, the model is developed in a training set (a subgroup of the study population) and validated in the remaining study population. This represents a form of internal (not external) validation and should not be considered to represent evidence that the models perform adequately in independent populations (Altman D et al 2009). These models were not considered further. The G score is unusual in the identified models in that it is identical to the Guys score, but the model beta co-efficients have been simplified to create a new model (Gompertz P et al 1994). The G score is therefore a modification of the Guys score rather than a model developed de novo and no further validation studies were identified. The Guys model has however been externally validated and as such, the G score was retained for further discussion.

Where included studies described the external validation of models, the papers describing model development were also retrieved (if these did not feature in the output from the original searches). Data extraction was therefore performed from papers describing seventeen prognostic models. Data extraction tables are included in appendix 7.2A-2***.*** Studies describing model development are grouped with subsequent validation studies. Table 2 offers a summary of the studies describing model development.

3.5 Discussion

Results are discussed according to each criterion on which the models were assessed. Models fulfilling criteria are then discussed further in terms of their statistical properties and performance. A brief overview of the modelling assumptions and statistical criteria against which the models were assessed is offered in section 3.5.6. Assessment of individual models against the initial criteria and subsequent assessment of statistical methods used in model development are summarised in Table 4 and Table 6. Complete data extraction tables are presented in appendix A-2.

3.5.1 Inception cohort

An inception cohort is a group of patients at the same point in the disease process – in the context of stroke an inception cohort is taken to mean a group of patients assessed or recruited into a study within (and preferably at) a specific time-period following their stroke (Altman D 2001; Counsell C et al 2001).

Four of the models identified in this review (Bristol, Edinburgh, Lincoln and Young) described cohorts assessed at greater than two weeks following the stroke event and were not considered further (Lincoln et al, 1990; Prescott RJ et al 1982; Wade DT et al 1983; Young J et al 2001). Measurement of variables for development of the Edinburgh model was at four weeks following acute stroke and this is likely to limit the utility of this model in the acute stroke setting (Prescott RJ et al 1982). The Orpington score is an adaptation of the Edinburgh score to include an assessment of cognition (Kalra L et al 1994; Kalra L et al 1993). Unlike the Edinburgh model, the Orpington score was developed on an adequate inception cohort and is therefore retained.

The models developed by Young et al (Young J et al 2001) and Lincoln et al (Lincoln et al, 1990) were developed using variables collected on admission to (or discharge from) a rehabilitation facility and therefore the inception cohort (time from stroke to assessment) was not uniform. The SSV model was developed using retrospective data (collected prospectively) from the OCSP cohort (Bamford J et al 1988; Counsell C et al 2002). Although in the original OCSP cohort about three quarters of assessments were performed within two weeks of the stroke event (median time to assessment 4 days) (Bamford J et al 1988), the SSV model was developed using data on the 86% of assessments performed up to 30 days following stroke(Counsell C et al 2002). This model was, however, retained as the proportion of assessments performed after 14 days was small.

Further discussion of models is restricted to the 13 models that are externally validated and developed on an adequate inception cohort. The NIHSS is also included for further discussion as a prognostic model as its use has been described as a predictor of outcome in studies identified through this review (Counsell C et al 2002; Lai SM et al 1998; Muir KW. et al 1996).

3.5.2 Sources of data for model development

Models should ideally be developed from prospective data – i.e. data that are collected according to a protocol with the express purpose of developing the model (Wyatt JC 1995). The convenience of data extracted from retrospective databases, may be offset by limitations in the data that are available, or its quality (Wyatt JC 1995). There were three main identified sources of data used for model development and validation in this review: prospective data collection for the purposes of model development, retrospective use of data collected within stroke registers, and the secondary use of data from previously conducted randomised controlled trials. Five of the remaining 13 models identified during this review were developed through studies where the primary purpose of the research was to develop the model (Belfast, G score, Guys, Orpington, Weimar_ICH) (Allen CMC 1984; Fullerton KJ et al 1988; Gompertz P et al 1994; Kalra L et al 1993; Weimar C et al 2006). These tended to be small studies (sample size 96-361, median 206).

The secondary use of retrospective data may introduce bias either due to inclusion and exclusion criteria of clinical trials, or to non-standardised methods of collection and definitions of prognostic variables (such as may be seen in the extraction of data from existing databases) (Wyatt JC 1995). Ideally, the external validation of models in independent datasets should also use data that is prospectively collected (to prevent any bias that could be introduced if prognostic information is recorded when the outcome is known) (Wyatt JC 1995). Models developed from retrospective data where cases are selected and extracted on the basis of concordance with inclusion criteria and complete outcomes data are particularly prone to selection bias as there may be systematic reasons why outcomes data are missing in certain patient groups. The reasons why certain types of patients might be lost to follow-up and their baseline characteristics should be examined and compared to cases with complete data to ensure that there is no systematic bias (Hayden JA. et al 2006).

Three models were developed using data extracted from stroke registries or prospective cohorts. These were the Weimar (ischaemic stroke) model (Weimar C et al 2002), the NIHSS+age (Weimar C et al 2004) and the SSV models (Counsell C et al 2002). Of these, two models were developed from prospectively collected data extracted from the German Stroke Database (Weimar and NIHSS+age) (Weimar C et al 2002; Weimar C et al 2004). Although only patients fulfilling inclusion criteria and with complete data were selected, baseline characteristics of patients with complete and incomplete outcomes data were compared during the development of two of the models and no statistically significant differences found (Weimar C et al 2002; Weimar C et al 2004).

The Six Simple Variable model (SSV) (Counsell C et al 2002) was developed retrospectively using prospectively collected data from a community cohort of stroke patients of whom about half were never admitted to hospital (Bamford J et al 1988). Patients excluded from model development included those who died before assessment or who were not assessed by a study neurologist within 30 days of the stroke event. Outcomes data for the remaining 530 patients was complete. The SSV model has, however, been subsequently externally validated using prospective data (Dennis et al 2006; Lewis S et al 2007; Reid J et al 2007), with collection of variables within a week of the stroke event. One further model (Uppsala) was developed using data extracted retrospectively from patient case-notes (Frithz G et al 1976).

Four models identified used data from previously conducted RCTs (Johnston, mNIHSS, NIHSS_8, Tilling) (Johnston KC et al 2000; Lyden PD et al 2001; Tilling K et al 2001a; Tirschwell DL et al 2002). RCTs performed on an intention to treat basis may ascribe the last available score, or worst outcome to patients lost to follow up or unable to complete assessments (Tirschwell DL et al 2002). Secondary use of data (from RCTs, databases or previously conduced cohort studies) means that patients with incomplete data (or those lost to follow up) may be excluded (Johnston KC et al 2000; Johnston et al 2003) with the risk of systematic bias. Moreover, inclusion or exclusion criteria of RCTs (e.g. exclusion of patients unable to transfer from bed to chair (Tilling K et al 2001a) or exclusion of patients with contraindications to thrombolysis (Tirschwell DL et al 2002)) may affect the ability of models developed from trial data to predict outcomes in the groups that were excluded from the training dataset. If validation studies were also performed in selected populations, the performance of models in empirical populations may remain untested and uncertain. Of four model development studies making secondary use of RCT data, only one (Johnston KC et al 2000) reported the number of patients excluded through incomplete outcomes data and none compared the characteristics of patients excluded through missing data with the study population (Table 2).

Table 2 Summary of studies describing construction of models identified in the review (Teale et al 2012)

	Reference	Sample size	Data source	Adequate inception cohort	Less than 10% loss to follow up	No systematic difference in patients lost to follow up?	Valid outcome measured at fixed time point	Modelling methods	Adequate EPV	Linearity assumptions tested and met	Control for collinearity
Belfast	Fullerton (1988)	206	P	+	+		-	CDA	0	-	-
Bristol	Wade (1983)	162	P	-	-	+	+	LinR	+	0	+
Edinburgh	Prescott (1982)	155	T	-	+		-	LinR	-	-	-
G score	Gompertz (1994)	361	P	+	-	0	+	N/A	+	0	0
Guys	Allen (1984)	148	P	+	+	0	-	S LR	-	-	+
Johnston	Johnston (2000)	256	T	+	-	0	+	LR	+	+	-
Lincoln	Lincoln (1990)	70	P	-	-	0	-	S LR	-	-	+
mNIHSS	Lyden (1999)	291	T	+	0	0	+	FA	NA	NA	NA
NIHSS_8	Tirschwell (2002)	233	T	+	0	0	+	S LR	-	-	+
NIHSS+age	Weimar (2004)	1079	D	+	+		+	S LR	+	+	+
Orpington	Kalra (1993)	96	P	+	+		+	LinR	+	-	-
SSV	Counsell (2002)	530	D	+	+		+	S LR	+	+	+
Tilling	Tilling (2001a)	299	T	+	0	0	+	MM	+	+	0
Uppsala	Frithz (1976)	344	CN	+	+		+	LR	+	0	+
Weimar	Weimar (2002)	1754	D	+	+		+	S LR	+	+	+
Weimar_ICH	Weimar (2006)	260	P	+	-	+	+	S LR	+	0	+
Young	Young (2001)	207	T	-	+		+	S LR	-	+	+

P = Prospective data collection, T = retrospective use of RCT data, D = Data extracted from database or cohort study,
CN= data extracted from case notes
(S) LR = (stepwise) logistic regression, LinR = linear regression, MM = multilevel modelling FA =factor analysis CDA Canonical Discriminant Analysis
+ = condition met, - = condition not met, 0 = unclear from study reports
Highlighted studies were not developed on an adequate inception cohort and are not considered further

In developing the NIHSS_8, Tirschwell et al extracted 223 cases with complete prognostic variable data from the placebo arms of three RCTs (239 patients randomised to placebo in these trials) (Tirschwell DL et al 2002).Two of these trials were analysed on an intention-to-treat basis with 'last observation carried forward' for patients who died (combined rate 27/191) or who were not followed-up (combined rate 8/191 patients). It is not clear if the patients who died but were ascribed the last available functional outcome score from these trials were coded as deaths during development of the shortened NIHSS models (Tirschwell DL et al 2002).

Johnson et al extracted data from both placebo and intervention arms of a therapeutic trial where no overall treatment effect was demonstrated (Johnston KC et al 2000). Patients with incomplete predictor or outcomes variables were excluded, and their characteristics were not compared to the baseline characteristics of the complete study sample (222/256 patients).

Tilling et al analysed all patients randomised into a trial of early supported discharge compared against usual care (Tilling K et al 2001a). To enable multilevel modelling of recovery trajectories, outcomes measurements were performed at a number of time points following stroke. All patients had at least one outcome measurement and were included in the model development. Mean Barthel Indices for patients in whom measurements were not made on all occasions were compared to patients with complete data.

Data for external validation studies of identified models were similarly obtained through secondary use of trial data, existing cohort data or gathered prospectively. The number of validation studies for individual models ranged from one (Belfast, mNIHSS, NIHSS_8, Tilling, Uppsala, Weimar and Weimar_ICH) to six (SSV), with cumulative validation sample sizes of 27 (mNIHSS) to 8964 (SSV), median 762 (see appendix 7.2A-2). Larger validation populations were generally those from databases and registries, whilst smaller sample sizes reflect studies with data collected prospectively to meet the a priori intention to validate a specific model. Some studies used the same study population to validate several models (Gladman et al 1992).

3.5.3 Clinically relevant prognostic variables

Statistical credibility of a prediction model in isolation is not useful unless the prognostic variables make clinical sense. Predictor variables in the identified models fell into three broad categories: markers of stroke severity at onset, possible confounding variables (e.g. age) and co-morbidities (Table 3). Generally, variables used to construct the identified models made clinical sense. However, some of the covariates are less convincing clinically e.g. the presence of 'non-specific ST or T wave changes' was included as a predictor in the Belfast model (Fullerton KJ et al 1988). During the development of this model, multiple univariate analyses of binary and categorical variables had been performed to identify candidate predictors for multivariable analysis. Some variables had several categories (up to 11), and variable selection was data driven (Fullerton KJ et al 1988). This highlights the importance of clinical, as well as statistical judgement during model development.

Table 3 Variables included in identified models (Teale et al 2012)

	Variables included in model
SSV	Age, living alone, independent pre stroke ,normal GCS verbal score ,able to lift both arms, able to walk
Tilling	Age, Sex, ethnicity, pre-stroke handicap, limb weakness, dysphasia, dysarthria, incontinence, conscious, swallowing deficit, stroke subtype
Johnston	Age, NIHSS score, small vessel stroke, previous stroke, diabetes, prestroke disability, infarct volume
Orpington	Arm power, proprioception, balance, cognition
Guys	Limb paralysis, higher cerebral dysfunction+ hemiparesis+ hemianopia, drowsy, age, unconscious at onset, uncomplicated hemiparesis
Belfast	Albert's test score, leg function, conscious level, arm power, weighted mental score, non-specific ECG changes
Uppsala	Adaptation of Mathew's score (0-100) Conscious level, orientation, dysphasia, conjugate gaze palsy, facial weakness, arm power, Performance Disability scale, reflexes, sensation
Weimar	Model 1: Neurological complications, fever, lacunar infarct, diabetes, previous stroke, sex, age, mRS, NIHSS score on admission Model 2: Fever, age, NIHSS score on admission
NIHSS_age	Age, NIHSS
Weimar_ICH	Age, NIHSS
NIHSS_8	NIHSS_15 items 1a, 2,3,4 6a&b 9, 10 conscious level, gaze visual fields, facial paresis and lower limb motor scores, language and dysarthria
mNIHSS	Items 1B, 1C, 2,3,5 a&b, 6 a&b, 8, 9, 11 from the NIHSS: Conscious level, gaze, visual fields, upper and lower limb power, sensory function , language and neglect

Counsell et al (2001) argue that the variables included in a prognostic model for stroke should include a marker of stroke severity (Counsell C et al 2001). It is possible, if not likely, that some clinical variables may act as proxy markers for stroke severity; e.g. patients with more severe strokes are more likely to develop new urinary incontinence. In this way, the presence of urinary incontinence may reflect constructs related to stroke severity such as mobility (difficulty in self-toileting), communication problems (difficulty in communicating the need for assistance with toileting) or conscious level. The potential for such collinearity between independent (predictor) variables should be examined and addressed during model development, but raises the possibility that more simple (univariate) case-mix adjustment may be possible. Indeed, it has been argued that multivariable prognostic models add little additional accuracy for prediction of discharge home over and above that of urinary incontinence alone (Barer et al 1989). A more recent examination of the role of urinary incontinence as a univariable predictor of outcome by Counsell et al found that although urinary continence was able to identify patients with good outcome (mRS ≤2), the specificity (correct identification of patients with unfavourable outcome) was poor (0.44, 0.40-0.48) (Counsell C et al 2004). This would tend to suggest that absence of urinary incontinence is a predictor of good outcome, rather than presence of urinary incontinence necessarily predicting poor outcome. This raises the question of whether there are particular subgroups of patients with new urinary incontinence following stroke that are more likely to have a poor outcome.

3.5.4 Feasibility of data collection at ward level

Three models require baseline data to be collected within six hours of admission (Johnston, NIHSS+age and Weimar_ICH models (Johnston KC et al 2000; Weimar C et al 2004; Weimar C et al 2006). A further three require variable collection within 24 hours (G score, mNIHSS, NIHSS_8) (Gompertz P et al 1994; Lyden PD et al 2001; Tirschwell DL et al 2002). One further model was developed using variables collected "on admission", although the exact time frame within which variables were collected is not specified (Uppsala) (Frithz G et al 1976).

The type of ward to which the patient is admitted also has implications for the types of data that may be collected to enter into prognostic models. The availability of staff trained to perform complex clinical assessments (e.g. the NIHSS) may limit the use of some models to specialist staff in stroke units. In addition, data collection is resource dependent. In funded research projects assessments are likely to differ from those that may be performed as part of routine care. Eight identified models (Johnson, Lincoln, Weimar & Weimar ICH models, Uppsala , Belfast, NIHSS+age, mNIHSS) (Frithz G et al 1976; Fullerton KJ et al 1988; Johnston KC et al 2000; Lincoln et al, 1990; Tirschwell DL et al 2002; Weimar C et al 2002; Weimar C et al 2004; Weimar C et al 2006) and one pre-existing severity score used to predict outcome (the NIHSS) (Muir KW. et al 1996) require complex clinical assessments for completion and are therefore unlikely to be feasible for collection in non-specialist settings or in routine care.

3.5.5 Assessment of valid and reliable outcomes at a fixed time point

The models identified through this review may be classified according to the outcomes that they were developed to predict. Some authors describe development of similar models to predict different outcomes, and these therefore counted more than once. Outcomes should be of proven validity and reliability in stroke populations. In addition, the outcome should be measured at a particular time point following the stroke event, such that time to measurement of outcome is standardised.

Two of the identified models predict the Barthel Index as an interval dependent variable (Orpington, Tilling) (Kalra L et al 1993; Tilling K et al 2001a). The Tilling model is a multilevel model and, as such, predicts average recovery trajectories (measured with Barthel Index) over time following a stroke event (Tilling K et al 2001a). The Orpington score was developed to predict the Barthel Index at three time points following stroke (Kalra L et al 1993). In a subsequent validation study of the Orpington score and the NIHSS (a pre-existing stroke severity scale), Lai et al predicted Barthel Index as an interval variable (Lai SM et al 1998).

The validity of this approach is dependent on some statistical assumptions which are discussed further in section 3.5.8.2.

Seven of the remaining 13 models predict dichotomised Barthel Index (Johnston, G score, Weimar ICH, Weimar, NIHSS+age, NIHSS_8, mNIHSS). Time to outcomes measurement for these models varied from two to six months.

Three models predict dichotomised modified Rankin Score (or Oxford Handicap Score) (SSV, mNIHSS, NIHSS_8) (Counsell C et al 2002; Lyden PD et al 2001; Tirschwell DL et al 2002), and two predict other dichotomised impairment scores (Johnston, NIHSS_8) (Johnston KC et al 2003; Johnston KC et al 2000; Tirschwell DL et al 2002)). The Johnston model predicts devastating outcome with the NIHSS, dichotomised BI or dichotomised Glasgow Outcomes Score whilst the NIHSS_8 model predicts a global outcome score (good/poor outcome) calculated from four other dichotomised outcomes: BI, NIHSS_15, mRS, and the Glasgow Outcomes Score. The authors of the Guys and Belfast models developed study specific impairment scales (Allen CMC 1984; Fullerton KJ et al 1988).

Four of the 13 models were developed to predict mortality – Uppsala*, S*SV, Weimar and NIHSS+age) (Allen CMC 1984; Counsell C et al 2002; Weimar C et al 2002; Weimar C et al 2004).

3.5.6 Statistical quality of studies describing model development

The majority of the models identified in this review comprise variables that are not feasible for collection in routine care (for reasons of time to assessment or complexity of assessment), require training for administration (NIHSS) or were developed on an inception cohort established more than two weeks following the stroke event (Table 2). No model was excluded solely on the basis of the characteristics of the cohort from which it was developed if there was evidence that it had been validated (and performed acceptably) in a more general post stroke population (e.g. the SSV model and Tilling models) (Counsell C et al 2004; Tilling K et al 2001a) (see data extraction tables in appendix 7.2A-2). Six models were therefore further scrutinised according to statistical criteria: the Tilling model, Orpington score, G score, Guys model, NIHSS_8 model and the Six Simple Variable model (see Table 4).

Five of these six remaining models use single level regression modelling for the prediction of outcomes. The sixth (Tilling model) uses multilevel modelling and is considered in 3.5.7.4 as a special case. Regression models are based on the 'generalised linear model' comprising a linear predictor (1) and random effects (ε) such that the general form of the equation to calculate the mean expected values of the dependent variable E(Y) from a linear model is given by (2) where μ_i =the predicted outcome for individual i, β = variable coefficient between limits i→ k, x= independent variable and ε= random effects. The generalised linear model (2) contains a 'link function' $f(\mu_i)$, dependent on the underlying distribution of the data (3). The inverse of which transforms the linear prediction to the probability distribution of the underlying data (Fox J, 1997 pp 487-488). For example, the logit function is used in logistic regression models to predict binomial probability distributions (dichotomous outcomes), and the transformation function is given by (3)(Fox J, 1997). The inverse of this function is therefore used to calculate the fitted values μ. The probability distribution of these fitted values should therefore follow a binomial distribution. For normally distributed linear outcomes, the function is f(μ) = μ.

$$(1)\quad \eta = \beta_0 + \beta_1 x_1 + \dots \beta_k x_k$$

$$(2)\quad And\ the\ fitted\ values\ are\ given\ by\ \mu_i = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon$$

$$(3)\quad then\ for\ the\ binomial\ distribution\ f(\mu_i) = log_e \frac{\mu_1}{1-\mu_i}$$

$$such\ that\ E(Y) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

3.5.7 Checking model assumptions – (Multiple) linear regression

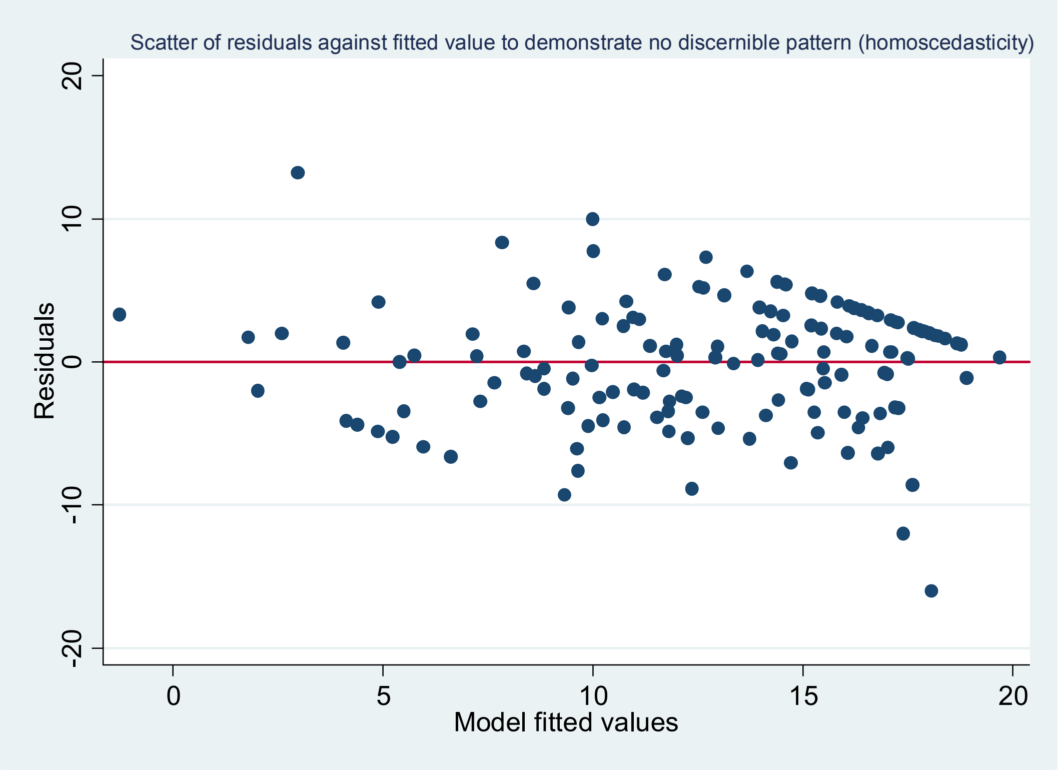
Multiple linear regression modelling requires a number of assumptions to be fulfilled in order for the approach to be valid (Fox J, 1997 p 113; Harrell FE et al 1996). Firstly, there must be an underlying linear relationship between the independent and dependent variables (or a transformation thereof). Secondly, the residuals (the difference in observed and predicted outcomes) must be normally distributed. Providing this assumption is met, individual continuous (or ordinal) variables entered into the model need not be normally distributed. Thirdly, the distribution of variance of predicted values of y should be inspected to ensure it is uniform across all values of the independent variables (homoscedasticity), i.e. there should be no pattern in a plot of model residuals against fitted values (Fox J, 1997). Finally, there must be no linear relationship (collinearity) between independent variables. A possible example of collinearity would be to include variables measuring leg weakness and mobility into a model: there is likely to be a strong relationship between these two variables. Entering highly correlated predictor variables into a model means that the individual effect of each variable is hidden within their combined effect such that the individual contribution of the variables cannot be discerned. Inclusion of collinear variables can overestimate the individual effects and result in inflated and unstable beta coefficients, i.e. estimates may vary widely with the addition or exclusion of individual cases, often reflected in wide confidence intervals around co-efficient estimates (Fox J, 1997 p 337). Collinearity between independent variables should be identified, through logical or clinical reasoning, and explored. Variables likely to be correlated should either be excluded from the model (if they add little to the explanation of outcome) or combined to form a sensible composite measure. Stepwise variable selection procedures in statistical software help to overcome collinearity through automatic exclusion of variables that are highly correlated (Concato J et al 1993).

Table 4 Models retained for statistical appraisal

Model	Citation	Inception less than 2 weeks	Clinically relevant prognostic factors	Feasible to collect at ward level	Prospective data collection (or validation in prospective population)	No systematic loss to follow up	Valid and reliable outcome measured at a fixed time point?	Generalisable to other stroke populations?	Retained for statistical appraisal?
Bristol	Wade (1983)	✗	✓	✓	✓	?	✓	✗	✗
Edinburgh	Prescott (1982)	?	✓	✓	✗	✗	✗	✗	✗
Lincoln	Lincoln (1990)	✗	✓	✗	✓	✓	✗	✗	✗
Young	Young (2001)	✗	✗	✓	✓	✓	✓	✗	✗
Belfast	Fullerton (1998)	✓	✓	✗	✓	?	✗	✓	✗
Johnson	Johnston (2000)	✓	✓	✗	✗	?	✓	?	✗
mNIHSS	Lyden (2001)	✓	✓	✗	✗	✓	✓	✗	✗
NIHSS	Lai (1998); Muir (1996)	NA	✓	✗	✓	✓	✓	✓	✗
NIHSS+age	Weimar (2004)	✓	✓	✗	✓	✓	✓	?	✗
Uppsala	Frithz (1976)	✓	✓	✗	✓	?	✓	✗	✗
Weimar	Weimar (2002)	✓	✓	✗	✓	?	✓	✗	✗
Weimar ICH	Weimar (2006)	✓	✓	✗	✓	✓	✓	✗	✗
G score	Gompertz (1994)	✓	✓	✓	✓	?	✓	?	✓
Guys	Allen (1984)	✓	✓	✓	✓	✓	✓	?	✓
NIHSS_8	Tirschwell (2002)	✓	✓	✓	✗	?	✓	✓	✓
Orpington	Kalra (1993)	✓	✓	✓	✓	?	✓	?	✓
SSV	Counsell (2002)	✓	✓	✓	✓	?	✓	?	✓
Tilling	Tilling (2001a)	✓	✓	✓	✗	?	✓	✗	✓

Model assumptions should be checked during development, and attempts made to overcome violations if they occur. Most assumption checks can be achieved through examination of simple scatter plots. Non-random distribution of model residuals in a scatter plot of residuals across values of individual independent variables may reveal a non-linear relationship between independent and dependent variables (Altman D 1999 p 346). In addition, a scatter plot of residuals against fitted values can demonstrate heteroscedasticity (non-uniform variance in residuals across fitted values of the dependent variable) which may indicate the omission from the model of an important factor exerting a systematic effect (Fox J, 1997 p302). Figure 6 is a plot of fitted values against model residuals and demonstrates homoscedascticity (constant variance across fitted values).

Figure 6 Model fitted values vs. residuals to demonstrate homoscedasticity



A histogram of (studentised) residuals can indicate a deviation from a normal distribution[1]. A (standardised) normal (Q-Q) plot can be used to examine the distribution of residuals against the normal distribution (Fox J, 1997 p42; Altman D, 1999 p 133), see also section 4.4.3.2.

Non-normally distributed residuals may be another indication of non-linear relationships between predictor and outcome. Transformations of the independent variable may overcome this non-linearity, and may also help to solve problems with non-uniform variance (Altman D, 1999 p 303; Royston P et al 2008).

[1] The theory behind the derivation of studentised and standardised residuals is discussed in appendix A-3

Independent variables in regression models may be continuous, ordinal or dichotomous. Prediction of dependent variables from categorical independent variables requires the creation of 'dummy variables'. A dummy is created for each level of the categorical variable apart from the 'reference category' coded 0 by convention (Fox J, 1997 142). Each dummy variable is then compared to the reference category in order to create a series of dichotomous pairs (Altman D, 1999 p 339). The beta coefficient of the dummy variable in the model is then equivalent to the difference in mean expected outcome for patients in the dummy category as compared with the reference category. As this could markedly increase the number of variables that need to be considered in the EPV calculation of sample size, the inclusion of categorical variables and creation of dummies should be considered during model development.

3.5.7.1 Model assumptions: Logistic regression models

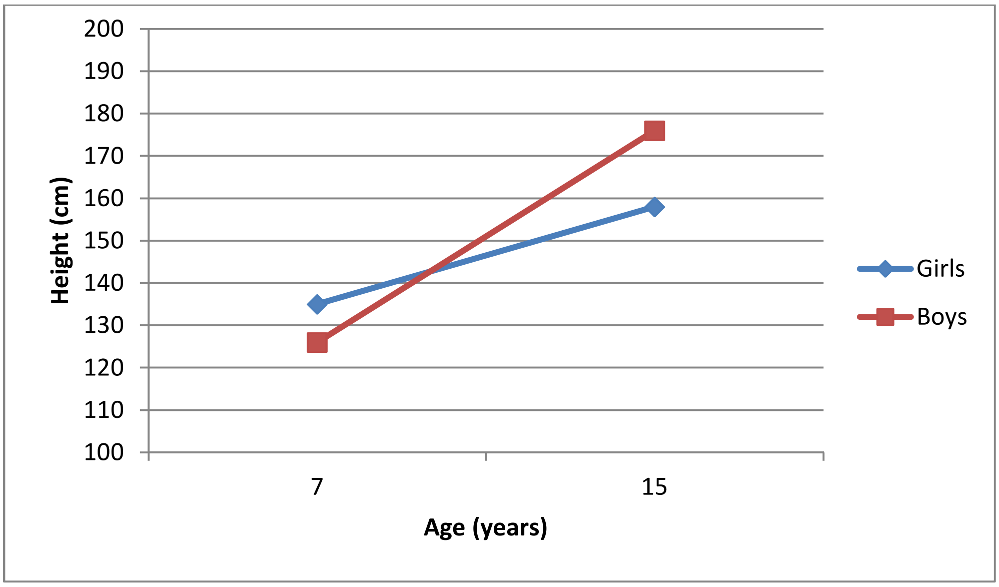
In the models identified in the review, continuous dependent variables were often dichotomised to circumvent some of the requirements for linear regression, e.g. dichotomised Barthel Index to reflect 'good/poor' outcome) (Gladman JR et al 1992; Gompertz P et al 1994; Johnston KC et al 2000; Lyden PD et al 2001; Weimar C et al 2004; Weimar C et al 2006). This allows a logistic regression model to be used to predict a binary outcome where the assumptions on the underlying distributions are less stringent. However, dichotomising continuous variables means that detailed information may be lost (Mallett S et al 2010a; Royston P et al 2008).

Logistic regression is used to predict a log transformation of the odds of a binary outcome (the 'logit function' see Equation (3) p 49) (Fox J, 1997 p 78). A linear relationship between the predictors and this logit function is assumed (Harrell FE et al 1996). There are no assumptions placed on the distributions of the independent variables but there must be no correlation (collinearity) between them (Bewick V et al 2005).

3.5.7.2 Interaction terms

Interaction is a problem with both linear and logistic regression analyses and occurs when the effect of one variable is mediated by the effect of another (Fox J, 1997 p 145). For example, the relationship between height and age would be mediated by gender if girls tend to be taller than their male counterparts when they are younger but relatively shorter as they grow older. This could be controlled for by entering an interaction term into a regression equation (as the product of the two terms). Interaction terms should be carefully considered through clinical reasoning, ideally *a priori*. In contrast with composite terms to control for collinearity, where the number of variables may be reduced, inclusion of an interaction term will increase the number of variables entered into a model (Harrell FE et al 1996). Interaction terms may be demonstrated by plotting independent against dependent variables at different levels of the mediating variable. If the two lines are not parallel, then an interaction term is likely (see example, Figure 7).

Figure 7 Theoretical example of an interaction term between age and height



3.5.7.3 Variable selection

Selection of variables to include in a logistic regression model may be data driven, or made through clinical reasoning. A common, but not recommended data driven approach is to perform multiple univariable analyses and discard variables which do not reach a pre-specified p-value (often 0.2) (Mallett S et al 2010a). This approach can result in selection bias (Royston P et al 2008) - predictors with larger co-efficients (perhaps through chance) are more likely to be statistically significant and therefore more likely to be retained in the model over variables with smaller co-efficients (Royston P et al 2008). Clinically unimportant variables may therefore be included, or relevant variables discarded on the basis of their p value (Mallett S et al 2010a; Royston P et al 2008). Predictor variables may also be selected for inclusion through automated (forwards or backwards) selection procedures with statistical software. Here, variables are selected on the basis of their influence on maximising the model R^2 statistic. The retention or rejection of variables previously entered or removed from the model is re-considered following the addition or removal of subsequent variables during the stepwise procedure (Fox J, 1997 p 356).

It has be argued that any clinically relevant predictor (or confounder) should be included in a model even if it does not reach statistical significance in univariable or multivariable analysis (Mallett S et al 2010a); variables may be excluded through chance and a model based solely on statistical criteria may lack generalizability due to the exclusion of these clinically important predictors (Rothwell PM 2008). Ideally the most parsimonious model that maximises explanation of the dependent variable should be developed through both clinical and statistical reasoning (Altman D et al 2000; Concato J et al 1993).

3.5.7.4 Multilevel (hierarchical) models

The Tilling model (Tilling K et al 2001a) was derived using multilevel modelling techniques. For this reason it is considered here as a special case. Multilevel modelling exploits the hierarchical nature of data by considering e.g. repeated Barthel Index measurement over time (Level 1) as a property of individuals (Level 2) (Tilling K et al 2001a). Consideration of higher levels (e.g. ward or hospital) may help to explain further residual variation in patient outcome. This clustering of data can be used to explain fixed and random effects at different levels of the model and to explore the interdependence of the measurements (Kline RB 2005 p 332), thereby providing additional information as to the structure of the data over and above that which is offered through single level regression modelling (Tilling et al 2001b). A multilevel modelling approach allowed Tilling et al to estimate average recovery trajectories based on baseline characteristics. From these, iterative calculations of an individual's outcome could be made at any time point conditional on both baseline characteristics and observed outcome trajectory (Tilling K et al 2001a).

3.5.8 Reporting of checks of model assumptions for models identified in the review

Checks of model assumptions are discussed in two parts: assumptions and checks during model development, and post-estimation checks of model assumption (i.e. after calculation of model beta-coefficients). Model fit and performance in external validation studies are discussed in subsequent sections.

3.5.8.1 Construction of models

The Guys score was developed by Allen et al and its utility re-examined in an independent cohort by Gladman et al (Allen CMC 1984; Gladman JR et al 1992). Twenty-nine candidate variables were identified for possible inclusion in the original model. Selection of variables for inclusion in a multivariable model was made through identification of univariate predictors where observed and expected frequencies were significantly different (at the 0.05 level) between patients with 'good' or 'poor' outcome (Chi-squared or t-test). At this significance level, examination of more than 20 variables makes it likely that at least one will reach statistical significance by chance. There was no control for collinearity or consideration of potential interaction terms in the development of the model (Allen CMC 1984).

The G score was developed from the Guy's score through simplification of the regression co-efficients to integers (Gompertz P et al 1994). Although this is, therefore, technically a new model, the method of variable selection is the same as in the original study (as described above).

The Six Simple Variable models (Counsell C et al 2002) were developed to predict dichotomised functional outcome (alive and independent vs. not, as measured with an Oxford Handicap Score <3), and survival at 30 days. Many of the issues surrounding variable selection were addressed explicitly. Variables were selected initially through clinical reasoning and feasibility of data collection. Remaining variables were separated into a core group of clinical variables ('set 1') and groups of additional variables of increasing complexity ('sets 2 & 3'). Eighteen clinical variables were included in set 1 to be entered into a forward stepwise regression model. Linearity of the relationship between the only continuous independent variable (age) and the dependent variables were tested and met. Interaction terms were tested between variables where interactions were clinically suspected (age, sex and previous disability), but none were found (Counsell C et al 2002).

The authors of the SSV models highlight that many regression analyses were performed in the development of the final SSV models and acknowledge that these multiple analyses may have led to the inclusion (or exclusion) of variables from the models by chance (Counsell C et al 2002).

The predictive accuracy of any regression model is dependent on the correct variables being included in the model and the stability of the beta co-efficient estimates. Too many variables included in the model may result in overfitting (Type 1 error). Here, the model has high predictive accuracy in the sample from which it was drawn but performs poorly in independent datasets (Peduzzi P et al 1996). Type 2 errors arise from a lack of power, e.g. where the effect of an individual predictor is too small to be detected given the sample size.

Sample size is also important in determining the number of variables that may be entered into a model. The number of variables that may be entered into a model to predict a binary outcome is determined by the ratio of observed events to the number of variables (including dummies) – the events per variable ratio (EPV). The number of events in this context applies to whichever is less frequent between the binary outcome pair. An EPV of greater than 10 has been suggested and widely accepted following a simulation study by Peduzzi et al (Peduzzi P et al 1996). Here, retrospective data from a study where binary outcomes had been predicted using a logistic regression model were used with a re-sampling technique (Monte Carlo) to simulate model variable co-efficients over a series of pre-specified EPV ratios (between 2 and 25). The distribution of estimated co-efficients was then examined and compared with the parameters derived from the original regression model. Below a cut-off EPV ratio of 10 the regression co-efficient estimates were unstable: there was lack of convergence (i.e. the simulated models did not 'settle' onto a value for the regression co-efficient), predicted co-efficients were not normally distributed and their confidence-intervals unacceptably wide (Peduzzi P et al 1996).

Thus, the number of variables that may be entered into a logistic regression model (but not necessarily retained (Rothwell PM 2008)) may be calculated by dividing the total number of *events* (e.g. number of deaths or dependent patients) by ten. If there are interaction terms, or dummy variables, these needed to be counted as separate variables for the purposes of the calculation. If data are to be collected prospectively, an estimation of the expected events should be calculated from previous studies or epidemiological data. Linear regression models should also meet an ‘events per variable’ ratio of ten, but here the number of events is the number of observations for the dependent variable.

Of the retained models in this review, the Guys score and Bristol model have an EPV of less than ten (Gladman JR et al 1992; Gompertz P et al 1994). The G-score, Orpington, Tilling and SSV (Set 1) models have adequate EPV (Counsell C et al 2002; Gompertz P et al 1994; Lai SM et al 1998; Tilling K et al 2001a).

3.5.8.2 Post-estimation checks of model assumptions

The Orpington model is a single level regression model that predicts the Barthel Index score (Lai SM et al 1998). During their examination of the predictive properties of the Orpington score, Lai et al entered individual items from the Orpington score into a linear regression model to predict the (ordinal) Barthel Index thereby treating the Barthel Index as a continuous variable. Although Lai et al acknowledge the potential problems due non-normally distributed ordinal data (and the question this raises regarding a linear relationship between predictor and outcome), normality assumptions (of model residuals) were not tested (Lai SM et al 1998). The variance of the Orpington score was noted to decrease over time (with successive measurements) (Lai SM et al 1998). This may suggest a non-linear relationship between predictor and dependent variable or perhaps some interaction between the score and time. This is not unexpected, as recovery trajectories are known to be non-linear (Jorgensen HS 1996; Tilling K et al 2001a). However, in the presence of this variability related to time, analyses should be cross-sectional as opposed to longitudinal.

Violation of the normality of residuals assumption in the Tilling model was felt, by the authors, to be due to the effect of 19 individual patients for whom the model did not fit well rather than the underlying distribution of the dependent variable (Tilling K et al 2001b). Thus, the dependent variable (Barthel Index) was treated as a normally distributed continuous variable. Tilling et al recognise that this assumption is violated due to the ceiling effect of the Barthel Index. Strategies to overcome this were explored through more complex modelling techniques that allow for censoring at the upper limit (ceiling) of the Barthel Index (Tilling K et al 2001b; Twisk J et al 2009). However, application of these techniques affected neither the estimates of model coefficients nor predicted Barthel Index values (Tilling K et al 2001b).

3.5.9 Model performance (external validation)

External validation (performance in independent population) for the models with acceptable properties as regards model development is considered in section 3.5.9.3 following a brief discussion of the methods used to measure model performance (sections 3.5.9.1 & 0). Two aspects of the models’ predictive function are discussed: discrimination (ability to distinguish between individuals with good and poor outcome) and calibration (accuracy of predicted outcomes).

3.5.9.1 Discrimination

Discrimination is a measure of how well a model is able to correctly distinguish patients with good over poor outcome (Harrell FE et al 1996), and is measured with the ‘c statistic’. This is calculated as the overall proportion of correct (good over poor) predictions across all non-concordant outcome pairs in the sample (Harrell FE et al 1996; Justice AC et al 1999). For binary outcomes, this is equivalent to the area under the Receiver Operating Characteristic (ROC) curve (Hanley et al 1982). Models with no discrimination (i.e. no better than chance) would be represented by a c-statistic, or area under ROC curve (AUC) of 0.5. Perfect discrimination is represented by a c statistic, or AUC of 1.0.

Sensitivity and specificity are often presented as a measure of how well a model predicts individual patient outcome. Patients predicted to have a poor outcome who are observed to have poor outcome are ‘true negatives’ whilst those with poor outcome predicted to have good outcome are ‘false positives’ (see Table 5). Acceptable values for sensitivity and specificity is dependent on the purposes of measurement, i.e. the tolerability of false positive and false negative rates depends on the clinical context (Altman D, 1999 p 418). For the purposes of this review, models were retained if sensitivity and specificity were both greater than 0.75.

Table 5 Contingency table

		Observed Outcome	
		0	1
Predicted outcome	0	True negative	False Negative
	1	False Positive	True Positive

3.5.9.2 Calibration

Calibration is a marker of how well a model can make correct predictions (e.g. patients actually have the outcome that is predicted). Calibration of a model can be examined through plotting proportions of predicted outcome in deciles against the proportion of patients with the observed outcome within each decile (Counsell C et al 2002). Perfect calibration is represented by a line y=x.

3.5.9.3 Predictive accuracy of included models

The performance of each model in external validation studies (external datasets) are given in the data extraction tables (appendix A-2).

Two SSV models predict either survival at 30 days (developed through a Cox proportional hazards model) or alive and independent at six months (logistic regression analysis) (Counsell C et al 2002). The AUC (equivalent to the c-statistic) for prediction survival was 0.84 in the external validation study by the developers of the model (Counsell C et al 2002) and 0.73 in an independent study (IST3 trialists 2008). Discrimination of the model to predict independent survival was greater than 0.75 in each of 4 external validation studies (Dennis et al 2003; Lewis S et al 2007; Reid J et al 2007; Weir N et al 2001).

The remaining models report sensitivity and specificity analysis to demonstrate model performance (see appendix A-2). Although the Guys has an acceptable specificity of 83% to predict death at three weeks (i.e. a low rate of false negatives), the sensitivity is poor (58%) representing a high rate of false positives (Gladman JR et al 1992). Conversely, the G score has a sensitivity of 72% but a low specificity (63% i.e. it will predict 37% of patients with an ultimately poor outcome as having good outcome) (Gompertz P et al 1994). A sensitivity analysis was performed for the Bristol model in a prospective observational study by Gladman et al (Gladman JR et al 1992). This aimed to ascertain the predictive ability of the Bristol score at one week to predict a BI>10 at 3 months. The sensitivity was found to be 100%, but the specificity 0% (highlighting the inverse relationship between these two parameters). The Tilling model predicted the Barthel Index to within 3 points on 49% of occasions. This was increased to 69% if the recovery trajectory (i.e. last BI score) was included in subsequent predictions (Tilling K et al 2001a). The average difference in predicted and observed outcome using the Tilling model was -0.4 with 90% limits of agreement between -7 and 6 (i.e. 90% of predicted values lie between -7 and +6 of the observed values) (Tilling K et al 2001a).

3.5.9.4 Reliability

The mNIHSS was derived from the NIHSS through removal of poorly performing and redundant items and through exploratory and confirmatory factor analysis using data from the two parts of the NINDS rtPA Stroke Trial (Lyden PD et al 2001; The National Institute of Neurological Disorders and Stroke rt-PA stroke study group 1995) as a means to simplify risk stratification and prognostication following stroke for the purposes of stratified randomisation in clinical trials (Lyden PD et al 2001). Correlations with the Barthel Index and mRS formed an assessment of concurrent validity. Substituting the modified scale into the original regression models yielded similar results to the parent NIHSS to predict a ‘global outcome score’ developed from four stroke outcome measures. To assess criterion validity, correlations between the mNIHSS and NIHSS were assessed. Inter-rater reliability for individual items was shown to be high between neurologists trained in the use of the NIHSS although reliability for non-trained practitioners remains untested (Lyden PD et al 2001). However, the mNIHSS may offer a simple and practical alternative to the NIHSS for risk adjustment if it can be shown to be reliable when used by non-stroke specialists.

3.6 Limitations of the review

This review provides a systematic overview of available externally validated prognostic models in stroke, updating previous reviews (Counsell C et al 2001; Kwakkel G et al 1996) to include more recent models and modelling methodologies. This review was based on a comprehensive and replicable search strategy producing a vast amount of literature for consideration. Despite this process, it is possible that relevant citations describing model development or validation of existing models have been overlooked. In addition, models that are yet to be externally validated and may yet prove to be good predictors of patient outcome may have been excluded from the review. Information regarding modelling techniques may not have been reported in detail in individual studies, and where this detail was lacking I have not attempted to obtain this information directly from authors. It is therefore possible that further robust models may have been excluded. Apparently poor performance of individual models in independent populations may reflect the methodology of external validation studies. It has not been possible to offer a quantitative summary of the performance of individual models in external populations due to the heterogeneity of external validation studies. Instead, validation studies have been presented individually to allow comparative assessment of their methodological quality and generalisability.

This review presented a number of methodological challenges. Firstly, a lack of universal criteria for scrutiny of prognostic research meant that the criteria against which the models were assessed are open to debate. Secondly, few of the included studies were based on data which was collected expressly for the purposes of model development. The secondary or retrospective use of data was common and models were often derived from available as opposed to desirable data. It is therefore possible that although the variables within the models make clinical sense, they are not necessarily the optimal factors to explain variability in patient outcome. Often, detailed descriptions of model development were lacking especially as regards checking of model assumptions and the characteristics of patients lost to follow up. Where such checks were not explicitly discussed, we have assumed that they did not occur, and this assumption may not be valid.

Table 6 Summary of statistical appraisal of models identified in the review

Model	Valid method of variable selection?		Control for Multicollinearity	Consideration of interaction terms	Events per variable >10?	linearity assumptions tested and met?	External Validation Acceptable discrimination (or sensitivity/specificity)	
Guys	✗	Multiple variables selected through identification of 'statistically significant' univariate predictors	✗	✗	✗	?	Sens Spec	0.83 0.58
G score	✗	Variables extracted from Guys model (simplified regression co-efficients to integers)	✗	✗	✓	?	Sens Spec	0.72 0.63
Bristol	?		✗	?	✗	✗	Sens Spec	1.00 0
SSV	✓	Use of stepwise variable selection and clinical reasoning	Stepwise variable selection	✓	✓	✓	✓	C statistic acceptable for prediction of alive and independent or dead/alive
Tilling	?		✗	?	✓	Tested; attempts to correct for censoring effects of Barthel Index did not affect the model	Predicts Barthel Index to within 3 points on 49% of occasions (increases to 69% if recovery history is included in the model). 90% limits of agreement -0.4 (-7, +6)	
Orpington	?		Stepwise variable selection	?	✓	✗	R^2 values used to assess model fit. Discrimination not tested	
Teale	✓	Variables selected through identification of important predictors in univariate analyses, regression trees and clinical reasoning	Stepwise variable selection	✗	✓	✓	Not externally validated	

3.7 Conclusions

Prognostic modelling has a number of applications in stroke. In the research setting prognostic models can be used in both randomised trials (for risk stratification) and observational studies (for case-mix adjustment) (Counsell C et al 2001). Potential applications in routine care range from the prediction of outcomes in individuals or groups of similar individuals (to facilitate treatment planning) to case-mix adjustment in the context of performance management within or between institutions. The usefulness of any prognostic model in these situations is likely to rely heavily on feasibility of data collection, and simplicity of application. In addition, any prognostic model should be robust in terms of the statistical methods used in its development, and in its predictive and discriminatory properties.

Of the six models that were subject to statistical scrutiny, only one (the SSV model) fulfilled all statistical criteria (Table 6 p 60). This model has been used in both randomised (Dennis MS et al 2003; IST3 trialists 2008) and observational (Reid J et al 2007) studies. Although the Tilling model has additional utility in terms of predicting individual recovery trajectories (Tilling K et al 2001a; Tilling K et al 2001b), the use of the Barthel Index as the predicted outcome is limiting due to its well documented ceiling effects (Salter K et al 2010).

This review aimed to update previous reviews (Counsell C et al 2001; Jongbloed L 1986; Kwakkel G et al 1996; Seenan P et al 2007) to identify risk adjustment models in light of the significant changes in stroke care that have occurred over the last decade. Twenty-three models were identified predicting a variety of outcomes following stroke. Of these, only six met quality criteria as regards the populations from which they were developed and the clinical utility of the covariates (Table 6 p 60). These factors are, to some degree, subjective and based around the specific requirements of a model for the CIMSS project. The exclusion of some models where the prognostic variables were felt to be too complex, or the time frames unrealistic for data capture as part of routine care by non-stroke specialists could be criticised for being over pessimistic. However, although there has been significant progress in direct admission to stroke units for patients presenting to hospital with stroke, this does not occur universally. Thus, any model which relies on specialist skills, or laboratory and radiological tests may result in systematic inconsistencies in data capture with the consequent introduction of bias.

With the exception of the Tilling model where multilevel modelling techniques were used, all the models included in this review were developed through single level regression modelling techniques. Alternative and more sophisticated modelling techniques exploit many of the conditions that make regression modelling difficult (e.g. the hierarchical or clustered nature of data) to provide more meaningful and clinically relevant models. Alternative modelling techniques such as latent class analysis, structural equation modelling or decision trees may offer additional explanations as to the relationships between case-mix and patient outcome and warrant exploration in the stroke setting.

Case-mix and risk adjustment is central to the validity of observational studies and also has utility for stratification in randomised trials. However, this review has not identified any new clinically useful and feasible model that can be used for these purposes since the Counsell model was developed ten years ago (Counsell C et al 2001). Despite advances in statistical modelling techniques, the available stroke risk adjusters are largely derived using regression modelling techniques with all their inherent problems. The use of more sophisticated techniques to develop robust case-mix adjustment models may increase confidence in the conclusions that may be drawn from observational studies of unselected populations.

Part II CIMSS research phase study

Chapter 4 Methods

To address the research questions posed in section 1.4 and to define a core stroke dataset for further testing, a prospective observational cohort study was performed (the CIMSS research phase study). The aims of this study were to test the feasibility of prospective data collection, to identify important (or redundant) data items, assess return and completion rates of postal questionnaires and explore case-mix adjusted relationships between care processes and patient reported outcomes.

4.1 Patient identification and recruitment

4.1.1 Study sites

Three study sites were selected as representative of the stroke services across Yorkshire and the Humber. Leeds Teaching Hospitals Trust (LTHT) is a large, multi-site teaching hospital and a tertiary referral centre within Yorkshire offering interventional radiological and neurosurgical services. Bradford Teaching Hospitals Foundation Trust (BTHFT) is a smaller teaching hospital with foundation status, and York Hospitals NHS Foundation Trust (YHFT) is a smaller foundation trust. All three sites have both acute and rehabilitation stroke units offering organised multidisciplinary acute and rehabilitation stroke services with the aim of direct admission to these units from the Emergency Department. The structure of stroke services at each trust is given in Table 7. During the study period, thrombolysis was offered at all sites Monday to Friday, during office hours. The main differences between the sites are the number of stroke beds, the provision of Early Supported Discharge (ESD) seven day rehabilitation services – both only available at LTHT. However, both BTHFT and YHFT offered ongoing community rehabilitation following discharge from hospital (Table 7).

4.1.2 Ethical and Research and Development (R&D) approvals

Ethical approval for the study was sought and obtained from the Bradford Regional Ethics Committee. R&D approvals were obtained individually from each of the three study sites.

4.1.3 Research staff

Researchers with a background in healthcare were employed to collect data in each of the three study sites. All the researchers underwent Good Clinical Practice (GCP) training and also attended an afternoon training session regarding the aims and objectives of the study, patient and carer recruitment and data collection processes. Specifically, training was provided on the processes for identification of potentially eligible patients and carers and obtaining informed consent. In addition, training was provided in the use of the data extraction forms (case report forms (CRFs)), use of the site file, creation of file notes and special processes for ‘unscheduled’ events such as patient death or withdrawal.

Table 7 Characteristics of stroke services at each study site

	Bradford	Leeds	York
Type of service	Hyperacute (Mon-Fri 9-5) Rehabilitation	Hyperacute (Mon-Fri 9-5) Rehabilitation	Hyperacute (Mon-Fri 9-5) Rehabilitation
Type of stroke unit	Acute Stroke Unit Rehabilitation Unit	2 Acute Stroke Units Rehabilitation Unit	Acute Stroke Unit Rehabilitation Unit
Total acute stroke beds	14	18, 15	15
Rehabilitation stroke beds	22	30	19
7 day rehabilitation?	No	Yes	No
ESD service available? Members of team	No	Yes CNS, SW, SLT, PT, OT, Dietician*	No
Community rehab team	Yes	Yes	Yes
Restrictions for access to SU	No	No	No
*CNS = Clinical Nurse Specialist (stroke), SW = Social Worker, SLT = Speech and Language Therapist, PT = Physiotherapist, OT = Occupational Therapist			

4.1.4 Screening data

Patients admitted to each Trust with stroke were identified by researchers through liaison with stroke care co-ordinators and stroke unit staff. Through this approach, patients that were admitted to wards other than the acute stroke unit were also identified.

Anonymous screening data were collected onto screening forms for all patients potentially eligible to take part in the study. This was in order to allow examination of the representativeness of the study sample through comparison of patients that consented to participate compared with the general post stroke population admitted to each site. Screening data comprising demographic details (age, sex and ethnicity) and a baseline functional score (Barthel Index (BI)) were collected on all patients admitted to participating centres with stroke during the study period. Reasons why patients were either not eligible or did not consent to participate in the study were also collected where this information was available.

4.1.5 Patient selection

Following screening, patients meeting eligibility criteria were approached for consent to participate in the study. Broad inclusion criteria were applied with the aim of recruitment of consecutive patients admitted to the study site with stroke. All patients were eligible for inclusion in the study if they had a primary diagnosis of stroke, and were recruited within a week of the onset of symptoms (or within two weeks if case-mix variable data could be extracted from the case-notes with respect to the week post stroke). Patients with subarachnoid haemorrhage and transient ischaemic attack were excluded, as were patients in whom it was clinically inappropriate to approach for consent (i.e. patients receiving palliative care). Patients were included if they provided informed consent to participate or, for patients unable to provide informed consent, only if they had a consultee (e.g. relative or carer) able to provide proxy consent. Patients unable to provide consent (i.e. patients who lacked capacity) were excluded from the study if they had no appropriate consultee. Therefore patients with cognitive impairment and dysphasia were not excluded from the study unless they lacked capacity and had no appropriate consultee. The main carer of patients (when available) was asked to provide consent to receive follow-up questionnaires regarding carer strain following stroke.

Consent was sought to extract data from patient case-notes and to send a questionnaire booklet to patients and carers at six months. A ‘tiered’ consent process was adopted whereby patients could consent to participate in certain parts of the study (e.g. baseline assessments), but withhold consent for e.g. follow-up.

Once patients had agreed to participate in the study and the consent forms were completed, the researchers telephoned the Academic Unit of Elderly Care and Rehabilitation in Bradford where a verbal eligibility and consent checklist was performed. Researchers were then given a study number which was added to all pages of the CRF.

Patients and carers were able to withdraw consent at any time during the study without offering a reason. Where a patient lost capacity during the course of the study, the decision to continue in the study was made by the main carer. Carers who had consented to participate in the study were withdrawn from further follow-up if the patient died between recruitment and follow-up.

4.1.6 Data collection

Data were extracted from case notes, therapy records and patient interview onto case report forms (CRFs) designed for the study. These included flowcharts for the completion and return of study paperwork, checklists to ascertain eligibility for the study and a patient and carer registration checklist. The CRFs were designed to reflect the patient pathway during their hospital stay with questions requiring data extraction at similar points in the pathway grouped together. A discharge checklist (including check of survival and discharge address) was completed and returned at patient discharge from hospital.

Researchers were asked to note any difficulties in recruitment, data extraction or particularly hard to collect data items for discussion in regular teleconferences. These teleconferences were also used to identify data items where differences in interpretation existed between researchers and sites. This information was used iteratively to improve standardisation of data collection processes and reduce any variability in application of data definitions that could impact negatively on the robustness of the study results.

Pages requiring collection of patient identifiable data contained no clinical data. Similarly, pages containing clinical data contained no patient identifiable data. Participant records were linked only by a unique identifier provided at registration.

4.1.7 Sample size

The study sample size was based on pragmatic consideration of the average number of patients admitted to each study site with stroke over a proposed recruitment period of six months. A formal power calculation was not performed as the 'treatment effect' of complex stroke care is difficult to quantify, and is likely to depend on process and care structure variables.

A conservative estimate of 30 patients per month per site admitted to each of the three sites was made based on the number of stroke admissions to the smaller study sites (Bradford Teaching Hospitals Trust and York Hospitals NHS Foundation Trust). Of the patients admitted with stroke it was assumed that one fifth would have suffered severe strokes and not be expected to survive until six month follow-up, and a further quarter would not be able to (or wish to) provide informed consent. A recruitment target of 300 (one hundred patients at each site) over six months was therefore set.

4.2 Development of the research dataset

The research dataset comprises four components: process markers, case-mix variables care structure variables and patient reported outcomes. The best available case-mix model and the outcome measures were defined through comprehensive examination of the literature as described in the following sections. The process variables included in the study were restricted to variables extracted from the RCP NSSA dataset and existing mandatory data requirements for stroke. Additional univariable case-mix variables were included in the study dataset based on the RCP dataset and clinical reasoning as described in section 4.2.3.

4.2.1 Patient reported outcomes dataset

A previously conducted systematic review of outcome indicators, valid and reliable for postal administration, examined as a thesis for a Master of Public Health was used to inform the choice of patient and carer outcomes questionnaires (Teale EA et al 2010). Six patient and three carer instruments with acceptable psychometric properties for self or proxy completion in physical, social and psychological domains (Table 8) were identified. Acceptable psychometric properties in terms of patient proxy agreement are particularly pertinent, as patients with dysphasia or cognitive problems were not excluded from study recruitment and some proxy completed questionnaires were returned.

Table 8 Patient and carer outcomes instruments identified in previously conducted systematic review (Teale EA et al 2010)

Patient outcomes instruments
Nottingham Extended Activities of Daily Living (NEADL)
Frenchay Activities Index (FAI)
Subjective Index of Physical and Social Outcome (SIPSO)
EuroQoL (EQ5D)
London Handicap Score (LHS)
London Stroke Satisfaction Questionnaire (LSSS)
Carer strain instruments
Carer Strain Index (CSI)
Carer Burden Score (CBS)
Bakas Carer Outcomes Score (BCOS)

In order to refine these instruments to create the battery of questionnaires for use in the CIMSS research study, consensus expert and consumer group consensus was sought. A stroke consumer group (Consumer Research Advisory Group (CRAG) for the Yorkshire Stroke Research Network) was consulted for views and opinions regarding utility of questionnaires including layout, wording and content.

A workshop with stroke clinicians and members of the stroke multidisciplinary team was also conducted and used group decision making techniques (nominal group theory) to rank the instruments identified through the postal stroke outcomes systematic review. Participants were asked to first list the important features of an outcomes measurement instrument in terms of utility (e.g. depth of questions, breadth of questions, important constructs to measure). These features were then ranked by all participants and the five most consistent important features identified. Participants were then asked to perform pairwise comparisons of all permutations of the outcomes instruments to create a ranking of all the identified instruments. These rankings were then combined to give an overall ranking of the individual outcomes measurement instruments. The instructions given to participants at the group decision making workshop is included in appendix B-1. Three patient questionnaires and one carer outcome questionnaire were identified through this process for inclusion in the outcomes datasets. These were the Nottingham Extended Activities of Daily Living (NEADL) (Lincoln et al 1992), the Subjective Index of Physical and Social Outcome (SIPSO) (Trigg et al 2000) and EuroQoL (EQ5D) (The EuroQoL Group 1990) patient outcomes questionnaires and for carers, the Carer Strain Index (Robinson B 1983). The use of visual analogue scales (VAS) has been shown to be unreliable in patients following stroke (Price et al 1999). For this reason, the EQ5D questionnaire was used, but the VAS was not included in the outcomes booklet.

The SIPSO instrument is provided in appendix B-3. It is a stroke specific scale in two subscales measuring physical and social reintegration following stroke (each comprising five questions with a five level response). The SIPSO has been shown to be well completed when administered by post (88-97% of returned questionnaires fully completed (Trigg 2000, 2003; Kersten et al 2004)) and to have acceptable patient-proxy agreement in total scores despite some variation in individual item agreement (Trigg et al 2003). A ceiling effect has been noted in the physical subscale in one validation study, although this study excluded dependent patients (Trigg et al 2000). The original validation studies suggest that the scores from the two subscales should be considered together (Trigg 2000), however, subsequent Rasch and Mokken analysis has suggested that the two subscores should be considered independently (Kersten 2010) (see also section 4.4.4).

The Nottingham Extended Activities of Daily Living instrument addresses four domains of functioning (see Table 9). The NEADL was completed at baseline (with respect to the week prior to completion) and again at six months. The instructions given with the questionnaire indicated that patients should “record what you have actually done over the last week”. At the time the baseline surveys were designed, it was anticipated that patients would be recruited within a few days of admission to hospital such that the previous week would relate to their pre-stroke function. However, as recruitment was often delayed, patients may have completed the baseline questionnaire with respect to their immediate post-stroke function. In future work, further clarification of the instruction to complete the baseline NEADL with respect to pre-stroke function will be required.

The Six Simple Variable case-mix adjuster model was developed to predict the dichotomised Oxford Handicap Scale (OHS) (Counsell C et al 2002). A postal version of the OHS questionnaire (as used in the FOOD trial (Dennis MS et al 2003)) was therefore included in the outcomes dataset to allow stratification of patients according to the SSV case-mix adjuster (see appendix B-2).

The systematic review of patient outcomes following stroke did not identify any measure of patient mood following stroke that was valid and reliable for postal administration. The GHQ_12 has been shown to be valid in patients following stroke, but lacks evidence of postal reliability (Teale EA et al 2010). In order to evaluate the reliability of the GHQ_12 collected by postal survey following stroke, a postal test-retest reliability study of the instrument was incorporated into the CIMSS research phase study.

In addition to these outcomes questionnaires, questions regarding return to work, return to driving and information provision were included in the outcomes questionnaire pack. These aspects of patient recovery are included as quality markers in the National Stroke Strategy (Department of Health 2007b) and were therefore collected to explore any relationship between these markers and processes of care.

Information regarding how the questionnaires were completed was also collected (self-completed, own answers but completed by carer or proxy responses). In all, six outcomes measurement instruments, and seven additional questions were included in the outcomes booklet. These are outlined in Table 9 (p 74) with a brief description of each instrument.

4.2.2 Process data set

A discussed in 1.2.1, best practice in stroke is described in national documents and there are several existing stroke process markers which have been developed to reflect the evidence base. The RCP NSSA audit dataset has evolved over the 12 years since its inception to reflect the emerging evidence base and consensus opinion on best practice (Intercollegiate Stroke Working Party 2011). Feasibility of retrospective extraction of these data from patient case-notes has been demonstrated through sequential audits. However, the audit datasets were not designed for prospective collection and, in addition, some of the markers are of unproven association with patient outcome. The process variables used in the CIMSS research phase were restricted to those used within the 2008 RCP NSSA dataset to allow specific exploration of case-mix adjusted process-outcome relationships. Through examination of systematically missing data, feasibility of prospective collection of RCP audit data will be tested. Components of the audit that were outwith the remit of the CIMSS research study were excluded (pre-hospital care and information regarding secondary prevention of stroke).

The latent traits of the SIPSO subscores are those of (physical and social) reintegration following stroke. These are complex constructs and are likely to rely not only on physical recovery, but also on psychosocial factors such as mood, social networks and community services. The process variables included in the study reflect these factors through recording assessment of impairments (e.g. speech and language assessments, occupational therapy and physiotherapy assessments); social care needs assessment, mood assessment and whether or not patients were able to return to their pre-admission address. The provision of community rehabilitation is likely to influence patients' functional and social reintegration following stroke. Collection of detailed data regarding post-hospital care was beyond the scope of the study. However, two markers post-discharge care delivery were collected: whether the patient was discharged to an intermediate care facility, and whether or not the patient received Early Supported Discharge (ESD) support. Whether or not a service included an ESD facility was also noted. For the purposes of the study, specific criteria as regards what comprises an ESD service were not stipulated, and ESD was said to be available providing a service was in place that facilitated early discharge from hospital with additional support and community therapy.

Times and dates of admission to hospital, admission to a stroke unit and imaging were collected to allow the derivation of metrics in line with the mandatory Integrated Performance Measures and Best Practice Tariff metrics (direct admission to a stroke unit, proportion of a patient's inpatient stay spent on a stroke unit and timeliness of brain imaging) (Department of Health 2008d; Department of Health 2010b). The complete process dataset is provided in Appendix B .

4.2.3 Case-mix adjustment variables

The best available case-mix adjuster was identified through a systematic review described in detail in Chapter 3. The Six Simple Variable case-mix adjustment model was used to adjust the study population for case-mix (Counsell C et al 2002). The SSV model was developed to predict independent survival with the dichotomised Oxford Handicap Score (OHS). The OHS is similar to the mRS, but with slight differences to the wording. A postal version of the OHS has also been developed (as used in the FOOD trial (Dennis et al 2006)). The postal OHS and mRS are included in Appendix 0 for comparison.

The SSV model was derived on an inception cohort of up to 30 days post-stroke onset (Counsell C et al 2002), although subsequent testing has shown that the model functions well if the variables are collected within a week of the stroke event (Dennis MS et al 2003). In order to define a discrete inception cohort, case-mix variables were therefore collected on patients within one week of stroke onset.

Additional case-mix adjustment variables were collected to investigate for a univariate predictor which may function as a simple case-mix adjuster. These variables were chosen as variables that either featured in the RCP NSSA dataset (e.g. reduced conscious level) or that have been postulated as predictors of post-stroke outcome (e.g. new urinary incontinence, or the Oxford Community Stroke Project (OCSP) classification of stroke (Bamford et al 1988)). Factors that may have a relationship with post-stroke function (e.g. the presence of speech or language deficits or the side of stroke), or that may confound the relationship between the baseline impairment and patient outcome (e.g. a previous disabling stroke or cognitive impairment) were also captured. For the purposes of the study, a pragmatic definition of drowsiness and confusion were applied. If there was documentation in the case-notes that there was evidence of the patient being drowsy or having a reduced conscious level between onset of stroke and recording of case-mix variables then patients were classified as having been drowsy. Similarly, if there was documented confusion (either through a narrative description or through more formal testing with a score such as the Abbreviated Mental Test or Mini Mental State Examination), patients were classified as having had confusion since the onset of the stroke. These pragmatic descriptions were applied in an attempt to reflect the ways in which these data may have been recorded during the course of routine patient care.

There is likely to be a degree of correlation between some of these univariable case-mix variables and the SSV model, as they are likely to be proxy markers for stroke severity. However, redundant markers (where there is significant collinearity) will be removed through construction of regression trees and stepwise variable selection during the modelling process. Case-mix variables collected during the study are tabulated in Appendix B.

There are likely to be a number of additional factors that confound the relationships between processes of care and patient outcome. For example, pain could act as a true confounder through restricting the delivery of specific care processes (e.g physiotherapy) and limiting functional outcome following stroke. However, the complexity of case-mix in the post-stroke population makes it unlikely that all of these factors could ever be accounted for. It is anticipated that the case-mix variables and process markers will act as summary measures or proxies for additional features of case-mix that are not measured explicitly (for example, a question regarding pain is included in the baseline EQ5D). An impression of how much variability in patient outcome is not explained by the variables in the model will be offered through examination of model fit. A poorly fitting model implies that there are important variables that have been excluded from the model. These could represent aspects of care process, organisational structure or case-mix.

4.2.4 Care-Structure

Organisational structure is likely to be an important mediator in the relationship between care process and patient outcome. The structure of stroke services in terms of staffing levels, capacity, patient monitoring, therapy time and specialist clinician input may all have an effect on patient outcome. However, in the CIMSS research study, it is unlikely that variability in organisational structure will be sufficiently diverse, nor the sample size large enough, to confidently attribute the effect of differences in patient outcome to variation in the organisation of stroke services.

However, information regarding the organisation of stroke services at each of the study sites was captured at the beginning and end of the data collection period according to the RCP NSSA organisational audit proforma (Royal College of Physicians 2009a), to ensure that there were no significant changes in the delivery of care over the data collection period that may present otherwise unmeasured confounding variables in the determination of patient outcome within or between sites.

4.3 Data collection processes

4.3.1 Baseline data

Following collection of screening data, patients and their carers were approached to provide informed consent. Patients receiving or likely to receive, palliative care (or their carers) were not approached to participate in the study. Once consent had been obtained, process data were extracted from patient case-notes. Data were extracted from existing records (case-notes and electronic hospital data systems) as far as possible in an attempt to mirror the capture of routine data. In this way, data items that are not routinely recorded (or that may require additional data extraction resource) were highlighted. Data were recorded onto Case Report Forms (CRFs) designed for the study. Case-mix data were extracted with respect to the week immediately following the stroke.

Patients (or their proxies) were asked to complete baseline outcome questionnaires (the Nottingham Extended Activities of Daily Living (NEADL), the General Health Questionnaire-12 (GHQ-12) and the EuroQoL). This was to allow the baseline assessments to be used to adjust for six month outcomes (i.e. to account for a change in the outcome score from baseline). The instructions for completion of the NEADL refer to activities actually performed in the week prior to questionnaire completion. The GHQ-12 and EuroQoL are completed with respect to the day of completion, and therefore represent measures of mood and quality of life in the immediate post stroke period. The environment in which these data are collected (i.e. the acute stroke unit) and the sudden change of circumstance in the immediate post-stroke period may make these measures difficult to interpret.

4.3.2 Data entry and verification of data

Baseline assessments and the CRFs were returned to the Academic Unit of Elderly Care and Rehabilitation in Bradford, and data entered into a bespoke web-based browser electronic data collection system. Double data entry was performed to flag and reduce data transcription errors. Attempts to obtain missing data identified at the data entry stage were made by data entry clerks.

4.3.3 Outcomes data

Follow-up questionnaire packs containing the outcomes instruments and instructions for completion were sent to surviving patients with a covering letter at six months post recruitment. In an attempt to maximise the return rates of the postal questionnaires, the outcomes packs were endorsed by the Stroke Association (TSA) and carried the Stroke Association logo (no additional funding for the study was provided by TSA). Checks on residency and survival were made through access to the “NHS spine portal” and through contacting patients’ General Practitioners. Participants who did not respond to the initial questionnaire were contacted by telephone and a further outcome pack sent if necessary. Outcomes packs were returned to the Academic Unit of Elderly Care and Rehabilitation in Bradford and entered into a bespoke electronic data collection system. Hard copies of identifiable and non-identifiable data were stored separately under a unique identifying study number. Patients who did not respond following reminders were deemed ‘non-responders’. The first 25 patients and carers at each site to respond to the outcomes questionnaire were sent a second (retest) questionnaire pack containing the postal version of the GHQ-12. The patient retest pack also contained the postal version of the modified Rankin Score.

Table 9 Outcomes questionnaires included in questionnaire packs

Patient pack	Number of items	Comments
"I would like more information about my stroke"	1 question	Quality marker from National Stroke Strategy.
Post-discharge review	1 question	Quality marker from National Stroke Strategy.
Return to work	2 questions	Quality marker from National Stroke Strategy.
"Two simple questions" (Lindley RI et al 1994)	2 questions	Two questions to place patients into one of three groups: independent completely recovered (1), independent some residual problems (2), residual problems requires at least daily assistance (3). These can be mapped onto the dichotomised OHS (1 = OHS of 0 or 1, 2 = OHS of 2, and 3 = OHS of at least 3)
Nottingham Extended Activities of Daily Living (Lincoln NB et al 1992)	22 questions in 4 domains, four level responses	Domains are: mobility, 'in the kitchen', domestic tasks, leisure activities. Questions are filled in with respect to what the patient has actually done in the last few weeks. Certain questions may be of limited relevance to some patients (do you write letters; do you manage your own garden; do you drive a car?)
Subjective Index of Physical and Social Outcome (Trigg et al 2003)	10 questions, 2 domains, 3 level responses	Physical and social subscores.
EuroQoL (The EuroQoL Group. 1990)	5 questions, 3 level responses	A measure of quality of life. A continuous utility score is calculated from which Quality Adjusted Life Years (QALYs) may be calculated.
GHQ-12 (Goldberg D et al 1988)	12 questions, 4 level responses	A screening tool for anxiety and depression. Lacks evidence of postal test-retest reliability in stroke populations and included in order to test this.
Postal Oxford Handicap Score (Dennis M et al 2006)	6 mutually exclusive questions	Scored from 0-5 to indicate level of dependency following stroke. An extra category (6) is often used to represent patients that have died. Often dichotomised <=2 and>=3 to represent independent vs. dependent survival
Proxy completion	3 questions	Respondents were asked to indicate if they completed the questionnaire unaided or with assistance
Carer pack		
Carer Strain Index (Robinson B 1983)	13 questions, yes/no responses	Questions relating to different aspects of caring and the effect on the carer
GHQ-12 (Goldberg D et al, 1988)	12 questions, four level responses	A screening tool for anxiety and depression. Lacks evidence of postal test-retest reliability in stroke populations and included in order to test this.

4.4 Statistical methods

A statistical plan designed to answer the research questions outlined in section 1.4 was developed a priori (see appendix B-4). Statistical support was offered by colleagues in the department of Biostatistics at Leeds University (Theresa Munyombwe, Brian Cattle and Robert West).

4.4.1 Data cleaning, outliers and missing data pattern analysis

Data were inspected for outliers and where these were identified, the original data were checked to ensure that there had not been data entry errors or data likely to reflect errors in recording data (for example, negative lengths of stay).

Tables were constructed (using STATA software) to examine the numbers of cases where there were missing data for individual item responses in patient reported questionnaires, and whether or not there were any patterns to this missingness.

4.4.2 Examination of return rates for outcomes questionnaire packs

The return rate for the six month questionnaire was calculated as the proportion of survivors to six month follow up who returned the questionnaire.

4.4.3 Descriptive statistics

4.4.3.1 Floor and ceiling effects of baseline and six month patient completed questionnaires

Floor and ceiling effects were identified through examination of histograms of patient reported questionnaires and presented as the percentage scoring minimum or maximum scores on each scale. Floor or ceiling effects were noted if questionnaires had more than 10% of respondents scoring at the extremes of the scale.

4.4.3.2 Tests of normality of continuous variables

Normality of continuous variables (and model residuals) was assessed through examination of histograms, and quantile normal (Q-Q) plots. In a normal (Q-Q) plot, the observed sample data is ranked in percentiles and plotted against the percentiles that would be expected if the data fitted a normal distribution. Deviation from a straight line in a Q-Q plot therefore indicates likely deviation from a normal distribution. Statistical significance at the 0.05 level on Shapiro-Wilk testing was used as a quantitative marker of deviation from a normal distribution.

4.4.3.2.1 Hypothesis testing

In order for continuous variables to be treated as parametric data for hypothesis testing (e.g. in the examination of representativeness of the study sample), they must approximate a normal distribution. If normality assumptions were not met, data were treated as non-parametric data.

4.4.3.2.2 Distributions of dependent regression model variables

The link function that is applied to the generalised linear model in order to derive the equation that fits a model is dependent on the underlying distribution of the outcome (dependent) variable (see section 3.5.6). Normally distributed continuous dependent variables may be modelled through linear regression and binary outcomes (binomial distribution) through logistic regression models. Examination of the distribution of dependent variables may identify that the variables are likely to fit an alternative distribution (e.g. a Poisson distribution for count data (Fox J, 1997)).

For the purposes of the linear regression modelling used in this study, continuous outcomes measurements should ideally be normally distributed. However, providing linearity assumptions (between individual independent predictors) and normality of residuals assumptions are met for any linear regression model, then non-normality of the dependent variable may be overlooked (see section 4.4.6.1.4). Where continuous outcomes variables are not normally distributed a variety of transformations have been explored to ascertain if the data can be normalised. This function is performed in STATA using the '*ladder*' and '*gladder*' commands. Lower (and statistically significant) chi-squared values in the STATA output suggest a better fit of the data to a normal distribution following the transformation. A Shapiro-Wilk test and p value are provided to indicate the confidence with which the null hypothesis (that data are normally distributed) may be accepted or rejected. Normal (Q-Q) plots are provided. Deviations at the tails of a distribution may represent the floor and ceiling effects of the measurement instruments. A mathematical function is unlikely to remove the presence of floor and ceiling effects, it is therefore unlikely that these distributions could be normalised through a simple transformation. An alternative model that accounts for censored data (such as a Tobit regression model) may be more appropriate in these instances (Twisk J et al 2009), although this is beyond the scope of this thesis.

4.4.3.2.3 Independent regression model variables

Normality of independent continuous variables in regression models is not essential, providing that model residuals are normally distributed (see 4.4.6.1.4). However, independent variables were examined to ascertain whether any transformation of the data improved normality, as this may improve model fit.

4.4.3.3 Examination of representativeness of study sample

Analyses were performed to ascertain whether there were significant differences in baseline variables between sites, between patients recruited and not recruited into the study (through examination of screening data) and between patients who responded to the six month survey and those who did not respond, who died or who withdrew from the study.

Examination of observed versus expected frequencies for categorical data were made with Chi squared tests (of Fischer's exact tests for contingency tables with cells containing less than five patients).

Identification of statistically significant differences between medians (non-parametric data) or means (parametric data) were made with Mann-Whitney U tests or independent sample t-tests respectively. Where these tests were across more than one group (e.g. comparisons across sites), a Kruskall-Wallis test (non-parametric) or oneway ANOVA (parametric) was performed. Pairwise examination of groups to identify the differences following a statistically significant Kruskall-Wallis or oneway ANOVA was then performed with Mann-Whitney U or independent sample t-tests.

4.4.4 Conversion of SIPSO to interval level data

The SIPSO outcome is an ordinal score and, as such, cannot be used in parametric statistical analyses. Rasch analysis is a statistical method whereby, providing the data fit the Rasch model, ordinal data can be converted to interval level data (thereby allowing mathematical manipulation and parametric analyses). Normality of the transformed score is not essential, providing that the assumptions of parametric analyses are met. The Rasch model generates a latent distribution of the probability of endorsement of an item in a scale (*p*) based on both person and question characteristics. In the context of the SIPSO, this means that patients with more of the latent trait (better social and physical reintegration) following their stroke are more likely to answer questions favourably, and that questions representing lower functioning are more likely to be endorsed by all (Kersten et al 2010).

Mokken analysis is a measure of the hierarchical nature of a scale. The hypothesis is that, patients will endorse items reflecting a level of function up to and including their actual function, but not items reflecting a greater level of function. Acceptable test statistics (Loevinger statistic >0.3) for this hypothesis, suggest that the scale is a valid, hierarchical scale (Kersten et al 2010).

The Rasch model is a logit function of the generalised linear equation containing two terms: person characteristics and item difficulty (equation (4)). The number of patients within each decile of logit (p) may then be plotted to give a histogram of the distribution of 'endorsement' (which should be approximately normal). If the item difficulty term is plotted against this histogram, it can be seen whether or not the items in the scale reflect the person characteristics (i.e. patients with more of the trait are more likely to endorse more difficult items and vice versa) and whether the scale items cover the whole of the distribution of the latent trait, or it there are areas that the scale does not address.

$$(4)\quad logit(p) = \beta_1 person\ characteristic \quad + \beta_2 item\ difficulty$$

The Rasch model relies on there being no interaction effects between person characteristics and item difficulty – i.e. older patients with a similar level of functioning should not answer questions differently to younger patients (differential item functioning). Differential item functioning should be explored to ensure that the scale properties do no vary according to baseline characteristics.

Rasch and Mokken analyses have been performed for the SIPSO in a population of younger stroke survivors (aged under 65) (Kersten et al 2010), and transformation factors provided such that the discrete scale may be transformed to interval level data measuring the latent trait. This Rasch analysis identified a two factor scale and confirmed unidimensionality of the two subscores. Mokken analysis confirmed that these two subscores behaved as valid ordinal scales (Kersten et al 2010). Differential Item Functioning (DIF) was observed for gender for some items in both subscores and this was dealt with by collapsing items such that both subscales conformed to the Rasch model. No DIF was observed for age, however, as the population excluded patients over 65, this is not surprising. For the purposes of the study, the SIPSO subscores were transformed using the transformation factors provided by Kersten et al (2010), with the caveat that the absence of DIF for age needs to be confirmed in an older population. This is, however, beyond the scope of this thesis and limitations based on the assumption that the transformations are valid in an older population are discussed in section 6.3.1.

4.4.5 Exploration of process-outcome linkages in the study population

4.4.5.1 Univariate (unadjusted) analyses

Unadjusted univariate analyses were performed to identify significant differences in patient outcome for patients that did, and did not receive specific process markers. Process markers coded “no”, “yes” and “no but”, would ideally be assessed with a oneway ANOVA. However, this relies upon normality assumptions being met for the outcome variables. Where these assumptions were not met, a Kruskall-Wallis test (non-parametric equivalent to the oneway ANOVA) was performed.

Variables reaching statistical significance at the 1% level were identified for inclusion in subsequent regression models providing their relationship with patient outcome made clinical sense. Conversely, variables failing to reach statistical significance in univariate analyses were entered into regression models if they were felt to be clinically important predictors. This helps to overcome problems with overfitting the model to the study population through inclusion or exclusion of variables on statistical rather than clinical grounds.

4.4.5.2 Construction of decision trees to predict CIMSS study outcomes to identify important predictors

Regression and classification trees are both types of decision tree and allow the graphical representation of the relative importance of independent variables in the prediction of the dependent outcome variable. Regression trees are used for the prediction of continuous outcomes, and classification trees for binary or categorical outcomes. As interval level outcomes have been used in the study (Rasch transformed SIPSO subscores), regression trees have been used. These have been constructed in R software (version 2.13.0) with no specification of the distribution of the dependent variable (i.e. a normal distribution of the outcome has not been assumed).

In the construction of regression trees, study participants are categorised into groups of predicted outcome based on their combination of predictor variable values. Starting with a full model (including all the predictor variables in the dataset), each predictor variable is considered in turn in order to identify the predictor which defines two groups between which the difference in mean outcome score is maximal. The value at which this split occurs is the cut point for that predictor and forms the first branch of the tree. This variable is the most important predictor in the dataset in terms of explaining diversity in outcome. The process is repeated, conditional on preceding branches such that, at the bottom of the tree, several outcome groups are created based on the tree algorithm defined from the dataset. Trees were 'pruned' (lower branches removed) in order to remove less influential variables and prevent over interpretation of the data. In interpreting regression trees, the left branch should be followed if the condition at the top of the branch is met (see Figure 95, Appendix C).

Regression trees do not rely on assumptions as regards the underlying distribution of the variables and there is no limit to the number or type of variables that may be entered into the equations to construct the trees (StatSoft Inc 2011). This allows the number of potential independent variables to be reduced before constructing final linear regression models. For the purposes of this study, the regression trees have been used to identify prominent variables (and therefore potentially important variables in the prediction of patient outcome) rather than for the prediction of absolute values of the SIPSO outcomes.

Two trees were created for each outcome. The first included baseline questionnaires (the Nottingham Extended Activities of Daily Living (NEADL), General Health Questionniare_12 (GHQ_12) and the EuroQoL utility score (EQ5D)). The second did not include these variables. Both models contained the Barthel Index as a marker of baseline stroke severity. The reason for excluding baseline assessments from one set of models was to ascertain whether outcome can be predicted without the need to collect baseline questionnaires, as collection of these data has implications in terms of resource and practicality in routine care. Variables entered into the regression tree models are given in Table 10 (p 81).

Table 10 Independent variables to be entered into regression tree models for prediction of the SIPSO subscores.

Demographic variables	Prognostic/severity variables	Patient movement	Process Variables	Baseline questionnaires
Gender	Length of stay	Admitted to stroke unit on same day, or day after admission	Scan within 24 hours of admission	Baseline Barthel Index
Ethnic group	Propensity score (calculated from age, independence pre-stroke, living circumstances alone pre-stroke, normal or abnormal verbal GCS score, ability to lift arms above head and ability to walk independently)	Ward type (ward patient first admitted to)	tPA given	Baseline NEADL
Study site	Pathological classification	No stroke unit care	Swallow screen in 24 hours	Baseline EQ5D
	Clinical classification (OCSP classification)	Early supported discharge	Aspirin in 48 hours	Baseline GHQ_12
	Weak side	Discharged same address	Physio in 48 hours	
	Dysphasia		OT in 4 days	
	Confusion at onset		MDT rehab goal setting	
	New urinary incontinence		Weighed during admission	
	Previous stroke		Mood assessment	
	Drowsy since presentation		Visual fields assessed	
			Sensory testing	
			Formal swallow in 72 hours	
			SLT communication assessment	
			Social worker assessment	
			Cognition screen	
			Malnutrition screen	
			Urinary incontinence care plan	
			Fluids within 24 hours of admission	
		Nutrition within 72 hours of admission		

4.4.6 Construction of linear regression models to predict SIPSO using important clinical variables and predictors identified in decision trees

The cut points defined in the regression trees are data driven – i.e. their absolute values are specific to the study data. The failure of important clinical predictors to feature in regression trees may represent peculiarities of the study dataset. The inclusion of these clinically important variables in the models may mediate the effect of variables which have been identified as important from the regression trees. As the focus of the study was the identification of potentially important predictors of patient outcome for further testing rather than the definition of prognostic models for external use, linear regression modelling was performed In order to explore the role of any clinically important predictors on variables identified through the data driven regression trees.

4.4.6.1 A priori model variable selection

4.4.6.1.1 Adjustment of the study sample using the SSV model

In observational studies, the propensity score is often referred to as the probability that a patient will have received a particular intervention on the basis of their characteristics. Instead, I have used the propensity score to denote the probability of the patient having a good or poor outcome (alive and independent vs. not as measured with the dichotomised OHS). The propensity score was calculated from the SSV model (probability of poor outcome as measured with the OHS) using the published, and externally validated beta coefficients (Counsell C et al 2002). Propensity score was added to regression tree equations and to linear regression models as an independent, continuous predictor. This approach has previously been adopted to adjust for case-mix in stroke studies (Bravata DM et al 2010).

The propensity score includes age and therefore age is not entered into the models as a separate variable (to avoid collinearity). Where propensity score does not feature in models, they have been re-run with age as an independent predictor as, in the absence of the propensity score, age may represent an important independent predictor of outcome. Additional case-mix or stroke severity variables are also added into the models to identify any further potentially important determinants of outcome that may be further investigated to see if they enhance the prognostic predictions of the SSV model.

4.4.6.1.2 Process variables

Independent variables included in each model are summarised in Table 11. These were identified for each SIPSO subscore, with and without baseline assessments, through clinical reasoning, the regression trees and through univariate analyses as described in sections 5.7 & 5.8.

The ‘no but’ codes of swallowing assessment, communication assessment and urinary incontinence variables indicate patients who do not require these assessments, either because their strokes are too mild, or too severe. Examination of twoway tables of association between individual process markers (receipt of communication or continence assessments) and the respective specific impairments (presence of dysphasia or incontinence) reveals strong correlations (Chi-squared tests, p=<0.001 see appendix E-1.1). As such, the process variables may act as proxy markers for the presence of these deficits, and the presence of the deficits is therefore not modelled explicitly.

Tests of linearity between individual predictors and the outcome were tested post model estimation as described in section 4.4.6.1.4. However, it was hypothesised a priori that the relationship between length of stay and physical outcome was likely to be non-linear, and this relationship was therefore explored prior to model development (see section 5.9.1).

Forwards and backwards stepwise automated variable selection procedures were applied to the variables identified through clinical reasoning, regression trees and variables that featured prominently in the univariate analyses. Model parameters were set such that variables reaching the 0.05 significance level were added to the model, and those consequently failing to reach significance at the 0.05 level were automatically removed. Models were also run with these parameters set at 0.5 to ensure that the statistically important predictors did not change appreciably when additional clinically (but not statistically) significant variables were included in the models. Variables where there is evidence of collinearity are automatically removed by the STATA software during stepwise variable selection procedures. Dummy variables were created automatically by the STATA software to represent levels of categorical data. Each dummy variable is entered as a dichotomous variable with respect to the reference variable which has been selected as the zero category for consistency. As the models created are linear, the beta co-efficients represent change in SIPSO subscore that would be expected for a one unit change in the independent variable when all other variables are held constant (Altman D, 1999 p 337)**.** For categorical (and dummy variables), the beta co-efficient represents the difference in mean SIPSO between the level of the variable and the reference variable with all other independent variables being held constant (Altman D, 1999 p 339). Variables within the model that reach statistical significance can either be identified through examination of the 95% confidence intervals (to see if they include zero implying non-significance) or through examination of the p value. Equations to predict the SIPSO score can be constructed from the beta-coefficients calculated through the modelling using the general linear equation (equation (2), page 49).

A table identifying statistically significant predictors, with beta co-efficients and confidence intervals is provided for each model. The adjusted R-squared value gives the ‘variance explained’ by the model (Altman D, 1999 p 345). A model with an R-squared of 0.4, therefore, would therefore explain 40% of the variation in patient outcome through the predictor variables. An F statistic that reaches statistical significance implies that the model explains a significant amount of variability in the dependent variable (Altman D, 1999 p 346). Each model is followed by tests to ensure that the final models meet linearity, normality of residuals, homoscedasticity and absence of collinearity assumptions (see section 4.4.6.1.4).

4.4.6.1.3 Pre-estimation checks

The study dataset contains several potential independent predictors of outcome. The STATA software will automatically exclude cases where there are missing data for independent variables – i.e. a complete case analysis is performed. There are therefore a number of cases that may be excluded from the analysis. Imputation techniques may be employed to overcome this difficulty, although this is beyond the scope of this thesis. It is important to consider whether these missing data may bias any analyses and in order to investigate this I compared the Barthel Indices of patients with complete data (that would be included in models) and those where data is incomplete (that would be automatically excluded). The baseline Barthel Index has been chosen for the comparison as there is only one missing case for this measure.

Pre-estimation checks of sample size for each model were performed, based on an event per variable (EPV) ratio of 10, as suggested by Peduzzi et al (Peduzzi P et al 1996). The number of variables entered into each linear model was limited to n/k where n=sample size and k=number of independent variables (including dummy variables).

Interaction effects between independent variables occur when the effect of one predictor on the dependent variable is mediated by the effect of another (see section 3.5.7.2). Inclusion of interaction terms (as the product of the two independent variables) into regression models as dummy variables accounts for these interaction effects. However, due to the size of the study dataset, the number of interactions that would need to be modelled and the potential reduction in EPV that would occur through inclusion of interaction terms, these have not been modelled explicitly.

Table 11 Dependent and independent variables included in each model. Total number of variables (including dummies) presented.

	Model								
	1	2	3	4	5	6	7	8	9
Dependent variable	Physical SIPSO	Physical SIPSO	Physical SIPSO	Physical SIPSO	Physical SIPSO	Social SIPSO	Social SIPSO	Social SIPSO	Social SIPSO
Description of model (independent variables)	Full model	Age instead of SSV	Influential cases removed	No baseline Ax	No baseline Ax, influential cases removed	Full model	Influential cases removed	No baseline Ax	No baseline Ax, influential cases removed
Independent variables	Number of variables including dummies								
Length of stay	1	1	1	1	1	1	1	1	1
Propensity score	1			1	1	1	1	1	1
Age		1	1						
Baseline NEADL	1	1	1			1	1		
Baseline EQ5D	1	1	1			1	1		
Baseline Barthel Index				1	1				
Admitted to stroke unit on day, or day after admission								1	1
Lacunar vs non-lacunar stroke								1	1
Early supported discharge								1	1
Imaging within 24 hours								1	1
Old stroke				1	1				
Formal swallowing assessment	2	2	2	2	2	2	2	2	2
Communication assessment	2	2	2	2	2	2	2	2	2
Social worker assessment	2	2	2	2	2	2	2	2	2
Urinary incontinence care plan	2	2	2	2	2	2	2	2	2
tPA given	2	2	2	2	2	2	2	2	2
First admitted to ward for hyperacute stroke care	1	1	1	1	1	1	1	1	1
Discharge to same address	1	1	1	1	1	1	1	1	1
Total variables (including dummies)	16	16	16	16	16	15	15	18	18

4.4.6.1.4 Post-estimation checks

a. Linearity assumptions

In order for the model to be valid, there must be a linear relationship between the continuous (or ordinal) predictors (independent variable) and the outcome (dependent variable) (Fox J, 1997 p 113). This may be assessed in different ways. The simplest way is to plot the model residuals against the individual predictors to identify obviously non-linear patterns (Chen X et al 2003). However, for this approach to be valid, there is an assumption that there is no relationship between the predictors in the model i.e. the presence of one predictor in a model does not affect the relationship between another predictor and the outcome. This is unlikely to be true for complex multivariable models (i.e. there is likely to be a degree of collinearity between variables). In order to circumvent this problem, a partial residual plot can be examined. Partial residuals are the component of variance attributable to a predictor having accounted for the variance due to other variables in the model. Post estimation 'augmented component plus residual plots' (acprplot) may be constructed easily using STATA software (Chen X et al 2003) and can identify more complex (e.g. polynomial) relationships between independent and dependent variables (Fox J, 1997 p 283).

An alternative approach to detect non-linearity is to categorise the independent variable and fit a model to predict the outcome. Comparison of the estimates from a univariable linear regression model, with those from a model using the categorised variable can identify whether the two models are significantly different (i.e. whether the model created from the categorised data, which allows a more complex relationship to be revealed, deviates significantly from a simple linear prediction – the likelihood ratio test) (UCLA: Academic Technology Services 2011). If there is no significant difference between the two models, the relationship may be assumed to be linear.

For the purposes of detecting non-linearity between predictors and outcome in the study data, I first plotted augmented component plus residual plots (acprplot). Where there is apparent deviation from linearity, I performed a likelihood ratio test to ascertain whether categorising the variable improves the model fit. However, it should be considered that categorisation of variables for the final models would result in the creation of dummy variables and this would therefore increase the number of variables which would need to be entered into the models.

Where linearity assumptions between continuous independent the dependent variables are not met, transformations that have been shown to improve the normality of the distributions (as outlined in section 5.3) have been substituted and the models re-run.

b. Normality of residuals

Residuals were estimated for each model constructed. These were tested for normality through Q-Q plots and Shapiro-Wilk testing (see 4.4.3.2)

c. Homoscedasticity

Homoscedasticity describes constant variance of model residuals across all fitted values. There should be no pattern in a scatter plot of fitted values against model residuals. Non-uniform variance may indicate an omitted variable exerting a systematic effect on the model (Fox J, 1997 p302). Homoscedasticity has been assessed through inspecting scatter plots of fitted values vs residuals and through quantitative hypothesis testing where rejection of the null hypothesis of homogeneity of variance occurs when the test reaches statistical significance at the 0.05 level (Breusch-Pagan test (Chen X et al 2003)).

d. Absence of collinearity

Entering independent predictor variables that are linearly related into regression models can lead to inflated or unstable beta co-efficients with wide confidence intervals. This can potentially result in poorly generalizable models where the relative importance of individual predictors is overestimated (Fox J, 1997 p337). Collinearity has been addressed a priori through application of clinical reasoning to model variable selection and during model construction through stepwise variable selection procedures, which reduces collinearity (Concaco J et al 1993). In addition, variance inflation factors (as a measure of any effect of collinearity on beta coefficients) were examined post-estimation (Fox J 1997). Variance inflation factors (VIF) greater than 10 are of concern and may indicate collinearity between independent predictors.

e. Influence and Leverage (DFBetas and Cook's D statistics)

The influence of individual cases on the model regression co-efficient or individual beta co-efficients depends on leverage (where a point lies relative to the distribution of the independent variable (X)), and it's residual. In simple terms, in the same way that torque is the product of distance from a pivot and force applied, the influence of a case on a regression line is a product of its leverage (distance from the centre of the distribution of X) and its residual (deviation of a point from the regression line for a given value of X).

High leverage points occur where individual cases occur at the extremes of the distribution of the independent variable (Fox J, 1997 p 268). Cases with unusual values for independent variables (at high leverage points) do not exert undue influence if the observed outcome is as predicted by the model (small residual) as they lie on, or near, the regression line (they are not regression outliers). Cases with large residuals exert less influence if the value of the independent variable is within the distribution of the variable for other cases (low leverage points). Conversely, cases with large residuals at high leverage points can exert considerable influence on the model co-efficients (Fox J 1997 p269).

A plot of leverage against the square of residuals can identify cases that are exerting particular influence on a regression model (Chen X et al 2003), and these have been provided for each specified model. The horizontal and vertical lines on the leverage vs. r-squared plot represent the mean leverage and r squared for all the points in the model.

Influence can be explored quantitatively through calculation of Cook's D statistic (D) - an overall marker of influence on the regression coefficient for each individual case (Fox J, 1997 p 277). The cut-off value of Cook's D above which individual points are likely to be exerting influence is determined as 4/n where n = the number of complete observations from which the model has been constructed (Chen X et al 2003). Cook's D statistics have been calculated for each model to identify particularly influential cases.

A measure of the effect of influential cases on individual beta-coefficients may be obtained through the calculation of DFBETA statistics. For each variable in a model, the beta co-efficients are calculated with all cases included, and then with each case excluded in turn. The modulus of the difference between these values is the DFBETA value for an individual case. This value is scaled by the standard error of the omitted co-efficient to enable the values to be compared on a single scatter plot (Fox J 1997 p276). Particularly influential cases are those where the magnitude of this difference is greater than 2/√n, (where n= the number of observations in the model (Chen X et al 2003). These limits may be presented the scatter diagram, such that outlying cases for particular variables can be seen.

It should be remembered that outliers do not necessarily represent 'wrong' data, but cases where outcomes are different to that which would be expected from the specified model.

4.4.7 Performance of the SSV case-mix adjuster to predict study outcomes

Utility of the SSV case-mix adjuster was explored through examination of its discriminatory properties (c-statistic) and calibration in the study dataset. These methods have been discussed in sections 3.5.9.1 and 3.5.9.2. In short, discrimination is the ability of a model to determine which, from of a pair of participants with incongruous outcomes, will have the outcome of interest (Harrell FE et al 1996). Calibration is the ability of a model to correctly predict outcome in the population of study participants.

In order to examine discrimination and calibration of the SSV and any identified univariable predictor to predict the SIPSO outcomes requires the SIPSO to be dichotomised to reflect 'good' over 'poor' outcome. Such a cut point for the SIPSO has not been determined in the literature. The cut point was therefore created at the level that represented a score of 3 on each of the individual SIPSO questionnaire items (representing a mild residual deficit that does not interfere appreciably with daily living), and the SIPSO subscores were dichotomised at 15. This was felt to be clinically comparable to an OHS dichotomised at <=2. Analyses were also performed using the data driven median of the SIPSO subscores to represent good over poor outcome. The calculations of c statistics were also performed first excluding, and then including and ascribing a score of zero, to patients who died.

4.4.7.1 Model discrimination (measured with c statistics)

For examination of the c statistics of the SSV model, covariates from the original published model were used to calculate the probability of outcome for each study participant (propensity score) using the generalised linear equation with a logit function (equation (3) p 49). These were then used against dichotomised study outcomes to plot Receiver Operating Curves (ROC).The area under a ROC curve for a binary outcome is equal to the c statistic. Confidence intervals were also calculated.

There is no value above which a c statistic is 'good' as this depends on both the clinical context and the purposes for which the model will be used. For the purposes of this study, the c statistics have therefore been used to examine the relative performance of the SSV model with any identified univariate predictor.

4.4.7.2 Calibration of the SSV in the CIMSS study population (calibration plots)

Within each decile of predicted probability (between 0 and 1), the proportion of patients (p) with observed good outcome (OHS <=2) was calculated and plotted (Counsell C et al 2002). Errors bars were created based on calculation of 95% confidence intervals for proportions, given by equation (5).

$$(5)\quad 95\%\ \text{confidence interval of p} = \text{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

Perfect calibration would be represented by all points falling on a line x=y (i.e. where predicted probability equals observed probability).

4.4.8 Exploration of potential univariate predictors of outcome that could be used in addition to, or instead of the SSV case-mix adjuster

Prominent variables in the models that may have utility as univariate case-mix adjusters were identified. In order to test their utility, they were entered as a single predictor into a logistic regression model to predict the dichotomised outcome. Fitted values from this model represent predicted probability of outcome using the univariable predictor. These fitted values were then used to plot ROC curves to examine discriminatory properties.

Calibration curves for variables identified through the study to predict the dichotomised outcomes have not been constructed as the predicted outcomes are derived from the observed outcome and these values would therefore be dependent on each other. In order to test calibration of univariable study variables, external validation in an independent dataset would be required.

4.4.9 Markov Chain MonteCarlo (MCMC) simulation iterations

In order to test the stability and convergence of model beta coefficients, single level regression models were recreated in MLWiN software and Markov Chain MonteCarlo iterations performed. A model is specified in the study dataset and the beta-coefficients and their standard errors are used to simulate latent distributions model beta co-efficients for each variable. This distribution is used to perform multiple automated calculations of estimates of the coefficient for one variable conditional on the other variables. The procedure is then repeated iteratively for each variable in turn resulting in estimations of model beta coefficients which 'settle' on an approximation of the true model co-efficients (central limit theory). If estimates of beta co-efficients fail to converge on a value after repeated iterations, the model is unstable. For the purposes of this study, 5000 MCMC post-estimation iterations were performed, with a 'burn in' of 50 iterations.

4.5 Data manipulation

Data were stored as comma separated variable (.csv) files within the data collection system and exported directly to statistical software. STATA version 11 was used for statistical analysis. Manipulation of variables was performed through creation of syntax files (.do files) to allow real time data exports to be used in data analysis.

Dates and times were converted from string variables to numerical variables, and categorical variables coded. For consistency, 'no' or 'false' was assigned a value of zero, and 'yes' or 'true' a value of one. 'No but' scores (processes that are either not indicated or contraindicated) are assigned a score of 2.

Where possible, durations were calculated from dates and times:

- Length of stay
- Time from hospital admission to scan
- Time to stroke unit admission
- Length of stay on stroke unit
- Length of stay post stroke unit discharge

Baseline and six month outcomes questionnaires were scored or coded as suggested in the literature, or by the authors of the instruments.

The NEADL and Barthel Index use a total summed score. Providing linearity and normality of residuals assumptions are met, these may be entered into regression models as continuous variables. If model assumptions are not met (see section 4.4.6.1.4.), they must be analysed as non-parametric data and treated as either ordinal data or categorised and treated as categorical data.

The General Health Questionnaire-12 (GHQ-12) has been scored using the dichotomised rather than Likert scoring system (Goldberg D et al, 1988). This system of GHQ-12 scoring ascribes a score of 0 for patients reporting absence of problems or no better/worse than usual and a score of 1 otherwise. A total GHQ-12 score is therefore out of a maximum of 12 (with higher scores indicating more problems). The GHQ-12 is again treated as continuous data unless violations to assumptions are encountered when it will be dichotomised. This is with the acceptance of the loss of information that this will incur.

Both the SIPSO and EuroQoL have conversion algorithms that allow the summed score to be converted to an interval score. These variables are therefore treated as continuous variables providing linearity and normality of residuals assumptions are met. Although performed on a population of younger patients (Kersten et al 2010), the output from the Rasch analysis of the SIPSO should not be dependent on the underlying population and the conversion should therefore be transferable. However, this is with the caveat that specific examination of differential item functioning of age has not been performed in an older population. Using 'time trade off (TTO) techniques and visual analogue scales based on 'value sets' the creators of the EQ5D have developed formulae to allow conversion of an individual's answers across the five EQ5D questions to a continuous score (between -1 and 1) to reflect perceived quality of life. These norms are country specific and the UK 'time trade off' values have been used for the purposes of this study (Rabin R et al 2011).

The SSV case-mix model is used to calculate the probability of good outcome using the beta coefficients from the equation created through the original logistic regression analysis (Counsell C et al 2002). This probability of outcome was then dichotomised at 0.8 to give the probability of good (≥0.8) over poor (<0.8) outcome. The value of 0.8 was chosen as this is the cut off that was used to stratify in the FOOD trial (M Dennis, personal communication), the data from which formed a large external validation study of the SSV model (Dennis MS et al 2003). In observational studies, a propensity score usually refers to the calculated likelihood of a patient receiving a specific treatment based on their characteristics, however, for the purposes of this thesis, the propensity score has been used to denote the probability of a patient having a good outcome (defined in this case as a dichotomised modified Rankin Score of less than three) based on their baseline characteristics.

Chapter 5 Results

5.1 Data cleaning

5.1.1 Outliers in continuous process data

Inspection of continuous process variables was performed to identify any outliers. This revealed some anomalies requiring further inspection of the variables ‘date and time of hospital admission’ and ‘date and time of admission to the stroke unit’. For four patients the date and time to admission to the stroke unit is before the date and time of admission to hospital. This may be possible for in-hospital strokes, but the differences in time are small (-4.3 to -0.8 hours). This is therefore more likely to reflect an error or inconsistency in the way that the time of admission was recorded. Discussion with researchers revealed that this variable was extracted from either ED records or from the Patient Administration System (PAS) where the ED records were not available. It is possible therefore, that the unreliability of the data stems from inconsistencies within the PAS database. This inconsistency has implications for the reliability of other variables that rely on time of hospital admission for calculation (e.g. time to scan). These variables have therefore been excluded from further analysis as the number of cases where there may be inconsistencies is not apparent. Variables that rely on date of admission (e.g. length of stay), are however unaffected by the time of admission and may therefore be calculated. The variable “scan within 24 hours of admission”, was recorded as a dichotomous yes/no response. Although the derivation of this variable requires knowledge of both the time of admission and time of scan, it does not rely on these times having been recorded in the CRF. There is an assumption in the use of this variable that the times of hospital admission and time of scan were available for the researchers to calculate whether or not the scan occurred within 24 hours of admission to hospital (and that this calculation was correct). However, this assumption may not be valid. A preferable approach would have been to record the primary data from which the variable was calculated rather than recording the derived variable. The relative merits and difficulties of recording data in this way are discussed in section 6.3.3.

There were marked inconsistencies between the length of stay on a stroke unit recorded as number of days by the stroke researchers and the calculated length of stay from dates of admission and discharge where these were available. The date of discharge from the stroke unit is missing in 203/298 (68%) of patients who received treatment on a stroke unit. As a consequence, the number of missing data for days on a stroke unit and proportion of stay spent on a stroke unit make the use of these variables unviable as the risk of systematic error is too great (39% each). Length of stay on a stroke unit has therefore not been used as a variable as the data were deemed to be unreliable.

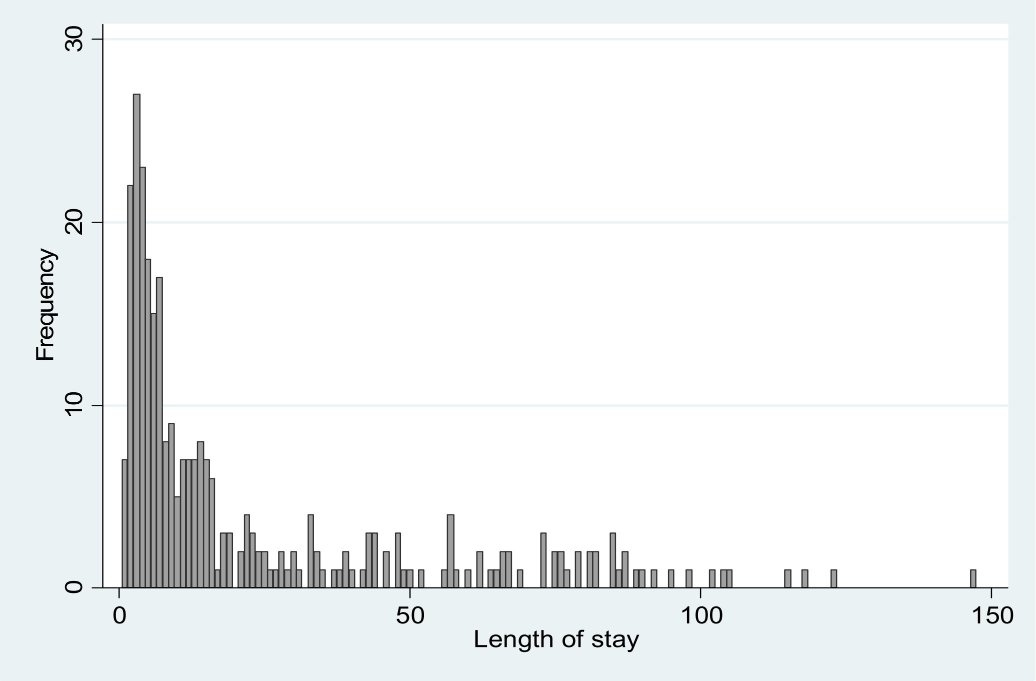
For the purposes of analysis therefore, admission to a stroke unit for any part of the inpatient spell (vs. no stroke unit care), admission to stroke unit on the same day or day after hospital admission and total length of hospital stay (calculated from date of hospital admission and date of hospital discharge) have been used as markers of timeliness of stroke unit admission.

Length of acute hospital stay in whole days was derived from the date of admission and date of discharge from the acute hospital, as recorded by the study researchers in the CRF. Additional post discharge lengths of stay in geographically distinct inpatient rehabilitation facilities, intermediate care facilities or any time spent under the care of post-discharge community rehabilitation or early supported discharge teams were not recorded. Patients discharged from the acute trust to receive further community therapy are likely to represent patients from a different subgroup of the post-stroke population to those patients that require protracted lengths of acute hospital care due to the severity of, or complications from, their stroke. It would therefore have been beneficial to measure and model the duration and nature of community rehabilitation separately from the acute hospital stay; the approach to the measurement of lengths of stay used in the study could be argued to be over simplistic, and to exclude important aspects of additional post-stroke rehabilitation. However, the study was not resourced to capture these post-acute hospital data.

Spurious data for length of stay were identified through examination of negative values and identification of cases where duration from hospital admission to stroke unit admission or stroke unit discharge to hospital discharge were particularly long (two cases) or where there were negative values for length of stay (one case). This identified three cases with spurious data which, when checked against the original CRF reflected data recording or data entry errors of exactly one month in either hospital admission or hospital discharge dates. These were assumed to be erroneous and were corrected.

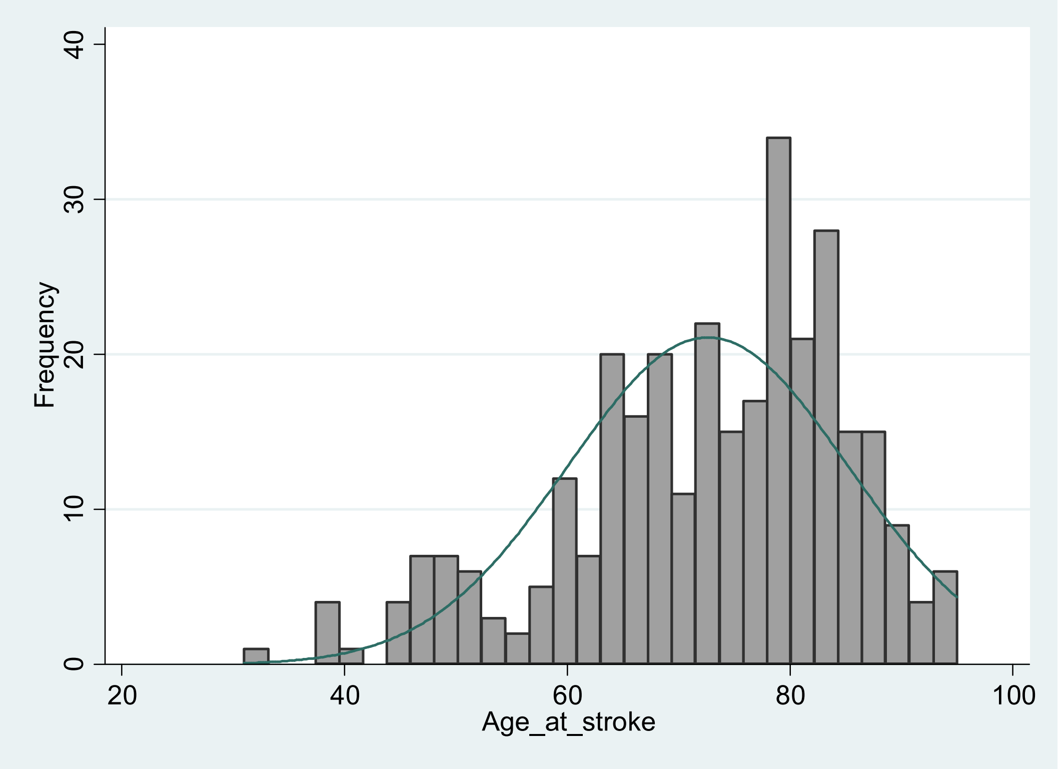
After correction for the spurious data, an examination of a histogram of length of stay reveals that there is still a very wide distribution. However, on examination of individual records for patients with lengths of stay greater than 100 days, these were felt to reflect true lengths of stay. They were therefore retained.

Figure 8 Distribution of length of stay in study population



The distribution of patient age at stroke demonstrates marked negative skew. This reflects the increased incidence of stroke with increasing age. Although there are two outliers markedly younger than the rest of the population, this is clinically feasible.

Figure 9 Distribution of age at stroke in study population



5.2 Missing data

5.2.1 Continuous process data

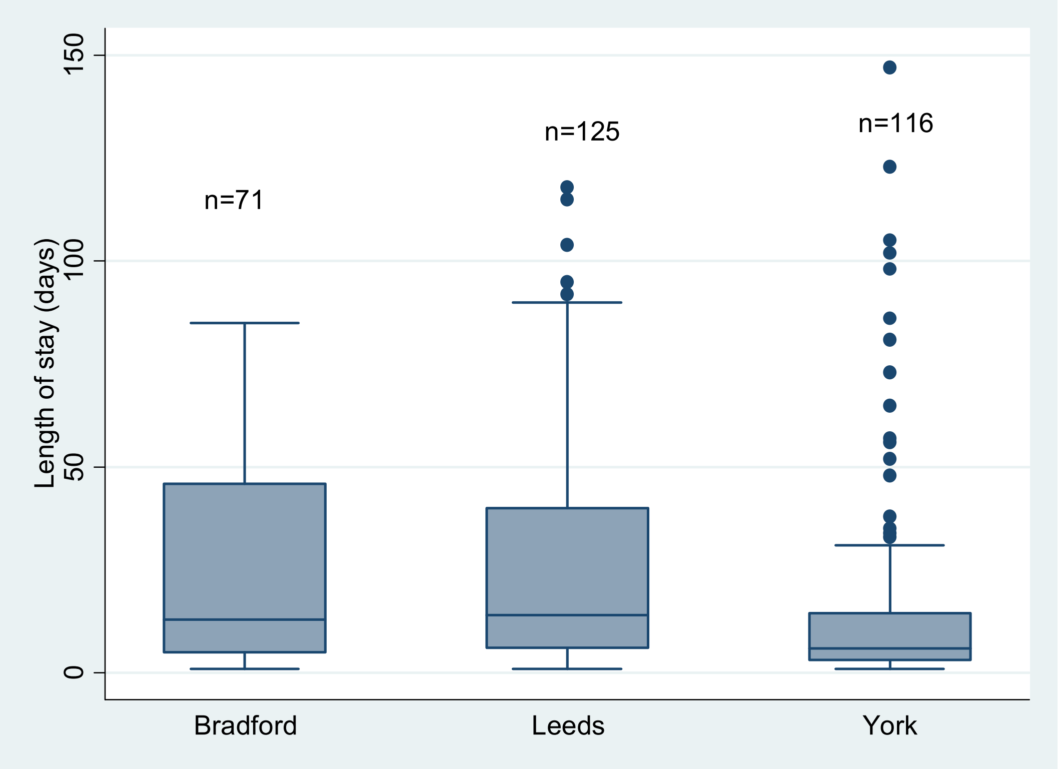
The process markers selected to represent length of stay and timeliness of stroke unit admission are generally well completed.

Table 12 Missing data regarding stroke unit treatment and length of stay by site

Variable	Site	Number recruited	Number missing data	Proportion of missing data
Treated on a Stroke Unit for part of inpatient stay	Bradford	71	1	1.5%
	Leeds	125	1	0.8
	York	116	0	-
Admission to SU same day, or day after admission	Bradford	71	3	4.2%
	Leeds	125	14	11.2%
	York	116	4	3.4%
Length of stay	Bradford	71	1	1.5%
	Leeds	125	11	9%
	York	116	4	3%

The rates of missing data for length of stay are small (16 cases in total, with the majority of missing data from Leeds). A Kruskall-Wallis (equivalence of populations test) between sites reveals that there is a significant difference in the median length of stay between sites (chi-squared (χ^2) = 21.1, degrees of freedom (ν) = 2, p<0.001).

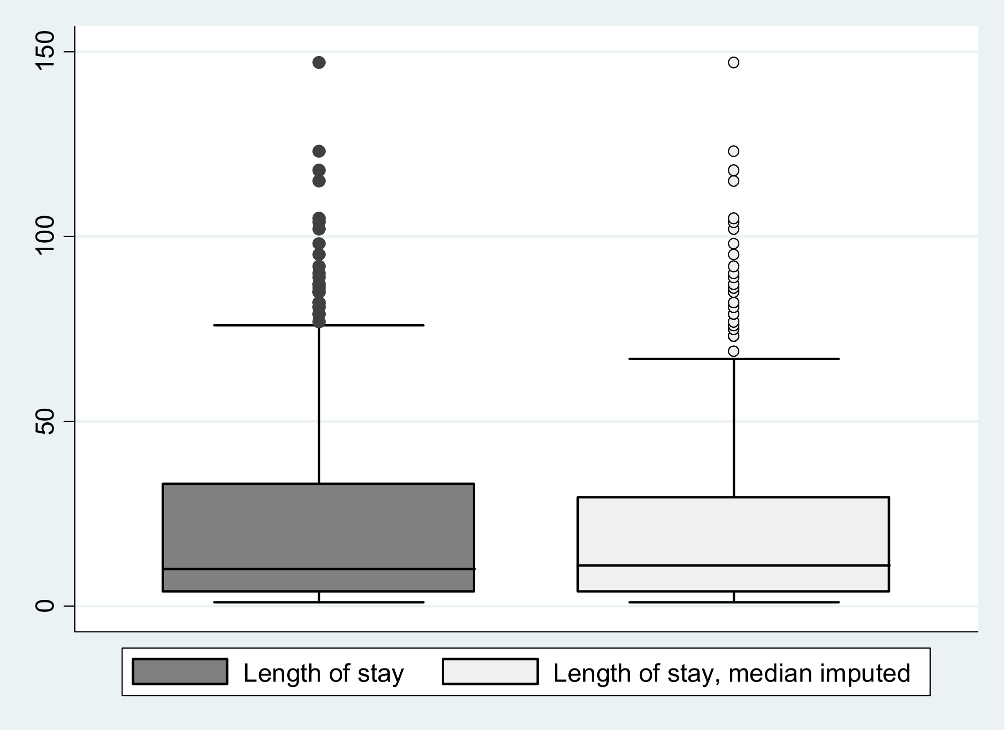
Figure 10 Length of stay by study site



If the median length of stay (for all sites) is imputed for missing values at each site (10 days), the difference in median length of stay between sites remains significant ($\chi^2 = 21.4$, $\nu = 2$, $p<0.001$). This difference also remains significant if the median length of stay at each site is imputed (Bradford 13 days, Leeds 14 days, York 6 days) rather than the median across sites ($\chi^2 = 21.5$, $\nu = 2$, $p<0.001$).

A box plot to examine the effect of imputation of the median length of stay (LOS_imput) on the distribution of length of stay reveals that this does not significantly change the median length of stay. Missing data for length of stay may therefore be ignored.

Figure 11 Box plot of length of stay across sites and with imputation of median length of stay



5.2.2 Categorical process data

Missing data for categorical data were infrequent (Table 12). Researchers reported particular difficulty in extracting the time of imaging as this was often not recorded in patient case notes. These data were therefore obtained from the electronic results servers at each trust.

Table 13 Missing categorical data

Variable	Missing (%)
Patient not treated on a stroke unit	2
Discharge address the same as admission address	2
Admitted to stroke unit on same day, or day after presentation	21 (7)
Type of ward patient first admitted to	3
Patient discharged with Early Supported Discharge team input	18 (6)
Radiological classification of stroke (infarct or haemorrhage)	2
OCSP classification of stroke	17 (5)
Side of weakness	1
Lived alone or cohabited pre-stroke, or admitted from nursing or residential care	10 (3)
Independent activities of daily living prior to stroke	0
Normal verbal GCS score	0
Able to lift arms above head (or MRC power score >=3) in week following stroke	1
Able to walk unaided in week following stroke	1
Drowsy since presentation to hospital	4
Evidence of dysphasia	0
Evidence of confusion	2
New urinary incontinence or newly catheterised since stroke	5
Previous disabling stroke	1
Imaging performed within 24 hours	5
Thrombolysis (rtPA) given	2
Swallowing screen performed within 24 hours of admission	2
Aspirin (or alternative antiplatelet) given within 48 hours of admission	1
Physiotherapy assessment within 48 hours of admission	1
Occupational therapy assessment within 4 working days of admission	4
Evidence of multidisciplinary team goal setting	1
Patient weighed during the admission	1
Evidence of an assessment of patient mood	7
Documented visual field assessment	2
Documented sensory assessment	3
Formal swallowing assessment (by Speech and Language therapist) within 72 hours	2
Formal communication assessment by Speech and Language therapist	1
Assessment by social worker	2
Assessment of cognitive function	3
Patient screened for malnutrition	0
Documented continence promotion plan	2
In receipt of fluids within 24 hours of admission	0
In receipt of nutrition within 72 hours of admission	0

As cases with missing data will be excluded automatically when entered as independent variables into regression models, the outcomes of those with missing data (and therefore excluded from the analysis) will be compared with those with complete data to ensure that there is no systematic difference between patients with complete data and those in whom data are missing.

5.2.3 Missing Baseline questionnaire data

Baseline questionnaire packs were not returned for ten patients. There was only one participant where a baseline Barthel Index was not available.

5.2.3.1 The Nottingham Extended Activities of Daily Living (NEADL) baseline questionnaire

The NEADL was fully completed by 90% of patients at baseline. Missing values were spread across seven variables (managing garden, writing letters, driving, going out socially, reading books, managing money and using the phone). Management of the garden was the most frequently omitted item (8 participants for whom baseline assessments were available excluded the item). Missing items tend to be from the 'leisure activities' subscale of the NEADL. The total number of missing items for each of the subscores is shown in Table 14, with the number of patients responsible for the missing data. The majority of the missing data are in the domestic tasks and leisure activities subscales of the questionnaire. Three participants missed the last 11 items which may reflect omitting (or overlooking) an entire page of the questionnaire.

Table 14 Missing data by domain for NEADL questionnaire (missing baseline packs excluded)

NEADL Subscore	Total number of missing data	Number of participants with missing data
Mobility	13	8
'In the Kitchen'	4	1
Domestic tasks	27	10
Leisure activities	42	16

5.2.3.2 The EuroQoL

The number of missing data items for each question of each questionnaire is presented. The 'missing value patterns' tables identify if there are patterns in the combinations of missing data by displaying number of patients with missing responses (indicated by a 1) for each question. In these tables a 1 that the item was missing. For example, in Table 15, 6 patients omitted only the question pertaining to ability to perform usual activities on the EQ5D, whilst three omitted two questions pertaining to pain and anxiety - highlighted in grey in the table.

Table 15 Missing data patterns by domain in the baseline EuroQoL (including ten missing baseline packs)

Number of patients	1	1	2	2	3	5	6	10	282	Total missing data
Mobility	0	0	0	0	0	0	0	1	0	10
Self-care	1	1	0	1	0	0	0	1	0	14
Usual activities	1	1	0	0	0	0	1	1	0	18
Pain/discomfort	0	1	1	0	1	0	0	1	0	16
Anxiety/depression	1	0	0	0	1	1	0	1	0	19

Nine returned baseline questionnaires had missing data for the anxiety question.

5.2.3.3 GHQ-12 baseline

The authors of the GHQ_12 instrument suggest that missing GHQ-12 items are replaced with the most pessimistic score (Goldberg D et al, 1988) . However, missing data analysis of the raw data from returned baseline questionnaires reveals missing data across all questions 0. Again, ten baseline assessments were not returned and these are included in the table of missing data. Three patients who returned a baseline questionnaire pack did not complete any of the GHQ-12 questionnaires.

Table 16 Key for Table 17 (questions of the GHQ-12)

Key	Question: Have you recently…
1	been able to concentrate on whatever you're doing
2	lost much sleep over worry?
3	felt that you are playing a useful part in things?
4	Felt capable of making decisions about things?
5	felt constantly under strain?
6	felt you couldn't overcome your difficulties?
7	been able to enjoy your normal day-to-day activities?
8	been able to face up to your problems?
9	been feeling unhappy and depressed?
10	been losing confidence in yourself?
11	been thinking of yourself as a worthless person?
12	been feeling reasonably happy, all things considered?

Table 17 Missing data patterns for baseline GHQ_12

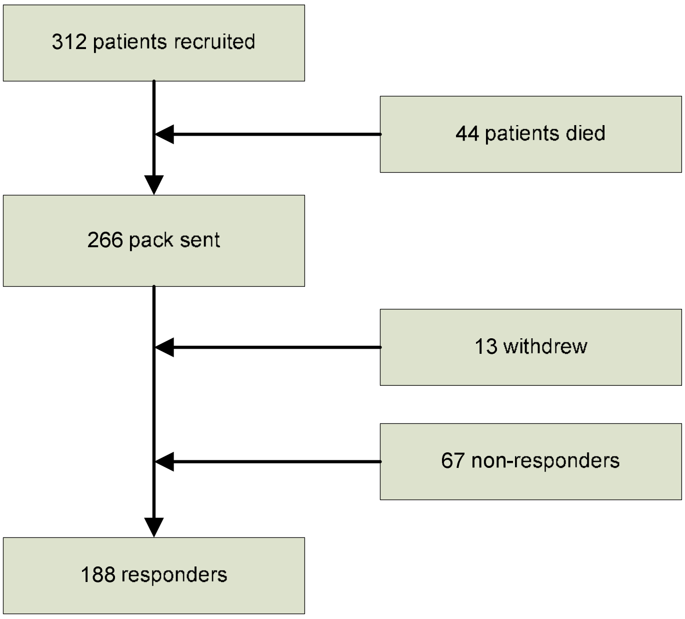
		Question number											
		1	2	3	4	5	6	7	8	9	10	11	12
Number of patients with missing item	13	1	1	1	1	1	1	1	1	1	1	1	1
	3	0	1	0	0	0	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	1	0	0
	2	0	0	0	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	1	0	0	0	0	0
	1	1	0	1	0	0	0	1	0	0	0	0	0
	1	1	0	0	1	0	0	0	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	1	1	0	0
	1	1	0	0	0	0	0	0	0	1	0	0	0
	1	0	0	0	0	1	1	1	1	1	1	1	1
	1	0	0	0	0	1	1	0	0	1	0	0	0
	1	0	0	0	0	1	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	1	1	0	0
	1	0	0	0	0	0	1	0	0	0	0	1	1
	1	0	0	0	0	0	1	0	0	0	0	0	1
	1	0	0	0	0	0	1	0	0	0	0	0	0
	1	0	0	0	0	0	0	0	1	0	0	0	0
	1	0	0	0	0	0	0	0	0	1	1	1	1
	1	0	0	0	0	0	0	0	0	1	1	1	0
	1	0	0	0	0	0	0	0	0	1	0	0	0
	1	0	0	0	0	0	0	0	0	0	0	0	1
	270	0	0	0	0	0	0	0	0	0	0	0	0
	Total	17	16	17	16	16	19	17	15	21	21	17	18

5.2.4 Outcomes data

5.2.4.1 Missing outcome packs

The flow of patients recruited into the study is shown in Figure 12. The overall response rate was calculated as the proportion of survivors responding to the questionnaire at six months (after reminders if these were required).

Figure 12 Questionnaire returns at six month follow up

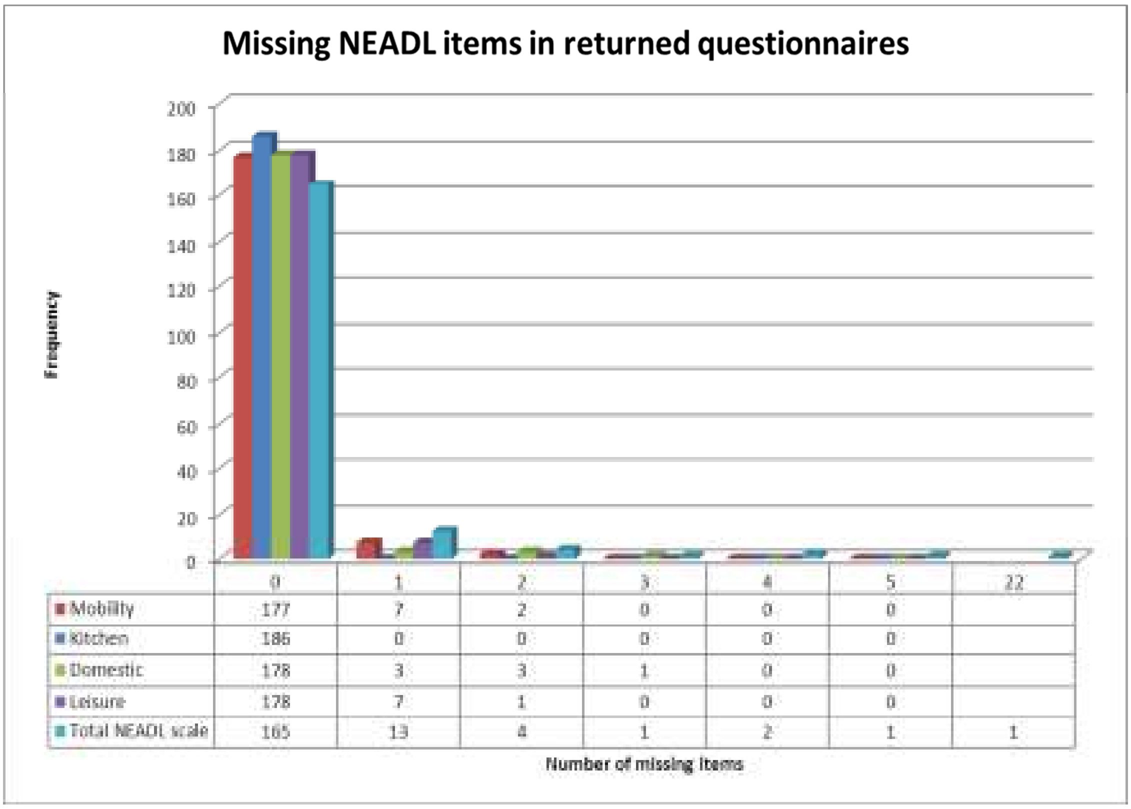


Response rate = 188 / 266 = 71%

5.2.4.2 Missing individual items

Figure 13 to Figure 18 represent the total number of missing questions for each outcome broken down by subscales where these apply (NEADL and SIPSO). For example, Figure 13 concerning the NEADL shows that seven respondents missed one item from the mobility subscale, one respondent missed three items from the domestic subscale and two respondents missed four items across all the NEADL subscales. There is no pattern to the missing items in the NEADL (i.e. missingness is spread across items).

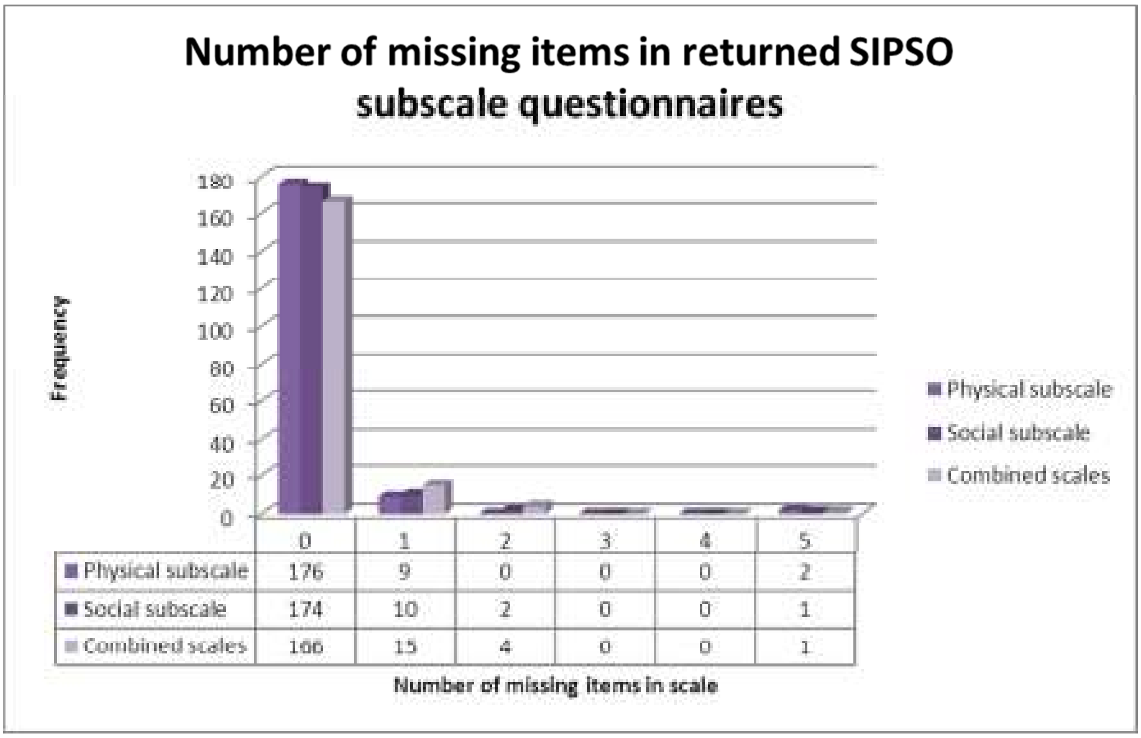
Figure 13 Frequency of missing NEADL data items in returned 6 month questionnaires



	0	1	2	3	4	5	22
■ Mobility	177	7	2	0	0	0	
■ Kitchen	186	0	0	0	0	0	
■ Domestic	178	3	3	1	0	0	
■ Leisure	178	7	1	0	0	0	
■ Total NEADL scale	165	13	4	1	2	1	1

There is no pattern to this missing data in terms of individual items that are not completed.

Figure 14 Frequency of missing SIPSO data items in returned 6 month questionnaires



	0	1	2	3	4	5
■ Physical subscale	176	9	0	0	0	2
■ Social subscale	174	10	2	0	0	1
■ Combined scales	166	15	4	0	0	1

Table 18 Key for table Table 19 (Physical subscore of SIPSO)

Key	**Question**
1	Since your stroke, how much difficulty do you have dressing yourself fully?
2	Since your stroke, how much difficulty do you have moving around *all* areas of the home?
3	Since your stroke, how satisfied are you with your overall ability to perform daily activities *in and around the home*?
4	Since your stroke, how much difficulty do you have shopping for and carrying *a few items* (1 bag of shopping or less) when at the shops?
5	Since your stroke, how independent are you in your ability to *move around your local neighbourhood*?

Table 19 Missing-value patterns in physical subscore of SIPSO

	Question					
Number of patients with missing item		1	2	3	4	5
	4	0	0	0	1	0
	2	0	0	0	0	1
	1	1	0	0	0	0
	1	0	1	0	0	0
	1	0	0	1	0	0
	Total	2	2	1	4	2

There is no pattern to the missing items on the physical subscore of the SIPSO. However, examination of the social subscore of the SIPSO reveals that five patients did not respond to the question “since your stroke, how do you feel about your appearance when out in public?”.

Table 20 Key for Table 21 (social subscore of the SIPSO)

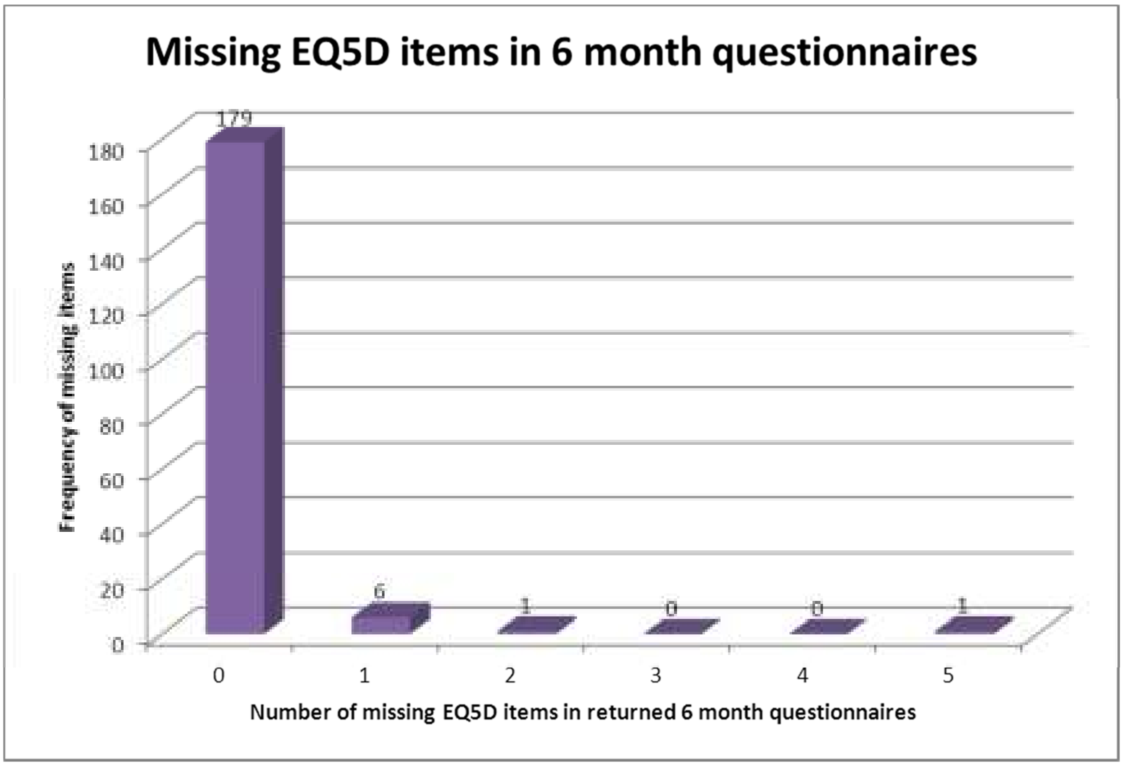
Key	**Question**
1	Since your stroke, how often do you feel bored with your free time at home?
2	Since your stroke, how would you describe the amount of communication between you and your friends/associates?
3	Since your stroke, how satisfied are you with the level of interests and activities you share with your friends/associates?
4	Since your stroke, how often do *you visit* friends/others?
5	Since your stroke, how do you feel about your appearance when out in public?

Table 21 Missing-value patterns in social subscore of the SIPSO

	Question					
Number of patients with missing item		1	2	3	4	5
	4	0	0	1	0	0
	3	0	0	0	0	1
	2	1	0	0	0	0
	1	1	0	0	0	1
	1	0	0	0	1	1
	1	0	0	0	1	0
	174	0	0	0	0	0
	Total	3	0	4	2	5

The Rasch analysis of the SIPSO suggests that the structure of the scale is such that the subscales should be considered separately (Kersten P et al 2010). Each SIPSO subscore is better completed than the total NEADL scale.

Figure 15 Frequency of missing EQ5D data items in returned 6 month questionnaires



No patterns were identified in the missing data for the six month EQ5D questions.

Figure 16 Frequency of missing GHQ_12 data items in returned 6 month questionnaires

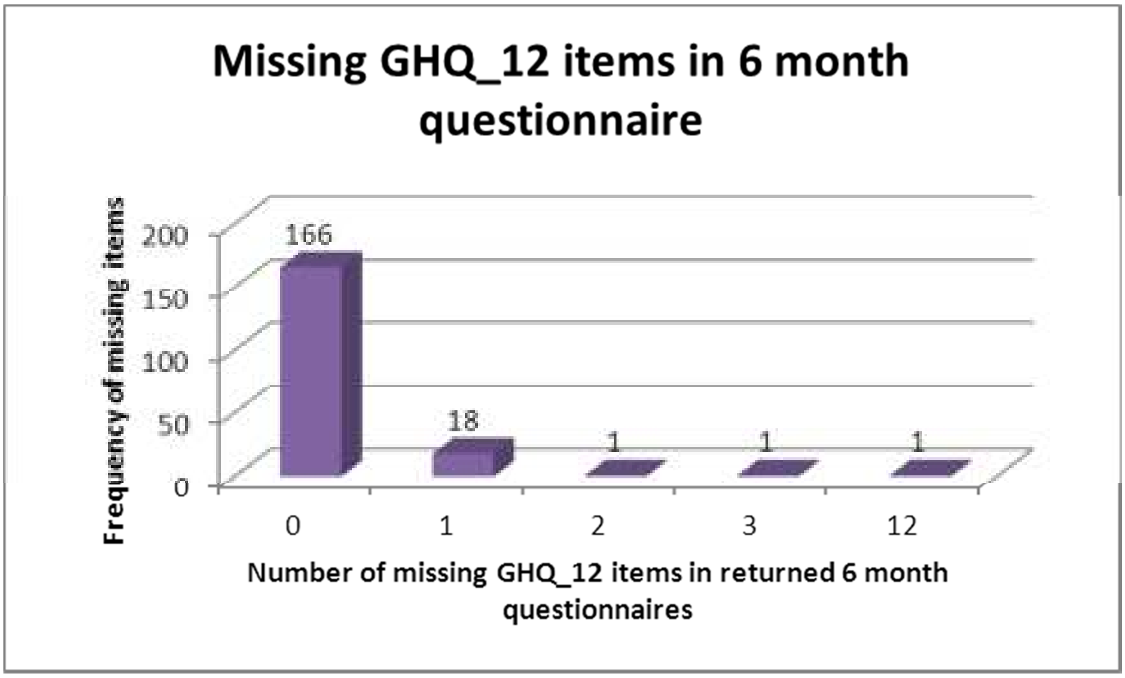


Table 22 Key for Table 23 (six month GHQ_12)

Key	Have you recently…
1	been able to concentrate on whatever you're doing
2	lost much sleep over worry?
3	felt that you are playing a useful part in things?
4	Felt capable of making decisions about things?
5	felt constantly under strain?
6	felt you couldn't overcome your difficulties?
7	been able to enjoy your normal day-to-day activities?
8	been able to face up to your problems?
9	been feeling unhappy and depressed?
10	been losing confidence in yourself?
11	been thinking of yourself as a worthless person?
12	been feeling reasonably happy, all things considered?

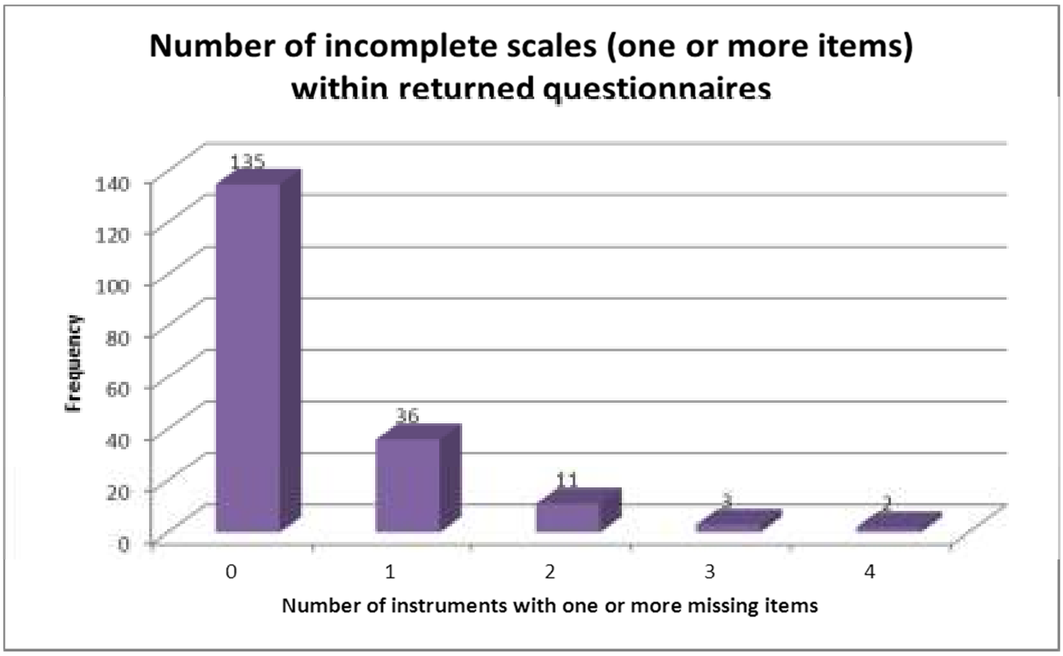
Table 23 Missing value patterns in six month GHQ_12

		Question number											
		1	2	3	4	5	6	7	8	9	10	11	12
Number of patients with missing item	5	0	0	0	0	0	0	0	1	0	0	0	0
	3	0	0	0	0	1	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	1	0	0	0	0	0	0
	2	0	0	0	0	0	0	1	0	0	0	0	0
	1	1	0	0	0	1	0	0	1	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0	0	0
	1	0	0	1	1	0	0	0	0	0	0	0	0
	1	0	0	0	1	0	0	0	0	0	0	0	0
	166	0	0	0	0	0	0	0	0	0	0	0	0
	Total	3	1	3	2	4	2	2	5	0	0	0	0

The 'do you feel able to face problems' question was the question omitted by five of the 18 respondents who only omitted one question from the GHQ-12.

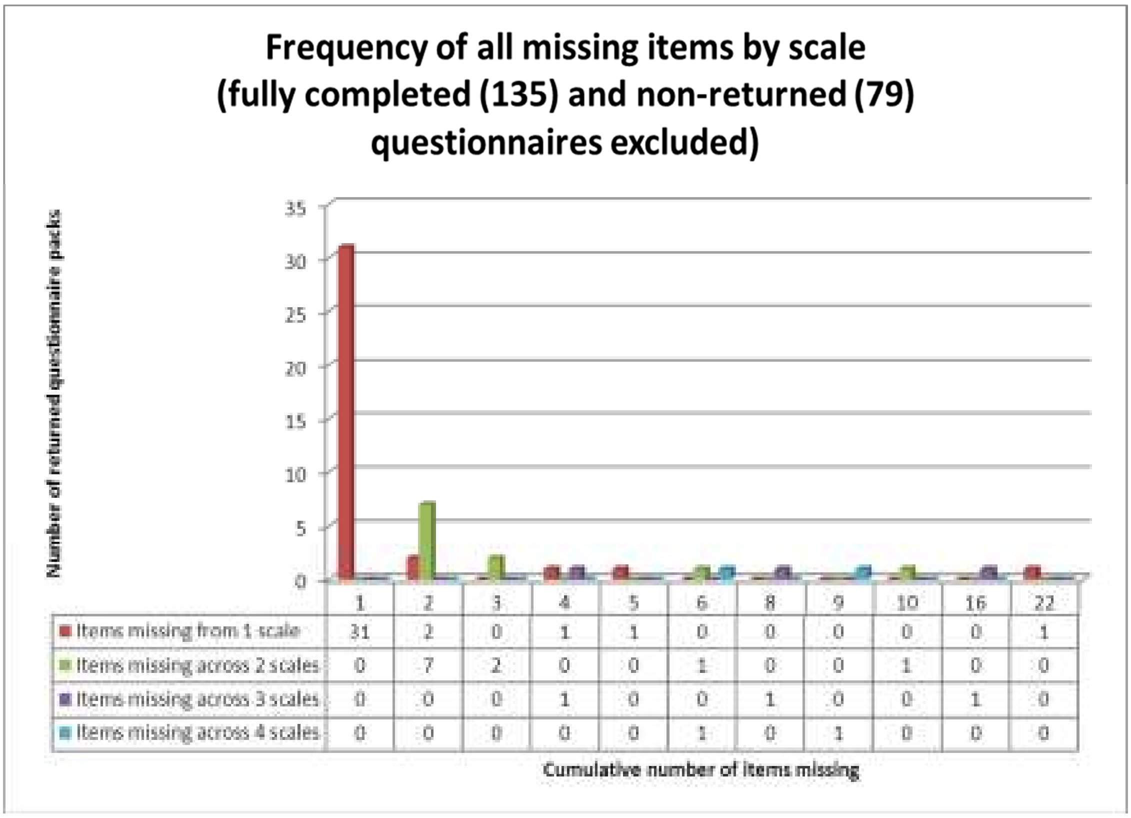
One hundred and thirty five patients completed all the questionnaires in their entirety. The numbers of incomplete questionnaires are presented in Figure 17 .

Figure 17	Number of incomplete scales within returned questionnaires (excluding OHS)



The majority of missing outcomes data is due to one item missing from 1 scale (31 cases, Figure 18). Data were most frequently missing from the NEADL and GHQ_12 scales.

Figure 18	Frequency of all missing items by scale



	1	2	3	4	5	6	8	9	10	16	22
🟥 Items missing from 1 scale	31	2	0	1	1	0	0	0	0	0	1
🟩 Items missing across 2 scales	0	7	2	0	0	1	0	0	1	0	0
🟪 Items missing across 3 scales	0	0	0	1	0	0	1	0	0	1	0
🟦 Items missing across 4 scales	0	0	0	0	0	1	0	1	0	0	0

5.2.4.3 Management of individual missing outcomes items

As with the baseline GHQ_12 (see section 5.2.3.3) the most pessimistic score (0) has been entered for patients with missing six month GHQ_12 data. For the most part, imputation of outcomes is not advisable and records with missing data will therefore be excluded from the analysis. The calculation of EQ5D scores and SIPSO subscales scores excludes records with missing data.

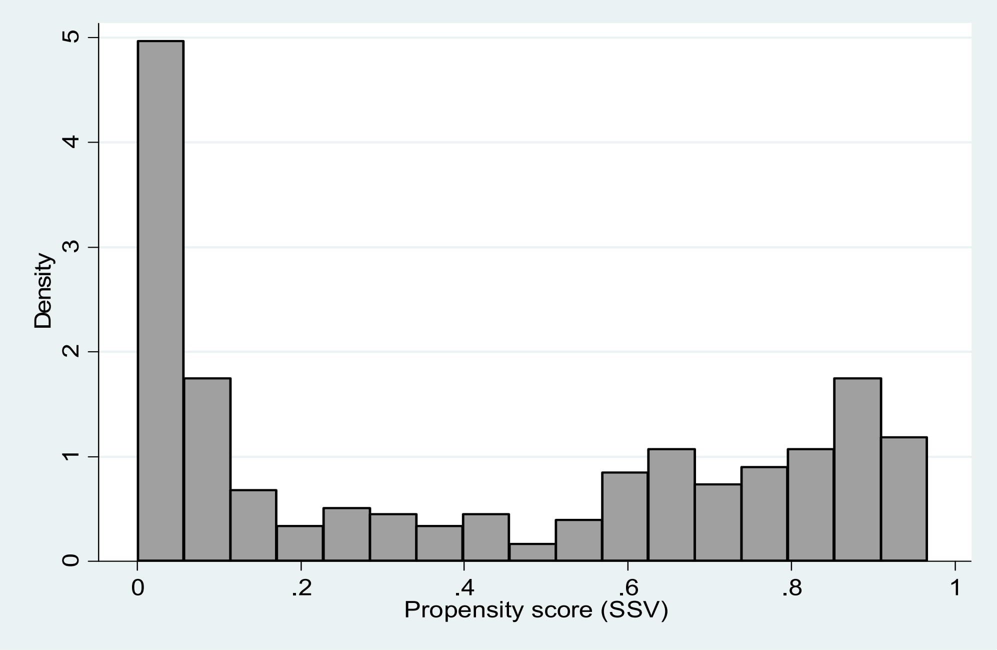
5.3 Distributions of continuous and ordinal variables

Examination of the distributions of continuous and ordinal variables can identify both significant floor and ceiling effects that may limit the sensitivity of the variable at the extremes of measurement, and unusual patterns that may require further exploration. If these are to be used as dependent variables, their distributions may affect the type of regression analysis that may be performed, or may cause problems with meeting linearity assumptions (see section 5.4, and also sections 4.4.6.1.4 & 3.5.7.1).

5.3.1 Distribution of the propensity score

The propensity score calculated from the SSV case-mix adjuster demonstrates a marked floor effect (prediction of poor outcome) with 111/312(35.6%) of patients in the lowest decile of predicted outcome. However, if the propensity score is entered into a model as an independent variable, providing the residuals from that model are normally distributed and that linearity assumptions are met, then this deviation from normality may be ignored. Failure to meet linearity assumptions may require transformation or categorisation of the variable.

Figure 19 Distribution of propensity score in the CIMSS study population



5.3.2 Length of stay

Length of stay demonstrates marked positive skew (see Figure 8, section 5.1.1)**.** It is likely, therefore that this variable will need to be transformed in order to linearize the relationship between length of stay and outcome and this will be explored further during model construction.

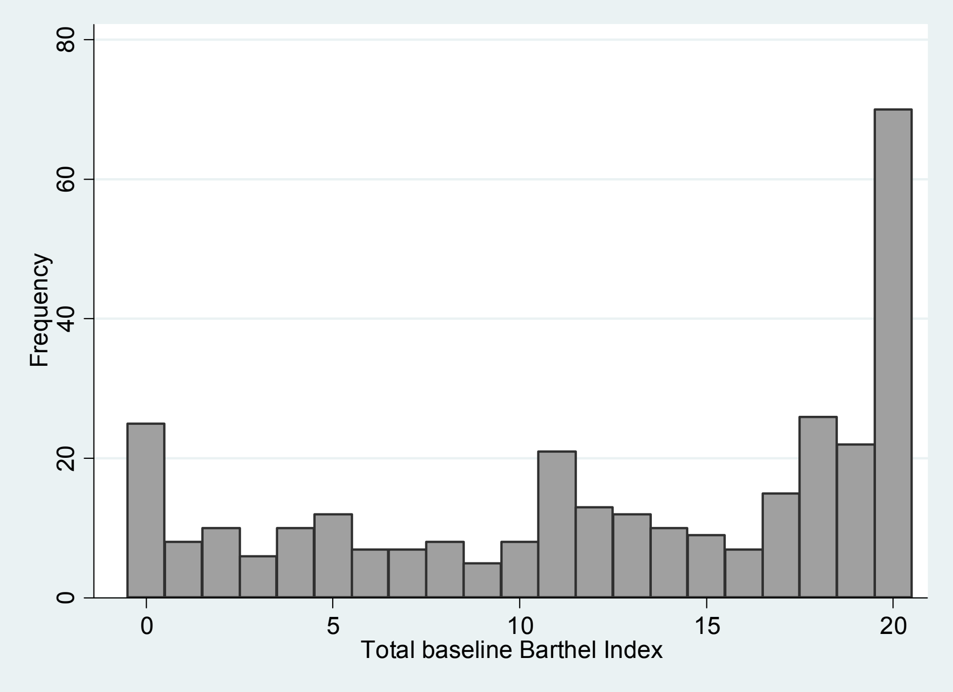
5.3.3 Distributions of baseline assessments

The presence of significant floor or ceiling effects (10% of patients scoring at extremes of the scale) suggests a lack of responsiveness of the instrument to detect change at the extremes of measurement. Large floor and ceiling effects in baseline functional assessments may reflect the immediacy of the assessment following the stroke event (i.e. reflection of immediate post-stroke disability) or the recruitment of patients that tended to have milder strokes.

Table 24 Floor and ceiling effects of baseline functional assessments

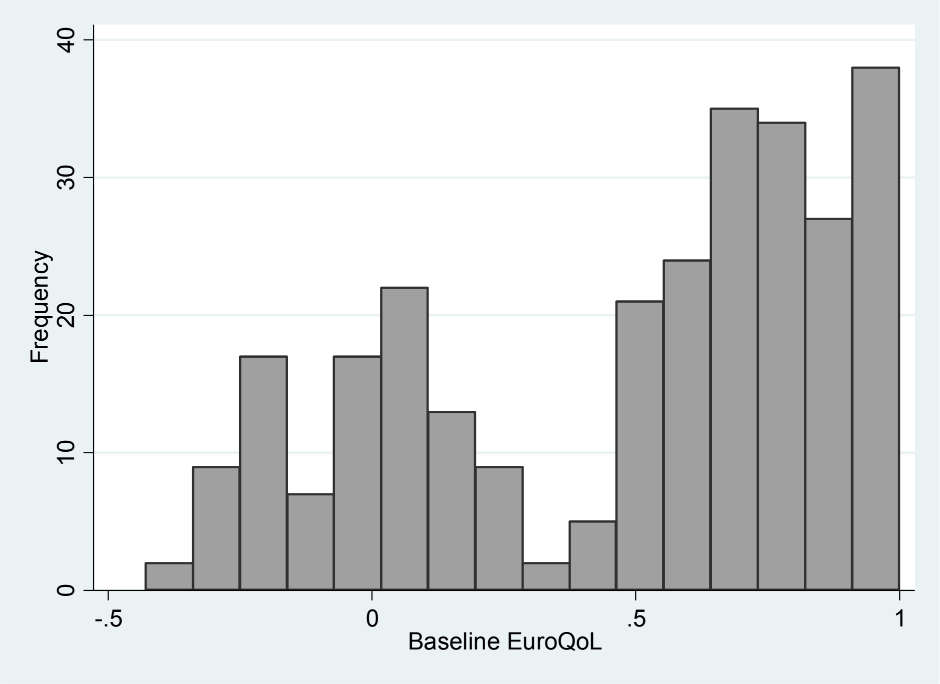
	Floor effect	Ceiling effect
Barthel Index	25/311 = 3.2%	70/311 = 23%
NEADL_baseline	14/302 = 4.6%	32/302=10.6%
EuroQoL_baseline	0/302 = 0%	38/302 = 12.6%
GHQ_12 baseline	125/302 = 41.4%	12/302 = 4.0%

Figure 20 Baseline Barthel Index



5.3.3.1 Distribution of the baseline EQ5D

Figure 21 Baseline EQ5D



The baseline EQ5D appears to be biphasic, with peaks around zero and 0.7. There is also a ceiling effect of 12.6%. This may reflect a correlation with baseline function, as three of the questions within the EuroQoL questionnaire relate directly to constructs measured with the Barthel Index (mobility, self-care and usual activities), and may therefore be acting as a proxy marker for the floor and ceiling effects seen with this instrument at baseline.

5.3.3.2 Distribution of the baseline NEADL

The NEADL score is filled out according to an individual's function over the last week. For some patients (in whom the time from stroke to recruitment was more than a week), this period will include the period since the onset of the stroke. For others, where time to recruitment was short, the NEADL is more likely to reflect pre-stroke function. The instructions for completion of the NEADL in the context of the study were therefore unclear. In addition to the factors that contribute to floor and ceiling effects of the Barthel Index, this timing of completion may account for the marked floor and ceiling effects of the baseline NEADL. The distribution of time from stroke (as reported by patients or their carers) to completion of baseline questionnaires are shown in Figure 22 (7 days to recruitment is marked with the black line). The lack of clarity as regards the instruction as to the period with which the NEADL should be completed with respect to may also account for the wide distribution of the baseline NEADL Figure 23.

Figure 22 Time (days) from stroke to completion of baseline questionnaires

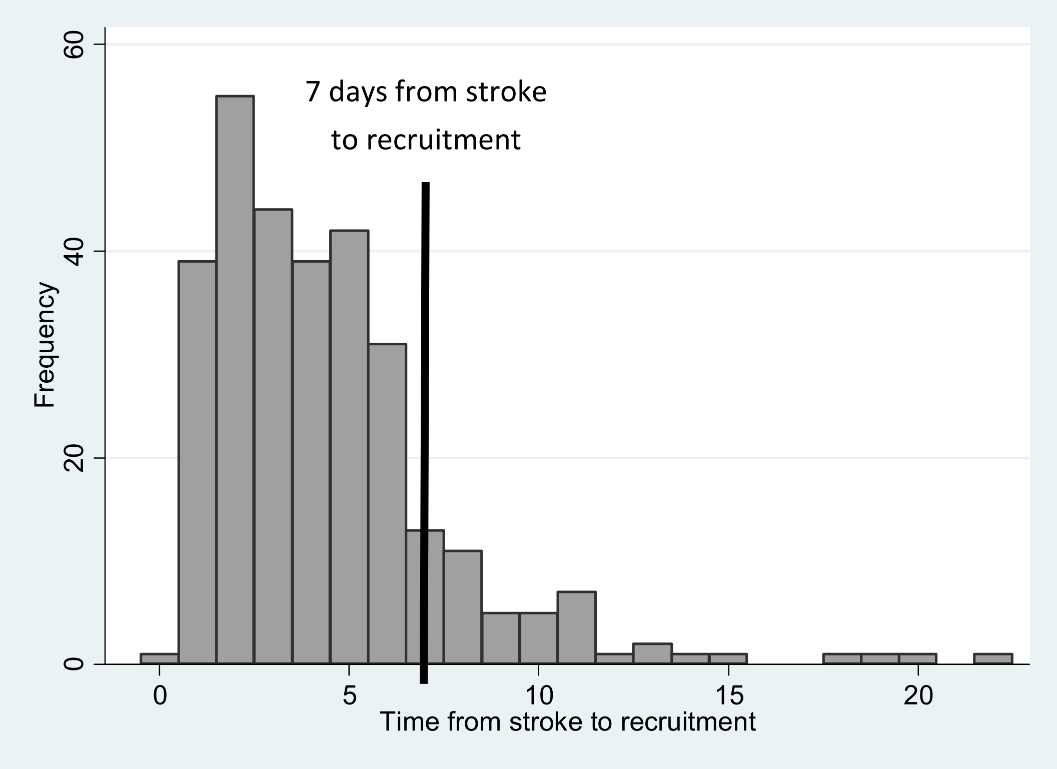
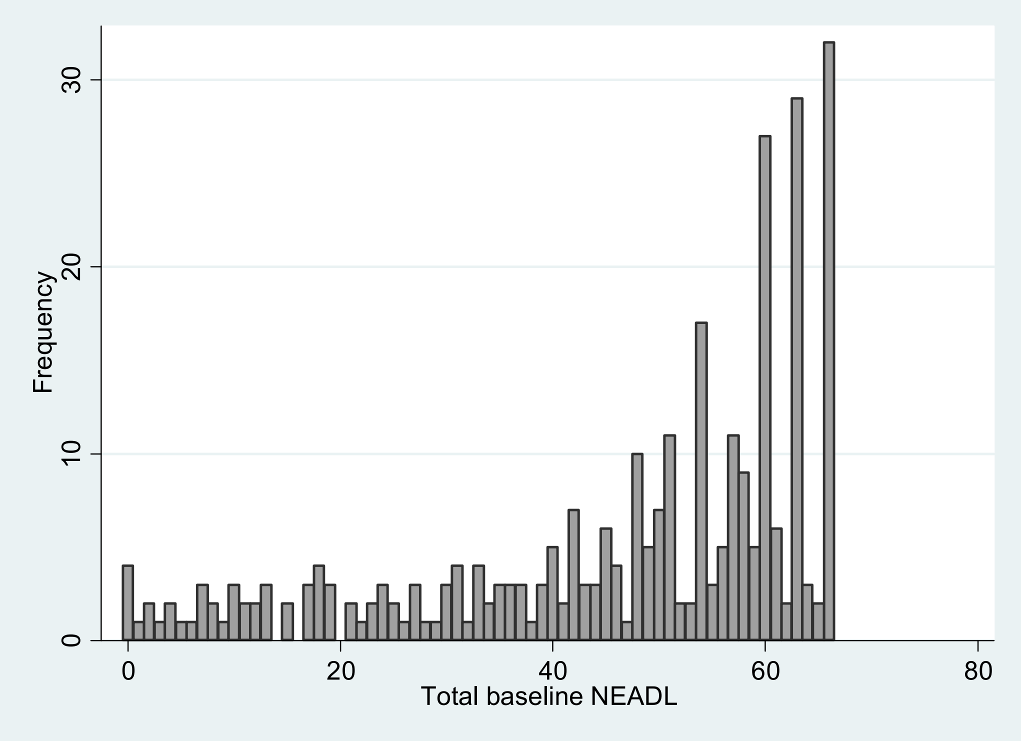


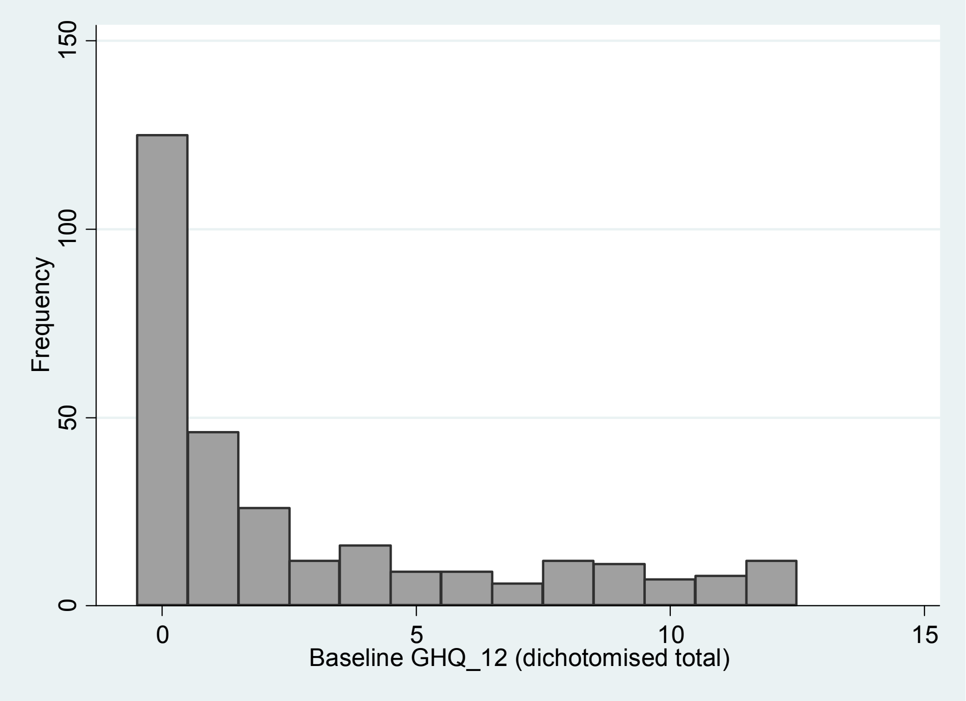
Figure 23 Distribution of responses on baseline NEADL



5.3.3.3 Distribution of the baseline GHQ-12

Marked floor effects are seen with the baseline GHQ_12 (46%). This may reflect understandable anxiety or low mood in the immediate post-stroke period (Figure 24).

Figure 24 Distribution of responses on the baseline GHQ_12



5.3.4 Floor and ceiling effects of outcomes instruments at six months

Table 25 Floor and ceiling effects in returned questionnaires

Outcomes instrument (fully completed six month questionnaires)	Floor effect	Ceiling effect
NEADL total	1/165 = 0.6%	16/165 = 9.7%
NEADL_mobility subscale	13/177 = 7.3%	57/177 = 32.2%
NEADL_kitchen subscale	9/186 = 4.8%	110/186 = 59.1%
NEADL_domestic subscale	28/178 = 15.7%	42/178 = 23.6%
NEADL_leisure subscale	8/178 = 4.5%	32/178 = 18.0%
SIPSO (physical)	10/176 = 5.7%	41/176 = 23.2%
SIPSO (social)	2/174 = 1.1%	19/174 =10.1%
EQ5D	2/179 = 1.1%	35/179 = 19.6%
GHQ_12	53/166 = 31.2%	13/166 = 7.8%

Ceiling effects are seen with the physical subscale of the SIPSO and the EQ5D. Marked floor effects are seen with the GHQ_12. The EQ5D and the SIPSO subscales only contain 5 items and are therefore more prone to ceiling and floor effects Consideration of the two SIPSO subscales together to give a total score eliminates the ceiling effect seen within the physical subscale (ceiling effect of total SIPSO score 41/176 = 9.0%). The subscores of the NEADL have significant ceiling effects (most pronounced in the kitchen subscore). In addition there are marked floor effects in the domestic subscale. This is likely to limit the use of the NEADL in the CIMSS population. These effects have not been noted in previous studies using the instrument, although floor effects in the mobility domain have previously been noted in more dependent patients (Gladman et al 1993; Gladman et al 1994).

5.4 Distributions of patient reported measures (six month questionnaires)

5.4.1 Conversion to continuous variables

Two of the patient reported instruments have scoring systems or conversion tables that allow them to be converted into interval level variables (the EuroQoL and the SIPSO). The EuroQoL is converted into a continuous utility score between -1 (lowest) and 1 (highest quality of life) (Rabin R et al, 2011). The SIPSO has been subject to Rasch and Mokken analyses which confirmed a two factor structure with two subscores which may be considered separately, but where the summed (total) score fails to meet scaling assumptions (Kersten P et al 2010). The two SIPSO subscales are therefore considered as two separate interval level outcomes. Limitations of this approach are discussed in sections 4.4.4 and 6.3.1.

Four of the patient reported measures are ordinal (the baseline and six month NEADL and GHQ_12, baseline Barthel Index and six month OHS). They may be treated as continuous data if linearity and normality of residuals assumptions are met, and models using these outcomes will be constructed first using parametric techniques (linear regression). Where linearity and normality of residuals assumptions are not met, the outcomes will be dichotomised and treated as non-parametric data with the caveat that this is with the loss of information.

5.4.2 Continuous patient reported measures

5.4.2.1 EuroQoL 6 months

The six month utility score for the EuroQoL deviates significantly from a normal distribution (Figure 25), and looks to follow a censored distribution. This cannot be normalised through transformation (Figure 26), and if the EQ5D were to be used as the dependent variable, a Tobit regression model (to account for the censored data) may be appropriate.

Figure 25 Distribution of the six month EQ5D

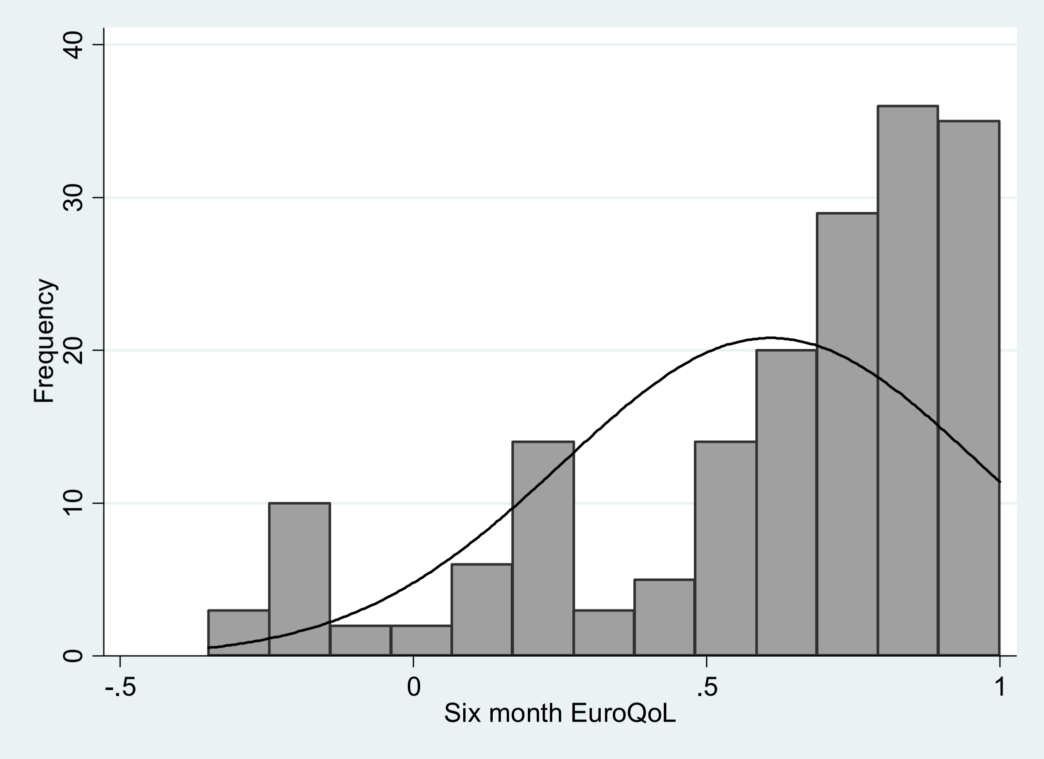
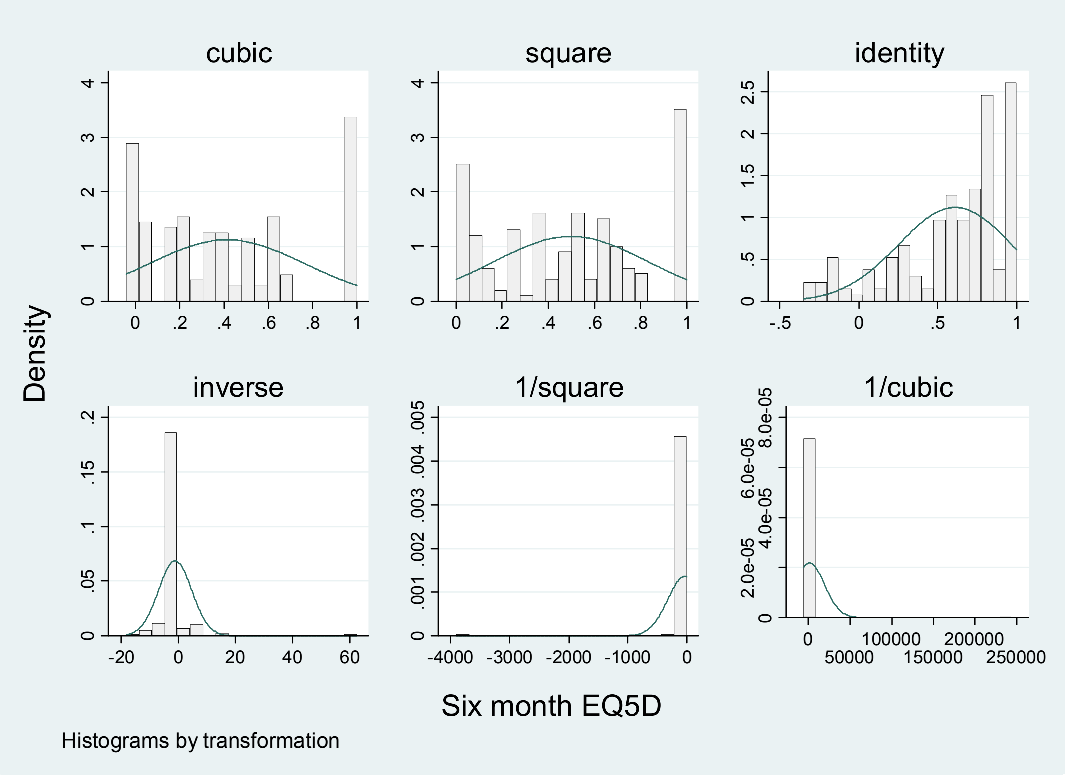


Figure 26 Transformations of the six-month EQ5D



5.4.2.2 SIPSO subscores at six months

A normal (Q-Q) plot deviates from the reference line; thereby suggesting that the physical subscore of the SIPSO is not normally distributed. The untransformed physical subscore of the SIPSO is highly statistically significant (p<0.01) on Shapiro-Wilk testing confirming that the variable cannot be assumed to be normally distributed. However, as no transformation appears to improve the distribution towards normality (Figure 28), the untransformed physical subscore of the SIPSO will be entered into the models. Linearity and normality of residuals assumptions must, however, still be met on testing of model diagnostics.

Figure 27 Normal (Q-Q) plot for untransformed six month SIPSO physical subscore

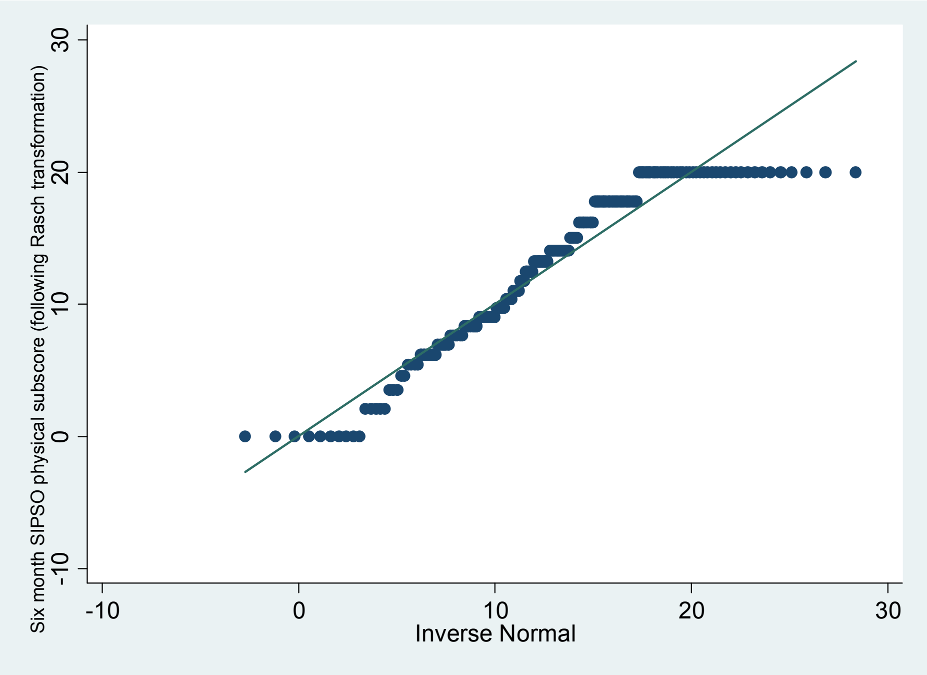
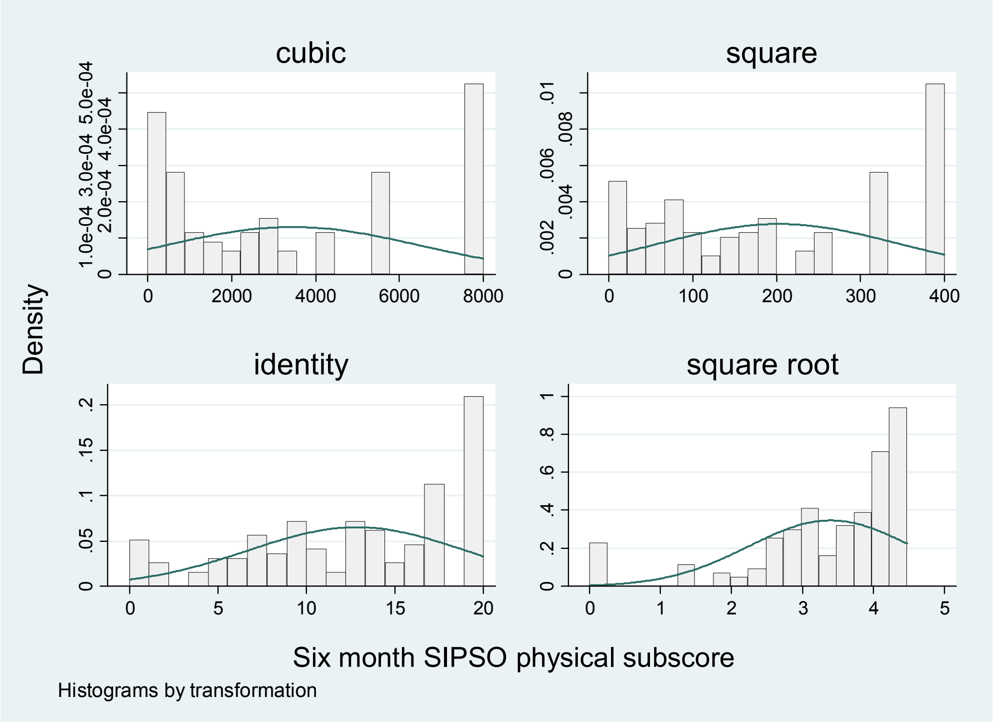
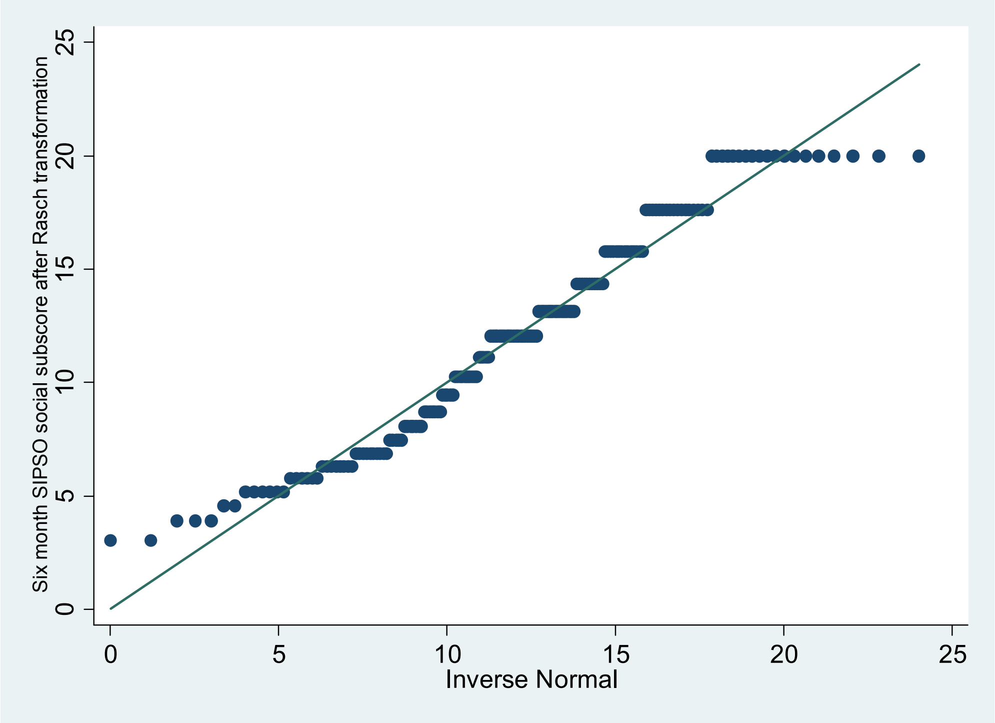


Figure 28 Transformations of the Rasch transformed six-month SIPSO physical subscale



A normal probability plot suggests that the social subscore of the SIPSO follows approximately a normal distribution, although there is some deviation at the tails. A Shapiro-Wilk test fails to reach significance (p=0.095), suggesting that a normal distribution can be assumed.

Figure 29 Normal (Q-Q) plot of SIPSO social subscore

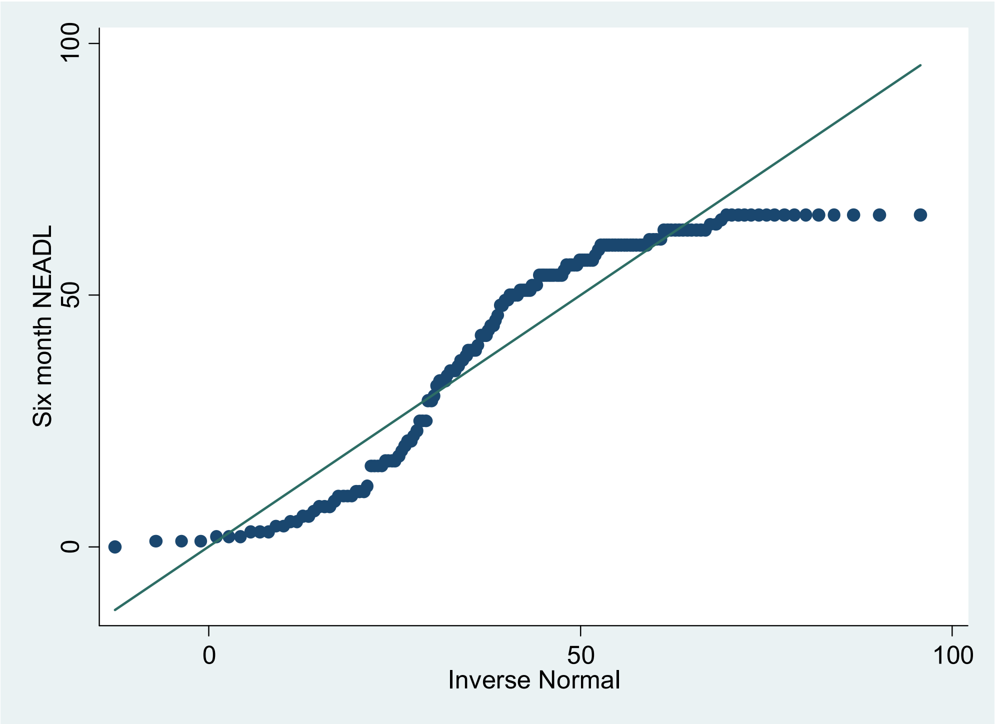


5.4.3 Ordinal patient reported measures

5.4.3.1 NEADL 6 months

The six month total NEADL score deviates markedly from a normal distribution as suggested by the normal plot (Figure 30) and confirmed by the highly significant Shapiro-Wilk statistic (p<0.001) . If the NEADL is to be used as an outcome in a regression model and residuals are not normally distributed the scale would need to be categorised and a multivariate ordinal regression, or logistic regression performed depending on the number of categories created.

Figure 30 Q-Q plot for the six month NEADL (total summed scores across subscales)



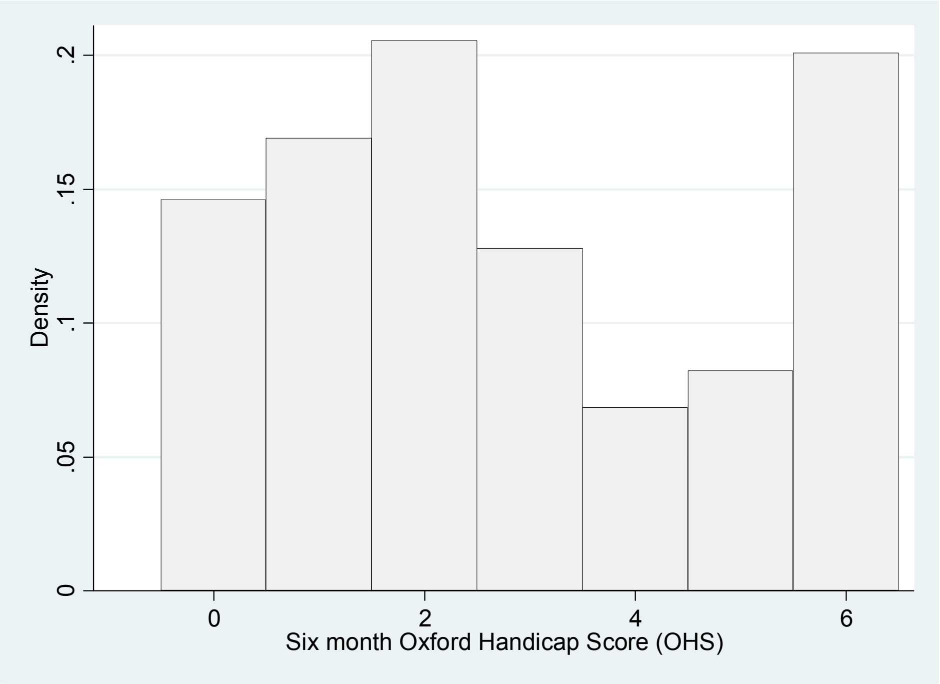
5.4.3.2 GHQ_12 baseline

The six month GHQ_12 can also not be considered to be normally distributed as there is marked deviation from the reference line on a Q-Q plot, and the Shapiro-Wilk statistic is again, highly statistically significant. For ease, if, the GHQ-12 is to be treated as a dependent outcome variable it will be dichotomised into ‘case’ and ‘non case’ and treated as a categorical variable in a logistic regression model (with a score of >=3 out of a total of 12 signifying a ‘case’ when using the dichotomised scoring system as described by the authors (Goldberg D et al, 1988)). This categorisation is not necessary for the creation of decision trees (with the baseline GHQ-12 being entered as a predictor, or the six month GHQ-12 as an outcome), as no assumptions are made regarding the underlying distribution of the data (see section 4.4.5.2).

5.4.3.3 Oxford Handicap Scale (six months)

The Oxford Handicap Scale (OHS) is an ordinal outcome, and will be dichotomised into good/poor outcome using the same cut of at <=2 as good outcome, 3 to 6 as poor outcome (which includes 6, dead) such that these classifications match those used in the development of the SSV case-mix adjuster (Counsell C et al 2002).

Figure 31 Distribution of responses on the postal modified Rankin Score



5.5 Choice of outcome measures for the CIMSS study

The GHQ-12 questionnaire was included in the outcomes questionnaire in order to examine its test-retest reliability when collected by post in stroke populations. It has not, therefore been examined as a primary outcome of the study for the purposes of this thesis. The postal version of the OHS was collected in order to calculate the SSV predicted probability of good outcome for the purposes of case-mix adjustment. The postal OHS will be used to ascertain if there are any univariable predictors which may perform as well as the SSV in terms of predicting patient six month OHS scores, and also to ascertain if there are any additional predictors which, when added to the SSV model, improve its utility in outcome prediction.

The distributions of the outcomes measures have important implications on the types of analyses that may be performed and the types of conclusions that may be drawn. The NEADL and SIPSO both measure aspects of post stroke function and activities of daily living. However, the property of the SIPSO that allows it to be treated as interval level variable offers significant advantages over the NEADL. Firstly, it is likely that it may be predicted through linear regression modelling whilst the NEADL is more likely to require categorisation and multivariable ordinal modelling. Linear modelling increases the number of predictor variables that may be entered into models, as the EPV reflects the number of observations and not the number of outcome events. Secondly, the SIPSO is less prone to the marked subscore ceiling and floor effects that are apparent with the NEADL and may, therefore, be a more sensitive instrument in patients nearing the extremes of the scale. Finally, and possibly due to its relative brevity as compared with the NEADL, the individual subscores of the SIPSO are well completed. However, the SIPSO does have drawbacks. It is less well externally validated than the NEADL, and the validation studies that have been performed are either by the authors of the instrument (Trigg R et al 2003), or have been performed on samples not necessarily transferrable to the wider stroke population (i.e. a population of younger stroke survivors (Kersten et al 2004)). This may have implications for the validity of the scale in reflecting the latent trait of reintegration in older stroke survivors (see sections 4.4.4 & 6.3.1) Finally, and importantly, the SIPSO is completed after the stroke event, and therefore does not allow direct comparisons with individual baseline function. The advantages of the SIPSO however, outweigh these drawbacks and it has been used at the primary outcome throughout the study analysis. Each subscore of the SIPSO is considered separately as an individual score (physical and social domains). This approach was encouraged in a recent study examining the scaling properties of the SIPSO (Kersten P et al 2010).

5.6 Descriptive statistics of study population and representativeness of sample

Descriptive statistics for the patients recruited into the study (study population) are compared with the stroke population screened at each site (as a marker of the wider stroke population) as a measure of sample representativeness. Descriptive statistics regarding process markers are provided for the study population. The characteristics of patients that respond to the six month questionnaire as compared with those that do not are also described to identify any systematic differences between those that did and did not respond to the outcomes questionnaires.

5.6.1 Barriers to data collection across study sites

Regular teleconferences were held with researchers in each site to identify and, where possible, resolve difficulties with patient recruitment and data collection. There were some barriers to recruitment that were common across the study sites. Identification of patients with stroke that were not admitted to the stroke unit was problematic, though close liaison with the stroke care co-ordinators at each site helped with both patient identification and tracking patients that had moved wards. It was not possible to include patients admitted and discharged during the course of a weekend as the researchers’ working week was Monday to Friday. The omission of these patients may have introduced bias to the study sample. Often patients and their carers were unwilling to discuss participation in a study soon after the stroke event and researchers expressed difficulties in identifying, recruiting and collecting case-mix data within a week of admission following stroke. Consequently, providing that case-mix data could be extracted retrospectively from case notes with respect to the week following admission, patients were recruited up to two weeks following admission.

It was found that carers were particularly difficult to recruit to the study, often because they visited the ward outside normal working hours. In York, the researcher liaised with the ward sister to allow carers to visit the ward at other times such that they could be approached for consent to participate in the study. In Leeds, the researcher worked flexibly and visited the ward during the early evening to obtain consent from carers.

5.6.2 Screening and recruitment

In total 656 patients were screened across the three study sites during the six month recruitment period. Initially 320 patients were consented, but 8 were not recruited as they were found not to meet eligibility criteria (one did not have information available regarding case-mix variables from the week following the stroke, one deteriorated and was receiving palliative care, one died prior to recruitment and four were found to have subarachnoid haemorrhage on imaging).

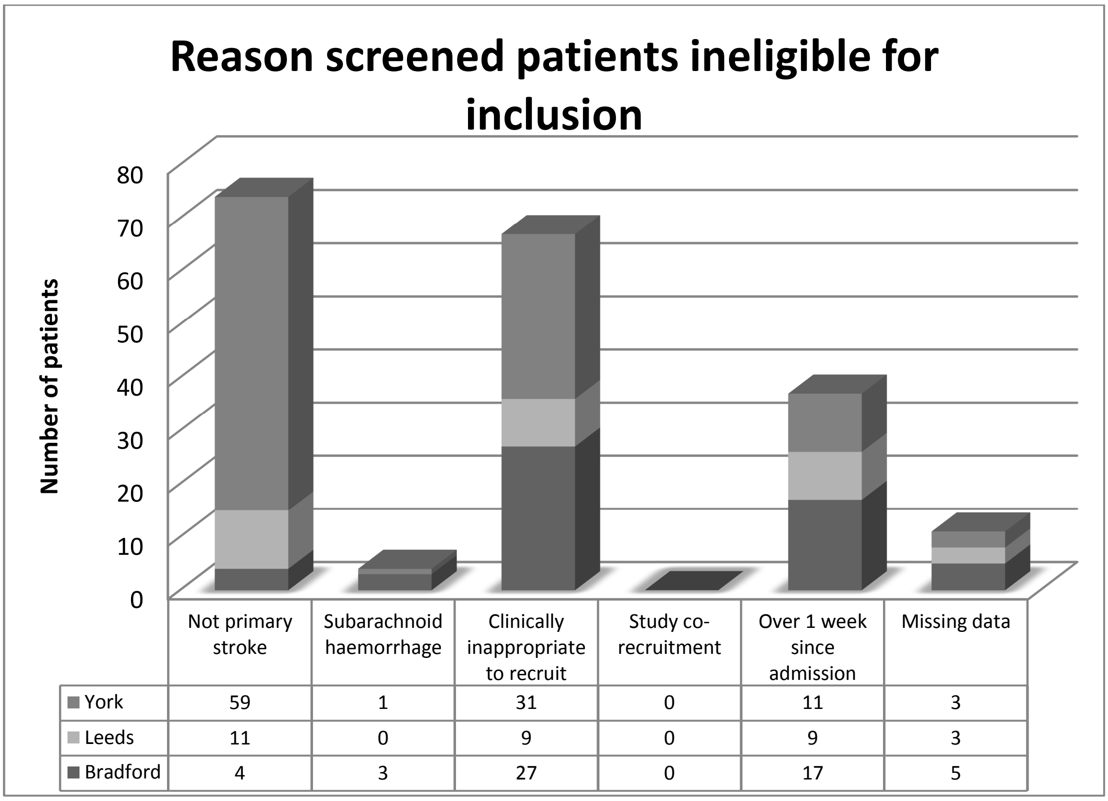
Table 26 outlines the absolute numbers and proportions of patients screened and recruited at each of the study sites. It can be seen that higher proportion of screened patients were recruited in Leeds than in both Bradford and York. Reasons why patients that were screened were not subsequently recruited (either through a failure to meet eligibility criteria, or through informed consent not being obtained) are shown in Figure 32 & Figure 33.

Table 26 Proportion of screened patients recruited into study by site

Site	Number screened	Number eligible patients recruited	Proportion recruited
Bradford	176	71	40%
Leeds	193	125	65%
York	287	116	40%
Total	**656**	**312**	

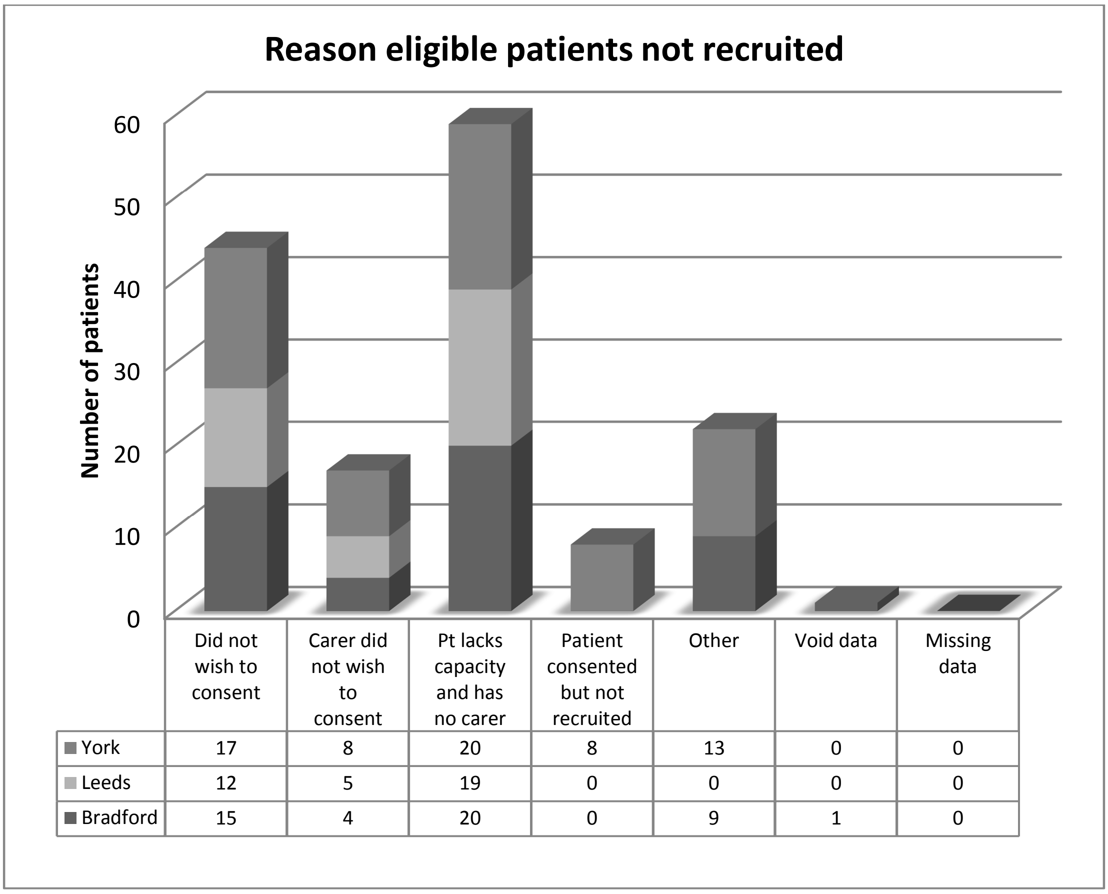
The most common reason for patients being ineligible for inclusion was that the diagnosis was not a primary stroke. The proportion of patients in whom this was the case is surprisingly high (74/193 = 38%). This may reflect the number of patients admitted with “query stroke” and commenced on the stroke care pathway prior to specialist assessment. The most common reason for eligible patients not being recruited was a lack of capacity and no available carer to provide consent to take part in the study.

Figure 32 Reasons screened patients not eligible for recruitment



	Not primary stroke	Subarachnoid haemorrhage	Clinically inappropriate to recruit	Study co-recruitment	Over 1 week since admission	Missing data
■ York	59	1	31	0	11	3
■ Leeds	11	0	9	0	9	3
■ Bradford	4	3	27	0	17	5

Figure 33 Reasons eligible patients not recruited into study



	Did not wish to consent	Carer did not wish to consent	Pt lacks capacity and has no carer	Patient consented but not recruited	Other	Void data	Missing data
■ York	17	8	20	8	13	0	0
■ Leeds	12	5	19	0	0	0	0
■ Bradford	15	4	20	0	9	1	0

5.6.3 Demographic data

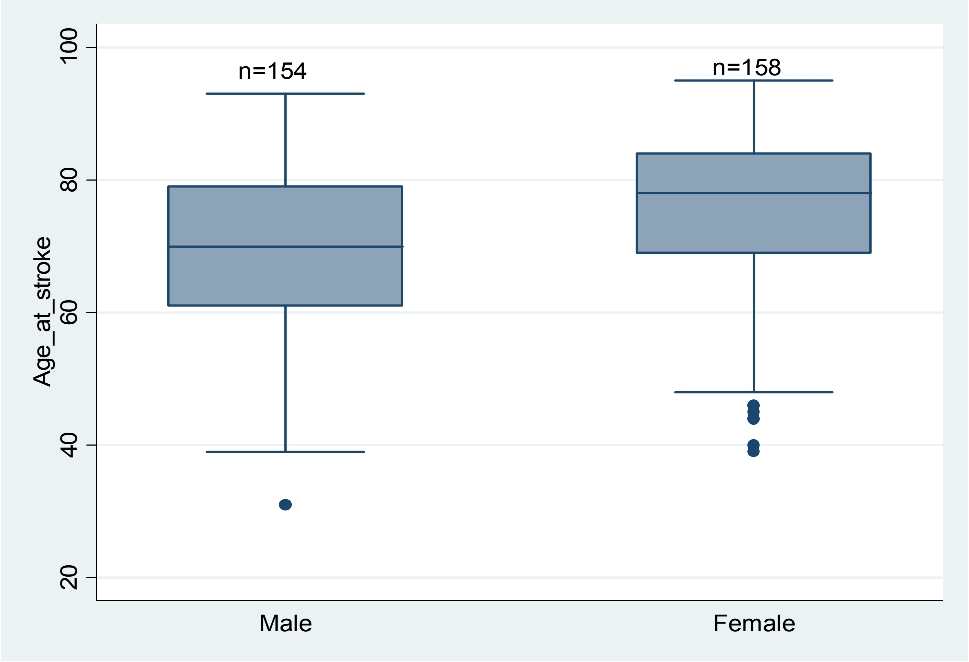
5.6.3.1 Differences in age at stroke and gender between screened and study population

The median age of patients recruited into the study was 74 (IQR 65-82), with a range of 31 to 95 years. Fifty-one percent of the study population were female.

An equivalence of proportions test reveals no difference in the proportion of women between screening and study populations (p=0.165 working shown in appendix D-1).

The distribution of age by sex in patients recruited into the study is shown in the boxplot (Figure 34)

Figure 34 Distribution of age by sex in study sample



This difference in age by sex is highly statistically significant (Mann-Whitney U test of equivalence of medians) and is likely to represent the longer life expectancy of women in the general population (working shown in appendix D-1.2).

Figure 35 reveals that patients who are not recruited into the study have a higher median age than those who are (the median is used as the data are negatively skewed). This difference is confirmed as statistically significant on a Mann Whitney U test (the non-parametric equivalent to a t-test – appendix D-1.3).

Figure 35 Distribution of age by recruitment into study

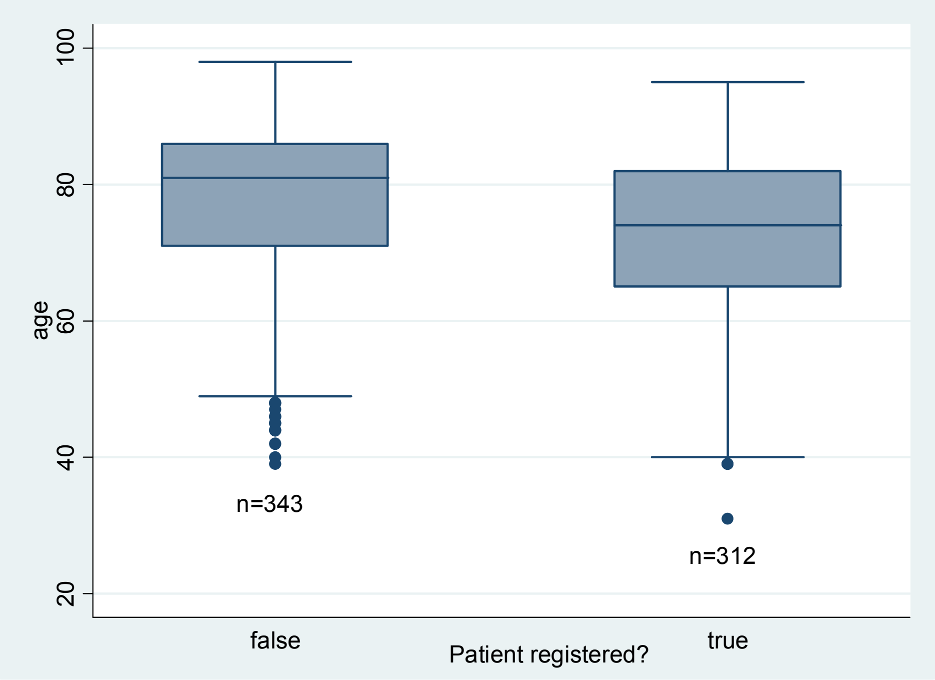


Table 27 Difference in median age between patients recruited and not recruited into study

	Range	Median	IQR
Recruited	31-95	74	65-82
Not-recruited	39-98	81	71-86

The difference in medians between patients recruited and not recruited is seen at each site (Table 28)

Table 28 Difference in median age between patients recruited and not recruited by site

Site	Median age recruited	Median age non recruited	Significance level for equivalence of medians (Mann Whitney U)
Bradford	72	79	0.012
Leeds	76	81.5	0.003
York	74	81	<0.001
Significance level for equality of population medians (Kruskall-Wallis[1] test)	**0.34**	**0.20**	

However, a difference is not seen in median age between sites for patients who are, and are not recruited into the study (Kruskall-Wallis tests for median age by site for patients recruited and not recruited into the study are not statistically significant).

[1] A Kruskall-Wallis test is the non-parametric equivalent to a oneway ANOVA

5.6.3.2 Ethnicity

The vast majority of patients recruited were white (see Table 29)

Table 29 Ethnicity of patients recruited and not recruited into the study

Ethnic Group	Screened not recruited (N (%))	Recruited (N (%))
White	311(90.4)	298(95.5)
Mixed White & Black Caribbean	2 (0.6)	0
Asian- Indian	3(0.9)	3(1.0)
Asian - Pakistani	11(3.2)	6(1.9)
Asian - Bangladeshi	2(0.6)	1(0.3)
Other Asian background	1(0.3)	1(0.3)
Black Caribbean	0(0)	3(1.0)
Chinese	2(0.6)	0
Missing	12(3.5)	0
Total	344	312(100)

There does not appear to be a difference in ethnicity between patients recruited and not recruited into the study (Table 29). Formal testing for association with a Fisher's exact test (to account for the low frequencies in some cells) confirms there is no association (p=0.31).

5.6.3.3 Availability of carer and living circumstances

The majority of patients recruited to the study cohabit (60%), 38% lived alone and 2% were admitted from nursing or residential care.

A two way measure of association (chi-squared test) showed that patients with carers available were no more likely to be recruited to the study than those without carers (p=0.06).

5.6.3.4 Baseline stroke severity

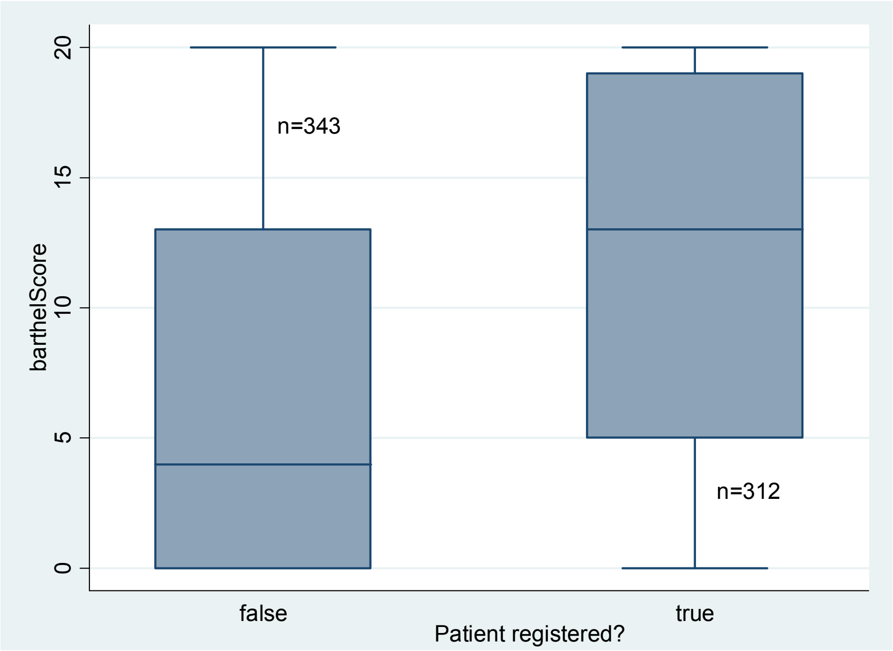
The descriptive statistics for the baseline Barthel Index in screened and recruited populations are shown in Table 30.

Table 30 Descriptive statistics for baseline Barthel Index

	Range	Median	IQR
Recruited	0-20	13	5-19
Screened	0-20	4	0-13

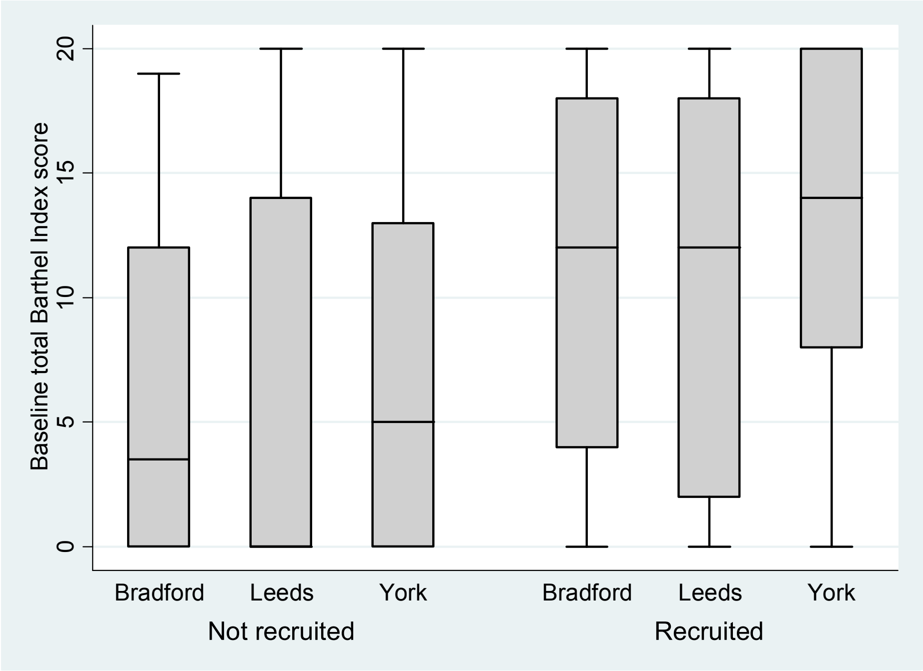
There is a marked difference in the Barthel Index scores between patients recruited into the study when compared with those that are not. An equivalence of medians test (Mann-Whitney U test), confirms a highly significant difference in baseline BI between patients recruited and not recruited into the study (p<0.001 – working shown in appendix D-2.1).

Figure 36 Difference in baseline Barthel Index between patients recruited and not recruited into the study



This difference in baseline disability is seen in all sites, and is most marked in Leeds. Patients with very severe strokes (i.e. those that were in receipt of, or likely to receive palliative care) were not recruited into the study and this is likely to have been reflected in this difference. The study sample is therefore more representative of a population that is more likely to survive to six month follow up rather than all strokes. For the purposes of the definition of a routine dataset with outcomes data being collected at six months, this should not be problematic.

Figure 37 Difference in baseline Barthel Index between patients recruited and not recruited into the study by site

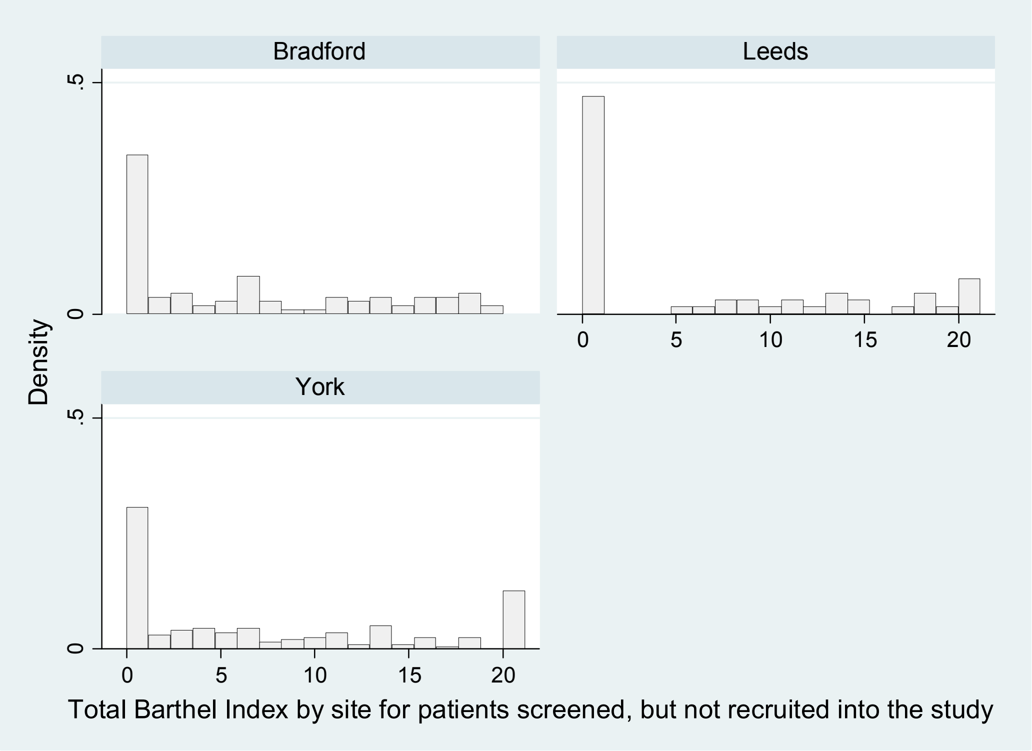


Kruskall-Wallis equivalence of medians tests for baseline Barthel for patients not recruited by site reveals no statistically significant difference in medians (appendix D-2.2). Inspection of the boxplot (Figure 37) would tend to suggest that the difference in medians between sites for patients not recruited into the study is marked, however inspection of the histograms by site reveals very large floor effects close to the median of 4 which may explain why the difference has not reached statistical significance (Figure 38). There is a statistically significant difference in median baseline Barthel Index across sites for patients recruited into the study, with York tending to recruit less disabled patients (median baseline 14 see Table 31). Two-way examination of medians in patients recruited into the study by site reveals this difference to be significant between Leeds and York (p=0.001) and of borderline statistical significance between Bradford and York (p=0.017) (appendix D-2.3). The difference in median baseline Barthel Index between patients recruited and not recruited is statistically significant at each site (Mann-Whitney U test (Table 31).

Table 31 Significant difference in median baseline Barthel Index between screened and recruited patients at each site

Site	Median BI recruited (N)	Median BI not recruited (N)	Sig level for equivalence of medians (Mann Whitney U)
Bradford	12 (71)	3.5 (175)	<0.001
Leeds	12 (125)	0 (68)	<0.001
York	14 (116)	5 (171)	<0.001
Sig level for equality of population medians (Kruskall-Wallis test)	**0.003**	**0.24**	

Figure 38 Distribution of baseline Barthel Index in patients screened but not recruited into study, by site



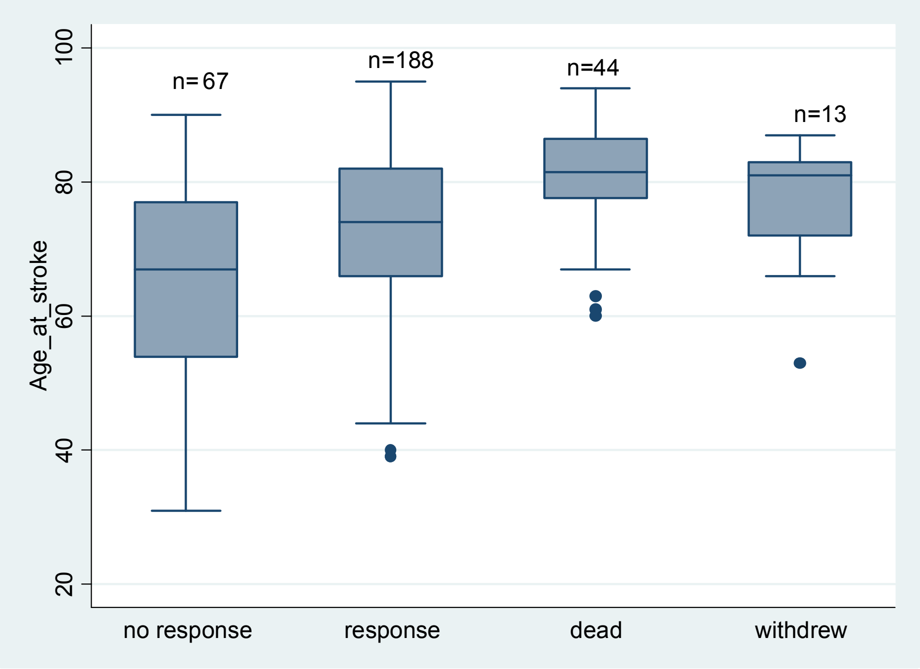
5.6.4 Admission data (patients recruited into study)

5.6.4.1 Differences in age at stroke and gender between responders and non-responders

A Kruskal-Wallis equivalence of medians test reveals a significant difference in age between groups that responded, died or withdrew from the study ($\chi2 = 38.4$, $\nu = 3$, $p<0.001$).

Two way examination of the difference in median age between these groups reveals that there are statistically significant differences in median age between patients that do not respond and each of those that respond, die or withdraw and also between patients who respond and those that die Figure 39.

Figure 39 Boxplot of age at stroke by response to six month questionnaire, death or withdrawal



5.6.5 Age and sex of patients who respond to six month questionnaires

Patients who responded to the questionnaire were significantly older (by 7 years) than those that did not respond (two sample t-test, p<0.001)

Table 32 Age of patients who responded to six month questionnaires

	Mean age (95% CI)
Response	72.9 (71.3-74.6)
No response	65.7 (62.2-69.1)

A Chi squared test confirms that there was no difference in the likelihood of questionnaires being returned from males or females (X^2 0.19 p=0.663).

5.6.6 Length of stay

The distribution of length of stay demonstrates marked positive skew. For this reason, the median as opposed to the mean has been used at the marker of central tendency. Median length of hospital stay is ten days (range 1-147). The length of stay varies markedly and significantly with study site (see Figure 10 in section 5.2.1, p 95).

Table 33 Length of stay by study site

	Median	Range	IQR
Bradford	13	1-85	5-46
Leeds	14	1-118	6-40
York	6	1-147	3-14.5

A Kruskall-Wallis equality of populations rank test is highly statistically significant p< 0.001, indicating that there is a significant difference in length of stay between at least two of the sites.

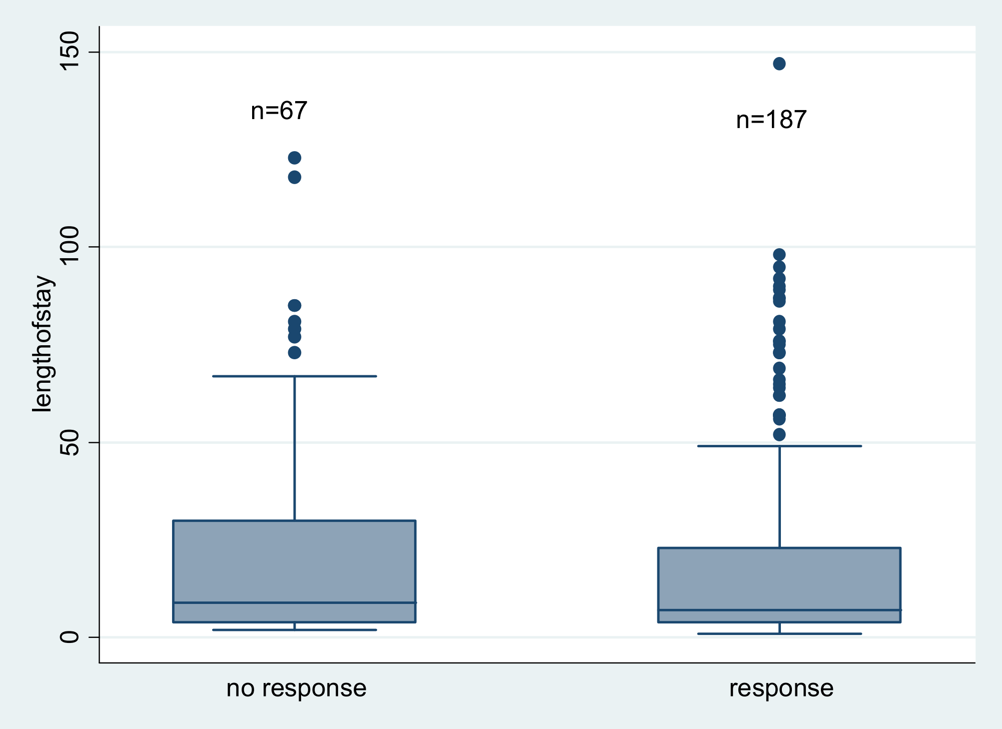
Pairwise examination of median length of stay (Mann-Whitney-U tests) between sites reveals the length of stay to be significantly shorter at York than the other two sites Table 34. This could reflect factors of organisational structure, but may also reflect the patients admitted to York had a higher baseline Barthel Index (i.e. were less disabled at baseline - see section 5.6.3.4.

Table 34 Pairwise comparison of length of stay across study sites

		Bradford	Leeds	York
	Median length of stay	13	14	6
Bradford	13			
Leeds	14	0.802		
York	6	0.001	<0.001	

There was no significant difference in length of stay for patients who returned six month questionnaire as compared with those that did not respond (Figure 40). This was confirmed on a Mann-Whitney U test (p=0.79).

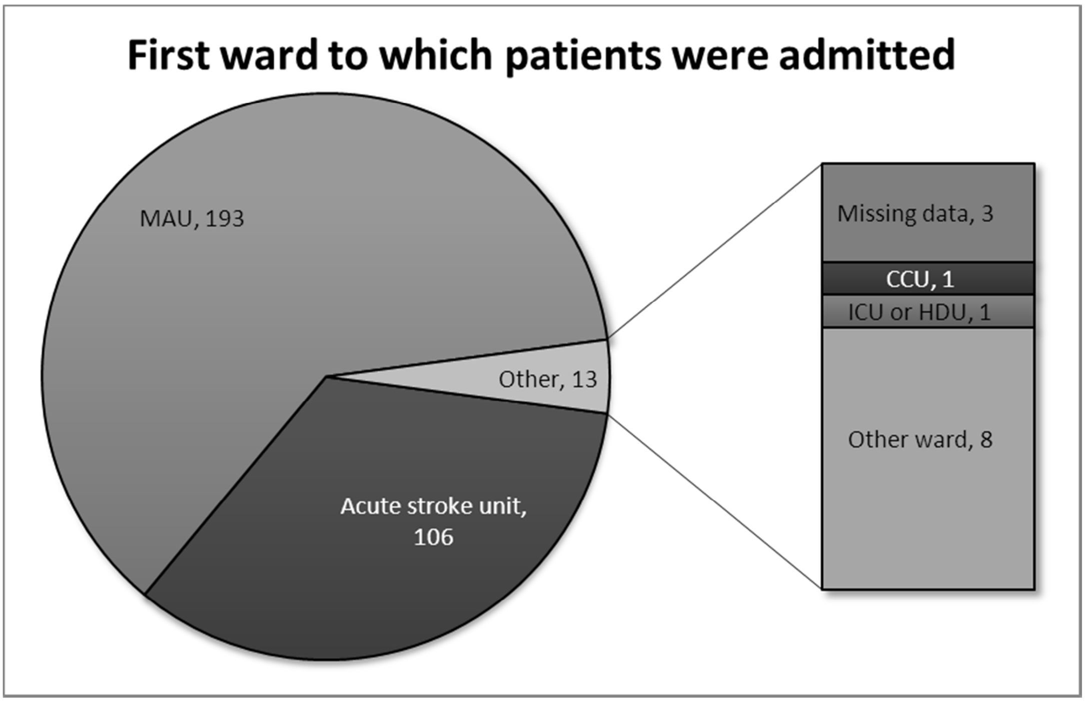
Figure 40 Length of stay in patients that did, and did not respond to the six month questionnaire



5.6.6.1 First ward to which patient was admitted

The majority of patients were admitted to a medical admissions unit, with just over a third (34%) being admitted onto a stroke unit, coronary care unit or intensive care/high dependency bed. Two hundred and thirty four of 291 patients with available data (80%) were admitted to a stroke unit on the same day, or day after presentation to hospital. Two hundred and ninety eight patients out of 310 patients with available data (96%) across the three sites spent at least some of their stay on an acute or rehabilitation stroke unit. Time of hospital admission has not been recorded with sufficient consistency to allow the time to stroke unit admission to be reliably calculated.

Figure 41 First ward to which patients were admitted



5.6.7 Stroke severity and case-mix variables

5.6.7.1 Stroke type

Two hundred and ninety five (94%) of the 312 patients enrolled in the study had a cerebral infarction, and 15 (5%) suffered haemorrhagic strokes. 48 patients (15%) presented with a recurrent stroke. Data on pathological stroke type was missing in one patient.

Clinical classification according the Oxford Community Stroke Project Classification of Stroke (Bamford J et al 1988) reveals just under a quarter of strokes to be total anterior circulation strokes (TACS) (23%), and a third partial anterior circulation strokes (PACS) (33%). Posterior circulation strokes (POCS) were least common at 14%, with the remainder (30%) lacunar strokes (LACS). A one way ANOVA (analysis of variance) reveals no significant difference in age (using a square transformation to normalise the data) between patients suffering different types of stroke.

Left sided weakness was more common (128/311) than right sided weakness (115/311). One patient had global weakness and 67 no weakness.

Table 35 shows other markers of stroke severity and their relative frequencies.

Table 35 Frequency of markers of stroke severity in the study population

Prognostic variable	Number (%)
Able to walk unaided at presentation	147/311 (47.2)
Dysphasia (speech or language deficit)	195/312 (62.5)
Confusion at presentation	58/310 (18.7)
New urinary incontinence	70/308 (22.7)

Chi-squared tests for association between response and markers of stroke severity revealed there to be no difference between responders and non-responders in the OCSP stroke subtype or in the presence of new urinary incontinence.

Patients who did not respond to the questionnaires were no more likely to have dysphasia than those that responded. Of the 69 patients who required assistance in completing questionnaires, 40 received help in recording their own responses, proxy answers were returned in 29. Twenty one of the 29 patients returning proxy responses had a speech or language disturbance at presentation (Table 36). A Chi-squared test of association between proxy response and dysphasia failed to reach significance, however, the severity of dysphasia was not recorded at baseline and it is possible that the patients for whom proxy responses were returned had more severe speech or language deficits.

Table 36 Association between dysphasia and proxy responses

		Proxy response	
		No	Yes
Dysphasia	No	109	8
	Yes	174	21

X^2 = 1.34 (p=0.25)

5.6.7.2 Differences in baseline Barthel Index between responders and non-responders

There is no significant difference in median baseline Barthel Scores between patients that respond to six month questionnaires and those that do not (Mann-Whitney U test). However, including deaths and withdrawals in this analysis to create four groups (no response, response, dead, withdrawn) revealed a highly significant difference between the groups.

Examination of pairwise combinations of these groups (using a Mann Whitney U test) reveals highly significant differences in baseline Barthel Index between patients who did not respond and patients who died; and patients who did respond and those that died (appendix D-2.4). There was, however, no significant difference in baseline Barthel Index between patients who withdrew and those who did not respond; died or did respond; or between responders and non-responders (see Table 37).

Table 37 Pairwise comparison of p values (Mann-Whitney U tests) for median baseline Barthel Indices (BI) by response

		Responder	Non-responder	Death	Withdrawal
	Median Baseline BI	17	13	1.5	12
Responder	17				
Non-responder	13	0.267			
Death	1.5	<0.001	<0.001		
Withdrawal	12	0.017	0.112	0.022	

5.6.7.3 Propensity score

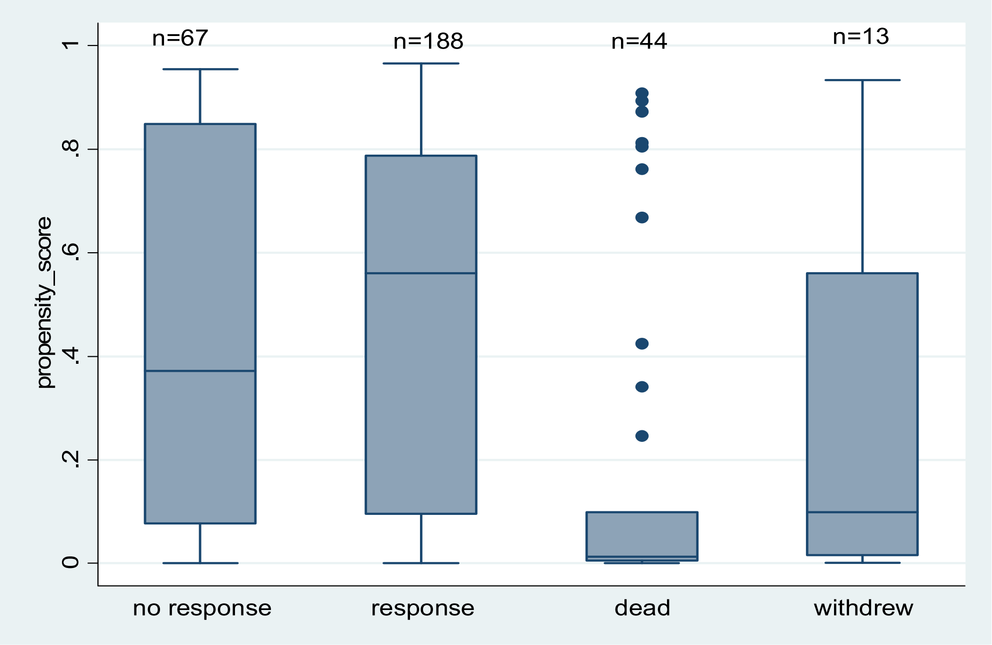
The propensity score was calculated from the six variables included in the SSV model (age at stroke, living alone pre-stroke, independent before stroke, able to walk independently within a week of admission, an MRC power score (arms) greater than 3 (i.e. able to lift arms against gravity), a normal verbal Glasgow Coma Score (orientated) (Counsell C et al 2002)). The covariates used to construct the original model were used to calculate this score (see equation (3)). The cut off for distinguishing good over poor outcome (as determined with a postal OHS ≤3) was set at 0.8. This value was chosen as it was the cut off used in a previous external validation of the SSV model (the FOOD trial, personal communication M Dennis).

The range of propensity scores was 0-0.96 (median 0.36, IQR 0.04-0.77). It can be seen, therefore, that the SSV predicts that the vast majority (243/312 = 78%) of patients in the study dataset to have a poor outcome (predicted dichotomised OHS of >=3) following their stroke.

Table 38 Differences in predicted outcome between responders and non-responders

	Range	Median	IQR
Non-responder	0.00074-0.95	0.37	0.078-0.85
Responder	0.00040-0.97	0.56	0.095-0.79
Dead	0.00051-0.91	0.013	0.006-0.10
Withdrawal	0.0016-0.93	0.10	0.015-0.56

Figure 42 Propensity score by response



Pairwise examinations of the differences between these groups (Mann-Whitney U tests) reveal that the propensity scores of both responders and non-responders are significantly higher than the median propensity score (probability of a good outcome) for patients who died (Table 39).

Table 39 Pairwise comparison of p values for differences in propensity score by response to six-month questionnaire (Mann-Whitney U tests)

		Responder	Non-responder	Death	Withdrawal
	Median propensity score	0.56	0.37	0.013	0.10
Responder	0.56				
Non-responder	0.37	0.95			
Death	0.013	<0.001	<0.001		
Withdrawal	0.10	0.077	0.169	0.048	

5.6.8 Process data

The proportion of patients receiving process markers, and the percentage of eligible patients in whom these were achieved are shown in Figure 44 (p 135). Specific aspects of care were delivered to patients in whom they were indicated (or not contraindicated) in over 80% of cases for fourteen of the nineteen processes shown. The care processes that were measured as part of the study (reflecting the indicators of the RCP NSSA (Royal College of Physicians 2009b)) are therefore often 'saturated' with little variability as regards receipt of specific aspects of care. It is therefore possible, if not likely, that these care processes will be poor discriminators of patient outcome.

There are exceptions, with some care processes being poorly achieved, for example, the proportion of eligible patients receiving a social worker assessment was particularly low (34%). Twenty five of all patients in the study received thrombolysis with recombinant tissue plasminogen activator (rtPA). Fifteen patients had a definite contraindication to thrombolysis (haemorrhagic stroke, two of whom had had a previous stroke event). The proportion receiving thrombolysis across sites was therefore 25/297 = 8.4%. A further 46 had had a previous ischaemic stroke which, in the presence of diabetes (which we have not recorded) is a further contraindication to thrombolysis for acute ischaemic stroke. It is therefore likely that the proportion of eligible patients receiving thrombolysis is higher than it appears in the study population. It should also be remembered that the CIMSS study population excluded patients with very severe stroke who were unlikely to survive to discharge. This is likely to have reduced the denominator such that the proportion of patients 'eligible' for thrombolysis within the study population is falsely elevated.

Much of the variability in whether patients received different aspects of care process is due to whether or not a particular care process is indicated. The variability in achievement of care processes in patients in whom they are indicated tends to decrease as the proportion of patients in whom a "no but" code is recorded decreases (moving left to right in Figure 44). This would tend to suggest that as a proportion of eligible patients, the care processes that are universally applicable are more readily achieved than processes that are not.

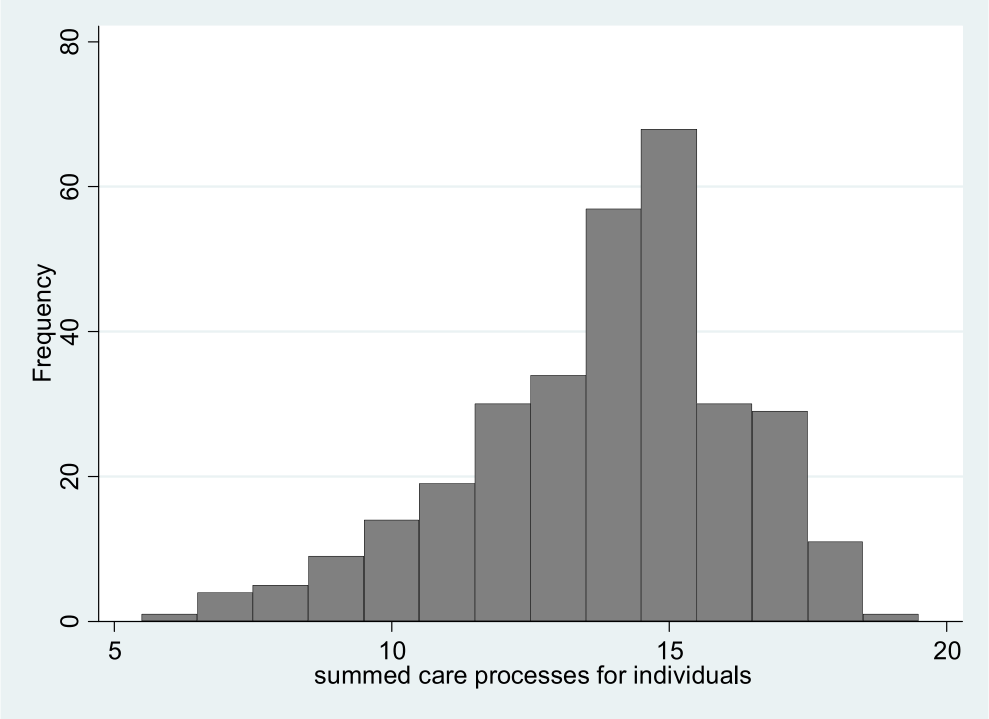
The exceptions to this are admission to a stroke unit the same day or day after the stroke, and an assessment of mood. Despite being relevant to all (stroke unit) or nearly all (mood assessment) patients admitted with stroke, the proportion actually achieving these processes was 77% and 81% respectively.

The reasons that specific processes are not indicated ("no but" codes) are either that the stroke is too mild or too severe (not possible or inappropriate to achieve). For example, patients with no speech deficit will not require a formal Speech and Language Therapist (SLT) communication assessment, and such an assessment would be inappropriate in some patients with very severe strokes (e.g. the drowsy or comatose). Patients with very severe stroke receiving or likely to require palliative care were excluded during recruitment for the study. It is possible, therefore, that in the study population the "no but" codes are more likely to reflect patients at the milder end of the spectrum of stroke severity. A summed score of the number of processes of care achieved for individual patients in whom they were indicated has been calculated using the 20 process indicators in Table 40 to give an overall picture of 'compliance' with the process markers measured in the study. This approach of summing process measures has been adopted by the RCP in the reporting of the NSSA data (Intercollegiate Stroke Working Party 2011). However, a summed process score is not useful either as a predictor of outcome or as a summary of process delivery, as this approach assumes that care processes are both additive and equally weighted. These assumptions are unlikely to be valid and would be particularly misleading if such a score were to be used as a single variable. For example, a simple summation of processes would fail to reflect that receipt of thrombolysis is likely to be a greater determinant of outcome than being weighed during the course of the admission. Moreover, this histogram is difficult to interpret, as the proportions of patients in whom particular processes are not indicated ("no but" codes) is not represented and as such it is difficult to appreciate what the maximum summed process score could be for individuals.

Table 40 Process markers measured in the study population

Process markers	
Admitted to stroke unit on day or day following admission	Visual fields assessed
Brain imaging within 24 hours of admission	Sensory testing
Patient given rtPA	Formal swallowing assessment within 72 hours
Swallowing screen within 24 hours	Speech and Language Therapy (SLT) communication assessment
Aspirin given within 48 hours	Social worker assessment
Physiotherapy assessment within 48 hours	Cognitive screening
Occupational therapy assessment within 48 hours	Malnutrition screening
MDT rehabilitation goal setting	Care plan for urinary incontinence
Weighed during admission	Fluids within 24 hours
Mood assessed during admission	Nutrition within 72 hours

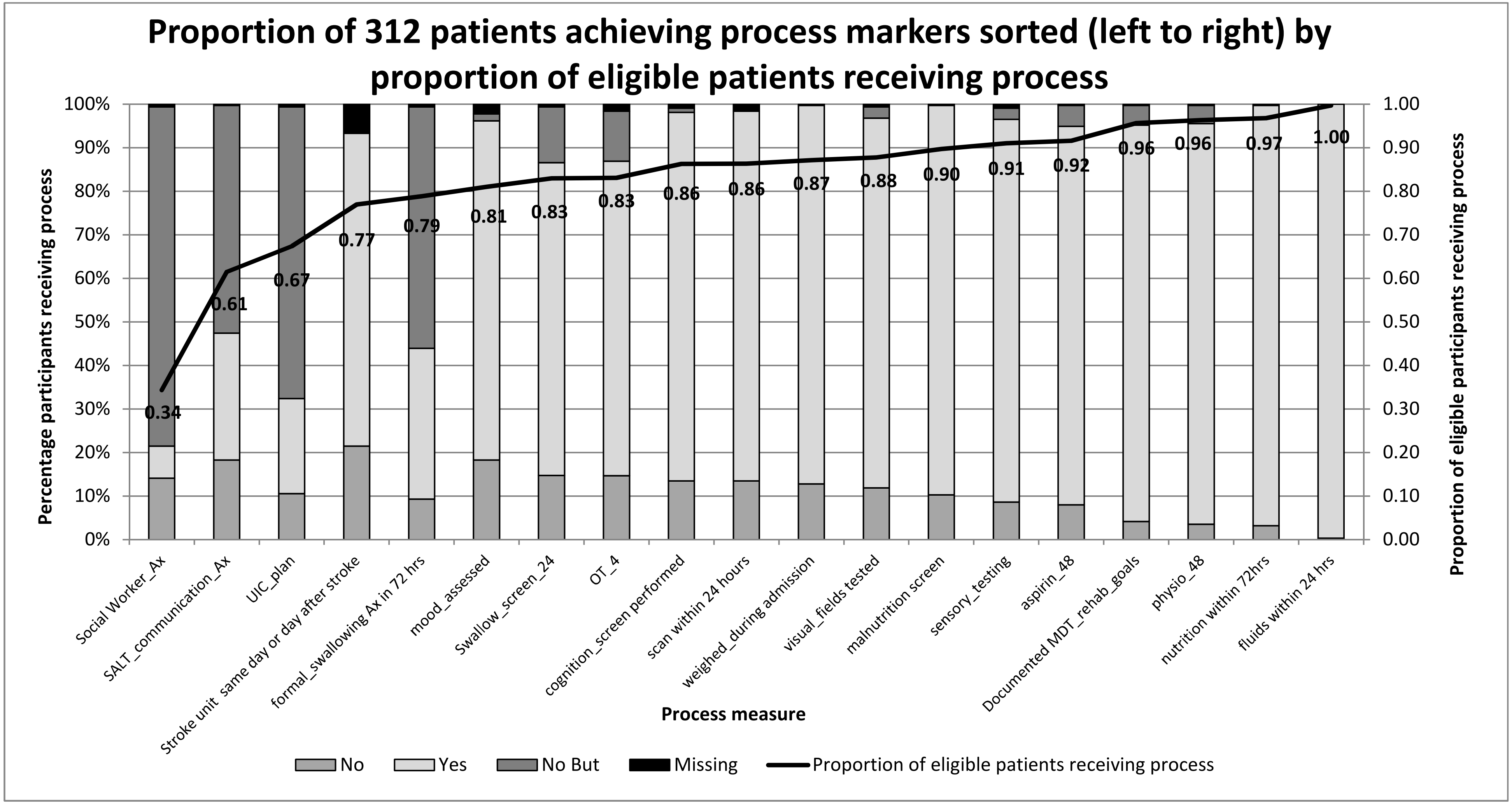
Figure 43 Total number of processes received in those for whom they were indicated



The median number of processes received was 14 (range 6-19 IQR 12-15)

Figure 44 presents the individual processes of care as the proportion of patients eligible for individual care processes that received them, and the proportion of patients in whom individual care processes were not indicated ("no but" codes). This offers a more useful summary of process delivery than presentation of summed process scores.

Figure 44 Proportion of patients receiving specific aspects of care by proportion eligible for individual care processes



5.6.9 Baseline questionnaires

The median, range and interquartile range for each of the baseline questionnaires is given in Table 41. These questionnaires offer a measure of stroke severity and baseline function. Scores on baseline questionnaires between patients who have responded and failed to respond to the six month questionnaire have been examined to ascertain if there are any systematic differences.

Table 41 Distribution of scores on baseline questionnaires

Instrument	Min-max score	Median	Range	IQR
Barthel Index	0 to 20	14	0-20	6-19
NEADL	0 to 66	53.5	0-66	36-60
EQ5D	-1 to 1	0.63	-0.429-1	0.082-0.814
GHQ-12 (dichotomised scoring)	0 to 12	1	0-12	0-4

5.6.9.1 Differences in baseline questionnaires between responders and non-responders

A Kruskall-Wallis test demonstrates that there are significant differences in both median NEADL and EQ5D scores at baseline between patients who responded and those that did not.

Figure 45 Baseline NEADL by response to six month questionnaire

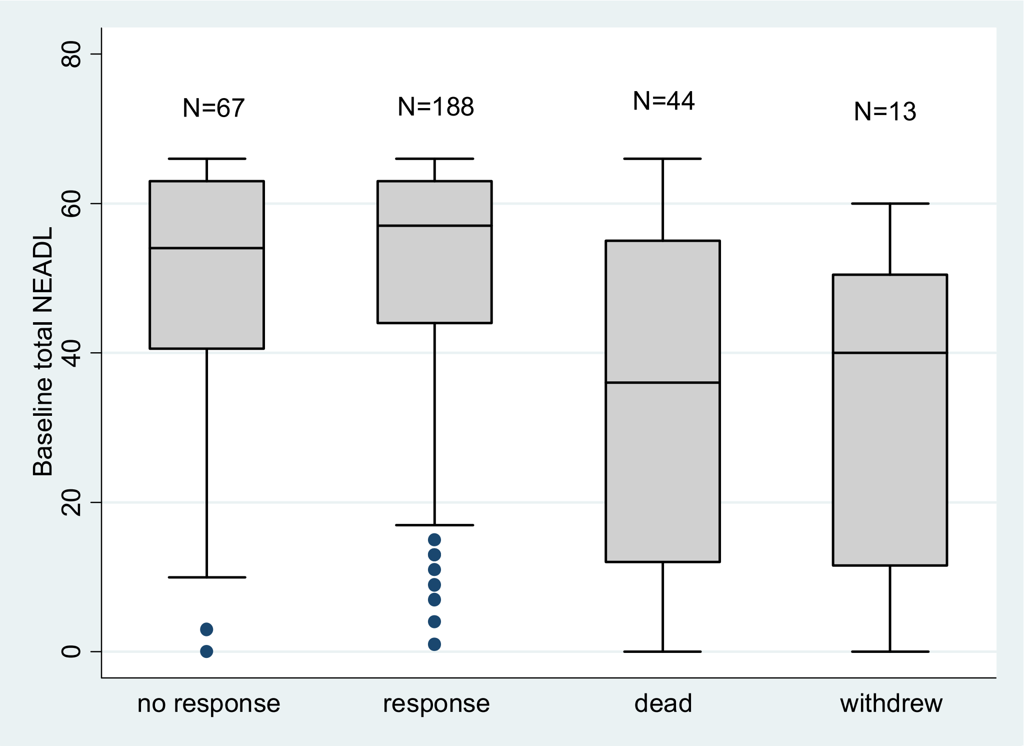


Table 42 Pairwise comparison of baseline NEADL by response (Mann-Whitney U tests)

		Responder	Non-responder	Death	Withdrawal
	Median Baseline NEADL	56	52.5	42	39
Responder	56				
Non-responder	52.5	0.215			
Death	42	<0.001	0.012		
Withdrawal	39	<0.001	0.011	0.466	

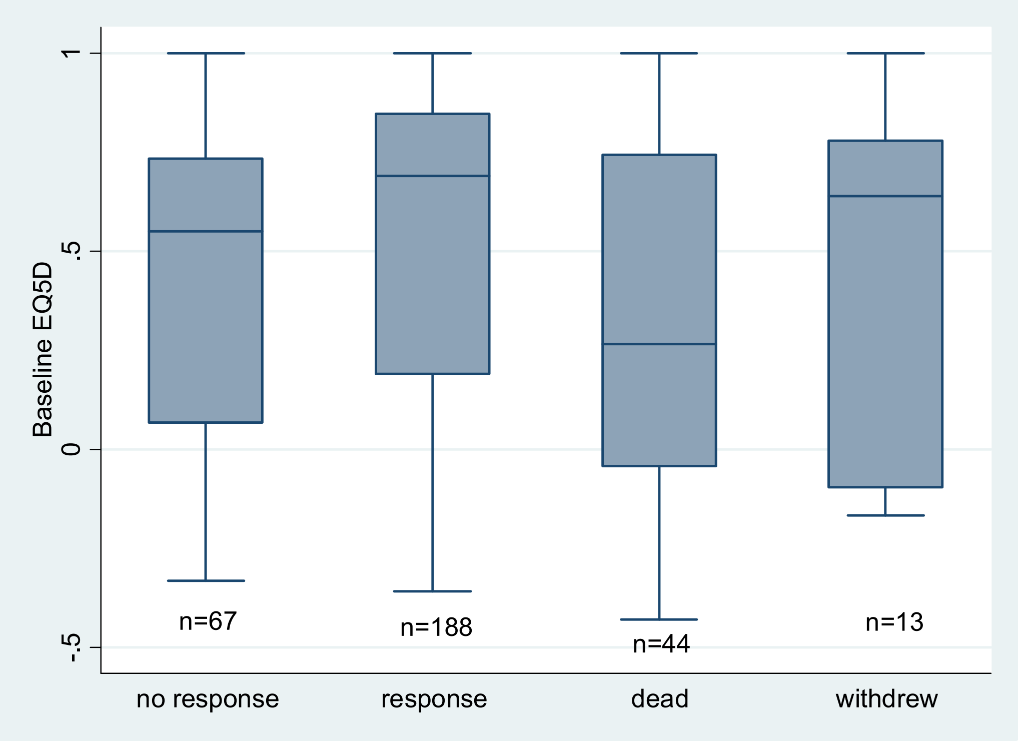
Patients who died or withdrew from the study had significantly lower baseline NEADL scores than those who remained in the study regardless of whether or not they responded to the six month questionnaire.

Due to the biphasic and non-normal distribution of the EQ5D, the median has been used as the measure of central tendency with non-parametric analyses. Patients who did not respond to the six month questionnaire had lower median quality of life scores than those who responded. Patients who subsequently died reported the lowest baseline quality of life scores, and this was significantly lower than patients who responded at six months.

Table 43 Pairwise comparison of baseline EQ5D by response (Mann-Whitney U tests)

		Responder	Non-responder	Death	Withdrawal
	Median Baseline EQ5D	0.691	0.551	0.267	0.640
Responder	0.691				
Non-responder	0.551	0.013			
Death	0.267	0.004	0.302		
Withdrawal	0.640	0.331	0.886	0.502	

Figure 46 Baseline EuroQoL by response to six month questionnaire



5.6.10 Six month outcomes questionnaires

The theoretical range of each outcomes questionnaire score, study median, observed range and interquartile range for returned questionnaires are given in Table 44.

Table 44 Descriptive statistics for individual outcomes questionnaires

Instrument	Returned completed questionnaires (N)	Min-max score	Median	Range	IQR
NEADL	165	0-66	47	0-66	21-60
EQ5D	179	-1 to 1	0.71	-0.349-1	0.414-0.85
SIPSO physical	176	0-20	13.2	0-20	8.36-17.8
SIPSO social	174	0-20	12.1	1.79-20	7.5-15.8
GHQ-12 (dichot scoring)	166	0-12	2	0-12	0-6
OHS	219 (includes 44 deaths)	0-6 (6=dead)	2	0-6	1-5

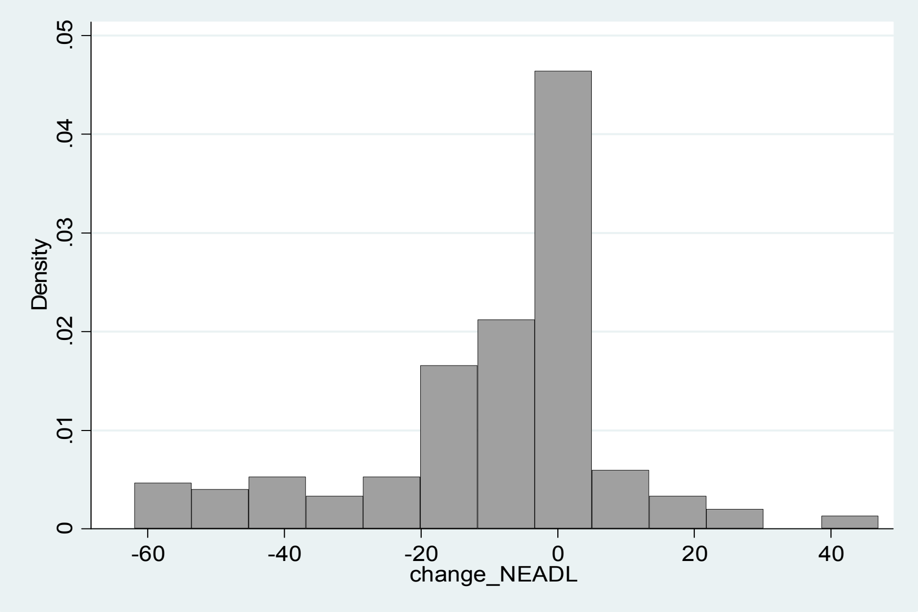
5.6.10.1 Proxy completion

Sixty two out of 175 (35%) participants in whom information was available on proxy completion of the outcomes questionnaires required some help in completing the questionnaire. In twenty nine returned questionnaires the responses were those of a proxy on behalf of the patient. In the remaining cases, the proxy recorded the patient's own responses.

5.6.11 Change in outcomes scores

For the Nottingham, GHQ_12 and EQ5D where assessments were made at baseline and at six months, a change in score may be plotted and this should be approximately normally distributed (data presented as histograms). 'Waterfall' plots are also presented which represent the change in scores between baseline and six-month assessments for individuals. This allows the proportion of patients with positive, unity and negative differences in scores to be seen.

Figure 47 Change in NEADL scores between baseline and six months



On inspection of the histogram, the change in NEADL score from baseline to six months is approximately normally distributed, but there is a negative skew to the distribution suggesting that the scores at six months were worse than the score at baseline. This is to be expected as the NEADL questionnaire is completed with respect to function and activity in the preceding week. Depending on the time of stroke relative to completion of the questionnaire, therefore, for some patients this may have represented function immediately before their stroke and for others soon afterwards.

Figure 48 Waterfall plot for change in NEADL between baseline and six months

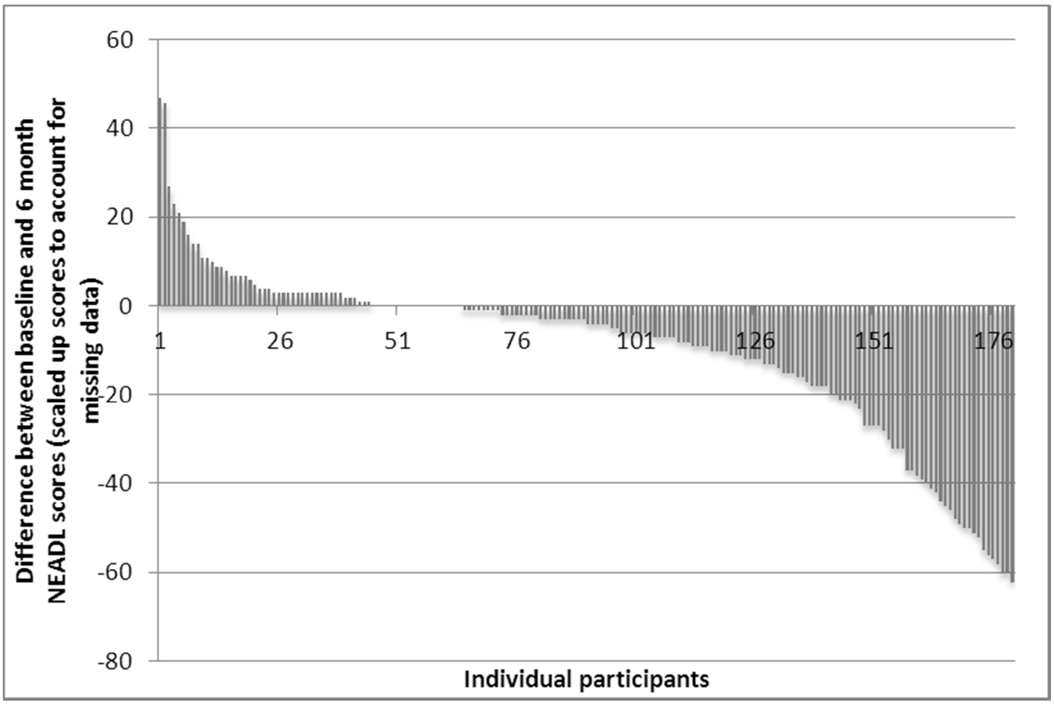
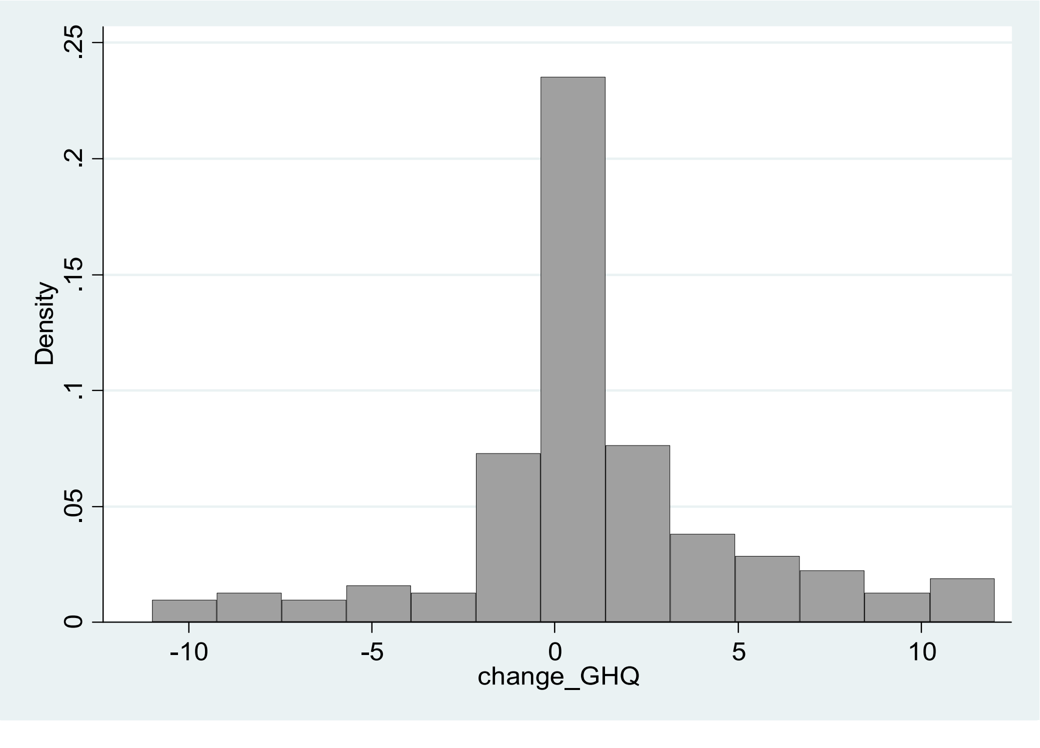


Figure 49 Change in GHQ_12 score



Median change in GHQ score between baseline and six months is zero, and the IQR is 0-3. Therefore changes in GHQ-12 between baseline and six months tend to be small, although there are a few outliers where there are marked changes in scores (in both positive and negative directions).

Figure 50 Waterfall plot of change in GHQ-12 score between baseline and six months

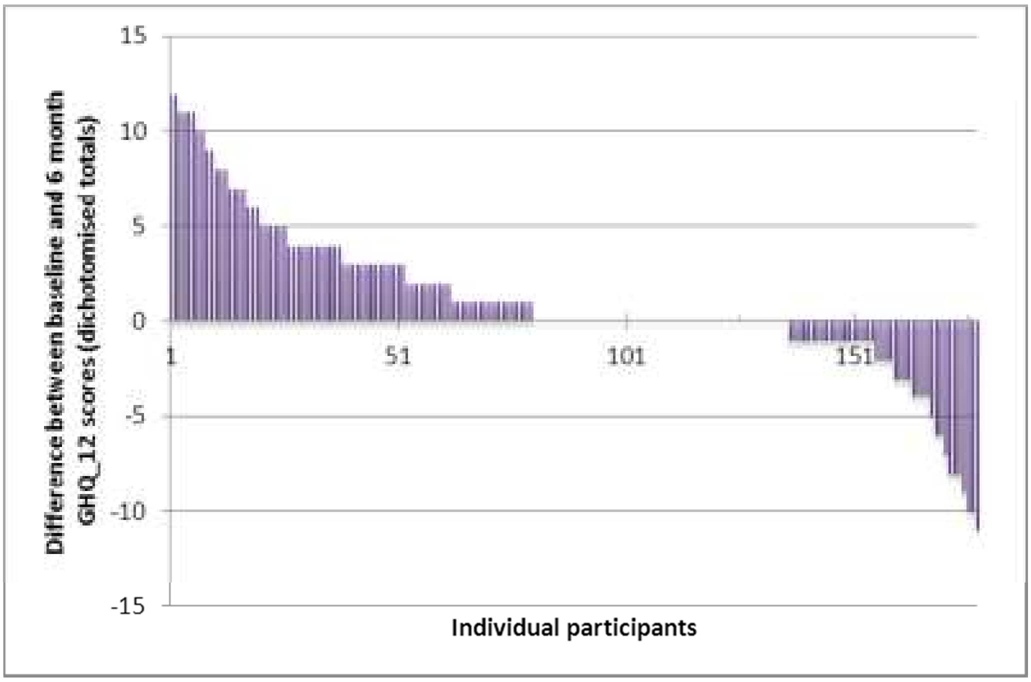
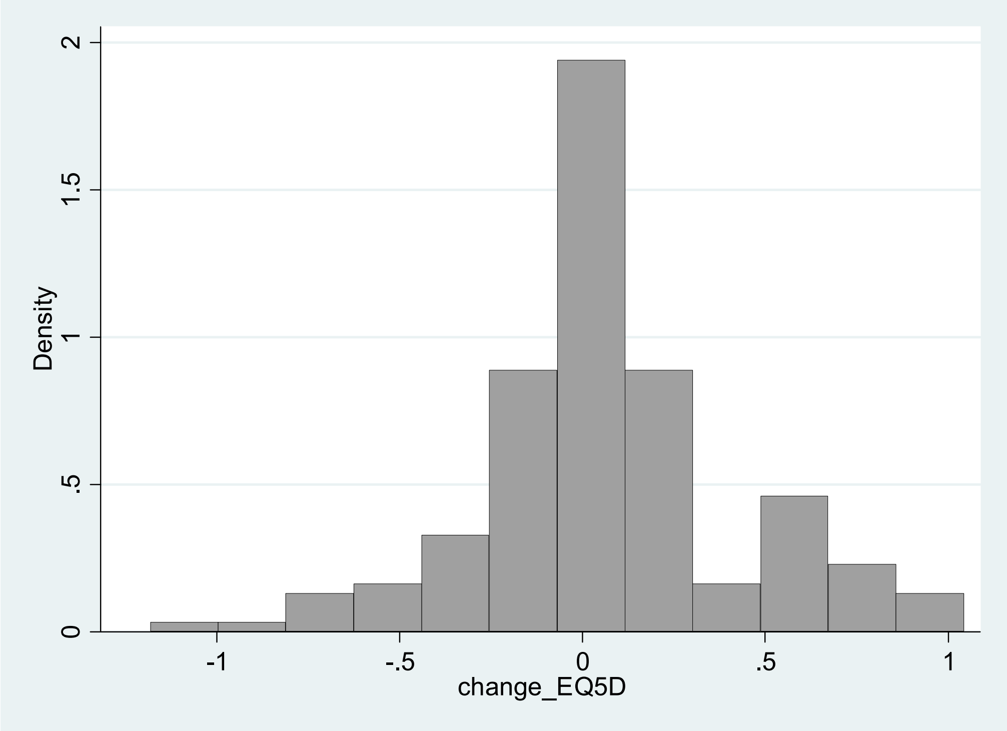
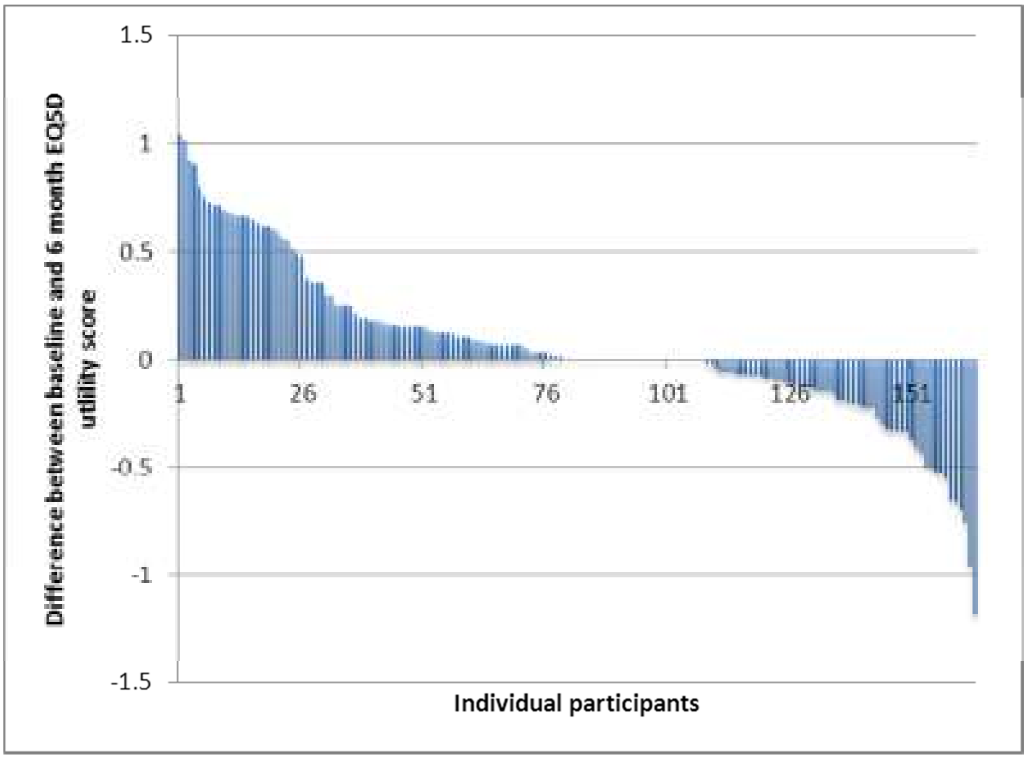


Figure 51 Change in EQ5D



Similarly for the quality of life score (EQ5D), the distribution is approximately normally distributed centred on zero. Large differences between baseline and six month EQ5D scores were therefore infrequent. The EQ5D is measured with respect to how an individual feels at the time of filling in the questionnaire.

Figure 52 Waterfall plot of change in EQ5D utility score between baseline and six months assessments.



5.7 Univariate analyses

The relationship between individual process markers and patient outcome in univariate analyses may offer an indication of important factors that contribute to patient outcome. However, interpretation of these relationships should be made with the caveat that there is no adjustment for confounding or mediating factors and markers which may appear to be important may cease to be so when other factors are taken into consideration. Entering univariate predictors that reach statistical significance into regression models without clinical reasoning may result in the inclusion of statistically, but not clinically, important predictors. Moreover, the risk of uncovering a statistically significant relationship (or refuting an important relationship) between process markers and outcome increases as the number of analyses increases, especially if the sample size is small. For example a 5% significance level means that if twenty analyses are performed, one is likely to be statistically significant through chance alone. Primary analyses have been performed using the SIPSO physical and social outcomes. This is due to the relatively superior properties in terms of absence of floor and ceiling effects when compared with the NEADL (see section 5.5) In addition, the Rasch analysis of the SIPSO that has been performed by previous authors allows the instrument to be considered as an interval scale (Kersten P et al 2010). The authors of this Rasch analysis argue that the population in which the Rasch analysis was performed should not affect the transformation of the ordinal scores to continuous scores in other populations, and the same transformation factors may be used (Kersten P et al 2010).). However, further exploration of the scale with respect to differential item functioning in an older population is required (see section 6.3.1).

Examination of the relationships between categorical data with two level responses (usually whether a process did, or did not occur) and the subscores of the SIPSO would ideally be analysed with a parametric test (ANOVA). However, as discussed in section 5.4.2.2, the physical subscore of the SIPSO is not normally distributed and this may lead to violation of the assumption of normality of residuals for an ANOVA. Univariate analyses for the physical subscore of the SIPSO have therefore been performed with the non-parametric equivalent to an ANOVA (Kruskall-Wallis test). ANOVAs have been used for the normally distributed social subscore of the SIPSO. Where an ANOVA demonstrates a statistically significant difference between groups, pairwise examination of the mean SIPSO scores (and confidence intervals) in each of the three levels of categorical outcome have been performed to identify where the differences lie.

5.7.1 Correlation of process variables with SIPSO physical subscore

Table 45 Univariate analyses of process measures and physical subscore of SIPSO

Care process	Kruskall Wallis test p value
Admitted to stroke unit on day or day after admission	0.25
Scan within 24 hours of admission	0.56
tPA given	0.31
Swallow screen in 24 hours	0.04
Aspirin in 48 hours	0.46
Physiotherapy within 48 hours	0.55
Occupational therapy assessment within four days	0.32
MDT rehab goals set	0.17
Weighed during the course of the admission	0.035
Mood assessed during admission	0.13
Visual fields assessed	0.08
Sensory testing	0.39
Formal swallow assessment by SLT within 72 hours	0.005
Communication assessment by SLT	0.001
Social worker assessment	0.0016
Cognition screen	0.15
Malnutrition screen	0.41
Urinary incontinence care plan	<0.001
In receipt of fluids within 24 hours of admission	N/A
In receipt of nutrition within 72 hours of admission	0.14

Variables failing to reach significance at the p ≤0.01 level were not explored further unless there were strong clinical reasons for doing so because of the small size of the data set and the number of univariate analyses (which increases the risk of spurious or chance correlations). Four variables (formal swallow and communication assessments by Speech and Language therapist, social worker assessment and a urinary incontinence care plan) all reached statistical significance at the 0.01 level for the prediction of the physical subscore of the SIPSO (highlighted in Table 45). In addition, a swallow screen within 24 hours of admission is a potentially clinically important process marker and will be considered further. Two-way examination of the relationships between these variables and the physical subscore of the SIPSO are highlighted in Table 46 with a Mann-Whitney U test. The levels of the variable between which there are statistically significant differences in six month SIPSO outcomes are presented in bold – for example, patients who require and receive a formal swallowing assessment ("Yes") have significantly lower SIPSO scores at six months than patients in whom such an assessment is not required ("No but"). This is also true for patients receiving a swallowing screen within 24 hours and those with a urinary incontinence care plan. For patients in whom assessments are indicated ("No" or "Yes"), there is no significant difference between those that do, and do not receive the assessments for any of the variables reaching significance at the p≤0.001 level the oneway ANOVA.

Table 46 Pairwise identification (Mann-Whitney U tests) of statistically significant differences in distributions of Rasch transformed physical SIPSO scores between levels of response for process variables significant at the 1% level on Kruskall-Wallis testing

Formal swallowing assessment within 72 hours			
	No	Yes	No But
No			
Yes	p=0.80		
No but	p=0.051	Medians (p=0.0032) Yes=11.4 No but=15.0	
SALT communication assessment			
	No	Yes	No But
No			
Yes	p=0.19		

No but	Medians(p=<0.001):
	No=9.1
	No but=15.0
		Medians (p=0.032):
		Yes=13.6
		No but=15.0	

Social worker assessment			
	No	Yes	No But
No			
Yes	p=0.25		

No but	Medians (p=<0.001):
	No=8.4
	No but=15.0
		p=0.21	

Urinary incontinence care plan			
	No	Yes	No But
No			
Yes	p=0.26		

No but	p=0.064	Medians (p=<0.001):
		Yes=8.4
		No but=14.0	

Swallow screen in 24 hours			
	No	Yes	No But
No			
Yes	p=0.55		
No but	Medians (P=0.009): No But=20 No=14	Medians (p=0.001): No But=20 Yes=12.8	

Further exploration of whether or not patients were weighed during their admission revealed that no patients received a “no but” code for this variable. There is therefore a statistically significant difference in the distributions of physical SIPSO scores between patients that were weighed and those that were not, with patients not being weighed having higher median SIPSO scores at six months than those that are not (medians Yes=13.2 No=16.2, z=2.13 p =0.033 Mann-Whitney U test).

Being weighed is used as a marker of process in the RCP NSSA audit. However, in this dataset there is an inverse relationship between being weighed and physical outcome. If this relationship was also evident in external datasets, the relevance of being weighed as a marker in the RCP summed process scores could be questioned – patients who are weighed have poorer six month physical SIPSO scores than those that were not. It is possible, that this phenomenon is a chance finding (there were only 22 patients that were not weighed during the course of their admission), but the difference in scores between the groups is large enough to be of some clinical significance (three points on SIPSO scale) if it were true. The characteristics of patients who were not weighed have therefore been explored further.

There is no significant difference in age or clinical classification of patients that were or were not weighed during their admission. However, the characteristics of patients who were not weighed differed significantly in terms of stroke severity variables as outlined in Table 47 below. Medians have been presented as the marker of central tendency due to the non-normal distributions of the variables.

Table 47 Characteristics of patients who were, and were not weighed during the course of their admission

Variable	Weighed (median{IQR})	Not weighed (mean[95%CI])
Propensity score	0.31[0.03-0.74]	0.69[0.21-0.83]
Length of stay	12[5-38]	4[2-9]
Baseline Barthel Index	13{6-19}	18{12.5-20}

Patients that have had more severe strokes are more likely to have longer hospital stays and therefore more opportunity to be weighed. It is likely that it is this relationship that results in the seemingly poorer outcomes in patients who are not weighed, i.e. being weighed is acting as a proxy marker of stroke severity. It is therefore unlikely that the apparent statistical significance of being weighed in univariate analysis would remain once stroke severity variables are controlled for in multivariable analysis.

5.7.2 SIPSO social subscore

A Q-Q plot and Shapiro Wilk test suggest that the social subscore of the SIPSO approximate a normal distribution such that parametric analyses may be performed (see Figure 29).

Table 48 Univariate relationships between processes of care and SIPSO social subscore

Care process	ANOVA p value
Admitted to stroke unit on day or day after admission	0.0073
Scan within 24 hours of admission	0.16
tPA given	0.26
Swallow screen in 24 hours	0.034
Aspirin in 48 hours	0.17
Physiotherapy within 48 hours	0.38
Occupational therapy assessment within four days	0.09
MDT rehab goals set	0.24
Weighed during the course of the admission	0.50
Mood assessed during admission	0.15
Visual fields assessed	0.13
Sensory testing	0.48
Formal swallow assessment by SLT within 72 hours	0.006
Communication assessment by SLT	<0.001
Social worker assessment	0.021
Cognition screen	0.036
Malnutrition screen	0.80
Urinary incontinence care plan	<0.001
In receipt of fluids within 24 hours of admission	N/A
In receipt of nutrition within 72 hours of admission	0.066

Table 49 Identification of statistically significant differences in mean social SIPSO scores between levels of response for process markers reaching significance at the 1% level in a oneway ANOVA

	mean social SIPSO scores [95% confidence interval]		
	No	Yes	No But
Admitted to stroke unit on day or day after admission	9.9 [8.1-11.6]	12.6[11.8-13.4]	N/A
Formal swallow assessment by SLT within 72 hours	10.9 [8.5-13.4]	10.6[9.2-12.0]	13.0 [12.1-13.8]
Communication assessment by SLT	10.2 [8.4-11.9]	10.3 [9.0-11.6]	13.5 [12.6-14.3]
Urinary incontinence care plan	10.7 [7.7-13.6]	9.1 [7.6-10.5]	13.0 [11.4-12.9]

The darkest shaded box for SLT communication assessment in Table 49 indicates a significant difference between "No"/"No But" and "Yes" / "No But". There is no significant difference between "No" and "Yes" responses.

Patients admitted to a stroke unit on the same day or the day after their presentation to hospital have better six month social SIPSO scores, although this just fails to reach statistical significance at the p=0.01 level. Patients who are formally assessed by speech and language therapists for both swallowing and communication have worse six month outcomes than patients that do not require such assessments ("no but" codes). This likely reflects patients without deficits rather than patients too unwell to undergo assessments. In addition, patients in whom a communication assessment is indicated but not performed also have significantly lower six month social SIPSO scores than patients in whom such assessments are not indicated, but they do not have worse outcomes than patients who receive formal communication assessments. This would tend to suggest that it is the presence of the deficit rather than the communication assessment itself that is correlated with the six month social outcome. In a similar way, patients who require and have a urinary incontinence care plan in place, have significantly worse social outcome scores at six months than patients in whom such an assessment is not required.

Consistently in univariate analysis, the significant differences are between the ‘no but’ group and the ‘no’ or ‘yes’ groups. This would tend to suggest that it is the requirement for, and not the receipt of, particular processes that is associated with outcome.

5.8 Regression trees

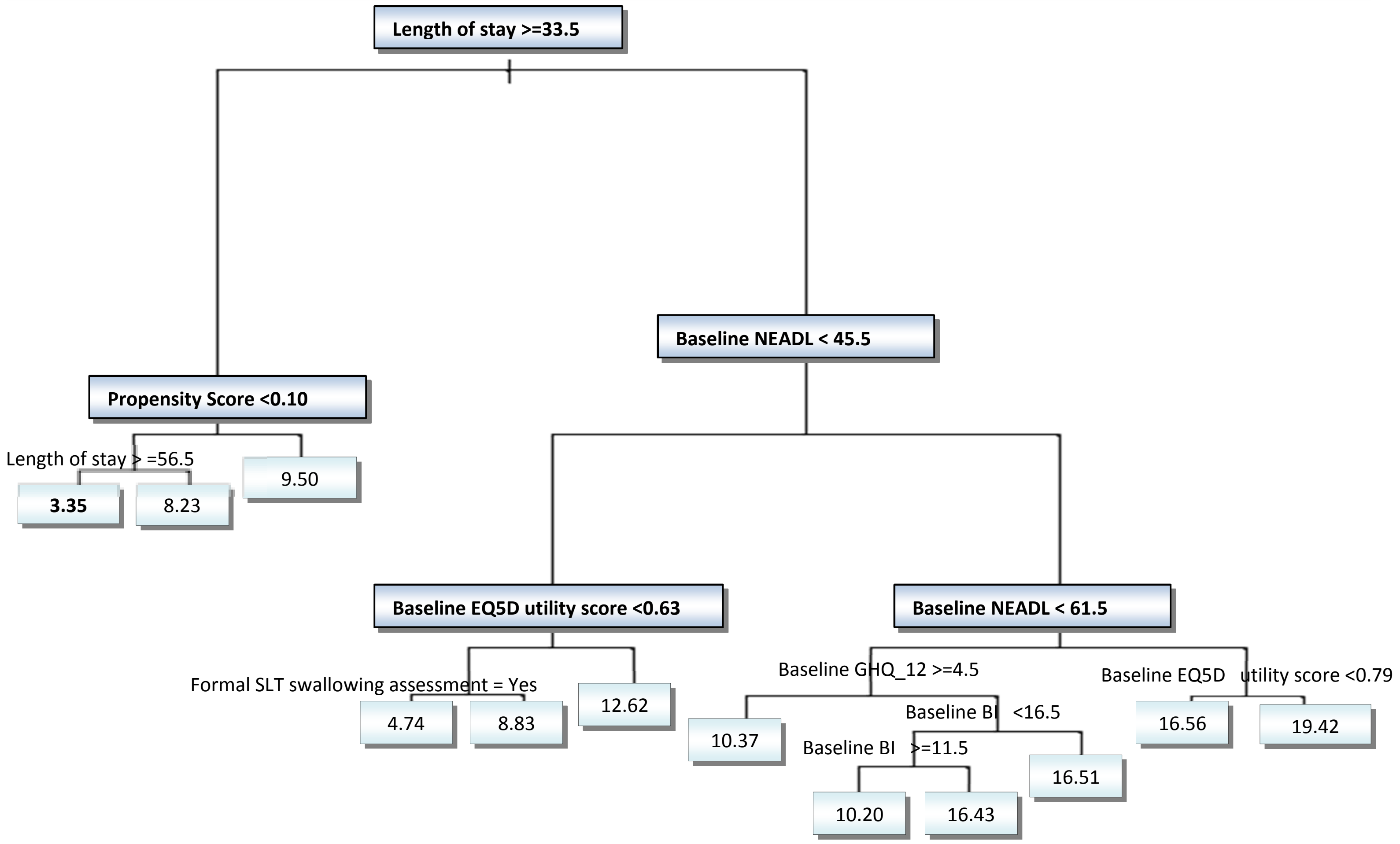
5.8.1 Prediction of Physical subscore of SIPSO

Entering the variables from Table 10 p 81 (including the baseline assessments) into a regression tree model to predict the Rasch transformed physical SIPSO subscore results in Figure 53. Pruning this tree to remove the variables that explain less of the SIPSO physical outcome gives the tree shown in Figure 54.

It can be seen from the regression tree that length of stay is the main determinant of physical SIPSO score at six months. Baseline NEADL is also an important predictor, and propensity score does not feature in the regression tree. It is likely that there is collinearity between the baseline EQ5D and the baseline NEADL variables. This however will be addressed further through both clinical reasoning and stepwise variable selection procedures during the construction of regression models.

Construction of the trees without the baseline assessments gives Figure 55 (unpruned) and Figure 56 (pruned tree). Inspection of these regression trees reveals length of hospital stay to be the most important predictor (see pruned tree Figure 54). It is likely that length of stay is acting as a proxy marker for stroke severity and there is therefore likely to be collinearity between length of stay and other prognostic or severity variables. Again, this will be considered explicitly during linear regression model development.

Figure 53 Regression tree of physical subscore of SIPSO on variables in Table 10 (including baseline assessments)



.

Figure 54 Pruned regression tree of physical subscore of SIPSO on variables in Table 10 (including baseline assessments)

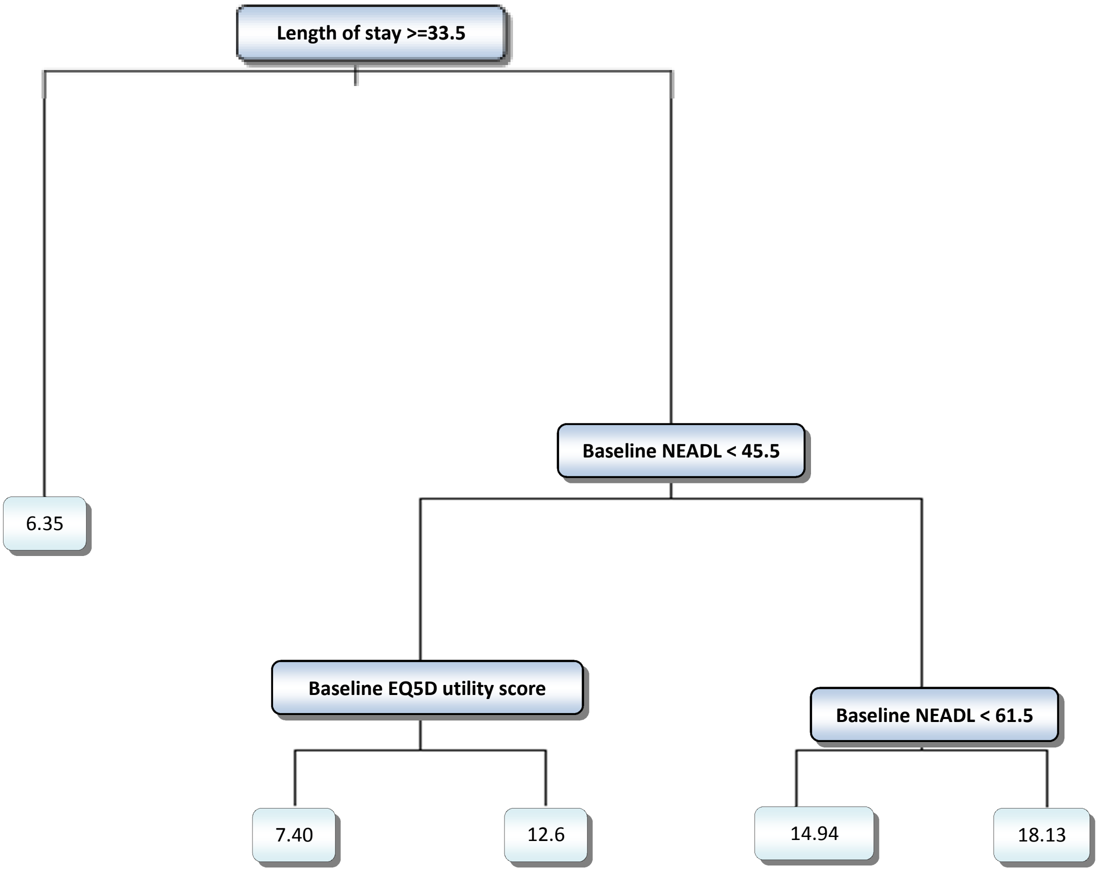


Figure 55 Regression tree of physical subscore of SIPSO on variables in Table 10 (excluding baseline assessments)

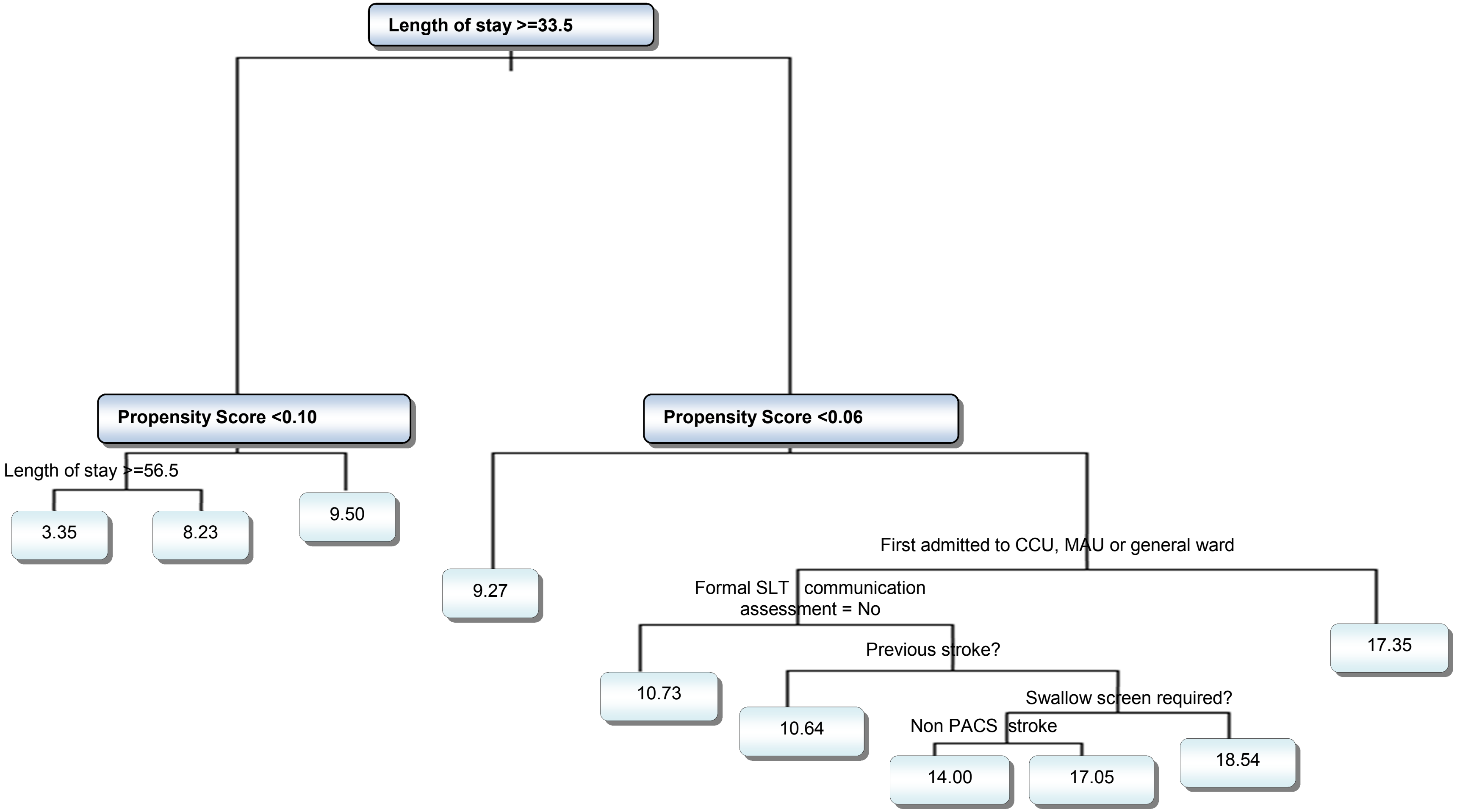
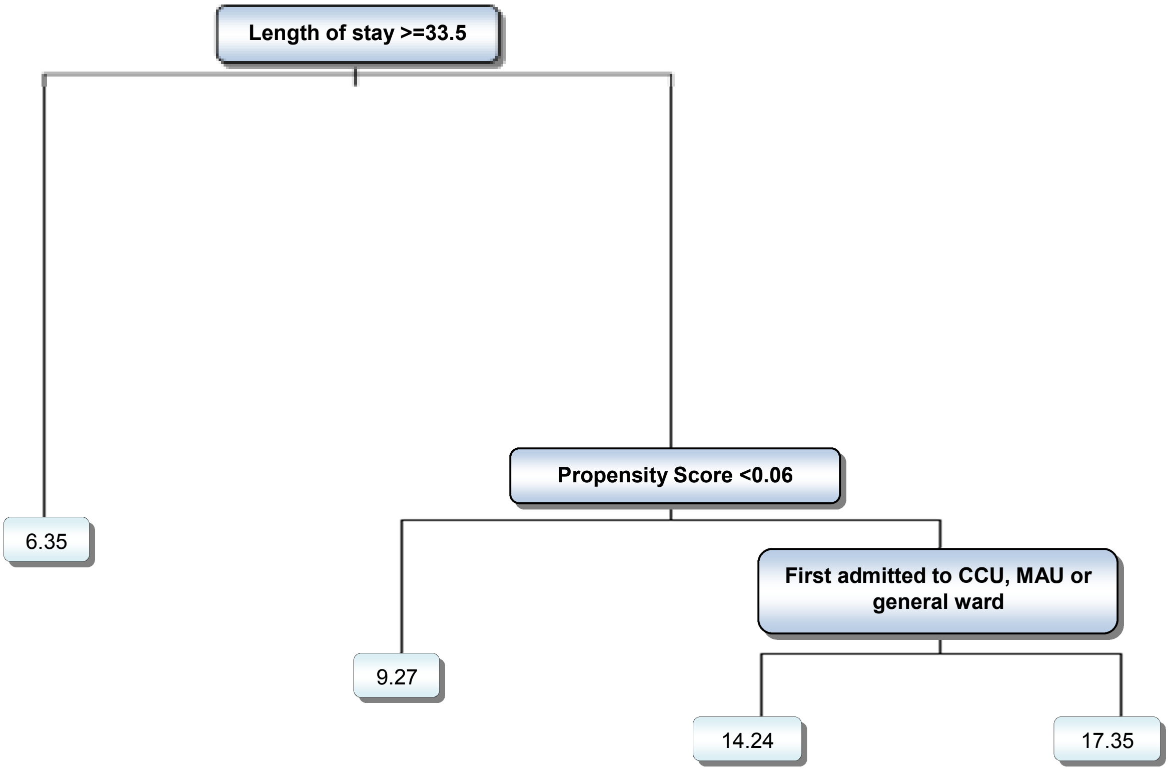


Figure 56 Pruned regression tree of physical subscore of SIPSO on variables in Table 10 (excluding baseline assessments)



5.8.2 Prediction of social subscore of SIPSO

Examination of the regression tree to predict the social subscore of the SIPSO that includes baseline questionnaires is shown in Figure 57. Here the major determinant of social outcome at six months is the baseline EQ5D. Pruning the tree reveals baseline EQ5D, the requirement for a formal speech and language assessment and the baseline NEADL as the most important predictors of outcome (Figure 58)

Figure 57 Regression tree of social subscore of SIPSO on variables in Table 10 (including baseline assessments)

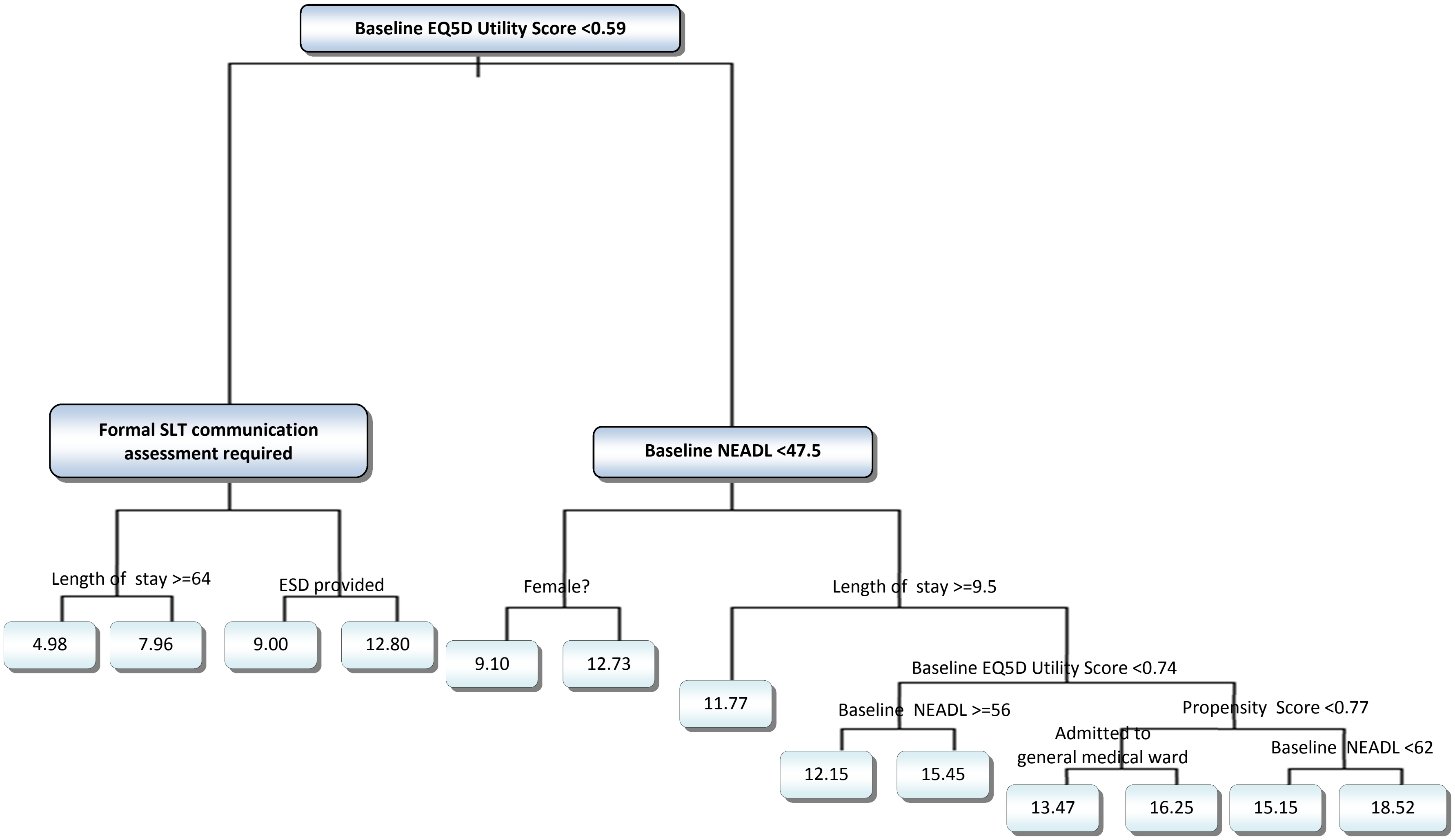
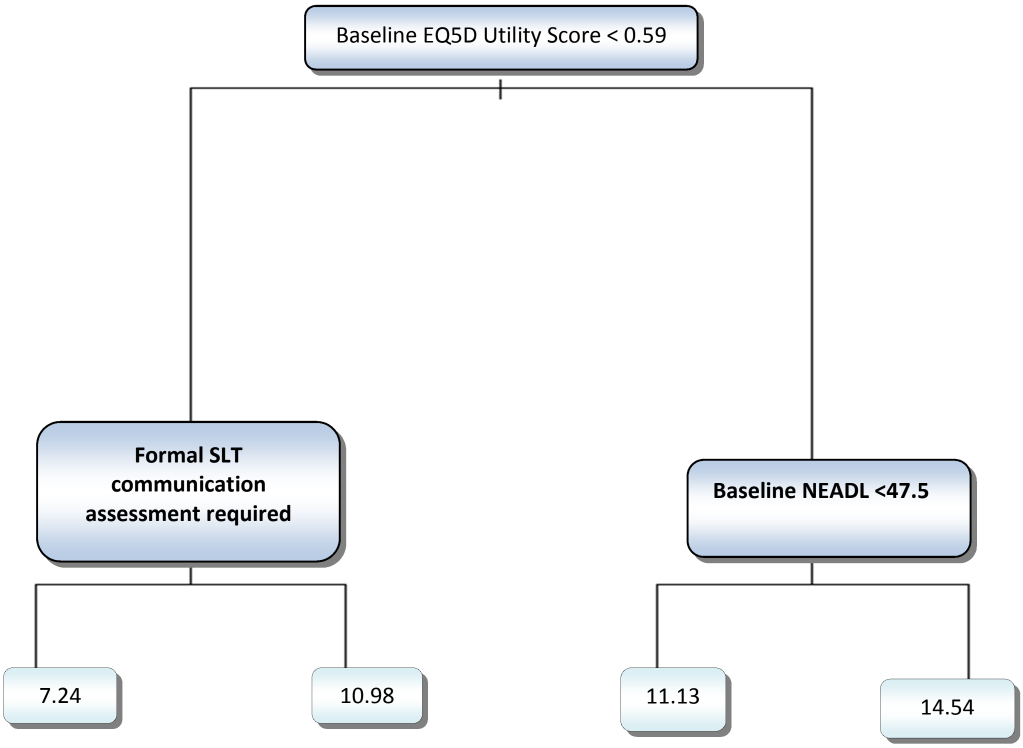


Figure 58 Pruned regression tree of social subscore of SIPSO on variables in Table 10 (including baseline assessments)


Baseline EQ5D Utility Score < 0.59
Formal SLT communication assessment required
Baseline NEADL <47.5
7.24
10.98
11.13
14.54

Length of stay is revealed as the most important predictor in the tree to predict social outcome that does not include the baseline assessments (Figure 59). Whether or not patients were treated on a stroke unit, stroke type (lacunar vs. other types of stroke) and imaging within 24 hours of admission were other important predictors, and these remained in the trees following pruning (Figure 60)

Categories of formal communication assessment by a speech and language therapist feature as predictors in three out of the four trees (Figure 53, Figure 55 & Figure 57), although it only features prominently in the model to predict the social SIPSO with baseline assessments (Figure 57)**.** This variable, the type of ward a patient is admitted to, are the only two process markers that feature prominently in the regression trees. The majority of variables which feature in the trees are markers of stroke severity and these overshadow the other variables. Some variables describing more organisational aspects of patient care do appear in the trees (e.g. whether a patient was first admitted to a ward capable of delivering hyperacute stroke care) and this may reflect local differences in service provision.

.

Figure 59 Regression tree of social subscore of SIPSO on variables in Table 10 (excluding baseline assessments)

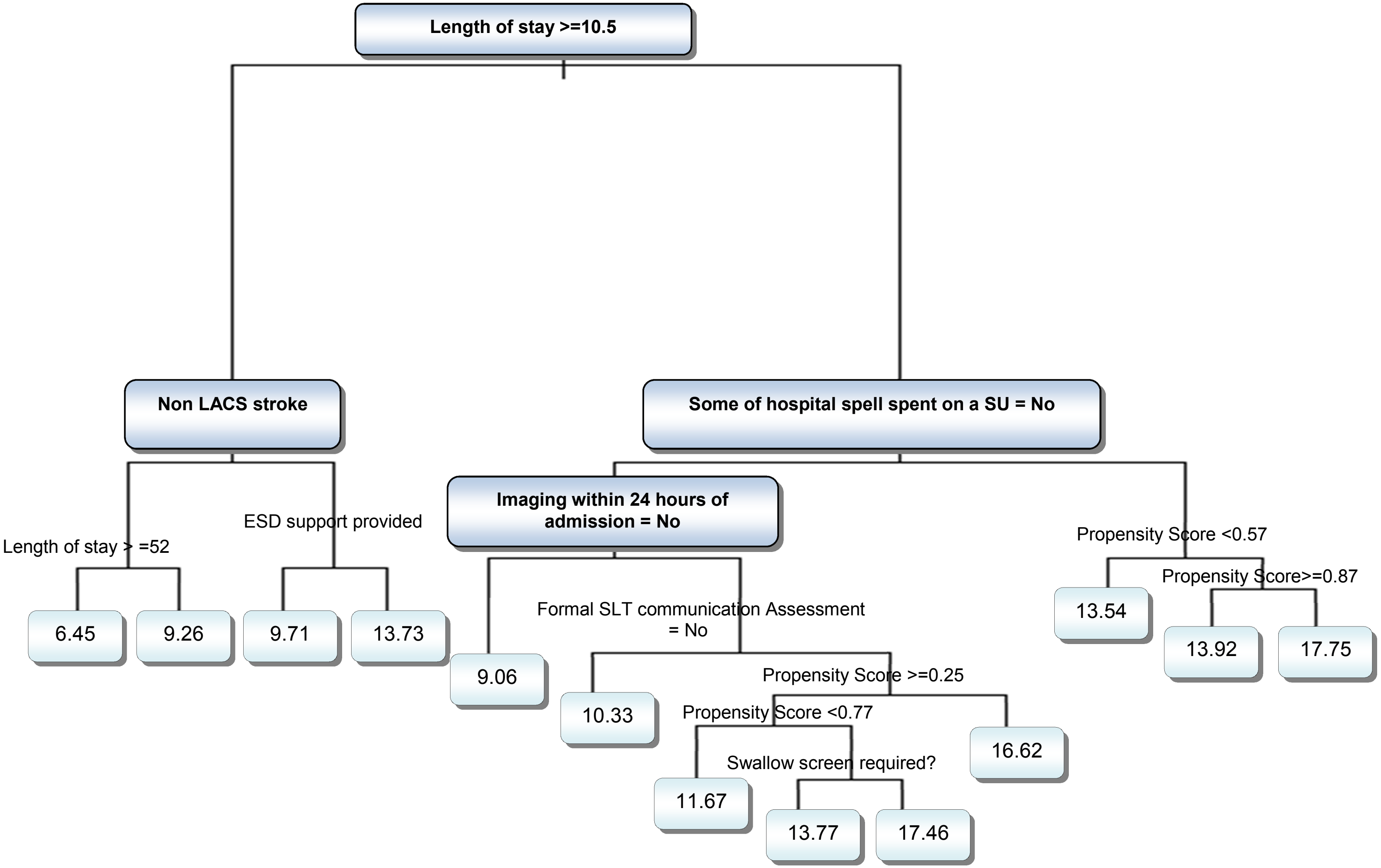
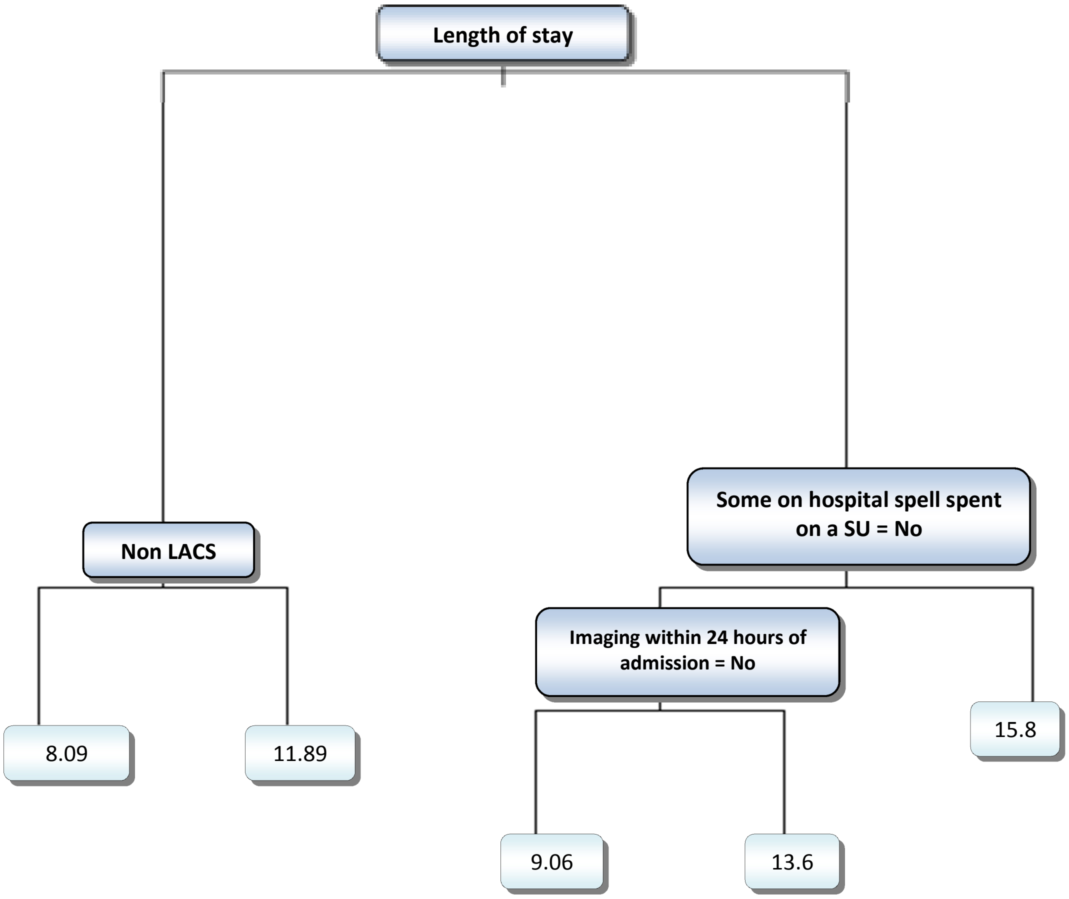


Figure 60 Pruned regression tree of social subscore of SIPSO on variables in Table 10 (excluding baseline assessments)


Length of stay
Non LACS
Some on hospital spell spent on a SU = No
Imaging within 24 hours of admission = No
8.09
11.89
15.8
9.06
13.6

5.9 Construction of linear regression models

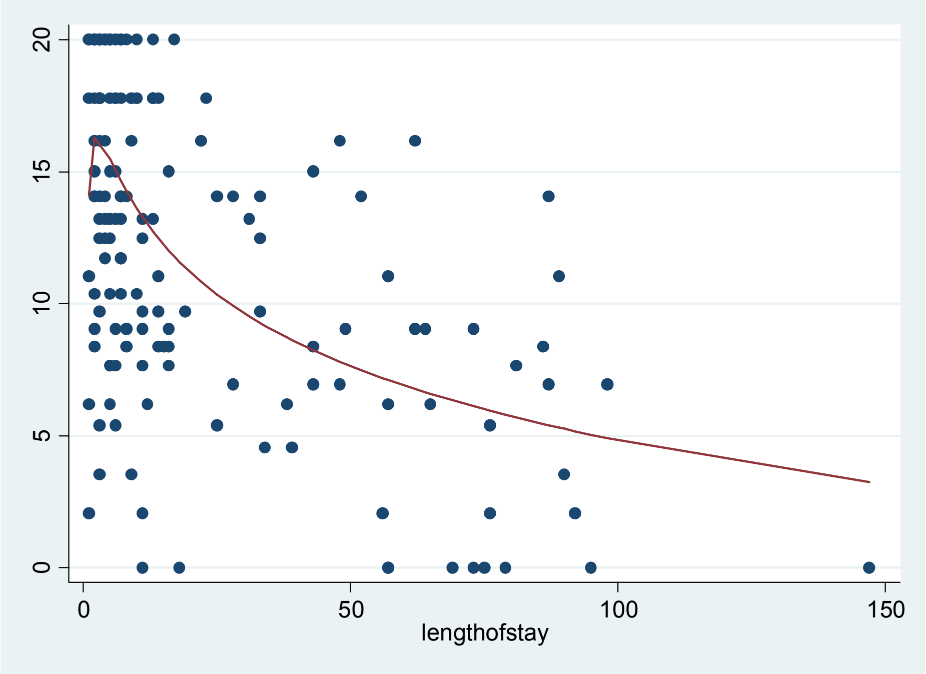
Linear regression models were constructed to explore the association between individual processes of care and the physical and social subscores of the SIPSO. Modelling methodology including post-estimation checks of assumptions is discussed in section 4.4.6. Two models were created for each outcome – one that includes and one that excludes baseline functional assessments. This was in order to ascertain whether there are prominent predictors of outcome in the absence of baseline assessment, as this would increase the utility of these predictors in routine care where the infrastructure to collect baseline assessments may be limited.

Models were re-run for each analysis with exclusion of influential cases to ascertain whether the prominent predictors changed (as a measure of model stability). However, post-estimation analyses were made on the full models.

5.9.1 Transformation of length of stay

A pre-estimation examination scatter plot of length of stay against SIPSO physical score revealed a likely logarithmic or reciprocal relationship (the line represents a fractional polynomial line of best fit calculated by STATA).

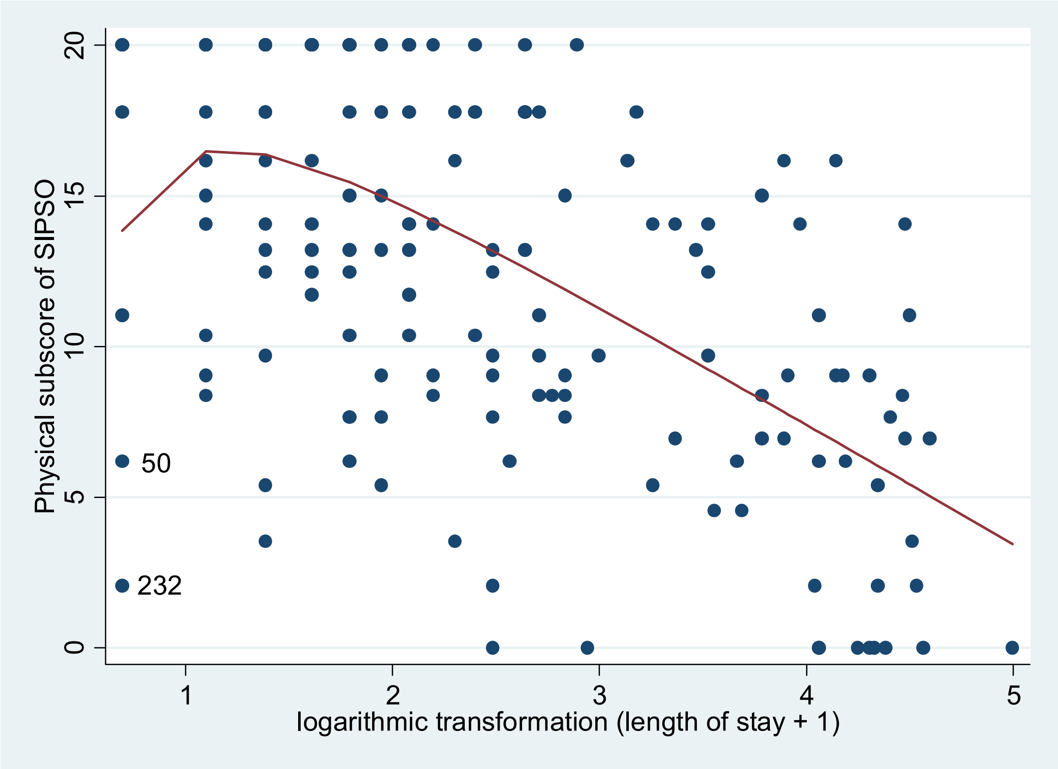
Table 50 Scatter plot of length of stay against SIPSO physical subscore with fractional polynomial line of best fit



A logarithmic transformation of the variable length of stay was performed in an attempt to improve the linearity of the relationship with the physical subscore of the SIPSO prior to modelling. A value of one was added to length of stay to ensure that there were no zero values prior to taking logarithms. However, as time of discharge had not been recorded in the study, cases where discharge was on the same calendar day of admission (discharge within 24 hours) had already been rounded up to one whole day (i.e. there were no zero values). A scatter plot of the logarithm of length of stay plus one against the physical subscore of the SIPSO reveals a linear relationship, but there is deviation from linearity at small values of length of stay. This may represent the rounding of length of stay, although it would be expected that patients discharged rapidly who survive to six month follow up (i.e. not discharged for palliative care) would have better outcomes than those with longer lengths of stay. It is possible, therefore, that this line of fit is being ‘pulled’ by influential cases (such as cases 232 and 50, highlighted on Figure 61), where outcomes are poorer than would be expected based on their length of stay.

A logarithmic transformation of length of stay plus one was used in all the models.

Figure 61 Scatter plot of logarithmic transformation of (length of stay + 1) against physical subscore of SIPSO with polynomial line of best fit demonstrating linearity



5.9.2 Prediction of physical subscore of the SIPSO with baseline assessments (Model 1)

Table 51 Independent variables to be entered into regression models for prediction of physical subscore of the SIPSO

	Variables	Number of variables (including dummy variables)
Variables identified from regression trees	Length of stay Baseline NEADL Baseline EQ5D	**1** **1** **1**
Variables identified from univariate analysis	Formal swallowing assessment Communication assessment Social Worker assessment Urinary incontinence care plan	**2** **2** **2** **2**
Probable important variables through clinical reasoning	tPA given	**2**
	First admitted to a stroke unit, CCU/HDU/ICU vs. general ward/MAU	**1**
	Propensity score (or age if propensity score removed)	**1**
	Discharge to same address	**1**
Total		**16**

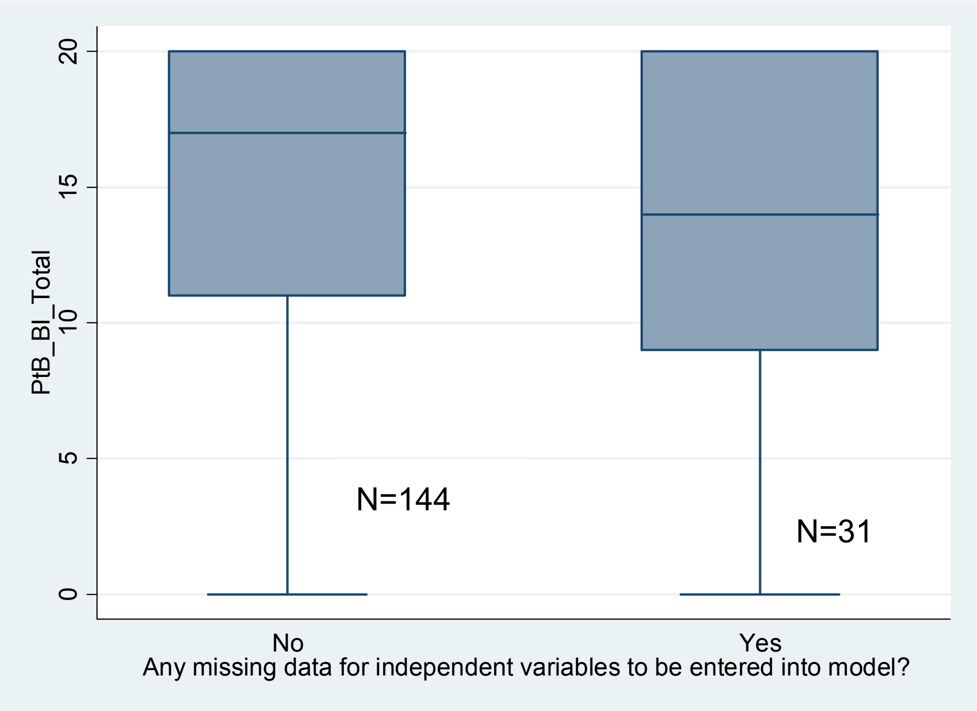
A variable was created to distinguish whether a patient was admitted to a ward for hyperacute stroke care (stroke unit (SU), coronary care unit (CCU), high dependency unit (HDU), or intensive care unit (ICU)) or to a medical admissions unit or general medical ward. The reasoning for the creation of this variable is discussed in section 5.9.3.1.

5.9.2.1 Effective sample size for Model 1

An EPV calculation has been performed using the number of completed outcomes There are 176 completed SIPSO physical subscore questionnaires, therefore 17 variables (including dummy variables) may be entered into the models in order to achieve 10 events per variable as recommended by Peduzzi et al (Peduzzi P et al 1996).

Figure 62 demonstrates that the baseline Barthel Index was lower in patients automatically excluded from the Model 1 due to incomplete predictor variables than in patients included in the model. However, a Mann-Whitney U test shows this difference is not statistically significant (p=0.555).

Figure 62 Difference in baseline Barthel Index for patients with complete and incomplete independent variables selected for entering into regression model to predict physical subscore of SIPSO (only cases with complete physical SIPSO shown).



Model 1 Linear regression of predictor variables (Table 51) on physical subscore of SIPSO, logarithmic transformation of length of stay

					R^2 = 0.54 Adj R^2 =0.52 N° Obs =145 F =31.99 P>|F|<0.001	
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Baseline NEADL	0.13	0.024	5.35	<0.001	0.08	0.17
Baseline EQ5D	4.43	1.23	3.58	<0.001	1.99	6.87
Log (length of stay +1)	-1.39	0.41	-3.40	0.001	-2.20	-0.58
Discharged to same address	3.73	1.24	3.00	0.003	1.27	6.19
Constant	3.92	2.14	1.83	0.069	-0.31	8.15

Propensity score has been automatically removed from this model, therefore the model has been re-run with age at stroke entered as an independent variable as it is likely to be an important predictor (Model 2), and is used to calculate the SSV.

Model 2 Linear regression of predictor variables (Table 51) on physical subscore of SIPSO, including age at stroke

					R^2 = 0.54 Adj R^2 =0.52 N^o Obs =145 F =31.99 P>|F|<0.001	
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Baseline NEADL	0.11	0.024	4.72	<0.001	0.07	0.16
Baseline EQ5D	4.73	1.23	3.86	<0.001	2.31	7.16
Log (length of stay +1)	-1.27	0.41	-3.10	0.002	-2.07	-0.46
Discharged to same address	3.49	1.23	2.84	0.005	1.06	5.92
Age at stroke onset	-0.07	0.33	-2.28	0.024	-0.14	-0.01
Constant	9.73	3.31	2.94	0.004	3.19	16.27

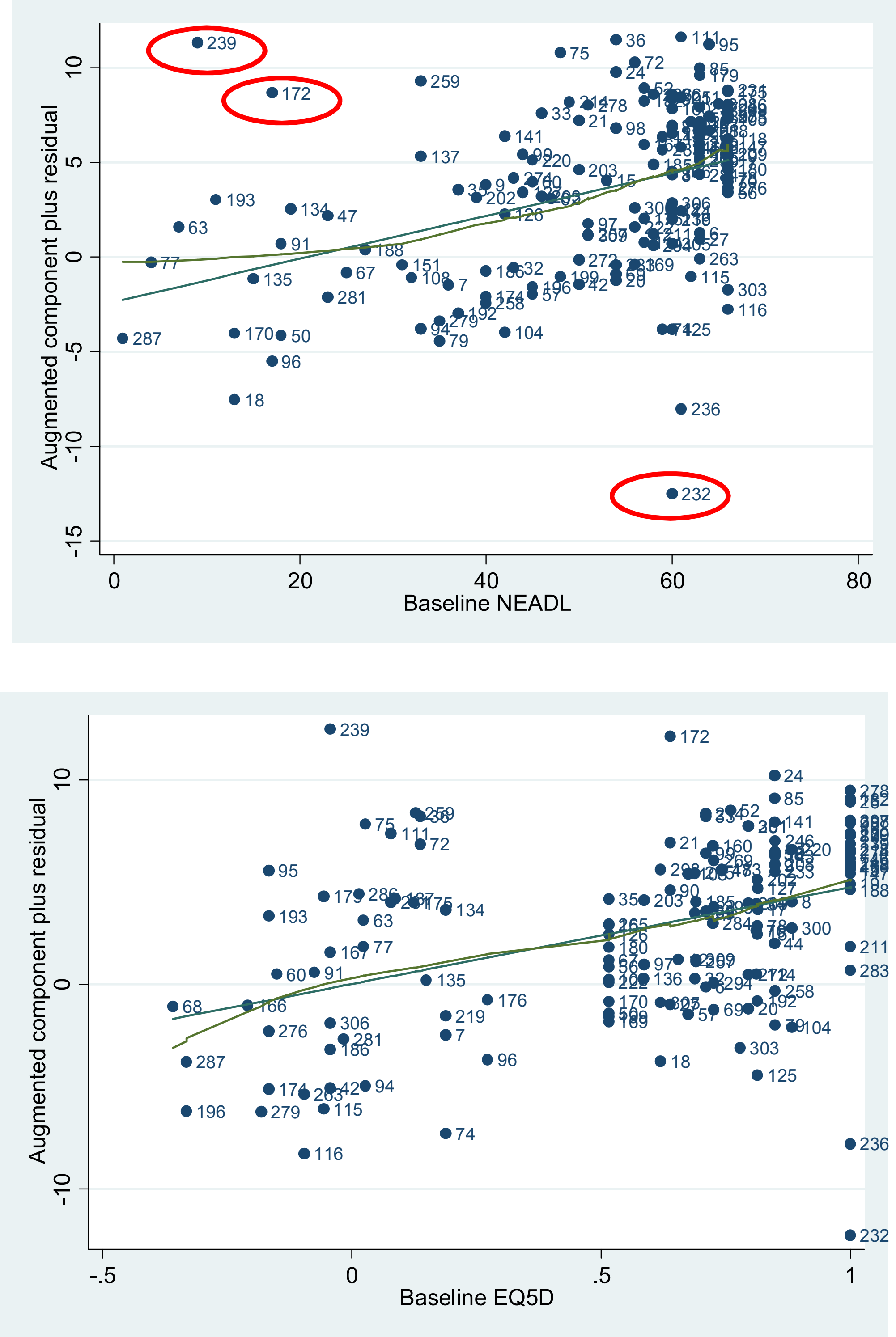
Age reaches statistical significance in this model and is therefore retained.

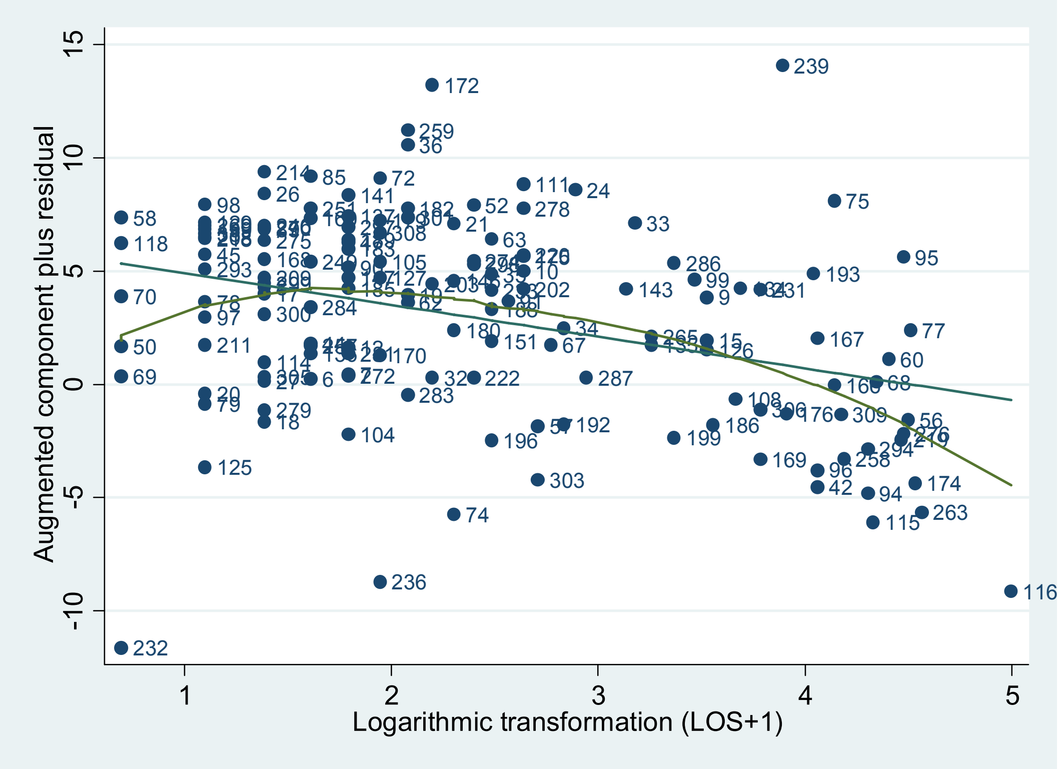
5.9.2.2 Post estimation checks

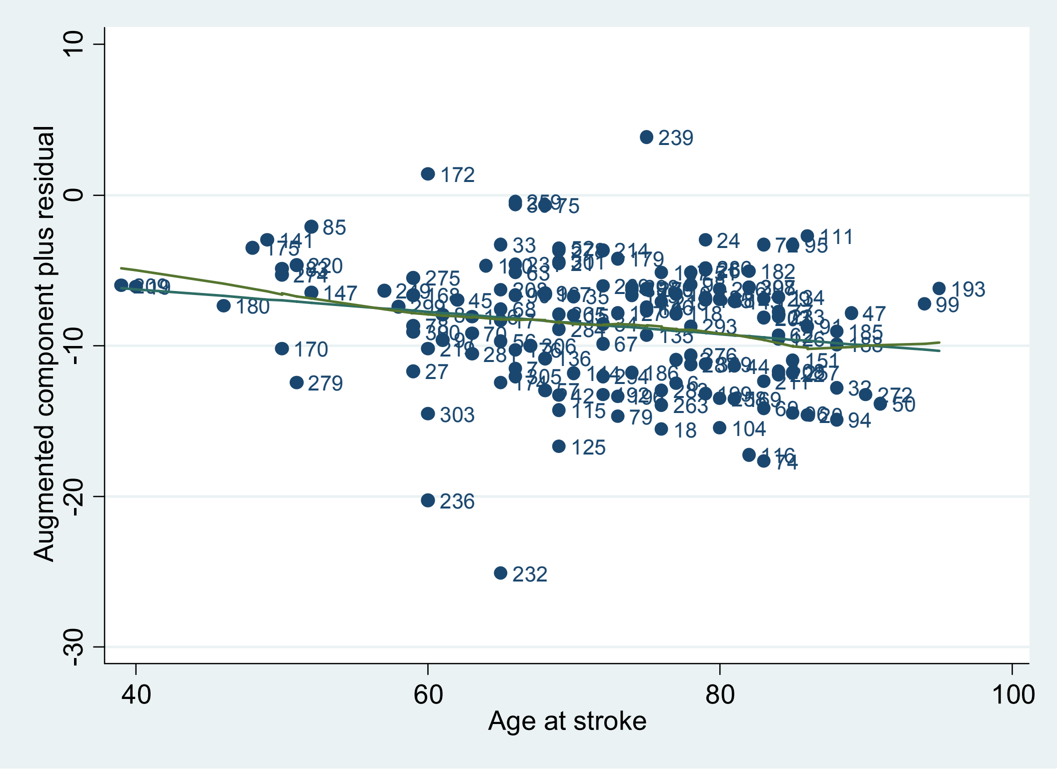
5.9.2.2.1 Linearity

Linearity assumptions were checked for continuous baseline predictors in the model. There are apparently non-linear relationships between the logarithmic transformation of length of stay and physical SIPSO subscore identified through augmented component plus residual plots (see section 4.4.6.1.4) (Figure 63). However, the estimates from a model where length of stay is divided into six categories and treated as a categorical variable data does not differ significantly from a linear prediction of physical SIPSO with length of stay treated as a continuous variable (likelihood ratio test - see appendix E-1.2).

Figure 63 Augmented component plus residual plots (acprplot) for continuous independent predictors in Model 2 (Baseline NEADL, baseline EQ5D, length of stay and age at stroke)



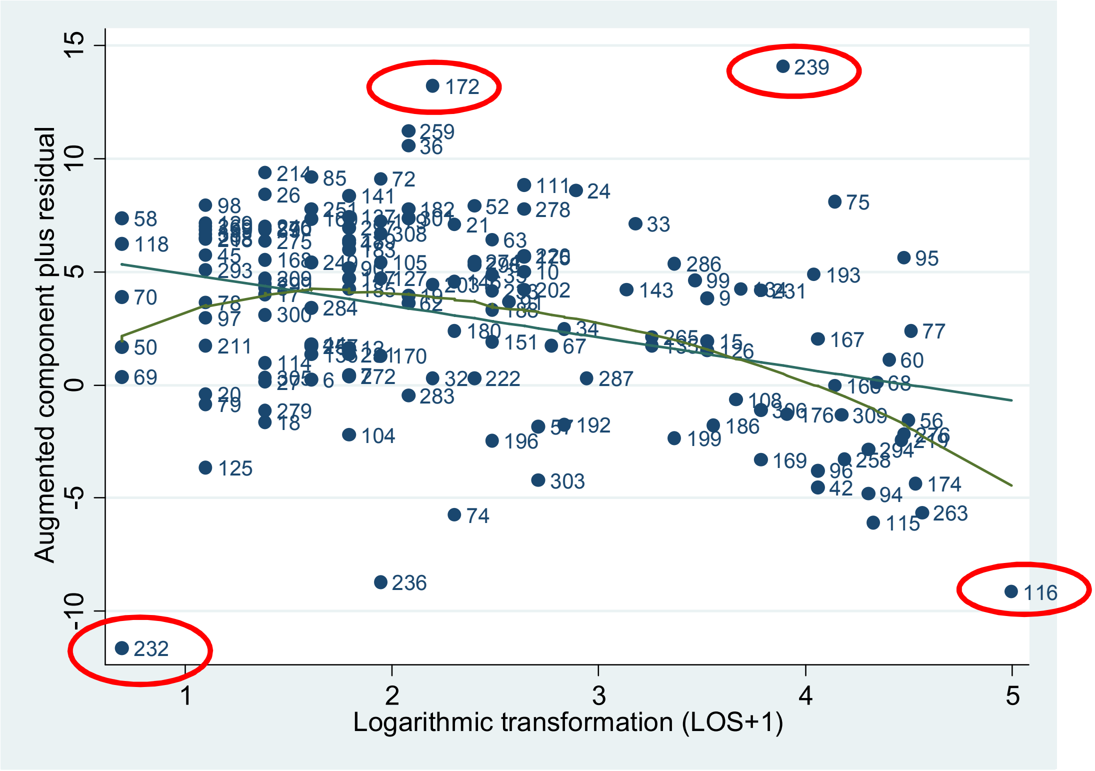

Augmented component plus residual
15
10
5
0
-5
-10
1
2
3
4
5
Logarithmic transformation (LOS+1)


Augmented component plus residual
10
0
-10
-20
-30
40
60
80
100
Age at stroke

5.9.2.2.2 *Influential cases and leverage*

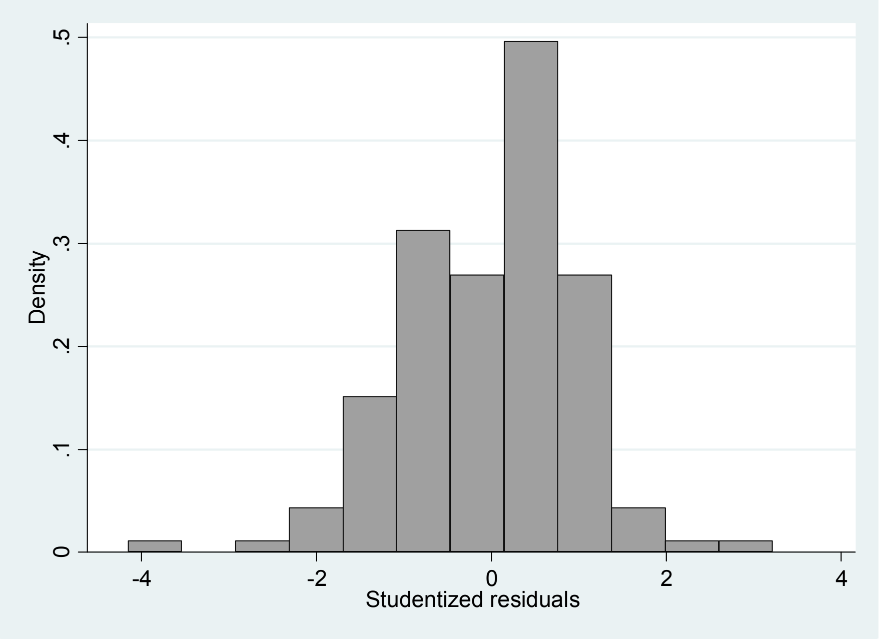
The deviation from linearity for both transformed length of stay and baseline NEADL seen on acprplots may be due to outlying cases with particularly large residuals at high leverage points (e.g. those highlighted in Figure 64 below).

Figure 64 Augmented component residual plots labelled by study number to identify likely cause of deviation from linearity



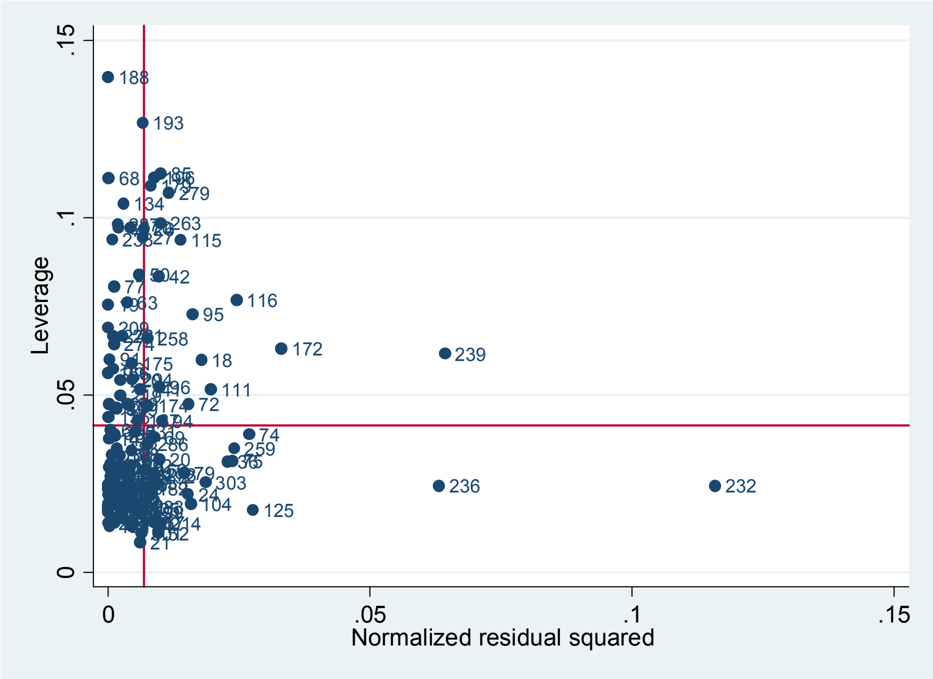
A histogram of studentised residuals for Model 2 reveals a few extreme outliers where observed outcome is different from that predicted from the model. Examination of the cases with studentised residuals ≥|3| reveals these to be cases 239 and 232.

Figure 65 Histogram of studentised residuals (Model 2)



When the augmented component residual plots for transformed length of stay and baseline NEADL are re-examined and individual points labelled (Figure 63), it can be seen that some of the cases identified in Figure 66 as likely to be exerting undue influence on the whole model may also be contributing to the deviations from linearity, i.e. these points may be distorting the relationship between individual covariates and the dependent variable.

Figure 66 Leverage vs. r squared plot for prediction of physical subscore of the SIPSO

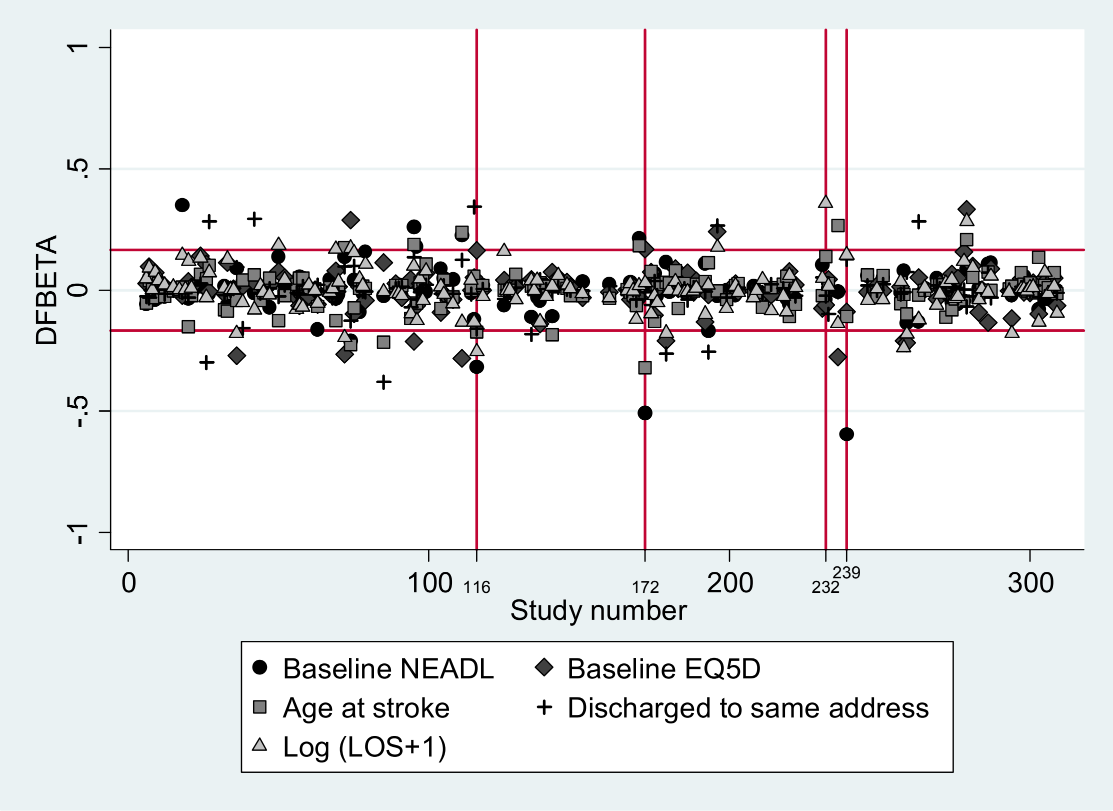


There are a number of cases of concern, study numbers 239, 172 and 116 are all outliers in both r-squared value and leverage and are therefore likely to exert influence on the model regression coefficient (i.e. the slope of the fitted regression line). Troublesome cases with a Cooks D statistic of >4/n = 0.028 (n=145 for this model) are shown below (see section 4.4.6.1.4). It can be seen that, as predicted, cases 239, 232, 172, 239 and 116 are of concern.

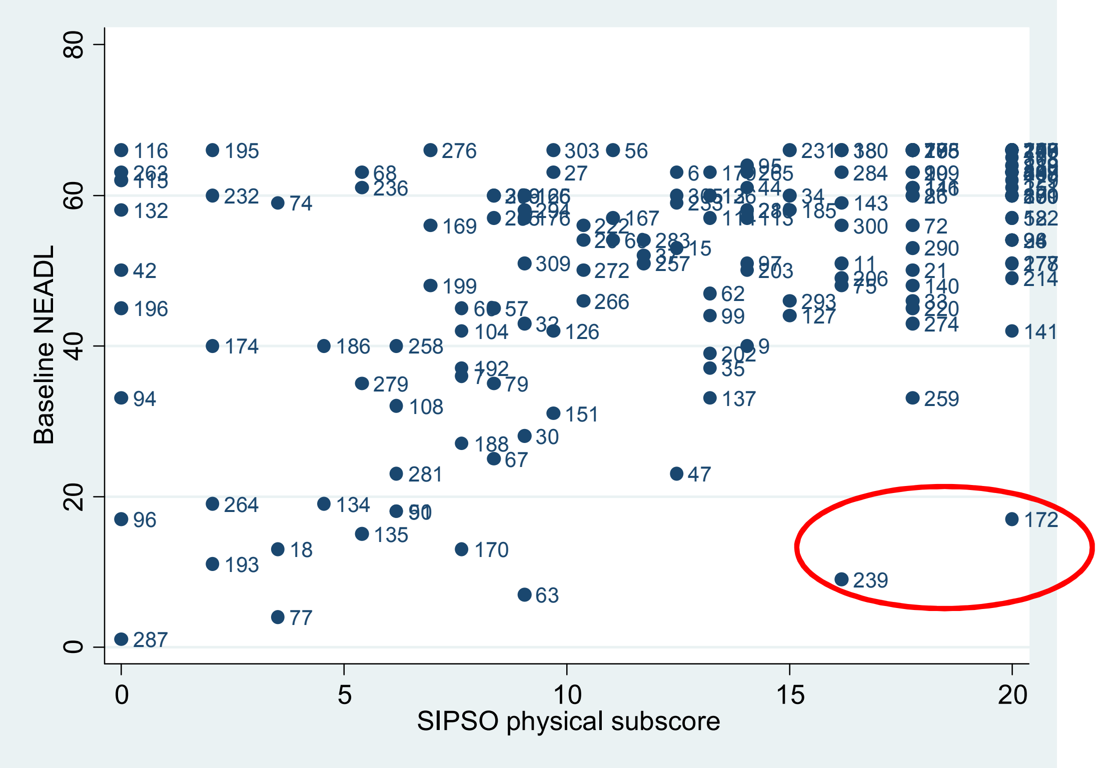
Study number	Cooks D
239	0.104
232	0.069
172	0.055
116	0.051
236	0.038

Examination of DFBETA for the independent predictors reveals that the cases identified on an acprplot as possible contributors to non-linearity between the baseline NEADL and physical SIPSO subscore are also particularly influential on the beta coefficient for the baseline NEADL in the model (identified with black circles on Figure 67 i.e. cases 239, 172). Similarly, for the length of stay (grey triangles on Figure 67), it is the points identified on the acprplot that appear to be exerting undue influence.

Figure 67 Scatter plot of DFBeta for independent variables across study numbers (prediction of physical subscore of the SIPSO with baseline assessments)



However, although none of these cases are extreme outliers in terms of their individual values for either baseline NEADL (172,239) or length of stay (232, 116), they represent cases with particularly good or poor outcome relative to that which would be predicted on the basis of the independent variable (i.e. they have high residuals). This can be seen from a simple scatter plot of the predictor against the SIPSO physical score. For example, cases 172 and 239 represent individuals that have much better physical scores at 6 months than would have been expected on the basis of their baseline NEADL scores.



Removing the 4 particularly influential cases from the regression model (232, 239, 172 and 116) makes a large difference to the R squared (variance explained) of the model and increases the proportion of the total variance that is explained by the model as opposed to the model residuals (Model 3), but this is to be expected as some of the residual variation has been artificially removed. However, omission of the influential cases does not alter which of the variables reach statistical significance in the model, nor are there large differences to the size or polarity of the beta co-efficients. For the purposes of identification of important predictor variables for the development of a dataset, the influential cases can remain in the model as small changes to the values of the individual beta co-efficients is of little importance. Moreover, removing the influential cases is likely to overfit the model as the identified cases do not represent ‘wrong’ data, but cases better or worse outcomes are observed than predicted with the model (Fox J, 1997 p 286). It may be useful, therefore, to examine these cases qualitatively to identify any salient features of management that may warrant further exploration.

Model 3 Linear regression of predictor variables (Table 51) on physical subscore of SIPSO, with influential cases removed

					R^2 = 0.63 Adj R^2 =0.62 N° Obs =141 F =45.85 P>|F|<0.001	
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Baseline NEADL	0.15	0.022	6.71	<0.001	0.11	0.19
Baseline EQ5D	4.50	1.08	4.16	<0.001	2.36	6.64
Log (length of stay +1)	-1.37	0.36	-3.79	<0.001	-2.09	-0.66
Discharged to same address	3.57	1.08	3.30	0.001	1.42	5.71
Age at stroke onset	-0.059	0.029	-2.00	0.05	-0.12	-0.001
Constant	7.05	3.00	2.35	0.02	1.12	12.99

5.9.2.2.3 Normality assumptions

Normality of residuals for the full model (with transformed length of stay and age at stroke included) (Model 2) is assessed through examination of standardised normal probability plots. There is some deviation from normality at the extremes of the distribution of the residuals (which may reflect the influence of the cases already highlighted). However, this deviation does not look too serious.

Figure 68 Normal probability (Q-Q plot) for Model 2

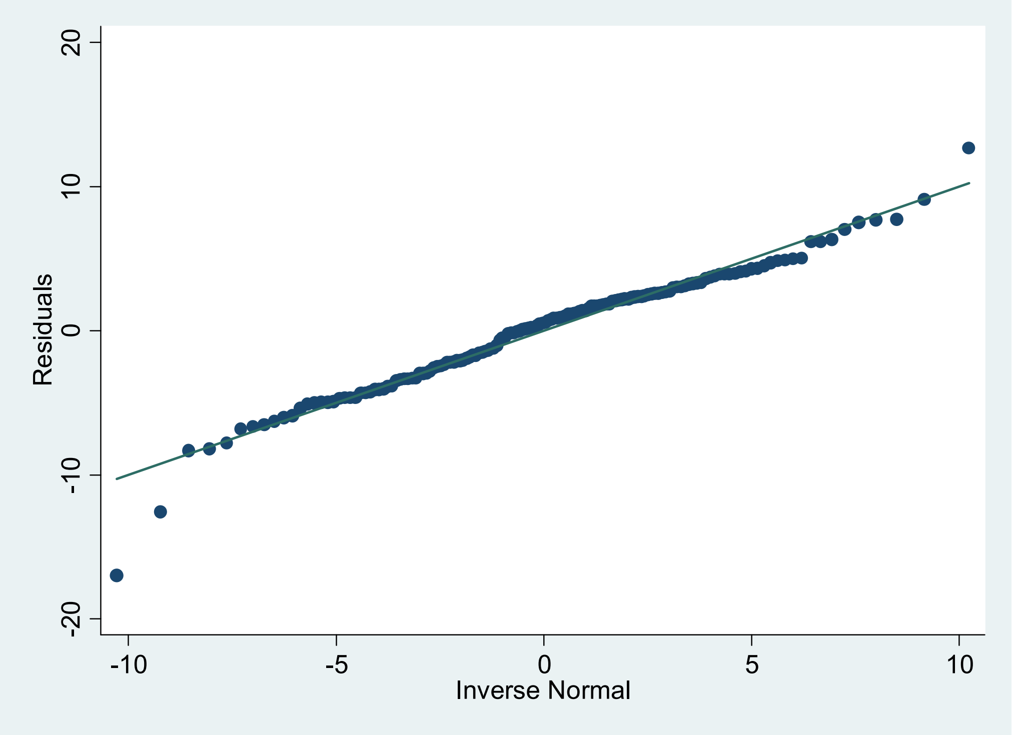
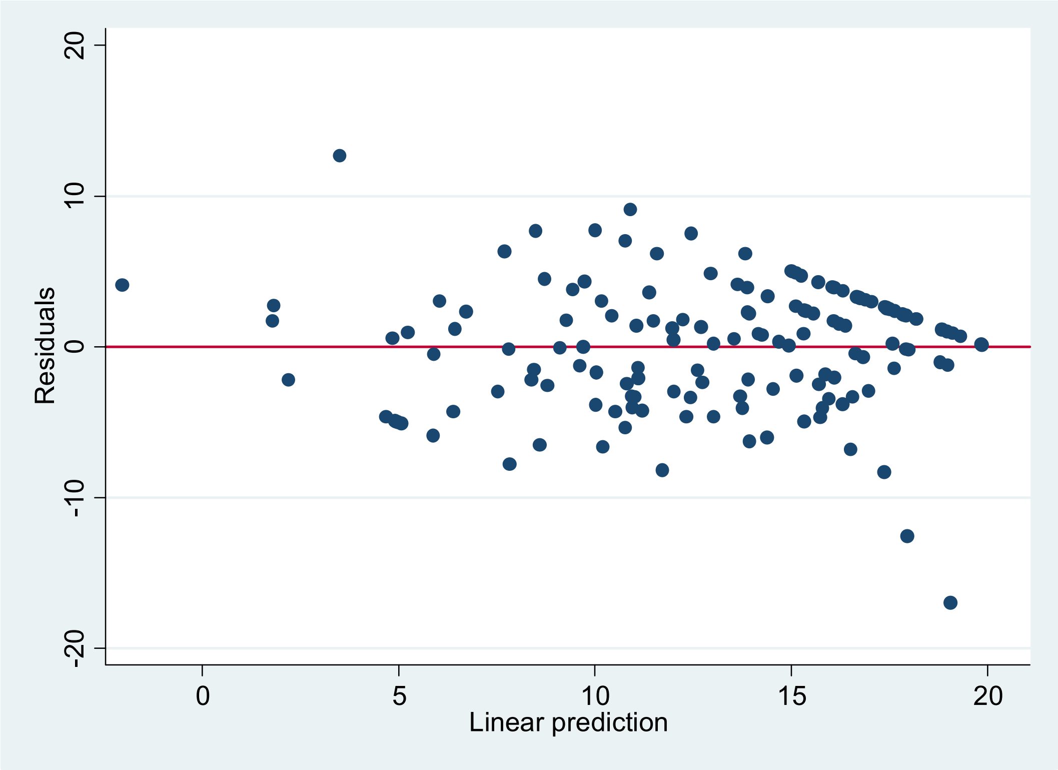


Figure 69 Scatter plot of fitted values vs. residuals demonstrating homoscedasticity for Model 2



There is no pattern in a plot of residuals against fitted values, so the variance of the model is assumed to be homoscedastic. Formal diagnostics reveal that the null hypothesis of constant variance can be accepted (p=0.46).

5.9.2.2.4 Tests for collinearity in Model 2

Variance inflation factors (VIF) were examined to identify any collinearity between predictor variables (see section 4.4.6.1.4). The variance inflation factors were all less than ten and therefore do not imply that there is any collinearity between the predictor variables. However, this is not surprising, as the model was constructed using stepwise variable selection such that collinear variables are automatically excluded from the model. A table of variance inflation factors is provided for Model 2 in appendix E-1.3; however these have not been repeated for subsequent models.

5.9.3 Prediction of physical subscore of SIPSO, no baseline assessments

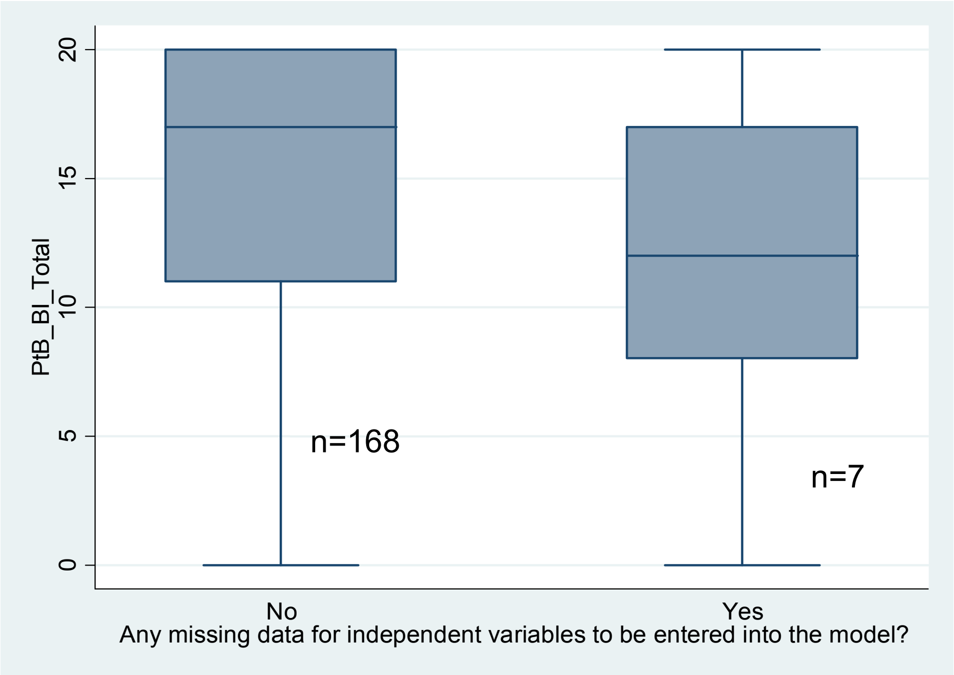
The variables identified through regression trees, univariate analysis and clinical reasoning for the prediction of physical SIPSO without baseline assessments are provided in Table 52.

Table 52 Independent variables to be entered into regression model to predict physical subscore of the SIPSO

	Variables	Number of variables (including dummy variables)
Variables identified from regression trees	Length of stay	**1**
	Propensity score	**1**
	Old stroke	**1**
	Ward type	**7**
	Communication assessment	**2**
Variables identified from univariate analysis	Formal swallowing assessment	**2**
	Social Worker assessment	**2**
	Urinary incontinence care plan	**2**
Probable important variables through clinical reasoning	tPA given	**2**
	Discharge to same address	**1**
	Age if propensity score removed	**0**
	Baseline Barthel Index	**1**
Total		**22**

Boxplots of baseline BI in patients with, and without complete predictor variables reveal that patients automatically excluded from the model to predict physical SIPSO subscore (without the baseline assessments) are more disabled than those in whom data are complete. However, this difference does not reach statistical significance in a Mann-Whitney U test (p=0.187)

Figure 70 Difference in baseline BI between patients with complete and incomplete model variables

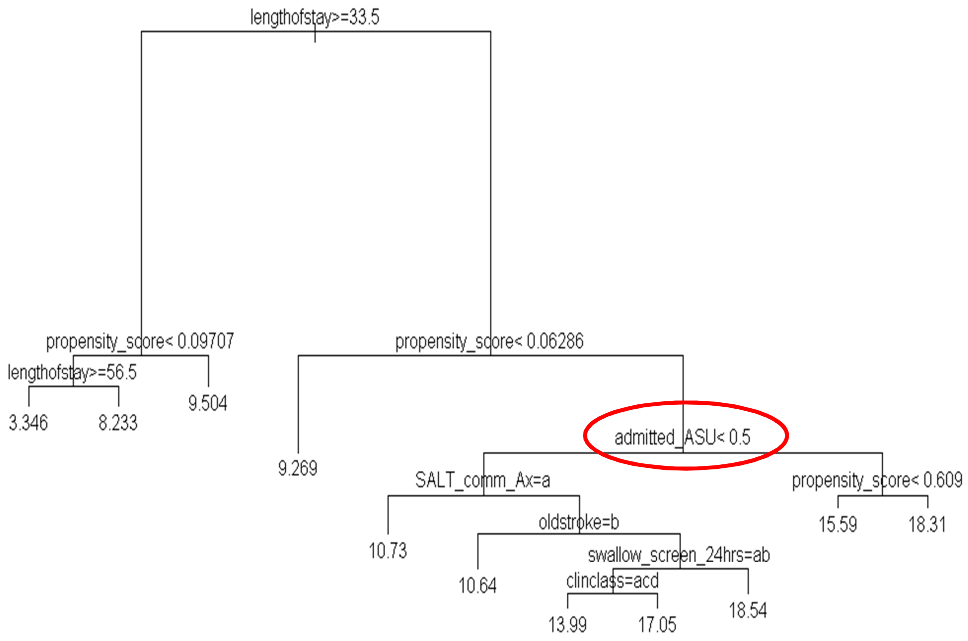


5.9.3.1 Sample size

If the predictors outlined in Table 52 were to be entered into the models as shown, the number of variables to be entered into this model would exceed 10 EPV, largely due to the large number of dummy variables associated with ward type. Examination of the regression trees and dummy variables for ward type reveals that the significant predictor is whether patients were admitted to an environment with facilities to provide hyperacute stroke care (coronary care unit (CCU), high dependency unit (HDU)/ intensive care unit (ITU), or acute stroke unit (ASU)). Although treatment on CCU appears alongside 'other ward' or 'medical admissions unit' in the regression tree to predict physical SIPSO without baseline assessments, this represents only one patient in the study. CCU has therefore been grouped with the other wards where hyperacute stroke care may be provided. The 'ward type' variable has therefore been re-categorised to reflect the receipt of specialist acute stroke care or not by combining CCU, HDU, ITU and ASU care into one category, and admissions unit and general ward care into another.

Entering the new variable (admitted ASU) into a regression tree reveals that it remains an important predictor of the physical subscore of the SIPSO Figure 71.

Figure 71 Regression tree to demonstrate prominence of composite variable of direct admission to a unit providing hyperacute stroke care on prediction of physical outcome



The number of variables entered into the model can therefore be reduced to 16. Length of stay will again be entered into the model as a logarithmic transformation.

Model 4 Linear regression of predictor variables (Table 52) on physical subscore of SIPSO, without patient reported baseline assessments

R^2 = 0.45
Adj R^2 =0.43
N^{o} Obs =167
F =18.53
P>\|F\|<0.001

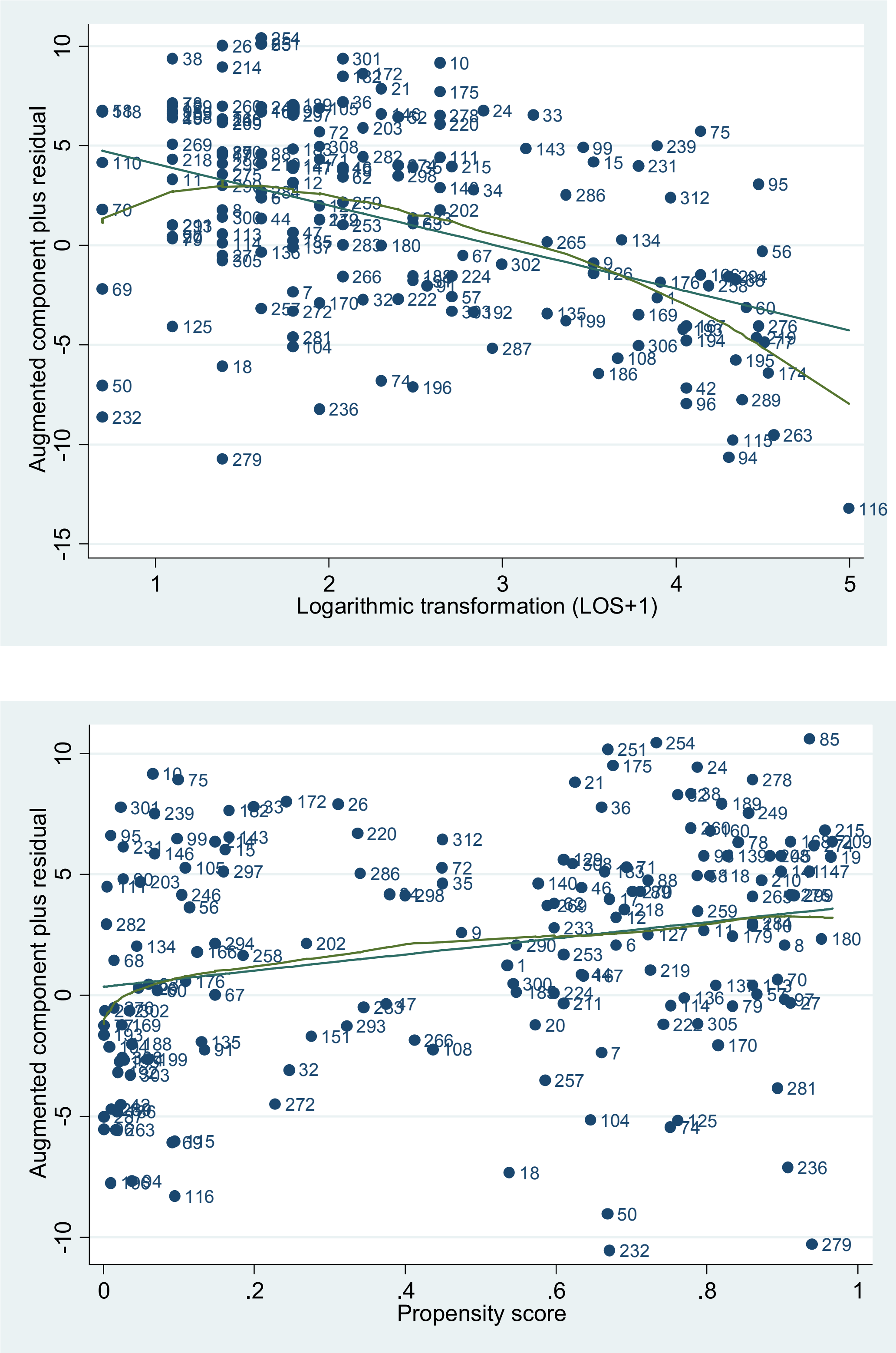
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Log (length of stay + 1)	-2.11	0.49	-4.92	<0.001	-2.96	-1.26
Propensity score (SSV)	3.18	1.43	2.22	0.028	-0.35	6.01
Admitted to hyperacute bed	2.11	0.78	2.75	0.007	0.60	3.62
SLT communication Ax “yes”	2.87	1.10	2.61	0.01	0.70	5.04
SLT communication Ax “no but”	2.76	0.97	2.85	0.005	0.84	4.67
Previous stroke	-2.61	1.06	-2.48	0.014	-4.70	-0.53
Discharged to same address	3.94	1.23	3.19	0.002	1.50	6.38
Constant	10.36	2.13	4.86	<0.001	6.16	14.57

5.9.3.2 Post estimation checks

5.9.3.2.1 Linearity assumptions

The augmented partial residual plots (acprplot) for the propensity score appears to be influenced by some points at the extremes, but this deviation is not serious. As before, there is deviation on the acprplot for length of stay which is likely to reflect the influential points at the extremes of the distribution. Linearity assumptions are therefore assumed to have been met.

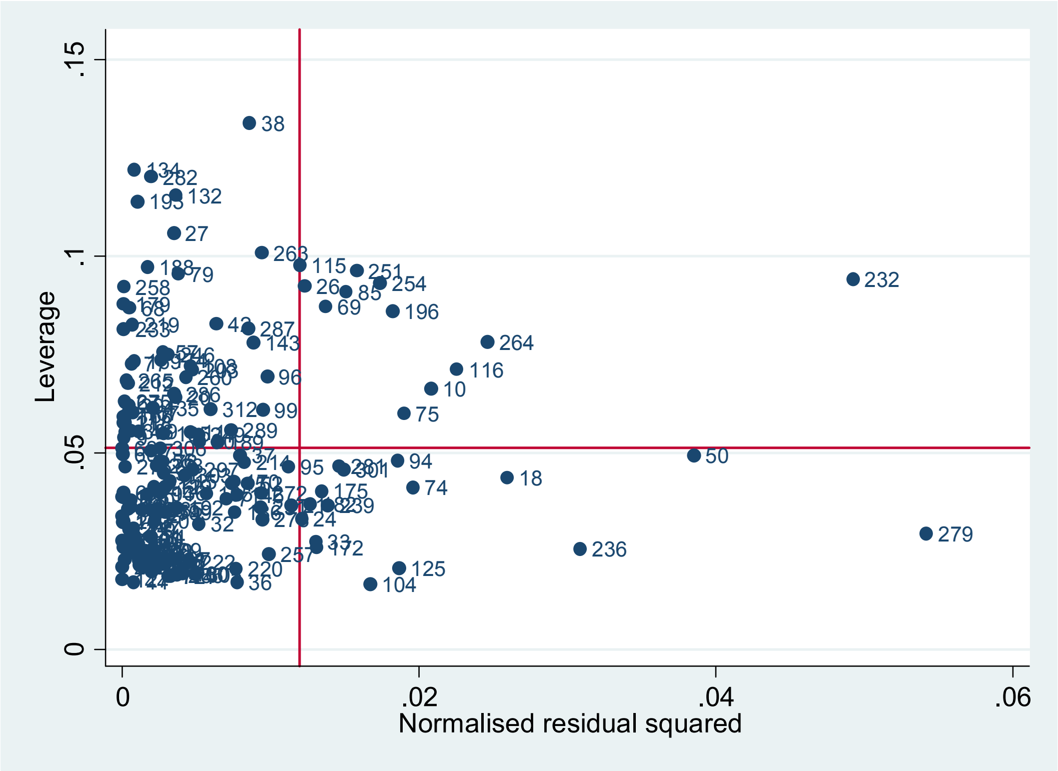
Figure 72 Augmented component plus residual plots for continuous predictors in Model 4



5.9.3.2.2 Leverage

The most influential cases in this model would be expected to be case 232, 264, 116, and 196. Examination of Cook's D statistic (as discussed in 4.4.6.1.4e) confirms the most influential points to be 232, 264, 50, 254, and 196.

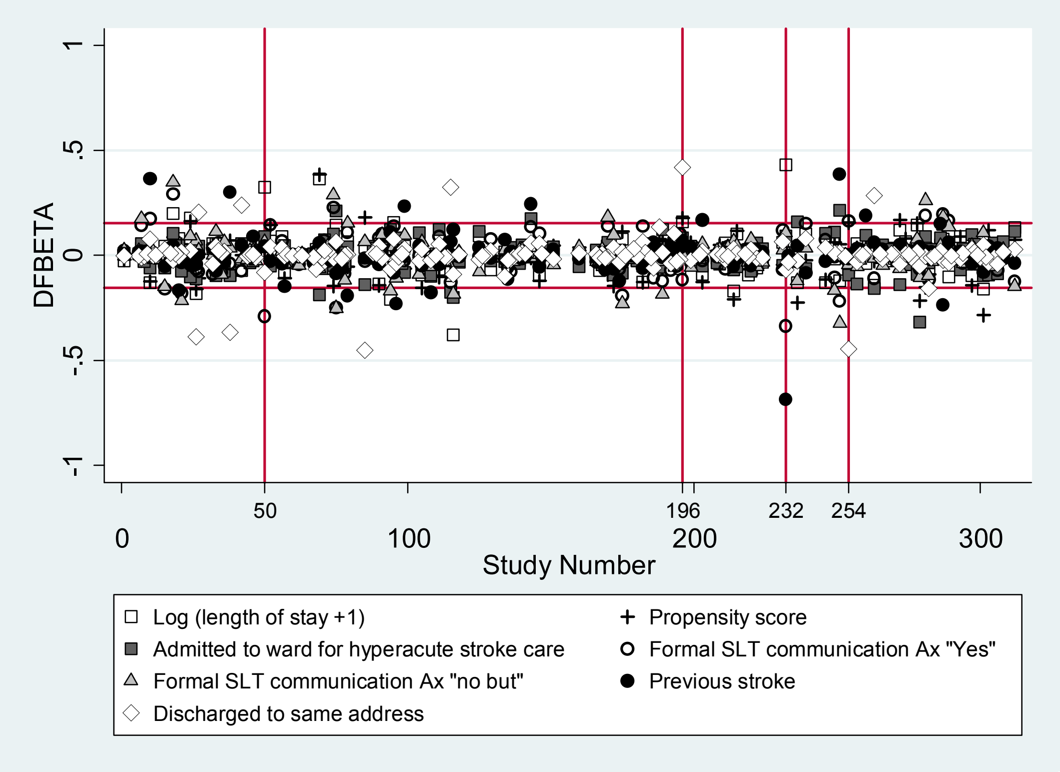
Figure 73 Leverage vs. r-squared plot for Model 4



Study number	Cooks D
232	0.12
264	0.047
50	0.043
254	0.041
196	0.039

The effect of individual cases on the beta co-efficients of each individual variable is shown in (Figure 74). It can be seen that the influential cases are similar to those that exert influence on the regression coefficient of the whole model.

Figure 74	DFBeta for independent variables across individual cases (prediction of physical subscore of SIPSO without baseline assessments) - Vertical lines identify potentially influential cases



Removing the influential cases and re-performing the regression analysis removes previous stroke as an important predictors of outcome. This may be due to the large DFBETA for case 232 for this variable. Again, this represents unexpected, but not spurious data and influential cases are retained for interpretation of the models.

Model 5 Linear regression of predictor variables (Table 52) on physical subscore of SIPSO, without baseline assessments and with influential cases removed

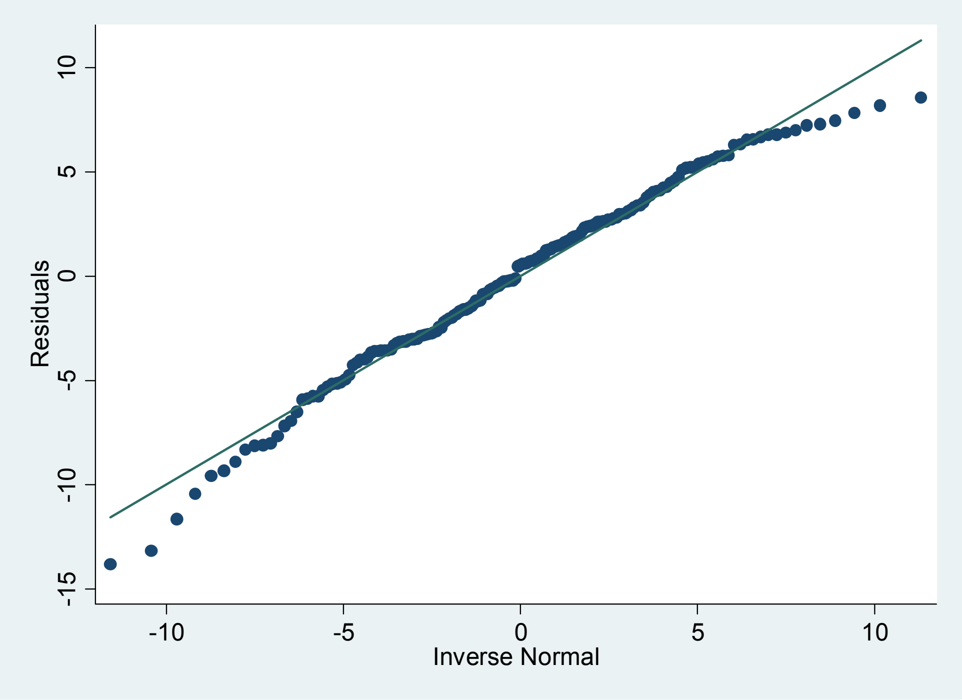
R^2 = 0.48
Adj R^2 =0.46
N° Obs =163
F =24.21
P>|F|<0.001

	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Log (length of stay + 1)	-2.56	0.42	-6.16	<0.001	-3.38	-1.74
Propensity score (SSV)	3.10	1.36	2.27	0.024	0.41	5.79
Admitted to hyperacute bed	1.98	0.73	2.71	0.01	0.54	3.42
SLT communication Ax "yes"	3.57	1.06	3.36	0.001	1.47	5.67
SLT communication Ax "no but"	2.53	0.92	2.76	0.007	0.72	4.34
Previous stroke						
Discharged to same address	4.01	1.24	3.24	0.001	1.56	6.46
Constant	11.21	2.08	5.37	<0.001	7.09	15.34

5.9.3.2.3 Normality of residuals

A standardised probability normal plot (for Model 4, with influential cases included) suggests there is deviation of the residuals from normality at the extremes of the distribution. This is confirmed on Shapiro-Wilk testing (z=1.89, p=0.03).

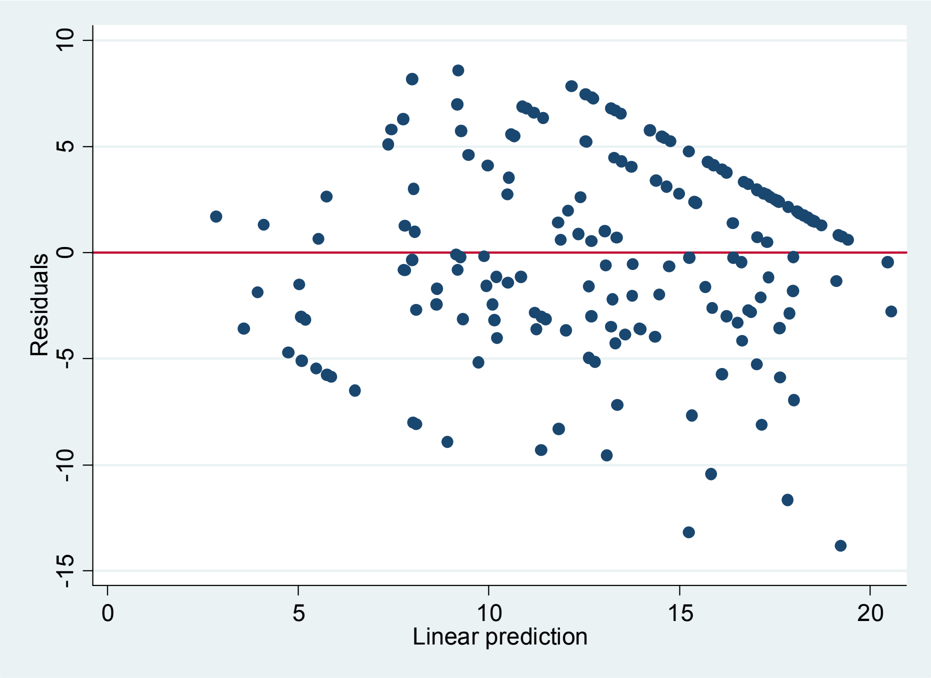
Figure 75 Normal probability plot for model residuals (Model 4)



5.9.3.2.4 *Homoscedasticity*

A plot of residuals against fitted values does suggest a downward linear trend in the data. This is again likely to represent a missing variable from the model. Diagnostics to assess for heteroscedasticity, however, fail to reach significance such that the null hypothesis of homogeneity of variance is accepted.

Figure 76 Scatter plot of model residuals against fitted values to identify heteroscedasticity (Model 4)



5.9.4 Prediction of social subscore of the SIPSO with baseline assessments

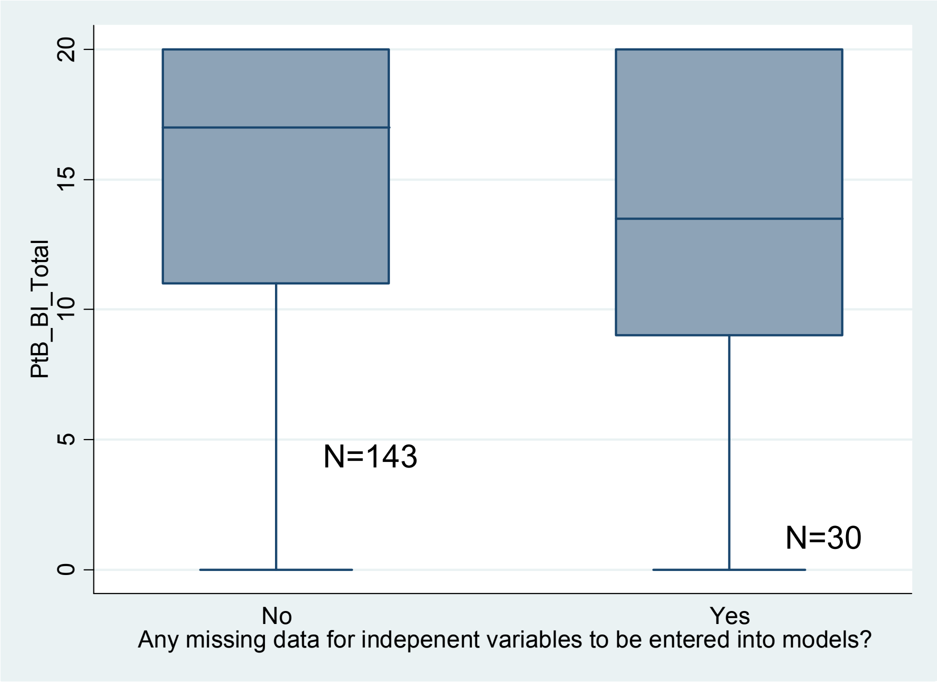
Variables to be entered into a model to predict the social subscore of the SIPSO (with baseline assessments) are shown in Table 53.

Table 53 Variables to be entered into regression model to predict social subscore (with baseline assessments)

	Variables	Number of variables
Variables identified from regression trees	Baseline EQ5D	**1**
	Baseline NEADL	**1**
	SALT communication assessment	**2**
	Length of stay	**1**
Variables identified from univariate analysis	Formal swallowing assessment	**1**
	Urinary incontinence care plan	**2**
Probable important variables through clinical reasoning	Propensity score (or age if propensity score excluded)	**1**
	First ward admitted to	**1**
	Discharged same address	**1**
	tPA given	**2**
	Social Worker assessment	**2**
Total		**15**

There were 144 completed and returned SIPSO subscore questionnaires, therefore entering these 15 variables into a model gives an EPV of 9.6.

Figure 77 Difference in baseline BI between patients with complete and incomplete model variables to predict social subscore of SIPSO



The apparent difference in the baseline BI between patients who are included in the models (complete predictor variables) vs. those who are automatically excluded (incomplete predictor variables) is demonstrated not to be statistically significant on Mann-Whitney U testing (p=0.504)

Model 6 Linear regression of predictor variables (Table 52) on social subscore of SIPSO (with baseline assessments)

					R^2 =0.40 Adj R^2 =0.39 N^o Obs =144 F =23.59 P>|F|<0.001	
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Baseline EQ5D	3.53	1.05	3.37	0.001	1.46	5.61
Baseline NEADL	0.06	0.02	2.89	0.004	0.02	0.10
Log (length of stay +1)	-0.81	0.37	-2.21	0.029	-1.53	-0.086
SLT communication Ax “no but”	2.11	0.65	3.22	0.02	0.81	3.40
Constant	7.78	1.67	4.65	<0.001	1.42	11.1

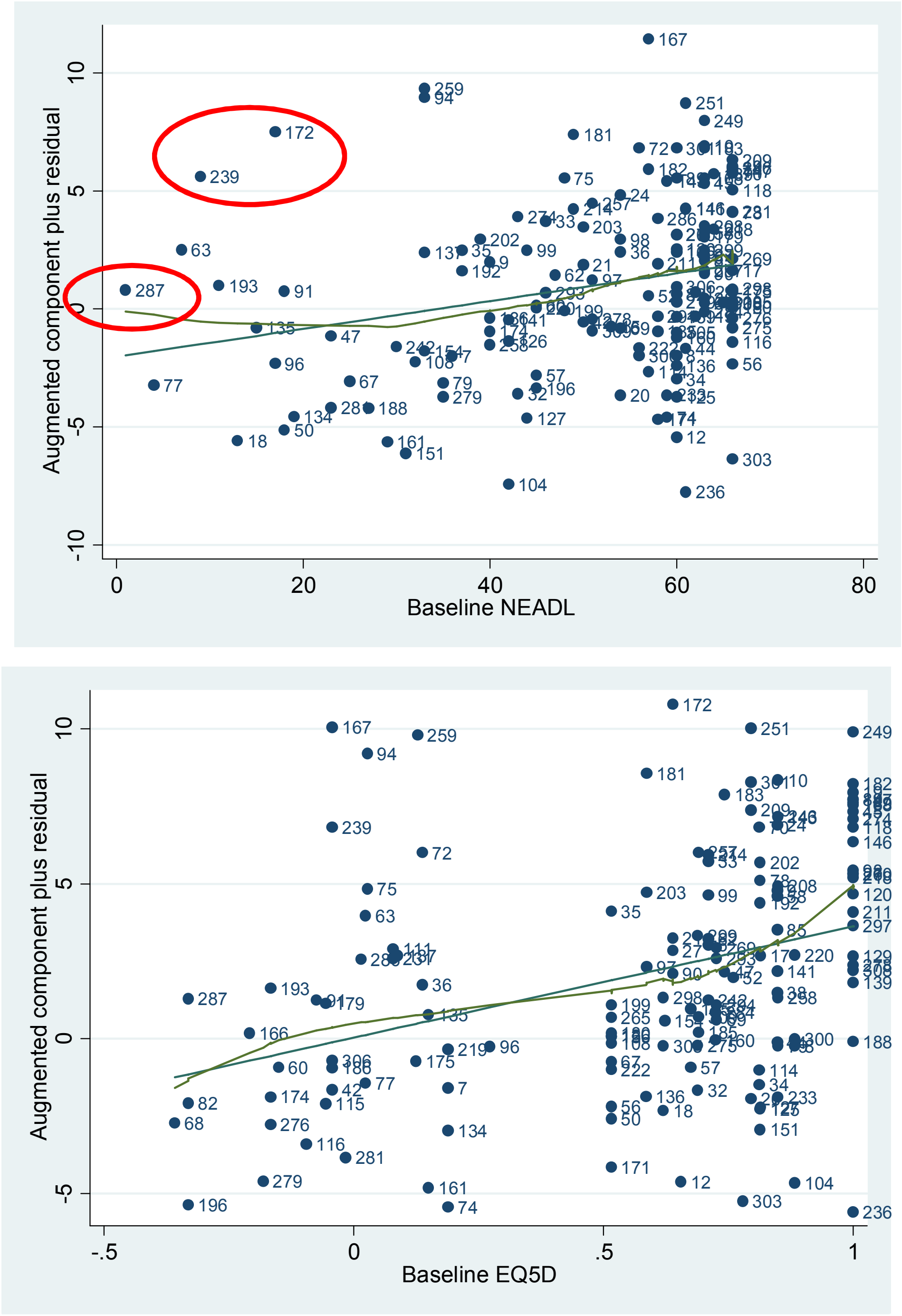
The propensity score again fails to reach significance in this model. Substitution for age instead of propensity also fails to reach significance, and make no real difference to either the beta-coefficients or total fit of the model. Propensity score is therefore retained in the model.

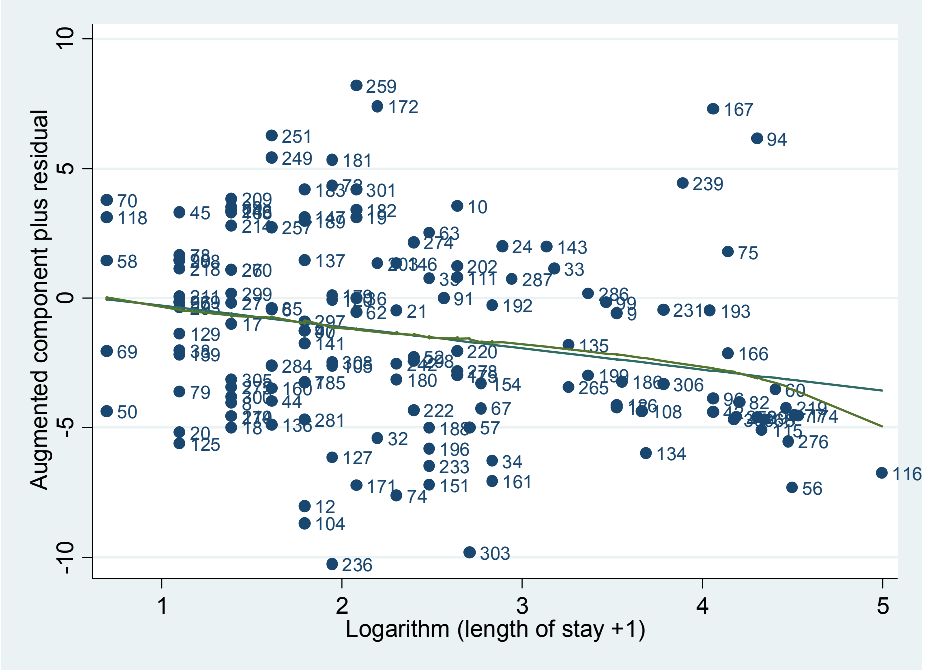
5.9.4.1 Post estimation checks

5.9.4.1.1 Linearity

Augmented component plus residual plots reveal that the largest deviation from linearity occurs with the baseline NEADL.

Figure 78 Augmented component plus residual plots of continuous variables entered into Model 6 (baseline NEADL, baseline EQ5D and length of stay)





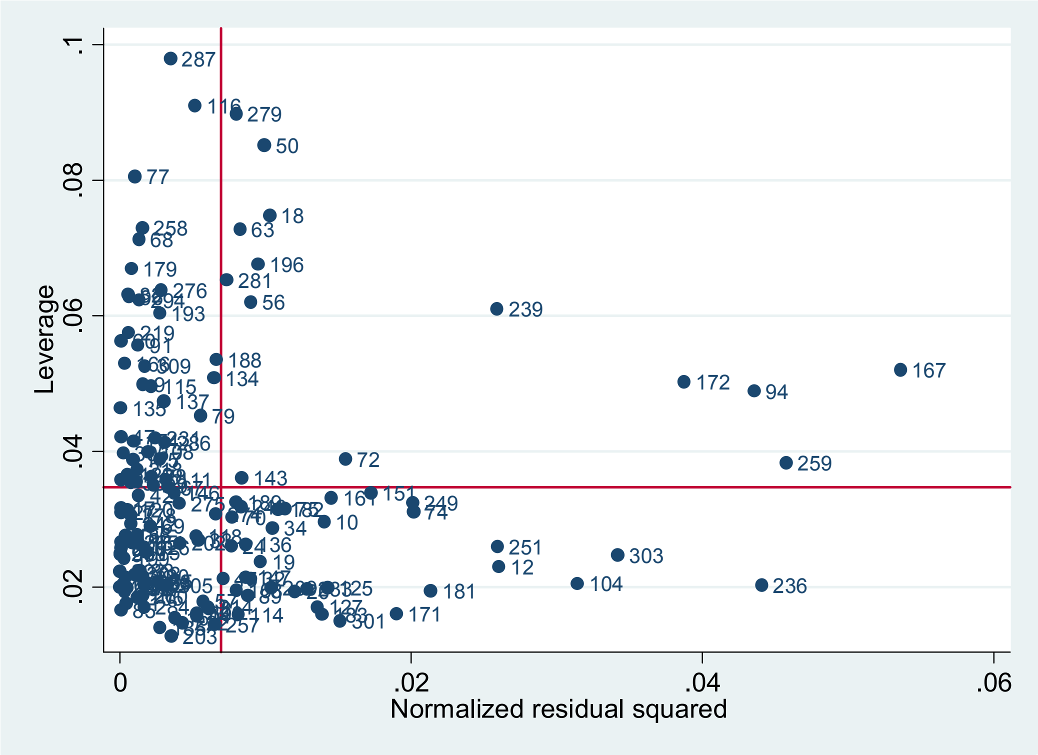
A likelihood ratio test for improved model fit using a continuous vs. a categorised NEADL variable fails to reach significance such that it can be assumed that categorising the NEADL will not significantly improve the model (p=0.332 working not shown).

As with the prediction of the physical subscore of the SIPSO, it is possible that some influential points are ‘pulling’ the fitted line away from linearity (large residuals). These are highlighted on the acprplots (Figure 78), and particularly influential points in terms of leverage will be explored further in the next section.

5.9.4.1.2 Influence and leverage

A plot of leverage against r squared for individual cases is shown below

Figure 79 Leverage against r squared Model 6



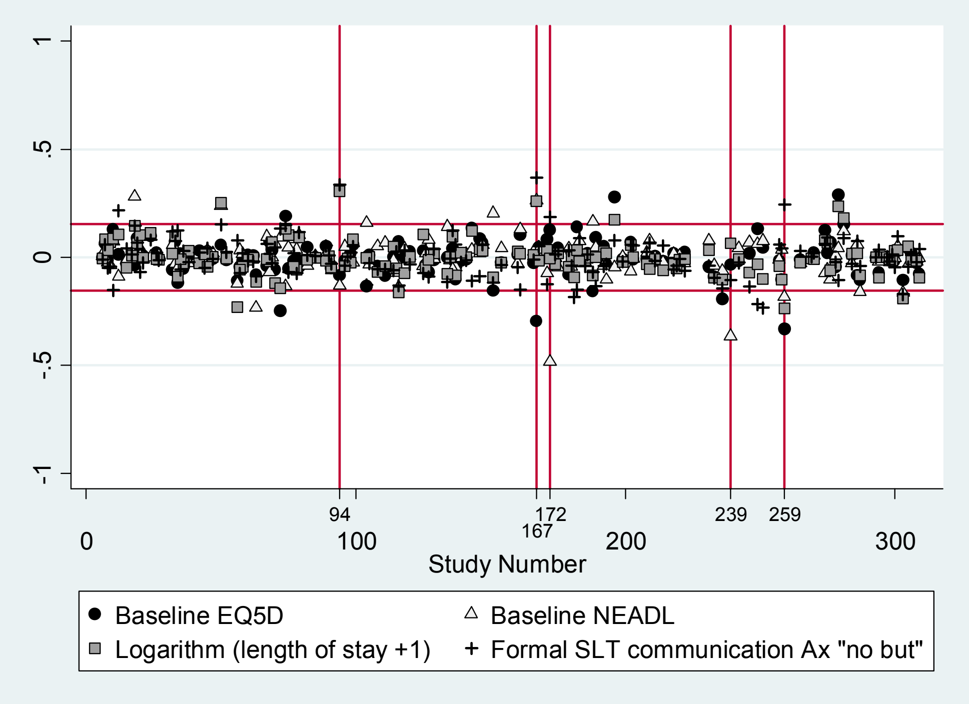
There are a few potentially influential cases highlighted on this plot. It is likely that cases 239, 172, 94, 167, 259 may be problematic.

Determination of Cook's D statistic (n=144) reveals that these are indeed the 5 most influential cases in terms of overall leverage.

Study number	Cooks D
167	0.09
94	0.07
172	0.06
259	0.05
239	0.05

A DFBETA plot (below) shows cases that are likely to be exerting undue influence on the beta coefficients for individual variables. Several cases are outwith the limits of 2/√144 (0.167). However, the cases which exert influence on individual variable beta coefficients (with high |DFBeta| values) tend to be the same cases as those which exert influence on the overall model regression co-efficient (high Cook's D).

Figure 80 Scatter plot of individual cases against DFBeta values for individual variables. Vertical lines represent cases likely to be of particular influence.



Re-running the regression models with cases with high Cook's D and with high DFBeta values removed (i.e. cases that are influential on the model regression coefficient and individual variable beta co-efficients) results in the propensity score entering the model reaching statistical significance at 0.05. The omission of this variable in the full model may therefore be due to the influential cases. However, as previously discussed, these cases represent unexpected, not spurious data and therefore are retained in the models to prevent overfitting.

Model 7 Linear regression of predictor variables (Table 52) on social subscore of SIPSO with influential cases removed

<table>
<tr><td colspan="5"></td><td colspan="2">R^2 =0.53
Adj R^2 =0.51
N^o Obs =139
F =29.45
P>|F|<0.001</td></tr>
<tr><td></td><td>Beta coefficient</td><td>Standard error</td><td>t</td><td>P>|t|</td><td colspan="2">95% confidence interval</td></tr>
<tr><td>Baseline EQ5D</td><td>4.48</td><td>0.93</td><td>4.82</td><td><0.001</td><td>2.64</td><td>6.32</td></tr>
<tr><td>Baseline NEADL</td><td>0.10</td><td>0.02</td><td>4.90</td><td><0.001</td><td>0.06</td><td>0.14</td></tr>
<tr><td>Log (length of stay +1)</td><td>-1.29</td><td>0.38</td><td>-3.39</td><td>0.001</td><td>-2.04</td><td>-0.54</td></tr>
<tr><td>SLT communication Ax "no but"</td><td>-1.86</td><td>0.70</td><td>-2.75</td><td>0.01</td><td>-3.21</td><td>-0.52</td></tr>
<tr><td>Propensity score</td><td>-2.25</td><td>1.13</td><td>-1.99</td><td>0.05</td><td>-4.49</td><td>-0.01</td></tr>
<tr><td>Constant</td><td>7.78</td><td>1.65</td><td>5.33</td><td><0.001</td><td>5.55</td><td>12.1</td></tr>
</table>

5.9.4.1.3 Normality assumptions (all cases included)

A normal probability plot is straight suggesting that residuals are normally distributed and this is confirmed on a Shapiro-Wilk test (z= 0.16, p=0.44)

Figure 81 Normal probability plot (Model 7)

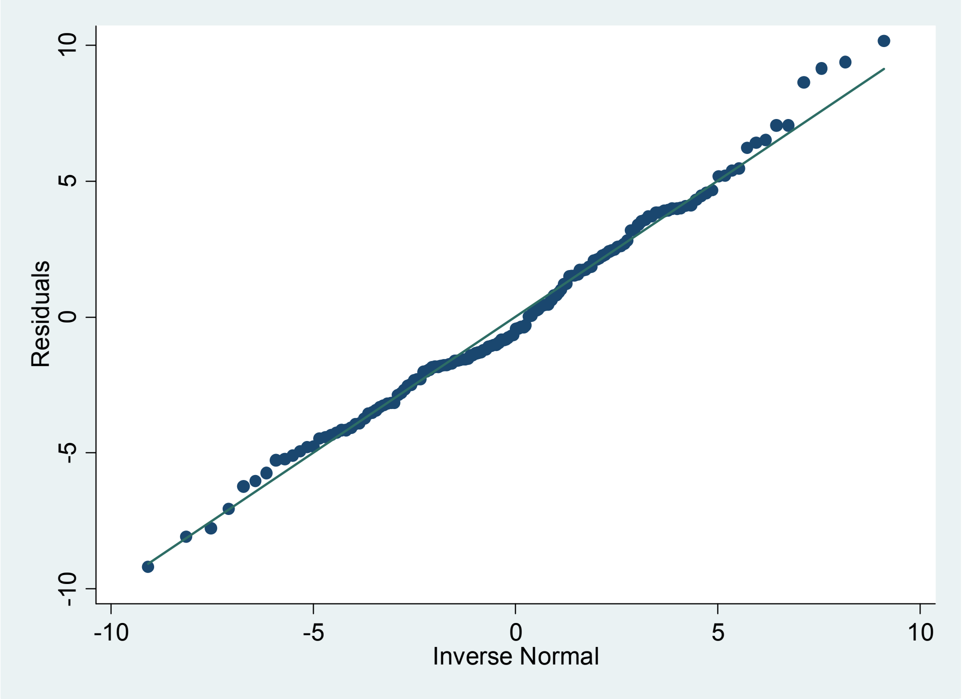
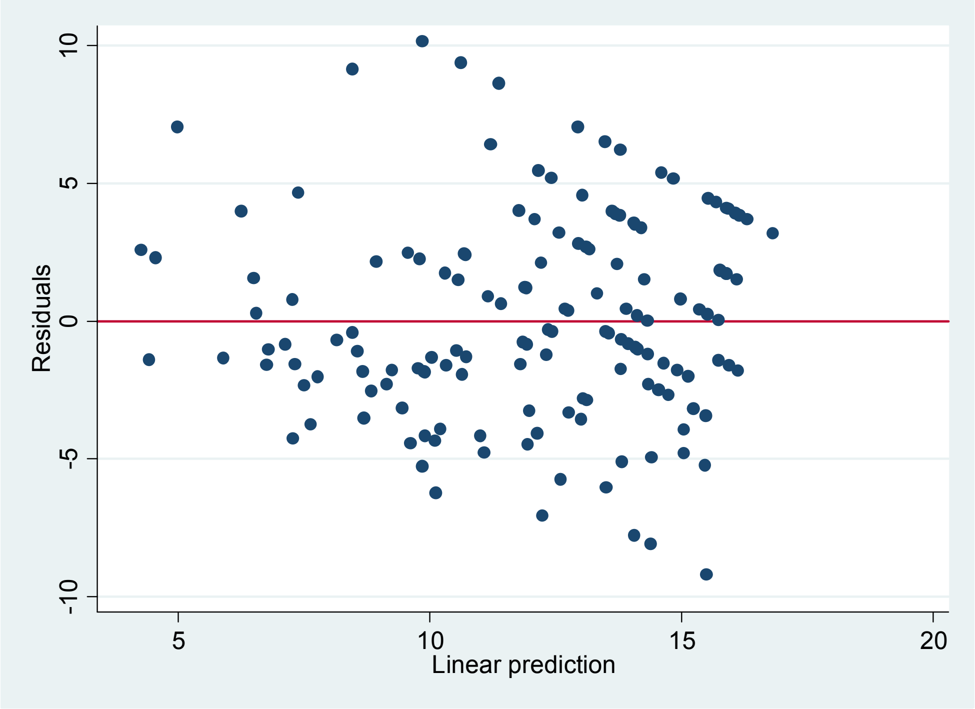


Figure 82 Fitted values against model residuals demonstrating homoscedasticity in Model 6



Therefore, although this model explains little in the way of variation in social SIPSO score at six months, modelling assumptions are met. The variance explained by the fixed (fitted) component of the model is considerably less than the proportion of variance explained by the random effects. It is therefore likely, that there are important factors that have significant influence on the social outcome of patients following a stroke other than the baseline assessments and process markers that have been recorded during the study.

5.9.5 Prediction of the social subscore of the SIPSO without baseline assessments

Variables to be entered into the model to predict the social subscore of the SIPSO without baseline assessments are presented in Table 54.

Table 54 Identified variables to be entered into linear regression models to predict the social subscore of the SIPSO without baseline assessments

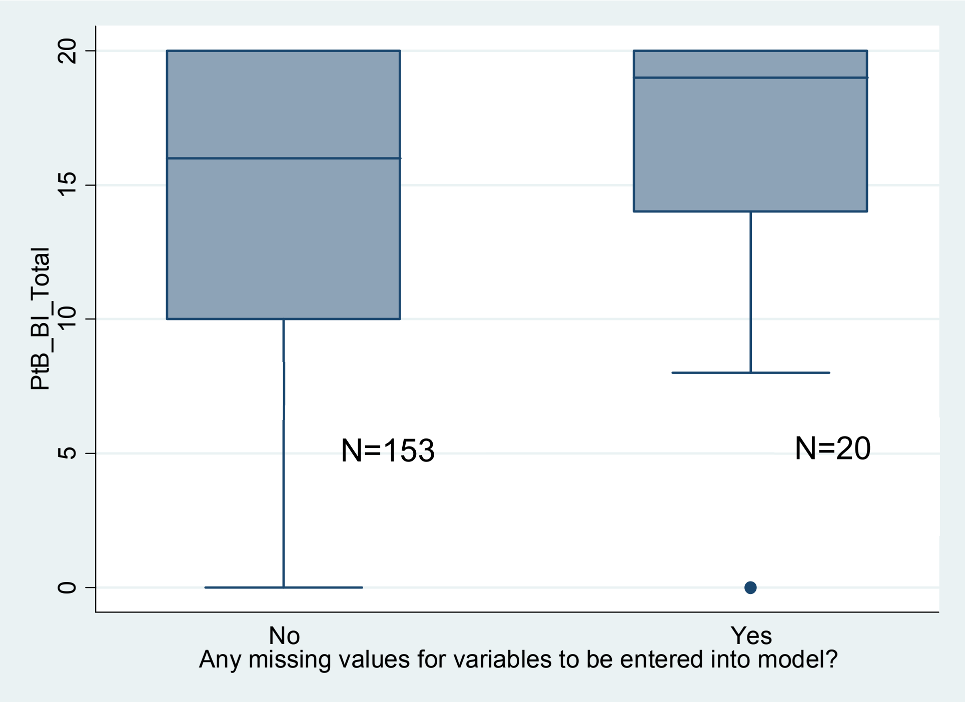
	Variables	Number of variables
Variables identified from regression trees	Length of stay	**1**
	Admitted to stroke unit on day, or day after admission	**1**
	Clinical classification	**3**
	Early supported discharge	**1**
	Imaging within 24 hours	**1**
	SALT communication assessment	**2**
Variables identified from univariate analysis	Formal swallowing assessment	
	Urinary incontinence care plan	**2**
		2
Probable important variables through clinical reasoning	Social Worker assessment	**2**
	Baseline Barthel Index	**1**
	tPA given	**2**
	Propensity score	**1**
	First admitted to a stroke unit, CCU/HDU/ICU vs. general ward/MAU	**1**
	Discharged to same address	**1**
Total		**21**

In the classification tree to predict the social subscore of the SIPSO without baseline assessments, the clinical classification of stroke ‘splits’ on partial anterior (PACS) vs. other type of stroke (see section 5.8.2, Figure 59). For the purposes of reducing the number of dummy variables to enter in the model, this variable will therefore be classified in this way. Admission to a stroke unit on the same day, or day after admission is likely to capture similar constructs to the first ward that the patient was admitted to. The latter variable will therefore be excluded from this analysis.

Table 55 Variables to be entered into the model to predict social subscore of the SIPSO without baseline assessments following refinements

Length of stay	Formal SLT swallowing assessment
Some of hospital spell spent on a SU	Urinary incontinence care plan
LACS vs. other stroke	Social Worker assessment
Early supported discharge	Baseline Barthel Index
Imaging within 24 hours	rtPA given
SALT communication assessment	Propensity score
Discharged same address	= 18 variables (including dummies)

Figure 83 Difference in baseline Barthel Index for patients with complete and incomplete independent variables selected for entering into regression model to predict social subscore of SIPSO without patient reported baseline assessments (only cases with complete social SIPSO shown)



Patients with missing predictor variable data (and therefore automatically excluded from the model) appear to have higher baseline BI than those with complete data (included in the models). However, Mann-Whitney U testing shows that this difference fails to reach statistical significance (p=0.116) such that the median BI between the two groups is assumed to be equal.

Model 8 Linear regression of predictor variables (Table 55) on social subscore of SIPSO, without baseline assessments

R^2 = 0.36 Adj R^2 =0.35 N^o Obs =153 F =21.02 P>|F|<0.001

	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Log(length of stay+1)	-1.69	0.31	-5.5	<0.001	-2.30	-1.08
SLT communication Ax "no but"	2.22	0.66	3.35	0.001	0.91	3.53
Discharged same address	2.67	1.18	2.26	0.03	0.34	5.00
Some of hospital spell on stroke unit	2.26	0.67	3.36	0.001	0.93	3.59
Constant	11.58	1.53	7.58	<0.001	8.56	14.61

Propensity score has been automatically removed from this model as it fails to reach significance at the $p<0.05$ level. If age at stroke is substituted for propensity score in the regression equation, it too is automatically removed from the final model. Neither age nor propensity score therefore feature as important predictors in this model.

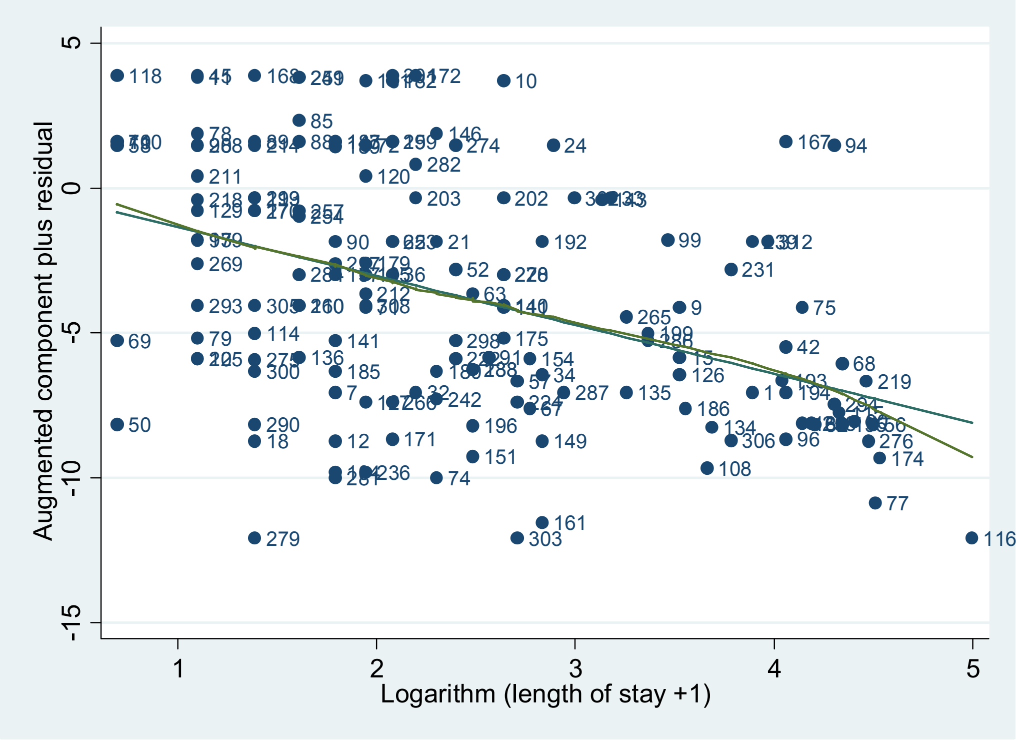
The model explains only 35% of the variance in patient outcome as measured with the SIPSO social subscale. In addition, the majority of this variance is attributable to the residuals rather than the fitted values of the model.

Post estimation checks

5.9.5.1.1 Linearity

Length of stay is the only non-categorical variable that appears in the model. The relationship between a logarithmic transformation of length of stay and the SIPSO social subscale is linear on an acprplot.

Figure 84 Augmented component plus residual plot for log transformed length of stay in Model 8



5.9.5.1.2 Leverage

Examination of studentised residuals reveals very few outlying values where |r| is >2.

Figure 85 Histogram of studentised residuals Model 8

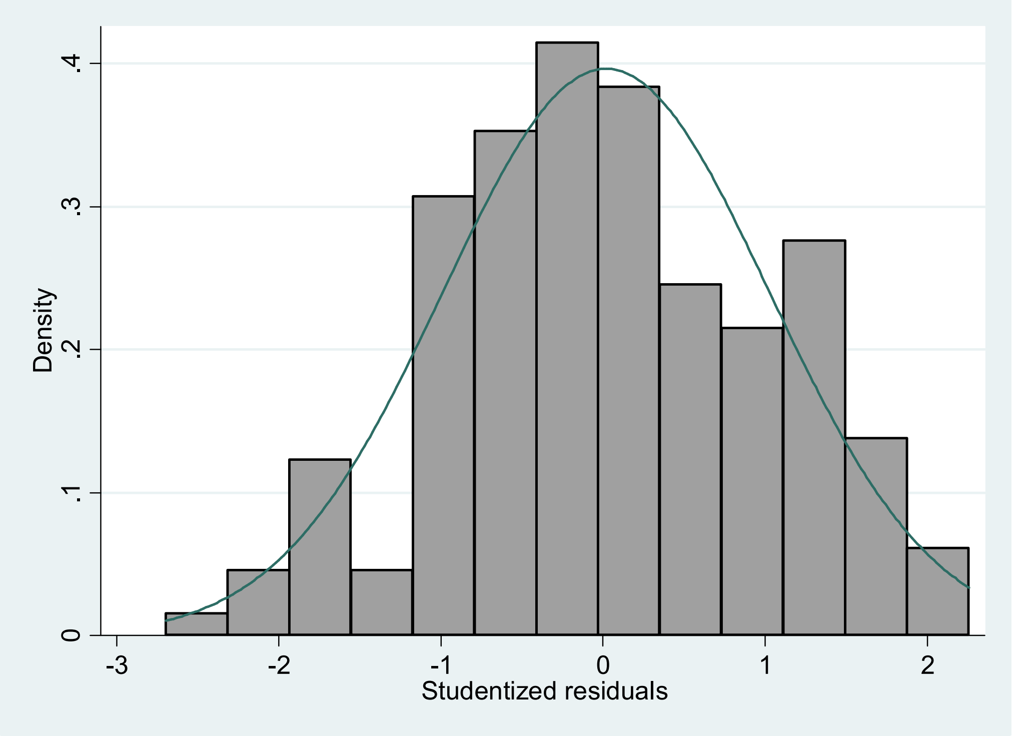
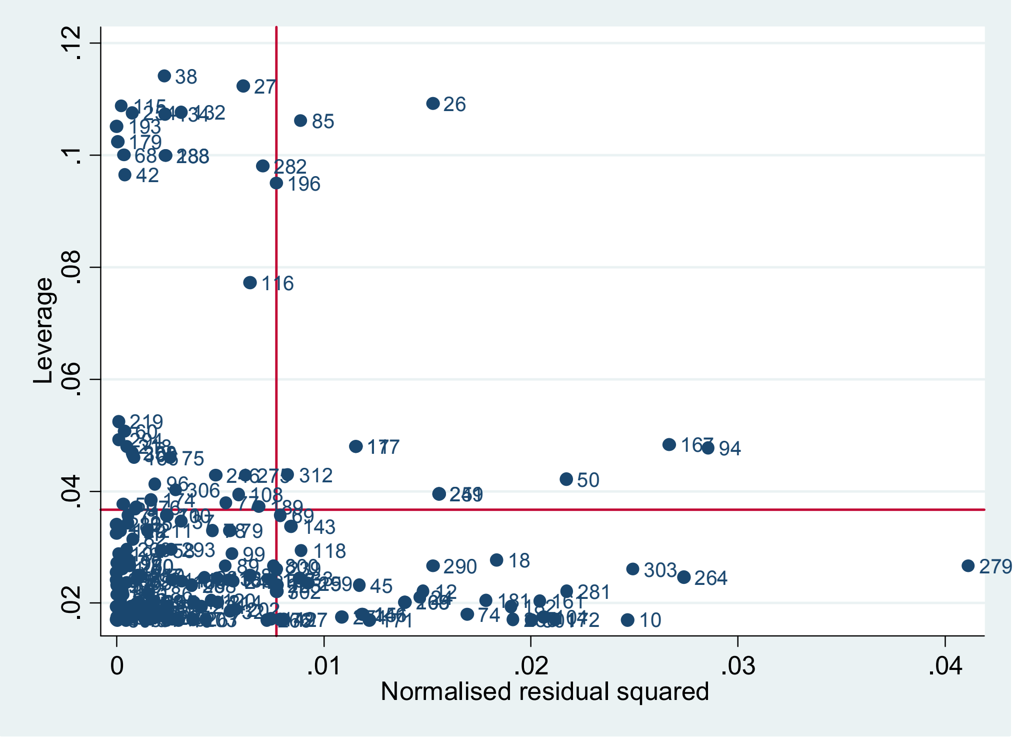


Figure 86 Leverage vs. r-squared plot Model 8

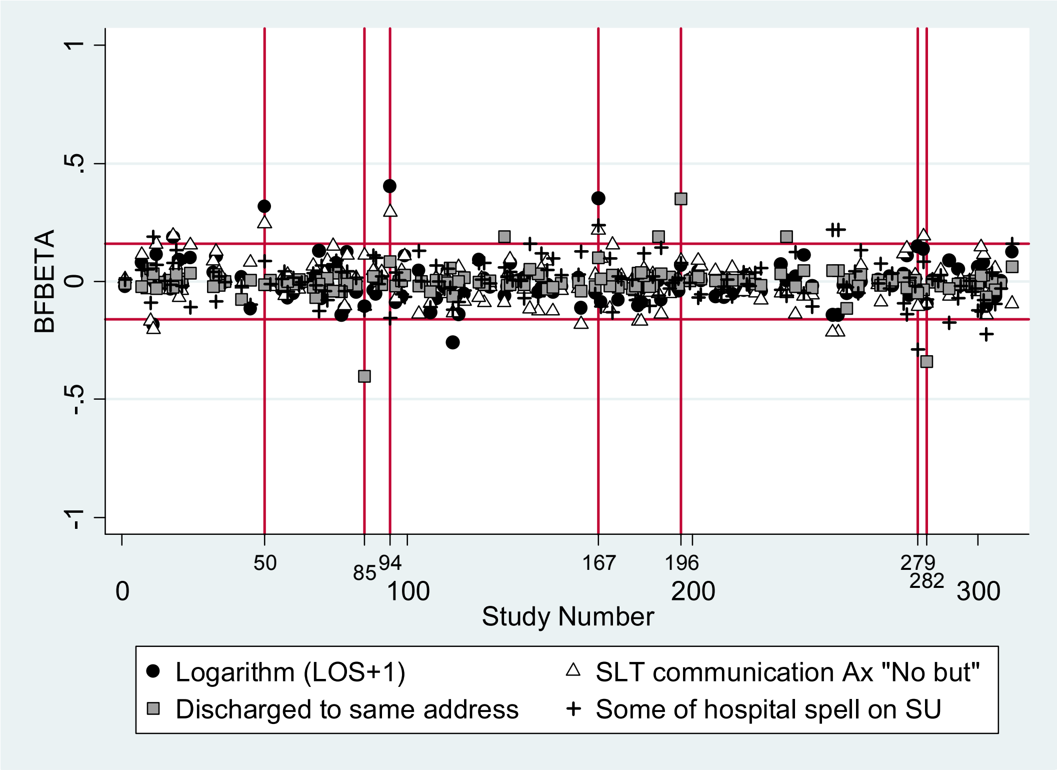


Cases 26, 167, 94, 85 appear as if they may exert undue influence on the beta-coefficients. Cook's D reveals the most influential case to be 85 (cut off value for d = 4/153 = 0.026).

Study number	Cooks D
26	0.07
94	0.05
167	0.05
85	0.04
279	0.04

Displaying DFBeta statistics with limits around 2/√153 reveals a number of cases outwith these limits across the variables. Removal of the most influential case (85) in terms of the regression coefficient for the whole model does not change appreciably the overall fit of the model. Removal of the cases with particularly high |DFBeta| values (50, 85, 94, 167, 196, 279) that are likely to exert influence on the individual beta co-efficients, results in the variable 'discharged to the same address no longer reaching statistical significance, and the entry of 'scan within 24 hours' into the model. There are minor changes to other beta-coefficients in the models, but no other variables change significance or polarity (Model 9).

Figure 87 Scatter plot to identify individual cases with particularly large DFBeta values across variables (Model 8)



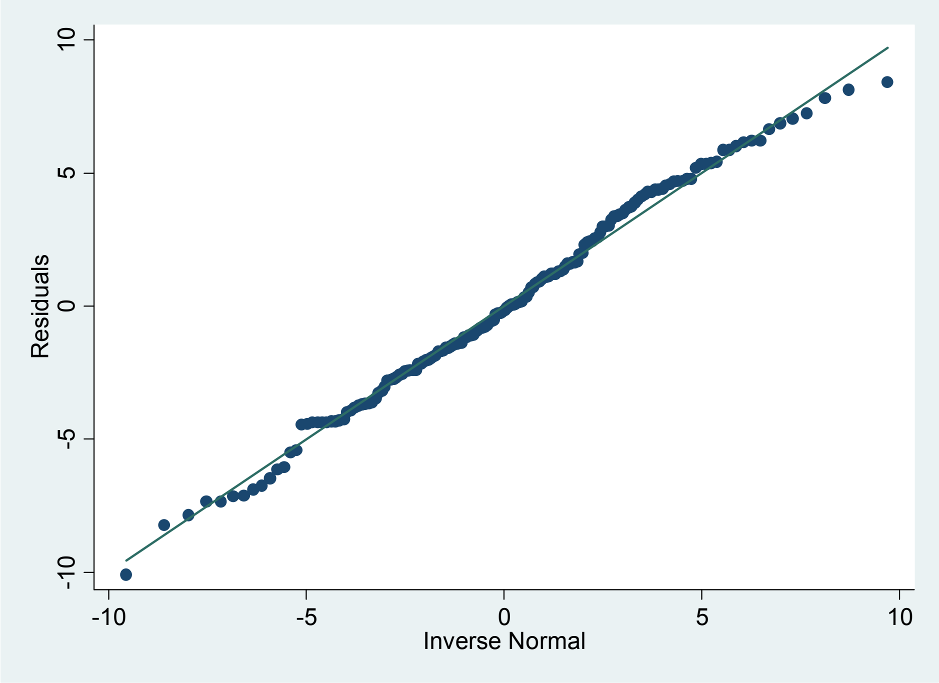
Model 9 Linear regression of predictor variables (Table 55) on social subscore of SIPSO, without baseline assessments and with the most influential cases removed (study numbers 50, 85, 94, 167, 196, 279)

					R^2 = 0.43 Adj R^2 =0.41 N^o Obs =147 F =26.57 P>|F|<0.001	
	Beta coefficient	Standard error	t	P>|t|	95% confidence interval	
Log(length of stay+1)	-2.64	0.30	-7.60	<0.001	-2.85	-1.68
SLT communication Ax “no but”	1.61	0.63	2.42	0.01	0.36	2.87
Discharged same address						
Scan within 24 hours of admission	1.99	0.82	2.42	0.02	0.36	3.62
Some of hospital spell on stroke unit	2.47	0.65	3.80	<0.001	1.19	3.76
Constant	11.95	1.53	7.82	<0.001	8.93	14.97

5.9.5.1.3 *Normality of residuals (Model 8)*

A normal probability plot and Shapiro-Wilk test confirm normality of residuals.

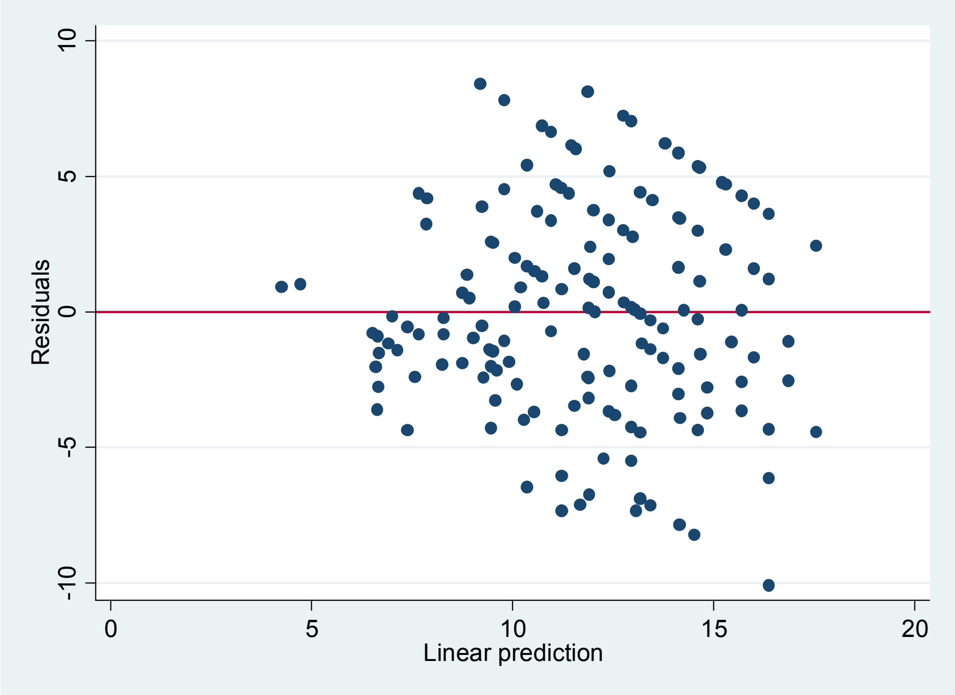
Figure 88 Normal probability plot Model 8



5.9.5.1.4 *Homoscedasticity*

There is no apparent pattern in the residuals vs. fitted values plot, and therefore homogeneity of variance across fitted values from Model 8 is assumed.

Figure 89 Fitted values vs. residuals demonstrating homoscedasticity



5.10 Stability of models

5.10.1 Markov Chain MonteCarlo iterations

The models for physical and social SIPSO outcomes, with and without baseline assessments (four models in total) were recreated in MLWiN software, and MCMC iterations performed to assess convergence of the beta coefficients (see section 4.4.9). Five thousand iterations were performed, with a 'burn in' of 50 iterations for each model. All variables in each model converged on values similar to those of the models generated through the linear regression modelling. Diagnostics of the iterations were acceptable suggesting that the model beta co-efficients are stable.

5.10.2 Significance level for stepwise selection procedures

Models were also run with the significance level for stepwise variable selection procedures set at 0.5 instead of 0.05. This allows examination of the effect of inclusion of clinically but not necessarily statistically significant variables in the models. This resulted in propensity score failing to reach statistical significance in model 3 (prediction of physical subscore of SIPSO without baseline predictors), and imaging within 24 hours reaching statistical significance in model 8 (prediction of social subscore of SIPSO without baseline assessments). This may reflect the relative inferiority of models where baseline assessments are not included (see 6.1.4).

5.11 Utility of the six simple variable case-mix adjuster in the study population

In order to test the utility of the SSV case-mix adjuster in the population, I will examine the discriminatory function (c statistic) and calibration.

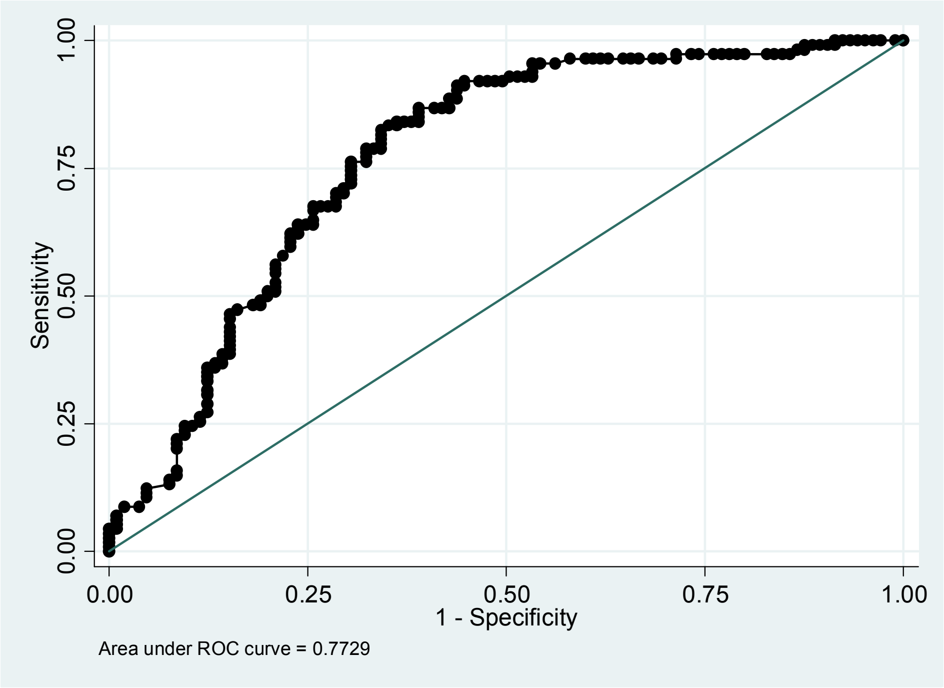
5.11.1 Discrimination

The three original SSV models were derived to predict survival, alive and independent (based on the OHS dichotomised at <=2) or alive and living at home. Of these three outcomes, alive and independent is the most relevant to the study, and so it is this model that has been selected to case-mix adjust the study population.

C statistics were calculated through creating receiver operating curves (ROC) of the propensity score (as calculated from the original SSV model to predict survival, alive and independent) against the observed outcome (OHS <=2). It can be seen that the AUC (which is equivalent to the c statistic for dichotomous outcomes) is 0.77 (95% CI 0.71-0.82), indicating that the SSV model has good discrimination in the study population to predict the outcome alive and independent (OHS<=2) at six months. The reference line represents discrimination no better than chance is given for comparison.

The ROC curve obtained is shown in Figure 90

Figure 90 Receiver Operating Curve (ROC) for propensity score against observed dichotomised OHS



C-statistic = 0.77 [95% CI: 0.71-0.84]

5.11.2 Calibration

Calibration is the ability of a model to correctly predict outcomes in patients that ultimately have the outcome. Deviation from the reference line (y=x) signifies over or under optimistic predictions as outlined in Figure 91.

Figure 91 Calibration of the SSV model to predict alive and independent at six months in the study population

192It has been previously observed that the proportion of patients with predictions of good outcome as determined with the SSV model tends to be over optimistic when compared with observed outcomes (Counsell C et al 2002). Conversely, the model tends to make over pessimistic predictions of the proportion of patients with poor outcome (death or inability to return to own home) (Dennis MS et al 2003). However in the CIMSS study population (where patients with very severe strokes are excluded from recruitment), the SSV makes both over pessimistic and over optimistic predictions in patients with predominantly good observed outcomes (Figure 91). However, it should be noted that some of these proportions are calculated from small absolute numbers of patients as reflected in the size of the error bars.

5.11.3 Utility of the SSV case-mix adjuster to predict the SIPSO outcomes.

In order to ascertain whether the SSV case-mix adjuster may be used to adjust for the SIPSO outcomes, it is necessary to determine that the SSV model can discriminate good over poor outcome and is reasonably calibrated for the SIPSO subscores (i.e. that it makes correct individual predictions). The SIPSO was dichotomised to reflect good over poor outcome as described in the statistical methods section (4.4.7).

5.11.3.1 Discrimination

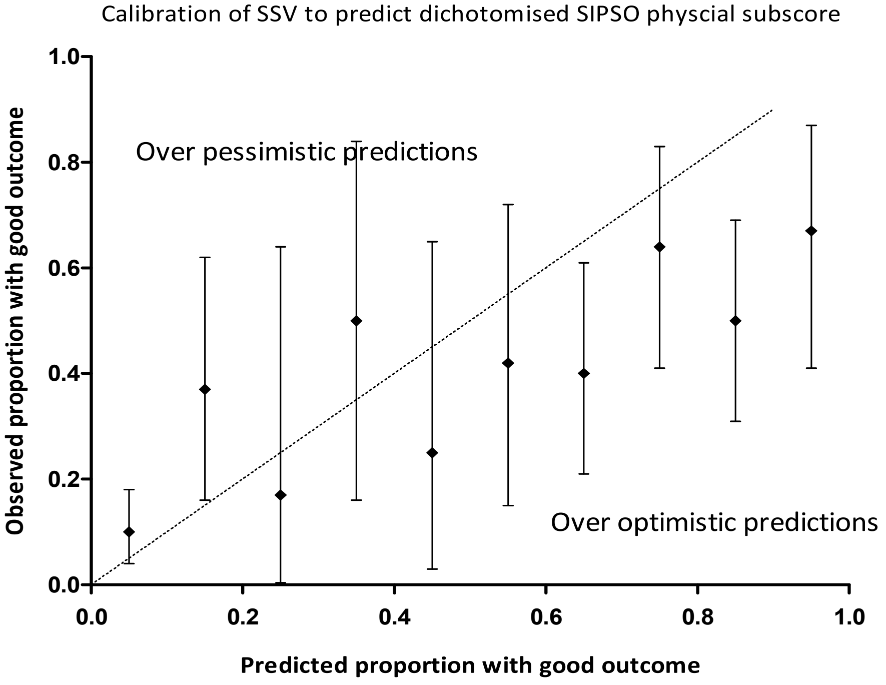
Table 56 C statistics for the SSV model to predict dichotomised study outcomes

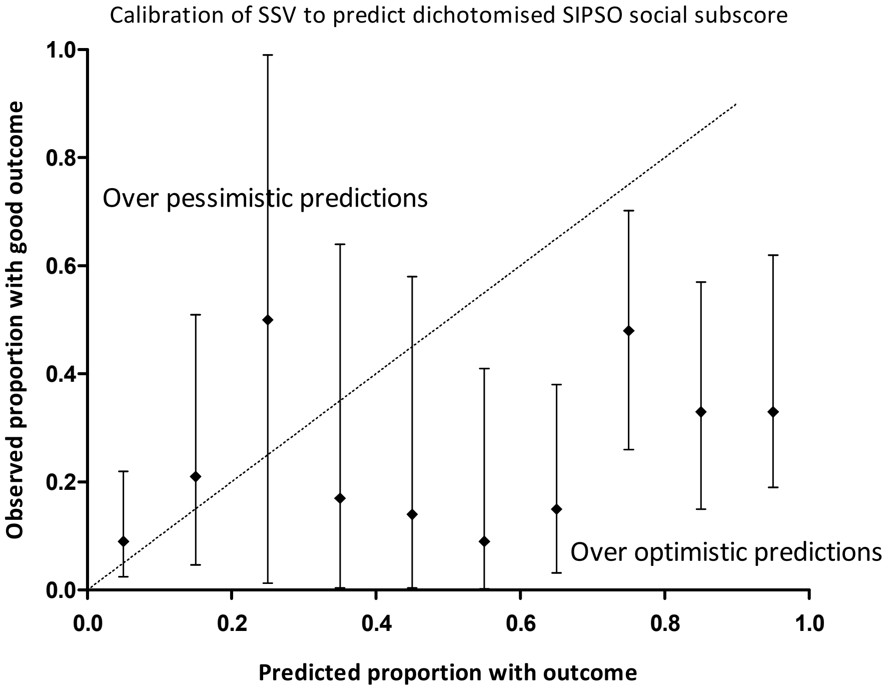
	Centile within which score of 15 lies	C statistic	95% confidence interval
SIPSO physical subscore			
Dichotomised at 15			
Dead patients excluded	60	0.73	0.65-0.79
Dead patients ascribed a score of zero	60	0.76	0.70-0.82
Dichotomised at median (data driven)			
Dead patients excluded (median 12.6)		0.75	0.68-0.81
Dead patients ascribed a score of 0 (median 10.2)		0.89	0.74-0.85
SIPSO social subscore			
Dichotomised at 15			
Dead patients excluded	70	0.66	0.58-0.72
Dead patients ascribed a score of zero	70	0.70	0.64-0.76
Dichotomised at median (data driven)			
Dead patients excluded (median 11.7)		0.70	0.63-0.76
Dead patients ascribed a score of zero (median 9.5)		0.75	0.70-0.81

C statistics were not significantly different with dead patients included (and ascribed a score of zero) or with them excluded, and patients who had died were therefore excluded from the further analysis. The SSV model performs poorly in the prediction of the social subscore of the SIPSO (if dead patients are excluded from the sample (c statistic 0.66).

5.11.3.2 Calibration

Figure 92 Calibration plot for prediction of the physical (top) and social (bottom) subscore of the SIPSO with the SSV model





In contrast to the prediction of survival in independent state (Figure 92), the SSV model tends to make over optimistic predictions of six month physical and social functioning as measured with the SSV case-mix adjuster (Figure 92). As before, small numbers of patients used to calculate some of these proportions (reflected in the wide confidence intervals) limit the conclusions that may be drawn from these graphs. However, if the SSV case-mix adjuster is not transferable to outcomes other than the OHS, its utility and generalisability in studies and populations where the OHS as not an endpoint may be limited.

5.11.4 Use of Length of stay to predict patient outcome

Length of stay has featured prominently in the regression models as a strong predictor of patient outcome. Moreover, the presence of length of stay has resulted in the SSV model not appearing in some models. It is therefore likely that length of stay is acting as a marker of stroke severity. The utility of length of stay as a univariable case-mix adjuster has been explored to determine whether or not it may offer a pragmatic alternative to more complex case-mix adjustment methods.

The discriminatory function of the length of patient stay can be examined through determining the predicted probability of a dichotomised (good/poor) outcome (as measured with the OHS or dichotomised SIPSO subscale scores) through a logistic regression model (see section 4.4.8). This predicted probability was then used to plot ROC curves for length of stay to predict each of the dichotomised study outcomes.

Table 57 C-statistics for SSV model and length of stay to predict dichotomised study outcomes with 95% confidence intervals

	C statistic [95% confidence interval]	
Dichotomised outcome	SSV model	Length of stay
Dichotomised OHS	0.77 [0.71-0.84]	0.79 [0.73-0.85]
Physical subscore of SIPSO (>15)	0.73 [0.65-0.79]	0.75 [0.68-0.81]
Social subscore of SIPSO (>15)	0.66[0.58-0.72]	0.73 [0.66-0.79]

It can be seen from the above table that there is a tendency for length of stay to be a better discriminator of good over poor outcome for all three of the outcome measures, although these differences are not significant. Length of stay would therefore appear to be non-inferior to the SSV model in terms of discrimination. The ROC curves for length of stay are provided appendix E-2.1.

5.12 Comparison of statistical validity of study models with case-mix adjusters identified through systematic review

Table 6 has been reproduced here, with an additional row to represent the statistical methods used in generating the models in this study. It can be seen that aside from the limitations due to a lack of external validation, and a failure to consider interaction terms (due to sample size), that the modelling methodology used here is robust

Table 58 Reproduction of Table 6 to compare statistical validity of ‘Teale’ models with models identified in systematic review of case-mix adjusters

Model	Valid method of variable selection?		Control for Multicollinearity	Consideration of interaction terms	Events per variable >10?	linearity assumptions tested and met?	External Validation Acceptable discrimination (or sensitivity/specificity)	
Guys	✗	Multiple variables selected through identification of ‘statistically significant’ univariate predictors	✗	✗	✗	?	Sens Spec	0.83 0.58
G score	✗	Variables extracted from Guys model (simplified regression co-efficients to integers)	✗	✗	✓	?	Sens Spec	0.72 0.63
Bristol	?		✗	?	✗	✗	Sens Spec	1.00 0
SSV	✓	Use of stepwise variable selection and clinical reasoning	Stepwise variable selection	✓	✓	✓	✓	C statistic acceptable for prediction of alive and independent or dead/alive
Tilling	?		✗	?	✓	Tested; attempts to correct for censoring effects of Barthel Index did not affect the model	Predicts Barthel Index to within 3 points on 49% of occasions (increases to 69% if recovery history is included in the model). 90% limits of agreement -0.4 (-7, +6)	
Orpington	?		Stepwise variable selection	?	✓	✗	R^2 values used to assess model fit. Discrimination not tested	
Teale	✓	Variables selected through identification of important predictors in univariate analyses, regression trees and clinical reasoning	Stepwise variable selection	✗	✓	✓	Not externally validated	

Chapter 6 Discussion

6.1 Identification of variables to be included in a routinely collected stroke dataset

6.1.1 Patient outcomes variables

In order to ascertain which combination of postal outcomes instruments best captures physical and social functioning following stroke, instruments identified in a previous review as valid and reliable for postal administration (Teale EA et al 2010) were further examined for utility and acceptability for patients and healthcare professionals. Discussion at a consumer group and a group decision making workshop identified the SIPSO as the preferred instrument, and this outcome was therefore selected as the primary outcome in the CIMSS study. The SIPSO was non-inferior to the NEADL in terms of missing data and problems with floor and ceiling effects were less pronounced. Moreover, the transformation of the SIPSO subscores to interval level data confers advantages over the NEADL in terms of the types of statistical analyses that may be performed (see section 5.5). However, Rasch analysis for the SIPSO has only been performed in a population of younger stroke survivors (under 65) (Kersten et al 2010). Although there were no interactions between age and the SIPSO items in this age group, differential item functioning has not been explored in older patients and this may limit the generalizability of the transformed scale to the current study.

The SIPSO physical and social subscores measure the underlying traits of reintegration following stroke. The conceptual relationship between the SIPSO subscores and patient care are complex, and are likely to be mediated by factors over and above delivered care processes. These mediating factors include the nature of specific impairments, recovery trajectories, mood, community rehabilitation and social networks.

6.1.2 Identification of important predictor (process) variables

Consideration of all the factors which may contribute to six month SIPSO scores is limited by the feasibility of capturing variables which represent them, and sample size. Process markers were identified through both statistical and clinical reasoning. In order to reduce the number of variables that were entered into the models, proxy or composite markers were chosen (i.e. SLT communication assessment represents whether or not an assessment was required (a proxy for dysphasia) and whether or not the assessment was performed). Other than whether or not ESD was planned, it has not been possible to capture detailed information regarding the period between discharge and the six month follow up questionnaire. This may limit the conclusions that may be drawn, as provision of post-discharge services and rehabilitation are likely to influence post-stroke reintegration.

In order to reduce the number of predictor variables to be entered into linear regression models such that the EPV was maintained below ten, three approaches were adopted. The first, univariate analysis, considers the association of individual process measures or predictors with patient outcome. Although this unadjusted approach is simple, there is no consideration for the mediating or confounding relationships of other factors. As such, although specific aspects of process may be identified for further exploration, the creation of models based solely on these univariate relationships is likely to include unimportant or exclude important predictors on the basis of chance alone (Altman D, 1999p 349). Moreover, the prominence of some of the relationships of individual process markers with outcome is removed when other factors are controlled for, e.g. although highly significant in univariate analysis, the presence of a urinary incontinence care plan failed to reach statistical significance as an important predictor of physical outcome in multivariable models.

The second approach to refining predictor variables was through the use of regression trees. These offer a simple and powerful visual representation of the statistically important factors in terms of prediction of patient SIPSO subscores. The benefits of this approach are that there is no assumption based on the distribution of either the independent or dependent variables and no limit to the number of variables that may be entered into the tree model. The importance of each predictor, having taken account of all other predictors is considered, and the 'split-point' is made at the value of the independent variable that maximises the diversity of outcome. However, the regression trees are 'data-driven' and require clinical interpretation. Important factors may not appear in the trees due to idiosyncrasies of the study dataset. This leads to the third approach for variable selection: the inclusion in regression models of any clinically important predictor that has not been identified through 'data-driven' approaches.

6.1.2.1 Process measures that are predictive of functional outcome in the study

The determination of sample size for specification of regression models depends not only on achieving sufficient 'events per variable' in order to ensure that the model is not overfitted (Peduzzi P et al 1996), but also in ensuring that the sample size is adequate for individual predictors to distinguish a clinically relevant difference in patient outcome. The failure of Davenport et al to detect an effect of stroke unit care on patient outcome despite an adequate EPV (Davenport RJ et al 1996) is likely have been due to the study being underpowered to detect the effect, as highlighted by Mant in his response to the article (Mant J et al 1996) (see also section 2.1.8).

If it is assumed that case-mix adjustment is sufficient to 'level the field' such that any residual variation in outcome is due to the delivery of specific care processes, the delivery of process measures must vary in order to detect the effect of deviation from process delivery on outcome. In Mant and Hick's simulation study describing delivery of care processes of proven association with outcome in myocardial infarction, the treatment effect of interventions were applied to theoretically identical populations to demonstrate the difference in sample size required to detect differences in mortality from myocardial infarction through measurement of process versus outcome (Mant J et al 1995). Here it was demonstrated that the higher the proportion of patients receiving a particular process or combination of processes in a particular hospital, the smaller the sample size required to detect deviations from care process delivery through measurement of both process and outcome (Mant J et al 1995). However, where the 'treatment effect' of specific interventions has not been determined and the relationship between process and outcome is not known as is the case with many stroke process measures, the larger the proportion of patients receiving a particular care process the harder it is likely to be to detect the effect of missing that process on patient outcome. For processes that near 100% saturation, the magnitude of the effect of these processes of care on patient outcome, if any, is unknown.

The process saturation in the study population Figure 44 (i.e. the lack of variability in patients that did and did not receive specific aspects of care process), is in concordance with the recent RCP NSSA audit where the median percentage achievement of the twelve key indicators across participating trusts in England, Wales and Northern Ireland was greater than 80% in all but 3 process markers audited (Intercollegiate Stroke Working Party 2011). Although some of the process markers entered into the study models represent 'best-practice' interventions for which there is good supporting randomised controlled trial evidence (e.g. treatment on a stroke unit (Stroke Unit Trialists' Collaboration 2007)), the demonstration that other interventions or processes of care are effective where such evidence is lacking is unlikely to be feasible in empirical post-stroke populations whilst there is such a degree of saturation of process markers; if the majority of the population receives an intervention routinely, it is difficult to discern the effect of not receiving that intervention on patient outcome. As the proportion of patients in whom a monitored process is achieved increases, the proportion in which it is not achieved, and therefore the variability, decreases. This is especially pertinent as the effect of individual processes on outcome is likely to be small (the effects of individual processes that typically occur on a stroke unit are unlikely to be larger than the overall treatment effect of stroke unit care over general ward care (an estimated ARR of 4.4%) (Sudlow C et al 2009; Stroke Unit Trialists' Collaboration 2007). A lack of variability in the delivery of specific processes means that distinguishing patients with good over poor outcome conditional on the achievement of a specific care process will require a larger sample (analogous to a randomised controlled trial to determine the benefit of an intervention with a small treatment effect). The likelihood of type 2 errors is high (falsely rejecting a hypothesis that is true) and potentially important predictors in the models may fail to reach significance due to a lack of power. Whether the difference in outcome observed between levels of a predictor reaches statistical significance is reflected in the model output by the‘t’ value and its significance level (which is the same as performing a t test between the model predictor and the reference value). The magnitude of the mean difference between the levels of the variable is represented by the beta co-efficient, the value by which the dependent value is increased for a unit increase in the independent variable. Standard errors for the beta co-efficients are provided in the STATA output from which the standard deviations may be calculated from the formula for the standard error of the difference between two sample means [s.e. = $\sqrt{}(s^2/n_1 + s^2/n_2)$] (where s.e. = standard error, s=standard deviation and n = sample size in each group (Altman D, 1999 p 160)). The power with which these t-tests have been performed during the modelling process can be calculated in STATA from the standard error, the observed difference in outcome between groups, the number of patients in each level of the variable and the α-significance level (set at 0.05). Using similar methodology to Mant in his criticism of the Davenport study (Mant J et al 1996; Davenport RJ et al 1996), I have used the example of the SLT communication assessment for the prediction of the social subscore of the SIPSO to demonstrate the large sample size that would be required to detect the difference in mean social SIPSO subscore between patients that do, and do not receive a SLT communication assessment resulting from Model 6. This process has some variability (“no” = 29/312 (9.3%), “yes” = 108/312 (34.6%), “no but” = 173/312 (55.4%)), although the absolute numbers of patients receiving the process is small compared with those in whom an assessment is not required (the “no but” dummy). Performing a power calculation in STATA reveals the probability of detecting a difference of 2 points on the physical SIPSO subscore between patients who do, and do not receive a SLT communication assessment (represented by the beta co-efficient for receipt of a communication assessment in the model) is just 33% (power 0.33). In order to detect such a difference with reasonable certainty (e.g. power 80%), would require 54 patients in the “no” group, 201 in the “yes” group and, accounting for those in whom an assessment is not appropriate (55% in the study population), a total sample size of 567 patients with complete data (working provided in Appendix 7.2E-1.4). Assuming that there are no missing data for process markers, and a return rate for outcome questionnaires of 70% (similar to that seen in the study), to detect the difference in SIPSO outcome between patients who do, and do not receive a SLT communication assessment with power of 80% would require a total sample size of ≈800 patients.

Previous studies to identify important aspects of stroke care that may determine patient outcome after adjustment for case-mix have tended to focus on the prediction of dichotomised outcomes of mortality and dependency (Bravata DM. et al 2010; Evans A et al 2001; Lingsma HF et al 2008; Mohammed MA et al 2005; Weir N et al 2001) although attempts have been made to explore relationships between processes of care and discharge home (Indredavik B et al 1999) and functional outcomes (McNaughton H et al 2003)**.** In many of these studies, variations in outcome between institutions and individuals are completely (Davenport RJ et al 1996; McNaughton H et al 2003) or partly (Lingsma HF et al 2008; Weir et al 2003) explained through differences in case-mix rather than differences in the delivery of care. Process markers which have been highlighted as potentially important predictors of outcome in previous studies include swallowing assessment (Bravata DM. et al 2010), measures to prevent aspiration (not further qualified) (Evans A et al 2001)**,** early feeding (Evans A et al 2001), organised stroke unit care (Evans A et al 2001; Weir N et al 2001), prophylaxis for venous thromboembolism (Bravata DM. et al 2010), treatment of all episodes of hypoxia with supplemental oxygen (Bravata DM. et al 2010), early mobilisation (Lingsma HF et al 2008) and antiplatelet therapy within 48 hours (Lingsma HF et al 2008). Use of some of these interventions are corroborated (stroke unit care (Stroke Unit Trialists' Collaboration 2007), antiplatelet therapy (Chen Z-M et al 1997; International Stroke Trial Collaborative Group 1997)) or questioned (use of graduated compression stockings (CLOTS Trial Collaboration 2009)) in randomised controlled trials, and some are subject to ongoing investigation (the use of supplemental oxygen (Roffe C 2011) and intermittent pneumatic compression devices (CLOTS Trial Collaboration 2011)). It has been postulated that the improved outcomes of patients admitted to acute stroke units are due to the prevention of complications of stroke, such as the prevention of infection (Govan et al 2007). Although many specific care processes which form existing stroke markers have not been linked to outcome in dedicated randomised trials, features of stroke unit care that are consistently provided in effective stroke units have been systematically identified from the Stroke Trialists' systematic review of organised stroke unit care (Langhorne P et al 2002; Stroke Unit Trialists' Collaboration 2007). These were recently summarised by McArthur et al (McArthur et al 2011) and are reproduced in Table 59.

Table 59 Important components of stroke unit care (from McArthur et al (2011), based on Langhorne P et al (2002))

Important components of stroke unit care	Potentially important components of stroke unit care	Components with no evidence of efficacy
Staff with a specialist interest in stroke care	Management of pyrexia, blood sugar, hypoxia, blood pressure, hydration, nutrition	Routine use of compression stockings (CLOTS Trial Collaboration 2009)
Early mobilisation	Mouth care	Early PEG feeding (Dennis M et al 2006)
Early investigation	Swallowing assessment	Routine use of nutritional supplements (Dennis et al 2006)
Prompt pharmacotherapy	Bladder and bowel care	
Physiological monitoring	Provision of information for patients and carers	
Discharge planning	Involvement of carers	
MDT goal setting		
Positioning		

Linear regression models were constructed from the variables identified through regression trees, univariate analysis and clinical reasoning. Important predictors featuring in these models are highlighted in Table 60. Only variables reaching significance at the p<0.05 level are included here, although it is possible, indeed likely that the variables in the models that have not reached statistical significance still represent important predictors due to the possibility of type 2 errors for individual predictors. For the purposes of definition of a routine dataset, any clinically and statistically important predictor should be included.

Table 60 Important predictors of outcome featuring in regression models

	With baseline assessment		**No baseline assessment**	
	Physical SIPSO	Social SIPSO	Physical SIPSO	Social SIPSO
Baseline NEADL	✓	✓		
Baseline EQ5D	✓	✓		
Propensity score			✓	
Age at stroke	✓			
Previous stroke			✓	
Length of stay	✓	✓	✓	✓
D/C to same address as admitted from	✓		✓	✓
SLT communication Ax		✓ (2[1])	✓ (1 or 2[1])	✓(2[1])
First ward ASU, CCU, HDU or ICU			✓	
Some time spent on a stroke unit during inpatient spell				✓

Three process markers appeared in one or more of the four linear regression models (a formal Speech and Language communication assessment, admission to a ward where hyperacute stroke care can be delivered, and admission to a stroke unit for some of the inpatient spell) (Table 60). This is in concordance with the important features of care process identified through the previous systematic review of stroke unit care (Langhorne P et al 2002; McArthur KS et al 2011). Of these, only formal SLT communication retains prominence across the models (featuring in 3 out of 4). Of particular note is that SLT communication assessment is one of the process markers with greatest variability across the three levels of the variable which may, in part, explain its prominence in the models (Figure 44). In accordance with the univariate analysis, patients who do not require a SLT assessment ("no but" code) have better physical and social SIPSO subscores at six months than those that require but do not receive an assessment. The caveat to this is that SLT communication assessment does not feature in the model to predict the physical subscore of the SIPSO where baseline assessments are also included in the model. This may reflect the possibility that the "no but" code is acting as a marker of case-mix but is overshadowed in the model to predict physical outcome where there are more explicit markers of baseline physical function present. In the model to predict physical subscore where these

[1] Where 1 = SLT communication assessment performed, and 2 = SLT communication assessment not required ("no but" code)

baseline functional assessments are excluded, not requiring an assessment is associated with intermediate physical outcomes scores (better than requiring and not receiving an assessment, but worse than if the assessment is performed). Here, it is likely that the outcomes of the heterogeneous group in whom “no but” codes are used (where assessments are either not required (mild strokes) or not appropriate (severe strokes)) fall, on average, between those in whom formal communication assessments are, or are not performed. These subtleties in the potential meaning of the prominence of different levels of the variables are important in their interpretation. Moreover, this represents an argument for the explicit capture of reasons why assessments are not indicated. This is particularly pertinent in routine care where, unless there is adequate and robust case-mix adjustment, the significance of different levels of the variable in terms of their relationship to outcome is difficult to interpret. In previous studies where markers from the RCP NSSA have been used, and in the report of the audit data from the RCP (Intercollegiate Stroke Working Party 2011), patients with a “no but” code are removed from the denominator (McNaughton H et al 2003; Weir N et al 2001) such that only patients who are eligible for interventions are included in the analysis.

The distinction between clinical and statistical significance of the predictors is key in terms of determining the relative importance of the difference predictors. As the models are linear, the beta-coefficient is interpreted to represent the difference in the mean outcome (i.e. physical or social SIPSO subscore) for a one unit change in the independent variable. For example, for a dichotomous predictor, the beta-coefficient represents the change in outcome score for one level of the predictor with respect to the other, with all other variables being held constant. For a continuous predictor (for example age at stroke), the outcome changes by the value of the beta-coefficient for each additional year. The magnitude of the change in outcome therefore needs to be interpreted in this context, taking into consideration the units of measurement and any transformations of the data that have occurred. Data transformations make the relative relationship of length of stay to outcome subscore difficult to interpret. However the mean difference in outcome score for patients staying for B days rather than A days can be calculated from the following equation:

$$(6)\quad \mathrm{M} = \beta_{\mathrm{log\_LOS\_plus\_one}} * [\log(\mathrm{lengthofstayA}+1) - \log(\mathrm{lengthofstayB}+1)]$$

$$= \beta_{\mathrm{log\_LOS\_plus\_one}} * \left[\log\frac{(\mathrm{lengthofstayA}+1)}{(\mathrm{lengthofstayB}+1)}\right]$$

Where M = mean difference in outcome score

If a length of stay of one day is taken as a 'reference' value, then the differences in outcome score dependent on changes in length of stay (with all other variables being held constant) are shown in Table 61.

It can be seen therefore, that the absolute difference in total SIPSO subscores attributable to each variable is very small (Table 62) and, although a highly statistically significant predictor of outcome, the differences in mean SIPSO subscore in the physical domain compared with a length of stay of one day is of questionable clinical significance (Table 61).

Table 61 Differences in average outcome subscore for length of stay compared with one day

	Compared with a length of stay of one day:			
	With baseline assessments		No baseline assessments	
Length of stay	Difference in average physical SIPSO subscore (1 dp)	Difference in average social SIPSO subscore (1 dp)	Difference in average physical SIPSO subscore (1 dp)	Difference in average SIPSO social subscore (1 dp)
1	0	0	0	0
3	-0.4	-0.6	-0.1	-0.5
5	-0.6	-1.0	-0.4	-0.8
10	-0.9	-1.6	-0.6	-1.3
30	-1.5	-2.5	-1.0	-2.0
90	-2.0	-3.5	-1.3	-2.8

Table 62 Beta co-efficients for statistically significant predictors in models, p<0.05 significance level – excluding length of stay (significant in all models)

	Change in SIPSO outcome per unit change in predictor			
	With baseline assessments		No baseline assessments	
Predictor	Physical subscore	Social subscore	Physical subscore	Social subscore
Baseline NEADL	0.11	0.06		
Baseline EQ5D	4.73	3.53		
Propensity score (SSV)			3.18	
Age at stroke	-0.07			
Previous stroke			-2.61	
Discharged to same address as admitted from	3.49		3.94	2.67
SLT communication Ax (assessment performed)			2.87	
SLT communication Ax (no but code)		2.11	2.76	2.22
First ward ASU, CCU, HDU or ICU			2.11	
Some time spent on a stroke unit during inpatient spell				2.26

6.1.2.2 Predominance of proxy markers of severity in models

Examination of the important predictor variables across the models reveals that markers of stroke severity predominate over markers of process. Two process markers feature prominently across the models as being associated with better outcomes: a “no but” code for a formal SLT assessment, and being discharged to the pre-admission address. No other process markers feature in more than one model. Discharge to the same address is associated with better physical outcomes, and a “no but” code for a communication assessment with a better social outcome. These markers remain in the models even when baseline assessments for severity are included (i.e. in the models that include baseline assessments). However, as discussed in section 6.1.2.1, it is likely that these process markers are actually acting as markers of stroke severity: discharge to the same address would usually reflect less physically impaired patients. It is likely that discharge to the same address is a proxy measure of independence, social support or ability to return to the pre-stroke address with a package of care as opposed to discharge to a continuing care facility. Although discharge home may be for palliative care, these patients would not usually be expected to survive until six month follow up and therefore would not have been included in the sample.

The models that do not contain baseline assessments are less explanatory of the variation in outcome than the models that do contain these assessments (adjusted R^2 for model to predict physical SIPSO with baseline assessment = 0.52, without baseline assessments = 0.43). Where process markers do feature in the models, these tend to reflect organisational processes: direct admission to a stroke unit or spending any part of the hospital spell on a stroke unit.

Unfortunately as detailed information regarding the movement of patients around the hospital was unreliable it has not been possible to extract the proportions of individual patient’s stay spent on a stroke unit, or to explore the optimal proportion of a hospital spell that should be spent on a stroke unit to optimise outcome. Stroke unit care is a complex intervention, and is likely to reflect may different aspects of patient care. Its presence in the models may reflect that these models are underpowered to detect the effect of individual processes of care that occur within a stroke unit, or that many of the care processes were saturated.

6.1.3 Length of stay as a marker of stroke severity

Length of stay was a prominent predictor of patient outcome as measured with the SIPSO and featured in all the models. It is likely that length of stay is acting as a marker of stroke severity, and its inclusion in the models overshadowed the ‘best’ existing case-mix adjustment model (the SSV model) such that the propensity score (probability of a good outcome as predicted with the SSV case-mix adjustment model) reached statistical significance in only one of the models (where physical subscore of the SIPSO was predicted without baseline assessments). There is a negative and highly significant correlation between the logarithm of length of stay and the propensity score r= -0.6, p<0.001), such that patients with higher propensity score (higher probability of good outcome) tend to have shorter lengths of stay. It is therefore likely that the two variables are collinear, hence the automated removal of propensity score from the model during stepwise variable selection procedures. If propensity score is forced back into the model, it is automatically removed by the STATA software due to collinearity.

As with other markers of severity, length of stay is a complex marker and is likely to be acting as a proxy measure for several different aspects of patient care. For example, length of stay is likely to reflect stroke severity (patients with more severe strokes are more likely to require longer spells in hospital), the requirement for increased social support at discharge, rehousing, equipment and complications of stroke (e.g. intercurrent illness, or the requirement for feeding via percutaneous gastroenterostomy (PEG)). However, length of stay may have utility as an overarching variable to adjust for the combined effect of these factors and as a crude marker of stroke severity. Length of stay is not available until after hospital discharge and therefore could not be used for stratified randomisation in trials, or to determine prognostic information for individuals from baseline data. However, for adjustment in observational cohort studies, or in routine data collections it may offer improvements to over simplistic approaches such as age-sex standardisation, whilst acting as a pragmatic alternative more complex case-mix adjustment models.

There are however, drawbacks to the approach. Reduction in length of stay is often targeted specifically as a positive outcome (Intercollegiate Stroke Working Party 2011). The introduction of Early Supported Discharge or community rehabilitation teams is a clear example of where length of stay may be shortened for certain groups of patients. However, patients who are ‘fit’ for these types of intervention are unlikely to be the same cohort as those with protracted lengths of stay due to severe strokes. Moreover, if necessary it would be possible to adjust for these interventions explicitly. Another possible factor confounding the relationship between stroke severity and length of stay would be discharge for palliative or nursing home care. Here a shorter length of stay may be associated with poorer outcome.

In-hospital deaths also spuriously reduce the length of stay. In routine care, there is likely to be a higher proportion of inpatient deaths than has been observed in this study where patients in receipt of palliative care were excluded. The utility of length of stay as a case-mix adjuster is for adjustment of outcomes in populations of survivors to hospital discharge. Inpatient deaths are therefore not included within this subgroup, although deaths between discharge and follow up could be included if death is considered as an outcome (e.g. independent survival).

Lengths of stay may also be affected by organisational factors (e.g. access to community rehabilitation facilities), and these would require specific consideration and additional adjustment. Length of stay could be used in this context as a marker of the efficiency of a stroke service – how rapidly patients are discharged conditional on their stroke severity.

Between institutions comparison may be complicated by inconsistent measurements of length of hospital stay across diverse healthcare systems. These may be mitigated through the application of precise definitions of what constitutes acute hospital care, and the start and end points of an acute hospital stay (e.g. exclusion of residential rehabilitation facilities from the length of hospital stay).

Providing external validity of length of stay as a univariate case-mix adjuster could be demonstrated in external datasets, application of consistent definitions to the routine recording of length of stay across healthcare systems could increase the feasibility and interpretability of large sets of observational data where outcomes are collected following discharge from hospital.

Length of stay has often been used as an endpoint (outcome) in stroke studies, but we are not aware of it previously being used as an independent variable to adjust post-stroke populations for case-mix.

6.1.3.1 Comparison of LOS with SSV case-mix adjustment model

The utility of the SSV case-mix adjustment model and length of stay as a univariate adjuster were directly compared. However, it should be noted that this represents external validation of the SSV model (to predict an outcome that it was not designed to predict), but internal derivation of length of stay as a case-mix adjuster. The relative performance of the two models is therefore biased to favour length of stay.

The SSV model was developed to predict probability of good over poor outcome as determined with the dichotomised modified Rankin Scale (alive and independent vs. not) (Counsell C et al 2002). When used to predict this outcome in the study population, the SSV model had good discriminatory function (see 5.11). However, it was poorly calibrated with a tendency to both over pessimistic and over optimistic predictions in patients with mild to moderate strokes. The c-statistic of length of stay to predict the same outcome was comparable (SSV c-statistic (0.77 [0.71-0.84]; LOS c-statistic (0.79 [0.73-0.85]). This represents internal validation of length of stay to predict this outcome. As such, length of stay as a prominent predictor is likely to be overfitted to the study data, and the c statistics for LOS are likely to be higher than would be expected in an external dataset.

In order to explore the utility of the SSV case-mix adjuster to predict the SIPSO outcomes, the SIPSO subscores were dichotomised to create two groups felt to reflect broadly the dichotomised OHS, i.e. some residual impairment, but not severe enough to interfere with daily living. It is perhaps unfair to expect the SSV model to be able to predict an outcome that is was not developed to predict, but as the best available case-mix adjuster, if the model lacks generalisability to outcomes other than the OHS, its utility in observational cohorts where functional outcomes are measured, is limited.

Despite this, the SSV model performed reasonably to predict the physical subscore of the SIPSO (c-statistic 0.73[0.65-0.79]), but less well to predict the social outcome (0.66[0.58-0.72]). Length of stay was non-inferior to the SSV model, with c-statistics of 0.75[0.68-0.81] and 0.73[0.66-0.79] to predict the physical and social subscores respectively.

6.1.4 Can outcome be predicted without the need for patient reported baseline assessments?

Recording patient reported assessments of function in routine care is costly in terms of resource and infrastructure. Such assessments also add a layer of complexity to the interpretation of routine data. However, the addition of these baseline assessments to the models greatly improved model fit and increased the amount of variation in patient outcome that was explained. The change in the variables reaching significance on exclusion of influential cases in models without baseline assessments indicates that these models are less stable than the models where baseline assessments were included. Indeed, for the models that included baseline functional assessments, more variance was explained through the model variables than through the residuals. It is a logical assumption that patient reported function at baseline will be linked to, and an important predictor of, patient reported function at six months.

It is also of note, that process measures are more prominent in the models which do not contain the baseline assessments, and that the Barthel Index (as a measure of objective baseline function) does not reach significance in the model to predict either physical or social subscore of the SIPSO. It is possible, that the presence of process markers in these models is a consequence of suboptimal case-mix adjustment in the absence of a baseline marker of severity rather than a true reflection of the importance of individual process markers. The marked improved fit of the models when baseline assessments are included provides evidence to support this possibility.

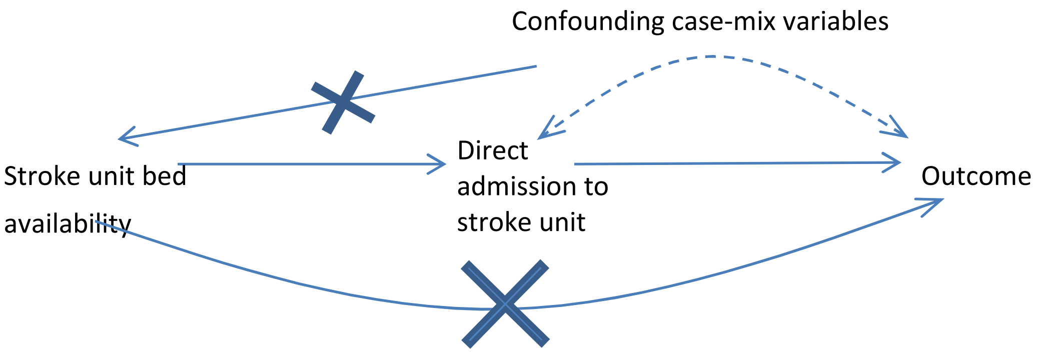
Baseline quality of life (as measured with the EuroQoL) was a particularly important predictor of patient outcome as measured with the SIPSO. This may be due to the EQ5D containing questions concerning mobility, self-care and ability to perform usual activities, constructs also contained within the SIPSO. Alternatively, the prominence of the baseline EQ5D may reflect the inclusion of a question pertaining to pain, a construct which is conceptually linked to patient outcome, but otherwise unmeasured in the study dataset. The value of a questionnaire to reflect quality of life in the week immediately following a stroke is debatable. This is likely to be a time of considerable emotional and physical stress, perhaps reflected in the perception of quality of life. Moreover, questions pertaining to ability to perform usual activities (contained within the EQ5D) may be difficult to answer in the inpatient hospital setting. However, as the baseline EQ5D is such a strong predictor of six month SIPSO score, it would be beneficial to include it in a routine dataset for further exploration, although a question specifically relating to pain at baseline may warrant further exploration as an alternative measure.

6.2 Alternative methodology for exploring the relationships between processes of care and patient outcome in observational cohorts

Conducting randomised controlled trials for processes of care that are established as ‘best practice’ would clearly be unethical and this approach would therefore be precluded. Exploration of the relationships between processes of care and outcome therefore rely on (prospective or retrospective) observational data.

An alternative approach to regression modelling to explore the effect of individual processes on patient outcome in observational cohorts could involve the use of instrumental variables. Instrumental variables are correlated with a covariate (process marker) but not the dependent variable (or any other variables which influence the dependent variable), such that any effect of the instrumental variable on the dependent variable is through its relationship with the covariate (Pearl J 2009 p 247; Newhouse et al 1998) see also Figure 93. Therefore, the receipt of a particular process is conditional on the instrumental variable. For example, availability of a bed on a stroke unit is likely to be highly correlated with direct admission to a stroke unit, but unlikely to have direct association with patient outcome other than through the association with early stroke unit treatment. Assuming that factors such as stroke severity have no association with transfer to a stroke bed if one is available, direct admission to a stroke bed is therefore dependent on stroke bed availability and the latter could be considered to be acting as a quasi-randomising variable (Figure 93, adapted from Newhouse et al 1998).

Figure 93 Stroke unit bed availability as a possible instrumental variable.


Confounding case-mix variables
Stroke unit bed availability
Direct admission to stroke unit
Outcome

Such an approach has been used in a stroke study by (Xian et al 2011) to compare outcomes of patients admitted to stroke centres vs. non-specialist hospitals using distance from the hospital as the instrumental variable. The use of instrumental variables may therefore be useful in observational cohorts in circumstances where randomisation is unethical, or not practicable. This approach, however, requires further research.

6.3 Study limitations

6.3.1 Statistical and methodological weaknesses

The linear regression models performed in this thesis are based on the assumptions that the latent trait of the SIPSO is conceptually linked to the predictor variables, and that the Rasch transformed SIPSO subscores represent the latent trait in an older post stroke population. It is possible that there are systematic differences in the way that SIPSO questions would be answered by older patients as compared to those of the younger post-stroke population – i.e. there may be variability in the performance of the SIPSO scale across baseline patient characteristics. Differential item functioning (DIF) of age with respect to the latent trait has not been performed in an older population, and if differences were to exist in the manner that patients of different ages answer SIPSO questions, this may limit the utility or validity of the SIPSO, and the conclusions that may be drawn. This therefore represents a significant limitation of the work.

Post estimation assumptions were generally met for the models generated in the study. The notable exception is the deviation from normality of residuals in model 4. This deviation is at the tails of the distribution suggesting it may be due to outliers. Re-running models without outliers did not improve normality, and resulted in previous stroke no longer featuring as an important predictor. The violation of normality assumptions may therefore represent the the absence of important predictors, and inferiority of models that do not include baseline assessments.

The ordinal NEADL has been entered into models as a summed score, an approach which makes an assumption that it may be treated as a continuous variable. This assumption may be responsible for the deviations from linearity and normality of residuals in models to predict the physical subscore of the SIPSO. However, these deviations may also reflect the effect of particularly influential cases (i.e. cases with large residuals at high leverage points). In models containing the NEADL at baseline, these influential cases may have arisen through inconsistencies as regards whether the NEADL was completed with respect to pre- or post-stroke function. In future studies, the instructions in this regard would need to be made more explicit.

The floor and ceiling effects of the NEADL and the SIPSO may limit the validity of the models. These effects cannot be mitigated through transformations of the data and modelling alternative distributions (e.g. using censored (Tobit) regression) may have helped to circumvent these problems.

Although attempts have been made to include clinically important variables in the models, the methodological approach to the identification of important predictors of patient outcome set out in this thesis is primarily data-driven. It is important in the development of any dataset for routine collection, that the choice of variables reflects clinical, as well as statistical reasoning. However, this is with the caveat that the dataset should be small enough to be feasible for routine collection.

6.3.2 Representativeness of study population

There is a potential for selection bias in the types of patient that were recruited as compared with the general post stroke population. The additional collection of anonymised screening data meant that an objective measurement of the representativeness of the study population was made. However, the number of patients screened across three sites in six months (656) is less than would be expected. A conservative estimate of a combined population of 1.5 million in Leeds, Bradford and York (Office for National Statistics 2011) and an annual UK stroke incidence of 1.3/1000 population per year (based on London Stroke Register incidence data) (Saka O et al 2009) would mean that the number of patients screened would represent about two thirds of expected strokes in a six month period. In addition the main reason for non-eligibility of screened patients was a non-stroke diagnosis (74/193 = 38%, with 59 of these cases in York), which would tend to imply that many patients admitted with a label of “query stroke” have an alternative diagnosis. Two thirds of the patients that were screened and eligible for the study (463) were recruited (312), with the main reasons for non-recruitment being the severity of stroke or its complications leading to the need for palliative care, patients not wishing to participate, or patients lacking capacity with no carer available for assent (Figure 33, p 120). When compared with the screened population, patients recruited into the study were younger (by seven years) and less disabled but no more likely to have a carer available. The difference in baseline Barthel Index (as a measure of disability) is likely to reflect patients in receipt or likely to receive palliative care being excluded from the study.

The study sample therefore represents a group of patients aged between 31-95, median 74, with equal sex distribution and baseline BI ranging from 0-20 (median 13) who were felt on admission, to be likely to survive to discharge. In this way, the study sample is reflective of a heterogeneous population of stroke survivors, rather than the general post-stroke population. Consideration of consecutive hospital admissions (as would occur with a stroke register) includes patients that die in hospital. These patients would not have a date of discharge but a date of death and this may result in spuriously short lengths of stay given their stroke severity. The proposed use of length of stay as a case-mix adjuster is in the routine adjustment of functional post stroke outcomes. Patients in whom these are available form a subgroup of the general post stroke population, not containing patients who do not survive to follow up. The exclusion from the study population of patients who were not felt to be likely to survive to discharge was therefore not unreasonable.

A further potential cause of selection bias is withdrawals or non-response to six-month questionnaires. Patients who responded to the questionnaire compared with those that did not respond were older (median difference 7 years), with no difference in sex distribution. There is no difference in baseline BI or predicted probability of good outcome (as calculated with the SSV case-mix adjuster) between responders and non-responders to the questionnaires. There are, however, significant differences in both BI and propensity score between patients who responded and those that died or withdrew from the study. As the regression models were constructed to explore important prognostic factors in patients who survive, these differences were not felt be problematic.

The number of study sites and the sample size are too small to draw conclusions regarding specific organisational or structural aspects of care and their relationship with patient outcome. However, there are specific features at individual sites that warrant specific consideration here as potential sources of bias. Firstly, the length of stay at York hospital was significantly shorter than at the other study sites. This may reflect the higher median BI at York, however, this may also reflect particular organisational structures at York that would warrant further examination (e.g. staffing levels for therapists or a well-established Early Supported Discharge team).

Bias due to exclusion of patients with dysphasia or cognitive impairment was reduced through the use of carer assent for recruitment. Over half of the patients included in the study were reported to have dysphasia at baseline. However, although the nature of the impairment (e.g. fluent vs non-fluent dysphasia) may have a bearing on ability to complete assessments, this was not specified. The presence of these impairments increases the likelihood of proxy responses to questionnaires. Although the proxy reliability of the SIPSO has been shown to be acceptable, this is likely to have affected the validity of the responses. Differential item functioning of the SIPSO items for patients with aphasia or cognitive impairment who self-complete the SIPSO has not been performed, and this may affect the scaling properties of the measure, and therefore the validity of results in subgroups with language or cognitive impairments. Future work to explore the scaling properties of the SIPSO in these patients would be useful.

6.3.3 Data completeness and quality

The validity of the models generated in this study is reliant on the quality of the data from which they are derived. Similarly, the quality of data collected routinely (as occurs for the purposes of prospective audits such as SINAP, or for remuneration), are reliant on data capture processes. The increase in routine stroke data reporting requirements in recent years has not necessarily resulted in improvements to data collection infrastructure (see section 1.2.1). Similar data capture methodology to that which was used in this study is often employed in routine data collection – i.e. data are extracted retrospectively from case-notes and existing hospital electronic records. Therefore the problems encountered during the study in terms of data quality and missingness are likely to be generalizable to routine data collections. Indeed, many of the problems encountered in attempting to obtain accurate data for the purposes of the research study are likely to be amplified in routine data collections where data collection resource may be scare, and motivation may be more focused on the volume, rather than the quality of data. In other words, if the rationale for capturing data is to meet mandatory data requirements for the purposes of remuneration or reporting, the emphasis is not necessarily on data quality and accuracy.

The key to accuracy in data collection is in the specification and application of explicit definitions of individual data items – i.e. there should be little or no scope for interpretation at the point at which the data are extracted. It is preferable, therefore, to collect 'hard' data from which further information may be derived – i.e. dates and times that specific events occurred (e.g. date and time of admission, and date and time of imaging) rather than a series of tick boxes to indicate dichotomous responses as to whether or not a particular process was performed within a time frame (e.g. imaging within 24 hours of admission). The latter approach lacks both standardisation and validation. However, it was seen during the study, that often, the dates that were extracted from case-notes and hospital electronic systems were not accurate resulting in spurious data – indeed, many fields were not used due to concerns regarding their accuracy (for example patient movement around the hospital, or time of admission to hospital). For the purposes of the study, this led to the requirement for creation of composite variables (e.g. patient admitted to hospital on the same day, or day after their stroke), with the consequent loss of information, and ongoing concerns regarding data accuracy.

The difficulties experienced in extracting accurate times of events from existing routine data sources highlights a major barrier to the use of these data to monitor routine care for the purposes of audit or remuneration. If these data are not recorded accurately, the information and conclusions that are derived from them are also not accurate. Future data collections should therefore be focused on collecting basic data (such as time of hospital admission) accurately, before more complex data are requested. Improvements to routine data collection may be made through electronic, point of care data capture – in this way, data may be captured according to explicit standard data definitions and recorded contemporaneously by those that create it. However, this approach requires both a significant change to patterns of current working and organisation-wide change in attitudes to data collection. This approach to data collection, adopting standard data definitions to collect stroke data electronically at the point of patient care forms the basis of the next phase of the CIMSS project.

Missing data form another significant barrier to the accuracy of the study models, and may result in non-representativeness of the sample. As baseline data were extracted from case-notes, each process or case-mix data item has several potential causes of missingness: process performed but not recorded, not performed, performed and recorded but missed (not extracted) by the researchers. I had hoped to be able to identify the different causes of missing data, although this proved not to be feasible due to the complexity and additional work involved in capturing these data. These problems are, in part, a symptom of retrospective data extraction, especially if paper records are not standardised and instead collected from the narrative entries of a patient's inpatient stay. This has specific implications for variables where "yes, no, no but" codes are required, as the distinction between "no" and "no but" may not be recorded explicitly. Recording of these data in a standardised format (e.g. on a stroke proforma) is becoming more widespread (all of the study sites in the study complete paper based stroke proformas during the course of the admission) and may help to overcome these problems. However, the presence of such a proforma is no guarantee that it is adequately and accurately completed. Moreover, attention to specific aspects of process on a proforma, may lead to saturation of these processes (missing process data were infrequent, and there was little variability in patients who did, and did not receive specific aspects of care process see Figure 44 and discussion in section 5.6.8). Patients with any missing process or baseline patient reported assessment data for the variables entered into the model were automatically excluded from the regression analysis by the STATA software. However, comparison of baseline BI between these patients and those with complete data did not reveal any significant differences.

The return rate of six months outcomes questionnaires of 71% is acceptable for a postal questionnaire (Teale EA et al 2010). Examination of baseline and six month patient completed questionnaires did not reveal any pattern to the missingness (i.e. there were no questions that were consistently missed in the returned questionnaires).

This study presents a pragmatic examination of the relationships between stroke care processes and patient reported outcomes following stroke, using methodology comparable to existing routine data collection infrastructure. The components of the study datasets have been determined through systematic examination of the stroke evidence base and, at the time of writing, include the best available case-mix adjustment model and the preferred patient reported postal outcomes instrument selected by expert and consumer groups. The process dataset comprised the markers from the 2008 RCP NSSA which, at the time of the development of the research datasets, was the most standardised and regular data collection in England, Northern Ireland and Wales (Royal College of Physicians 2009b).

Chapter 7 Conclusions

7.1 Definition of a minimum dataset based on study findings

The aims of this study were to identify which patient reported outcome measure(s), case-mix adjuster and care process markers should be included in a routinely collected stroke dataset. A previous systematic review (Teale EA et al 2010) had identified candidate outcome instruments on the basis of their validity and reliability for postal collection following stroke. These were refined using group decision making techniques to the two preferred instruments of a group of consumers and stroke experts (the NEADL and the SIPSO). The SIPSO was chosen as the primary endpoint in the study due to its relatively superior properties in terms of completion rates and fewer floor and ceiling effects. However, in order to confirm the SIPSO instrument reflects the latent trait of reintegration in older stroke survivors, testing of differential item functioning with respect to age in an older population is required.

Routine collection of patient outcomes following stroke is not currently performed in England, Northern Ireland and Wales. However, this is likely to change through the Outcomes Framework (Department of Health 2010e). An open competition held by the Department of Health to identify a marker of stroke recovery (the 'Innovation in Outcomes competition') has resulted in the modified Rankin Score at six months post stroke being incorporated into the Outcomes Framework (Department of Health 2011c). It is also possible that the Patient Reported Outcome Framework (PROMs) (Department of Health 2008a) will expand to include stroke. A robust case-mix adjustment method for routine stroke outcomes data is therefore desirable.

A systematic review performed as part of this thesis has identified the SSV case-mix adjustment model as the most clinically feasible and statistically robust model for use in routine stroke care for the prediction of dichotomised OHS (see Chapter 3) (Counsell C et al 2002). This study has shown that despite being useful for stratified randomisation in clinical trials, the SSV adjustment model may lack generalizability, and therefore utility to predict outcomes other than the dichotomised OHS in routine empirical populations. This study identified that length of stay was a prominent predictor of both physical and social subscores of the SIPSO. Discriminatory properties of LOS in this (internal derivation) study showed that LOS was non-inferior to the SSV model in prediction of dichotomised OHS. This requires external validation.

Adjusting for length of stay as a proxy for stroke severity may offer a pragmatic alternative to more complex case-mix adjustment methods in observational cohorts of survivors to hospital discharge. However, it is possible that LOS may be more useful as a measure of service efficiency rather than as a proxy for stroke severity and this requires further investigation.

The main determinants of post stroke outcome in this study have been identified as direct or proxy markers of stroke severity. This study does not add additional convincing evidence that the current (RCP NSSA) process markers are associated with improved patient outcomes. However, as the majority of these process markers are near saturation, demonstration of their benefits may be limited by a lack of variability, and by the sample size of the study. Moreover, process markers are masked by (or act as) case-mix or stroke severity variables in both regression trees and linear regression models. Where process markers do feature in models, this is largely as a reflection of organisational structure rather than the delivery of particular processes of care. The exception to this observation may be a formal communication assessment in patients in whom it is indicated which was associated with clinically significant better physical outcome in one model. However, this may represent a phenomenon particular to the study dataset and would require further examination and verification in external datasets.

Markers of quality in stroke care have been changing rapidly and erratically since the publication of the National Stroke Strategy (Department of Health 2007b) (see also 1.2.1), often with little or no strong evidence to support the relationship between individual process markers and patient outcome. In the absence of this understanding of these relationships, information may be misinterpreted, service development may misdirect resources and healthcare provider institutions be unfairly sanctioned on the basis of poorly comparable data (Lilford RJ et al 2004). Moreover, concerns regarding the accuracy and quality of these routinely collected data may further limit their utility. This study offers preliminary data as regards important core predictors of functional patient outcomes that may be used as the basis of a minimum dataset for further testing.

7.2 Future work

This thesis raises a number of unanswered questions which require further exploration and verification. Firstly, is length of stay a feasible and valid alternative to more complex methods of case-mix adjustment for routine and observational post-stroke cohorts? Further testing of this hypothesis through secondary use of the FOOD trial data is planned to determine the external validity of length of stay as a univariate case-mix adjuster. The FOOD trials comprised three international multicentre randomised controlled trials of early feeding (via PEG or nasogastric tube) versus ordinary diet with the primary endpoint of dichotomised OHS at 6 months post-randomisation. Randomisation was stratified according to the SSV case-mix adjustment model. Eligibility criteria were broad, comprising patients where consent was given (or obtained from a relative), admitted to hospital within 7 days of stroke (or inpatient stroke), where the responsible clinician was unclear as to the best method of feeding. Patients with subarachnoid haemorrhage or where supplementary feeding was unlikely to be beneficial (e.g. TIA) or contraindicated (e.g. the morbidly obese, unconscious, or imminently dying) were excluded (Dennis M et al 2006). The FOOD trial data therefore represents a heterogeneous post-stroke population, and offers an opportunity to externally validate LOS against the SSV, and to test the calibration of LOS as a univariate predictor of dichotomised OHS in an external dataset.

The Rasch analysis that has been performed on the SIPSO outcome measure was performed in a population of younger stroke survivors, with consequent uncertainty regarding the properties of the scale in older patients. The CIMSS study offers an opportunity to repeat this Rasch analysis in an older population, and to examine whether there is differential item functioning for baseline patient characteristics, especially age.

A communication assessment performed in patients who required one, was the only variable featuring in the models that was likely to represent a true marker of care process. It is unclear from the current study whether it is the assessment, or any consequent therapy that afforded better outcomes in these patients. This therefore requires further verification, and exploration in an external dataset.

In line with the direction of travel from Connecting for Health (Department of Health 2011a), the development of a core stroke dataset for electronic collection should be based on the principles of robust data definitions (a data dictionary) and a standard way of combining data to derive metrics (a standard data model). Based on the findings from this study, the key fields for inclusion in such a dataset are outlined in Table 65, along with explicit data definitions. From these 17 fields, the important predictors, and case-mix variables identified in the study models may be derived. As data collection infrastructure improves, particular aspects of stroke care that are clinically important, where individual clinicians or services have particular data requirements, or where there are areas that require further research could then be added onto this core dataset in a modular and incremental fashion to describe further aspects of patient care. For example, the addition of a field to capture start and finish times of individual therapy sessions as a repeated measure would allow exploration of the optimal time frame within which a patient should be assessed by a therapist, patterns in delivery of therapy, total duration of therapy and possible ceiling effects of interventions. Moreover, if mandatory data reporting requirements were to change (for example to physiotherapy assessment within 24 hours), the individual data items that are collected need not change in order for the new metric to be derived. Of additional benefit is that this approach circumvents many of the problems with saturation of process markers through allowing the creation of continuous 'time to event' variables. Using the data in this way also allows overlap in existing data requirements to be exploited (as many fields are common to different markers and metrics) and offers reassurance that derived metrics are comparable.

The next phases of the CIMSS CLAHRC study focus on the Information Technology solution and behavioural change aspects of implementation of point of care (electronic) data capture in hospitals across West Yorkshire. The dataset that is embedded within these hospitals is based on the findings from this study, with additional fields to allow trusts to produce reports to meet existing mandatory and voluntary data requirements. Once these data collection processes are embedded, future work may examine the feasibility of prospective point of care data capture and the data dictionary approach to stroke data.

List of References

Allen CMC. Predicting the outcome of acute stroke: a prognostic score. *Journal of Neurology, Neurosurgery & Psychiatry* 1984; 47: 475-480.

Altman D. Comparability of randomised groups. Journal of the Royal Statistical Society, Series D (The Statistician) 1985; 34: 125-136.

Altman D (1999). *Practical Statistics for Medical Research* (2nd ed.) London: Chapman & Hall.

Altman D. Systematic reviews in health care: Systematic reviews of evaluations of prognostic variables. *British Medical Journal* 2001; 323: 224-248.

Altman D, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; 19: 453-473.

Altman D, Vergouwe Y, & Moons K. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal* 2009; 338: 1432-1435.

American Stroke Association's Task Force on the Development of Stroke Systems. Recommendations for the establishment of stroke systems of care. *Stroke* 2005; 36: 690-703.

Anderson CS, Jamrozik KD, Broadhurst RJ, & Stewart-Wynne EG. Predicting survival for 1 year among different subtypes of stroke: Results from the Perth community stroke study. *Stroke* 1994; 25(10): 1935-44.

Asplund K and RIKS stroke register (2011). *The RIKS stroke register*. Retrieved 23-09-2011, from www.riks-stroke.org/index.php?content=&lang=eng&text=

Australian Stroke Clinical Registry (2011). *The Australian Stroke Clincial Registry (AuSCR)*. Retrieved 26-09-2011, from www.auscr.com.au/

Bamford J, Sandercock P, & Dennis M. A prospective study of acute cerebrovascular disease in the community: the Oxfordshire Community Stroke Project 1981-86. *Journal of Neurology, Neurosurgery & Psychiatry* 1988; 51: 1373-1380.

Barer DH, Mitchell JRA. Predicting the Outcome of Acute Stroke: Do Multivariate Models Help? *Quarterly Journal of Medicine* 1989; 70(1): 27-39.

Barton A, Mulley G. History of the development of geriatric medicine in the UK. *Postgraduate Medical Journal* 2003; 79(930): 229-234.

Bevan G, Hood C. What's measured is what matters: targets and gaming in the English Public Health Care system. *Public Administration* 2006; 84(3): 517-538.

Bewick V, Cheek L, & Ball J. Statistics review 14: Logistic regression. *Critical care* 2005; 9(1): 112-118.

Black N. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal* 1996; 312(7040): 1215-1218.

Boehringer Ingelheim (2009). *Actilyse: Summary of product characteristics*. Retrieved 24-02-2010, from emc.medicines.org.uk/medicine/308/SPC/Actilyse/

Bravata DM., Wells CK, Lo AC, Nadeau SE, Melillo J, Chodkowski D et al. Processes of Care Associated With Acute Stroke Outcomes. *Archives of Internal Medicine* 2010; 170(9): 804-810.

Buetow SA, Roland SO. Clinical governance: bridging the gap between managerial and clinical approaches to quality of care. *Quality in Health Care* 1999; 8: 184-190.

Campbell SM, Roland SO, & Buetow SA. Defining Quality of Care. *Social Science and Medicine* 2000; 51: 1611-1625.

Centre for Reviews and Dissemination (2009). *Systematic Reviews: CRD's guidance for undertaking reviews in healthcare*. University of York. Retrieved 24-10-2011 from www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf

Chen X, Ender P, Mitchell M, and Wells C (2003). *Regression with Stata.* Retrieved 15-08-2011, from www.ats.ucla.edu/stat/stata/webbooks/reg/default.htm

Chen Z-M, Chinese Acute Stroke Trial Collaborative Group. CAST: randomised placebo-controlled trial of early aspirin use in 20000 patients with acute ischaemic stroke. *The Lancet* 1997; 349: 1641-1649.

CLOTS Trial Collaboration. Effectiveness of thigh-length graduated compression stockings to reduce the risk of deep vein thrombosis after stroke (CLOTS trial 1): a multicentre, randomised controlled trial. *The Lancet* 2009; 373(9679): 1958-1965.

CLOTS Trial Collaboration (2011). *Clots in Legs or sTockings after Stroke*. Retrieved 21-09-2011, from www.dcn.ed.ac.uk/clots/

Cochrane Database of Systematic Reviews Stroke Review Group (2009). *Stroke search strategy*. Retrieved 10-06-2009 onlinelibrary.wiley.com/o/cochrane/clabout/articles /STROKE/frame.html .

Concato J, Feinstein AR, & Holford TR. The risk of determining risk with multivariable models. *Annals of Internal Medicine* 1993; 118: 201-210.

Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovascular Diseases* 2001; 12(3): 159-170.

Counsell C, Dennis M, & McDowall M. Predicting functional outcome in acute stroke: comparison of a simple six variable model with other predictive systems and informal clinical prediction. *Journal of Neurology, Neurosurgery and Psychiatry* 2004; 75: 401-405.

Counsell C, Dennis M, McDowall M, & Warlow C. Predicting Outcome After Acute and Subacute Stroke: Development and Validation of New Prognostic Models. *Stroke* 2002; 33(4): 1041-1047.

Darzi A (2008). *High Quality Care for All*. London: Crown Copyright. Retrieved 24-02-2010 from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_085828.pdf.

Davenport RJ, Dennis MS, & Warlow CP. Effect of correcting outcome data for case mix: an example from stroke medicine. *British Medical Journal* 1996; 312: 1503-1505.

Davies HTO (2005). *Measuring and reporting the quality of health care: issues and evidence from the international research literature*. NHS Quality Improvement Scotland. Retrieved 23-09-2010 from www.clinicalgovernance.scot.nhs.uk/documents/Davies%20Paper.pdf

Davies HTO, Crombie IK. Interpreting health outcomes. *Journal of Evaluation in Clinical Practice* 1997; 3(3): 187-199.

Dennis M and Scottish Stroke Care Audit Steering Committee (2011). *The Scottish Stroke Care Audit*. Retrieved 23-09-2011 from www.strokeaudit.scot.nhs.uk/

Dennis M, Lewis S, Cranswick G, & Forbes J. FOOD: A multicentre randomized trial evaluating feeding policies in patients admitted to hospital with a recent stroke. *Health Technology Assessment* 2006; 10(2): 1-91.

Dennis MS, Cranswick G, Fraser A, Grant S, Gunkel A, Hunter J et al. Performance of a statistical model to predict stroke outcome in the context of a large, simple, randomized, controlled trial of feeding. *Stroke* 2003; 34(1): 127-133.

Department of Health (1999). *Quality in the New NHS*. Retrieved 09-02-2009, from www.dh.gov.uk/en/Publicationsandstatistics/Lettersandcirculars/Healthservicecirculars/DH_4004883.

Department of Health (2007a). *Payment by Results for stroke and TIA services*. Retrieved 24-02-2010 from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_086860.pdf

Department of Health (2007b). *The National Stroke Strategy*. Retrieved 24-02-2010 from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_081059.pdf

Department of Health (2007c). *The New NHS; modern, dependable*. Retrieved 22-07-2011, from www.archive.official-documents.co.uk/document/doh/newnhs/wpaper3.htm

Department of Health (2007d). *Now I feel Tall.* Retrieved 22-02-2012, from http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/Browsable/DH_5575267

Department of Health (2008a). *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)*. Retrieved 23-09-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_092625.pdf

Department of Health (2008b). *Operational Plans 2008/9-2010/11 (implementing the 2008/09 framework)*. Retrieved 24-02-2010, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_082560.pdf

Department of Health (2008c). *Using the Commissioning for Quality and Innovation (CQUIN) payment framework: For the NHS in England 2009/10*. Retrieved 24-02-2010, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_091435.pdf

Department of Health (2008d). *Vital Signs Monitoring Return - Health Improvement and Reducing Health Inequalities*. Retrieved 24-02-2010, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_090849.pdf

Department of Health (2009). *NHS 2010-2015: from good to great. Preventative, people-centred, productive*. Retrieved 25-09-2011 from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/@sta/@perf/documents/digitalasset/dh_109887.pdf

Department of Health (2009b). *The Operating Framework for the NHS in England 2010/11*. Retrieved 24-02-2010, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/@sta/@perf/documents/digitalasset/dh_110159.pdf

Department of Health (2010a). *Equity and Excellence: Liberating the NHS*. Retrieved 25-09-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_117794.pdf

Department of Health (2010b). *Payment by Results: Guidance for 2010/11*. Retrieved 15-10-2010, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_112970.pdf

Department of Health (2010c). *Revision to the Operating Framework for the NHS in England 2010/11*. Retrieved 29-04-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_116860.pdf

Department of Health (2010d). *The National Quality Board*. Retrieved 07-01-2011, from www.dh.gov.uk/en/Healthcare/NationalQualityBoard/index.htm

Department of Health (2010e). *The NHS Outcomes Framework 2011/12*. Retrieved 29-04-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/@ps/documents/digitalasset/dh_123138.pdf

Department of Health (2011a). *An Information Revolution: a consultation on proposals: Launched 18/10/2010*. Retrieved 07-01-2011, from www.dh.gov.uk/en/Consultations/Liveconsultations/DH_120080

Department of Health (2011b). *Patient Reported Outcome Measures (PROMs) in England: a methodology for identifying potential outliers*. Retrieved 25-09-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_128447.pdf

Department of Health (2011c). *"Winners of Innovation in Outcomes competition announced"*. Retrieved 21-03-2012 from http://www.dh.gov.uk/health/2011/10/winners-of-innovation-in-outcomes-competition-announced/

Donabedian A. Evaluating the Quality of Medical Care. *The Milbank Memorial Fund Quarterly* 1966; 44(3) Part 2: 166-203.

Dr Foster (2010). *Hospital Guide 2010 "Test Results"*. Retrieved 28-09-2011, from www.drfosterhealth.co.uk/docs/hospital-guide-2010.pdf

Dubois RW, Rogers WH, Moxley JH, Draper D, & Brook RH. Hospital inpatient mortality. Is it a predictor of quality? *New England Journal of Medicine* 1987; 317(26): 1674-1680.

Early Supported Discharge Trialists. Services for reducing duration of hospital care for acute stroke patients. *Cochrane Database of Systematic Reviews* 2005; Issue 2. Art. No.: CD000443. DOI: 10.1002/14651858.CD000443.pub2.

Evans A, Perez I, Harraf F, Melbourn A, Steadman J, Donaldson N et al. Can differences in management processes explain different outcomes between stroke unit and stroke-team care? *Lancet* 2001; 358: 1586-1592.

Fox J (1997). Applied regression analysis, linear models, and related methods. (1st ed.) Thousand Oaks: Sage Publications.

Frenk J. Obituary: Avedis Donabedian. *Bulletin of the World Health Organisation* 2000; 78(12): 1475.

Frithz G, Werner I. Studies on cerebrovacsular strokes II Clinical findings and short-term prognosis in a stroke material. *Acta Medica Scandinavica* 1976; 199(1-6): 133-140.

Fullerton KJ, MacKenzie G, & Stout RW. Prognostic indices in stroke. *Quarterly Journal of Medicine* 1988; 66(250): 147-162.

Gale CP, Roberts AP, Batin PD, & Hall AS. Funnel plots, performance variation and the Myocardial Infarction National Sudit Project 2003-2004. *BMC Cardiovascular disorders* 2006; 6(34).

Gladman JR, Harwood DM, & Barer DH. Predicting the outcome of acute stroke: prospective evaluation of five multivariate models and comparison with simple methods. *Journal of Neurology, Neurosurgery & Psychiatry* 1992; 55(5): 347-51.

Gladman JR, Lincoln NB. Follow-up of a controlled trial of domiciliary stroke rehabilitation (DOMINO Study). *Age & Ageing* 1994; 23(1): 9-13.

Gladman JR, Lincoln NB, & Adams SA. Use of the extended ADL scale with stroke patients. *Age and Ageing* 1993; 22(6): 419-424.

Goldberg D & Williams P (1988). *A user's guide to the General Health Questionnaire*. Windsor, UK: NFER-Nelson.

Gompertz P, Pound P, & Ebrahim S. Predicting stroke outcome: Guy's prognostic score in practice. *Journal of Neurology, Neurosurgery & Psychiatry* 1994; 57(8): 932-5.

Govan L, Langhorne P, Weir CJ, & the Stroke Unit Trialists Collaboratio. Does the Prevention of Complications Explain the Survival Benefit of Organized Inpatient (Stroke Unit) Care? *Stroke* 2007; 38(9): 2536-2540.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143(1): 29-36.

Harrell FE, Lee KL, & Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; 15(4): 361-387.

Hatano S, on behalf of the participants in the WHO Collaborative Study on the Control of Stroke in the Community. Experience from a multicentre stroke register: a preliminary report. *Bullitin of the World Health Organisation* 1976; 54(3550): 541-553.

Hayden JA., Côté P, & Bombardier C. Evaluation of the Quality of Prognosis Studies in Systematic Reviews. *Annals of Internal Medicine* 2006; 144(6): 427-437.

Health Quality Improvement Partnership (2011) *Sentinel Stroke National Audit Programme (SSNAP)*. Retrieved 24-09-2011, from www.hqip.org.uk/procurement

Healthcare Quality Improvement Partnership (HQIP) (2011). *National Clinical Audits for inclusion in Quality Accounts 2011-12*. Retrieved 21-09-2011, from www.hqip.org.uk/national-clinical-audits-for-inclusion-in-quality-accounts-2011-201/

Heberden W (1892). Commentaries on the history and cure of diseases. (Special edition for the Classics of Medicine Library). Birmingham, Alabama: Leslie B Adams Jr.

Hemingway H, Riley RD, & Altman D. Ten steps towards improving prognosis research. *British Medical Journal* 2010; 339(b4181): 410-414.

Henderson GR, Mead SE, van Dijke ML, Ramsay S, McDowall MA, & Dennis M. Use of statistical process control charts in stroke medicine to determine if clinical evidence and changes in service delivery were associated wtih improvements in the quality of care. *Quality and Safety in Healthcare* 2008; 17(301): 306.

Hier HB, Edelstein G. Deriving clinical prediction rules fom stroke outcome research. *Stroke* 1991; 22: 1431-1436.

Horn SD, DeJong G, Ryser D, Veazie PJ, & Teraoka J. Another look at observational studies in rehabilitation research: going beyond the holy grail of the Randomised Controlled Trial. *Archives of Physical Medicine and Rehabilitation* 2005; 86(Suppl 2): S8-S15.

Hospital Episode Statistice (HESonline) (2011). *PROMs data*. Retrieved 25-09-2011, from www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1295

Indredavik B, Bakke F, Slordahl SA, Rosketh R, & Haheim LL. Treatment in a combined acute and rehabilitation Stroke Unit: Which aspects are most important? *Stroke* 1999; 30(917): 923.

Institute for Innovation and Improvement (2010). *Commissioning for Quality and Innovation (CQUIN) payment framework*. Retrieved 01-03-2010, from www.institute.nhs.uk/world_class_commissioning/pct_portal/cquin.html

Intercollegiate Stroke Working Party (2008). *National clinical guideline for stroke 3rd ed.* Retrieved 03-04-2011, from bookshop.rcplondon.ac.uk/contents/6ad05aab-8400-494c-8cf4-9772d1d5301b.pdf

Intercollegiate Stroke Working Party (2010). *National Sentinel Stroke Audit Organisational Audit*. Retrieved 23-09-2011, from http://www.rcplondon.ac.uk/sites/default/files/2010-stroke-public-report.pdf

Intercollegiate Stroke Working Party (2011). *National Sentinel Stroke Clinical Audit 2010, Round 7, Public Report for England, Wales and Northern Ireland*. Royal College of Physicians. Retrieved 23-6-2011, from http://www.rcplondon.ac.uk/sites/default /files/national-sentinel-stroke-audit-2010-public-report-and-appendices_0.pdf .

International Stroke Trial Collaborative Group. The International Stroke trial (IST): a randomised trial of aspirin,subcutaneous heparin,both,or neither among 19 435 patients with acute ischaemic stroke. *The Lancet* 1997; 349(9065): 1569-1581.

Ipsos MORI (2011). *Public perceptions of the NHS and social care*. Retrieved 28-09-2011, from www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_126705.pdf

IST3 trialists. Predicting outcome in hyper-acute stroke: validation of a prognostic model in the Third International Stroke Trial (IST3). *Journal of Neurology, Neurosurgery & Psychiatry* 2008; 79(4): 397-400.

Jenkinson C, Gibbons E, and Fitzpatrick R (2009). *A structured review of patient-reported outcome measures in relation to stroke*. Retrieved 24-09-2011, from http://phi.uhce.ox.ac.uk/pdf/PROMs_Oxford_Stroke_17092010.pdf

Johnston KC, Connors AF, Jr., Wagner DP, & Haley EC, Jr. Predicting outcome in ischemic stroke: external validation of predictive risk models. *Stroke* 2003; 34(1): 200-202.

Johnston KC, Connors A, Wagner D, Knaus W, Wang X, & Clarke HE. A predictive risk model for outcomes of ischemic stroke. *Stroke* 2000; 31(2): 448-55.

Jongbloed L. Prediction of function after stroke: a critical review. *Stroke* 1986; 17 765-776.

Jorgensen HS. The Copenhagen Stroke Study Experience. *Journal of Stroke and Cerebrovascular Diseases* 1996; 6(1): 5-16.

Justice AC, Covinsky KE, & Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; 130: 515-524.

Kalra L, Dale P, & Crome P. Evaluation of a Clinical Score for Prognostic Stratification of Elderly Stroke Patients. *Age and Ageing* 1994; 23(6): 492-498.

Kalra L, Crome P. The role of prognostic scores in targeting stroke rehabilitation in elderly patients. *Journal of the American Geriatrics Society* 1993; 41(4): 396-400.

Kersten P, George S, Low J, Ashburn A, & McLellan L. The subjective index of physical and social outcome: Its usefulness in a younger stroke population. *Int J Rehabil Res* 2004; 27: 59-63.

Kersten P, Ashburn A, George S, & Low J. The Subjective Index for Physical and Social Outcome (SIPSO) in Stroke: investigation of its subscale structure. *BMC Neurology* 2010; 10(1): 26.

Kline RB (2005). *Principles and Practice of Structural Equation Modelling*. (2nd ed.) New York: The Guilford Press.

Kwakkel G, Wagenaar RC, Kollen BJ, & Lankhorst GJ. Predicting disability in stroke - A critical review of the literature. *Age and Ageing* 1996; 25(6): 479-89.

Lai SM, Duncan PW, & Keighley J. Prediction of functional outcome after stroke: comparison of the Orpington Prognostic Scale and the NIH Stroke Scale. *Stroke* 1998; 29(9): 1838-1842.

Langhorne P, Sandercock P, & Prasad K. Evidence-based practice for stroke. *The Lancet Neurology* 2009; 8: 308-309.

Langhorne P, Pollock A, & in Conjunction with The Stroke Unit Trialists' Collaboration. What are the components of effective stroke unit care? *Age and Ageing* 2002; 31(5): 365-371.

Laupacis A, Sekar N, & Stiell IG. Clinical prediction rules: A review and suggested modifications of methodological standards. *The Journal of the American Medical Association* 1997; 277: 488-494.

Lewis S, Dennis M, & Sandercock P. Predicting outcome in hyper-acute stroke - validation of a prognostic model in the Third International Stroke Trial (IST3). *Journal of Neurology, Neurosurgery & Psychiatry* 2007.

Lilford RJ, Brown CA, & Nicholl J. Use of process measures to monitor the quality of clinical practice. *British Medical Journal* 2007; 335: 648-650.

Lilford RJ, Mohammed MA, Speigelhalter D, & Thompson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *The Lancet* 2004; 363: 1147-1154.

Lilford RJ, Pronovost P. Using hospital mortality rates to judge hospital performane: a bad idea that just won't go away. *British Medical Journal* 2010; 340: 955-957.

Lilford RJ, Mohammed MA, Spiegelhater D, & Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care; avoiding institutional stigma. *Lancet* 2004; 363: 1147-1154.

Lincoln N, Jackson JM, Edmans JA, Walker MF, Farrow VM, Latham A et al. The accuracy of predictions about progress of patients on a stroke unit. Journal of Neurology, Neurosurgery & Psychiatry (1990); 53: 972-975.

Lincoln NB, Gladman JRF. The Extended Activities of Daily Living scale: A further validation. *Disability and Rehabilitation* 1992; 14(1): 41-43.

Lindley RI, Waddell F, & Livingstone M. Can simple questions assess outcome after stroke? *Cerebrovascular diseases* 1994; 4: 314-324.

Lindsay MP, Gubitz G, Bayley M, Hill MD, Davies-Schinkel C, Singh S et al. (2010). The Canadian Best Practice Recommentations for Stroke Care (Update 2010). On behalf of the Canadian Stroke Strategy Best Practices and Standards Writing Group. Ottawa, Ontario Canada: Canadian Stroke Network.

Lingsma HF, Dippel DW, Hoeks SE, Steyerberg EW, Franke CL, van Oostenbrugge RJ et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *Journal of Neurology, Neurosurgery & Psychiatry* 2008; 79(8): 888-94.

Lyden PD, Lu M, Levine SR, Brott TG, Broderick J, & Ninds rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. *Stroke* 2001; 32(6): 1310-1317.

Mallett S, Royston P, Dutton S, Waters R, & Altman D. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine* 2010a; 8 20.

Mallett S, Royston P, Waters R, Dutton S, & Altman D. Reporting performance of prognostic models in cancer: a review. *BMC Medicine* 2010b; 8 21.

Mant J. Process versus outcome indicators in the assessment of quality of health care. *International Journal for Quality in Health Care* 2001; 13(6): 475-480.

Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *British Medical Journal* 1995; 311: 793-796.

Mant J, Hicks NR, & Fletcher J. Letter: study should have had more patients of longer time scale. *British Medical Journal* 1996; 313: 1006.

Masiero S, Avesani R, Armani M, Verena P, & Ermani M. Predictive factors for ambulation in stroke patients in the rehabilitation setting: a multivariate analysis. *Clinical Neurology & Neurosurgery* 2007; 109(9): 763-9.

McArthur KS, Quinn TJ, Higgins P, & Langhorne P. Post-acute care and secondary prevention after ischaemic stroke. *British Medical Journal* 2011; 342: 861-867.

McLaughlin V, Leatherman S, Fletcher M, & Wyn Owen J. Improving performance using indicators. Recent experiences in the United States, the United Kingdom, and Australia. *International Journal for Quality in Healthcare* 2001; 13(6): 455-467.

McNaughton H, McPherson K, Taylor W, & Weatherall M. Relationship between process and outcome in stroke care. *Stroke* 2003; 34: 713-717.

Mears A, Webley P. Gaming of performance measurement in health care: parallels with tax compliance. *Journal of Health Services Research & Policy* 2010; 15(4): 236-242.

Meijer R, Ihnenfeldt D, van LJ, Vermeulen M, & de HR. Prognostic factors in the subacute phase after stroke for the future residence after six months to one year. A systematic review of the literature. *Clinical Rehabilitation* 2003a; 17(5): 512-20.

Meijer R, Ihnenfeldt D, de Groot I, van Limbeek J, Vermeulen M, & de Haan R. Prognostic factors for ambulation and activities of daily living in the subacute phase after stroke. A systematic review of the literature. *Clinical Rehabilitation Vol* 2003b; 17(2): 119-29.

Meijer R, van Limbeek J, Kriek B, Ihnenfeldt D, Vermeulen M, & de Haan R. Prognostic social factors in the subacute phase after a stroke for the discharge destination from the hospital stroke-unit. A systematic review of the literature. *Disability and Rehabilitation: An International, Multidisciplinary Journal Vol* 2004; 26(4): 191-97.

Mohammed MA. Bristol, Shipman and clinical governance: Shewhart's forgotten lessons. *The Lancet* 2001; 357: 463-467.

Mohammed MA, Mant J, Bentham L, & Raferty J. Comparing processes of stroke care in high and low-mortality hospital in the West Midlands, UK. *International Journal for Quality in Healthcare* 2005; 17(1): 31-36.

Muir KW., Weir CJ, Murray GD., Povey C, & Lees KR. Comparison of Neurological Scales and Scoring Systems for Acute Stroke Prognosis. *Stroke* 1996; 27(10): 1817-1820.

National Audit Office. Reducing Brain Damage: Faster access to better stroke care. 2005.

National Audit Office (2010). *Progress in improving stroke care*. London: UK Stationary Office. Retrieved 15-10-2010, from www.nao.org.uk/publications/0910/stroke.aspx

National Institute for Health and Clinical Excellence (2008). Stroke: Diagnosis and initial management of acute stroke and transient ischaemic attack (TIA). Developed by the National Collaborating Centre for Chronic Conditions. Retrieved 03-04-2011, from www.nice.org.uk/nicemedia/live/12018/41331/41331.pdf

National Institute for Health and Clinical Excellence (2010a). *NICE quality standards*. Retrieved 07-01-2011a, from www.nice.org.uk/aboutnice/qualitystandards /qualitystandards.jsp

National Institute for Health and Clinical Excellence (2010b). *Stroke Quality Standard*. Retrieved 07-01-2011b, from www.nice.org.uk/media/7EC/67/StrokeQualityStandard.pdf

Newhouse JP, McClellan M. Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health* 1998;19: 17-34

NHS Connecting for Health (2011). *What is data quality*. Retrieved 31-01-2011, from www.connectingforhealth.nhs.uk/systemsandservices/data/dataquality/whatisdq

NHS Improvement (2010a). *Accelerating Stroke Improvement*. Retrieved 07-01-2011a, from http://www.improvement.nhs.uk/stroke/MeasuringforImprovement/tabid/185/Default.aspx

NHS Improvement (2010b). *The Stroke Improvement Programme*. Retrieved 15-10-2010b, from www.improvement.nhs.uk/stroke/StrokeCareNetworks/tabid/55/Default.aspx

Nicholl J. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *Journal of Epidemiology and Community Health* 2007; 61(11): 1010-1013.

O'Brien S, DeLong E, & Peterson ED. Impact of case volume on hospital performance assessment. *Archives of Internal Medicine* 2008; 168(12): 1277-1284.

Office for National Statistics (2011). *Population estimates for UK, England and Wales, Scotland and Northern Ireland, mid 2010 popultation estimates*. Retrieved 25-09-2011, from www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-231847

Pearl J (2009). *Causality: Models, reasoning and inference*. (2nd ed.) New York: Cambridge University Press.

Peduzzi P, Concato J, Kemper E, Holford TR, & Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; 49(12): 1373-1379.

Perel P, Edwards P, Wentz R, & Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC medical informatics and decision making* 2006; 6 38.

Petersdorf RG, Adams RD, Braunwalk E, Isselbacher KJ, Martin JB, & Wilson JD (1983). *Harrison's Principles of Internal Medicine*. (10th ed.) McGraw Hill.

Pitches DW, Mohammed MA, and Lilford RJ. What is the empirical evidence that hosptials with higher risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Services Research* 2007;7:91 www.biomedcentral.com/1472-6963/7/91.

Pound P, Bury M, & Ebrahim S. From apoplexy to stroke. *Age and Ageing* 1997; 26 331-337.

Prescott RJ, Garraway WM, & Akhtar AJ. Predicting functional outcome following acute stroke using a standard clinical examination. *Stroke* 1982; 13(5): 641-647.

Price CIM, Curless RH, & Rodgers H. Can stroke patients use visual analogue scales? *Stroke* 1999; 30(7): 1357-1361.

Rabin R, Oemar M, & Oppe M (2011). *EQ-5D 3L User Guide on behalf of the EuroQoL Group*. Version 4.0 Retrieved 30-09-2011 from www.euroqol.org/fileadmin/user_upload /Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-3L.pdf

Reid J, Gubitz GJ, Dai D, Reidy Y, Christian C, Counsell C et al. External validation of a six simple variable model of stroke outcome and verification in hyper-acute stroke. *Journal of Neurology, Neurosurgery & Psychiatry* 2007; 78(12): 1390-1391.

Robinson B. Validation of a Caregiver Strain Index. *Journal of Gerontology* 1983; 38 344-348.

Roffe C (2011). *Stroke Oxygen Study*. Retrieved 21-09-2011, from www.so2s.co.uk/protocol.shtml

Rothwell PM. Prognostic models. *Practical Neurology* 2008; 8: 242-253.

Royal College of Physicians (2009a). *National Sentinel Stroke Audit Phase 1 Organisational audit 2008: Report for England, Wales and Northern Ireland*. London. Retrieved 24-02-2010 from www.rcplondon.ac.uk/clinical-standards/ceeu/Current-work/Documents/Public%20 organisational%20report2008.pdf .

Royal College of Physicians (2009b). *National Sentinel Stroke Audit Phase II: (clinical audit) 2008*. Retrieved 24-02-2010b, from www.rcplondon.ac.uk/clinical-standards/ceeu/Current-work/stroke/Documents/stroke-audit-report-2008.pdf

Royal College of Physicians (2011). *Stroke Improvement National Audit Programme (SINAP) public results*. Retrieved 23-09-2011, from www.rcplondon.ac.uk/sites/default/files/sinap-public-results-1april-2011-to-june-2011-admissions.pdf

Royal College of Physicians Stroke Programme (2010). *SINAP (Stroke Improvement National Audit Programme)*. Retrieved 10-04-2010, from www.rcplondon.ac.uk/clinical-standards/ceeu/Current-work/stroke/Pages/SINAP.aspx

Royston P, Moons KGM, Altman D, & Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *British Medical Journal* 2008; 338: 1373-1377.

Rubin HR, Pronovost P, Diette GB. The advantages and disadvantages of process-based measures of health care quality. *International Journal of Quality in Healthcare.* 2001; 13(6): 469-474

Saka O, McGuire A, & Wolfe C. Cost of stroke in the United Kingdom. *Age and Ageing* 2009; 38: 27-32.

Salter K, Jutai J, Zettler L, Moses M, McClure A, Foley N, and Teasell R (2010). *Evidence based review of stroke rehabilitation: Outcome measures in stroke rehabilitation*. Retrieved 20-06-2011, from www.ebrsr.com/uploads/Outcome-Assessment-SREBR-13.pdf

Scott IA, Ward M. Public reporting of hospital outcomes based on administrative data: risks and opportunities. *Medical Journal of Australia* 2006; 184(4): 571-575.

Scottish Intercollegiate Guidelines Network (SIGN) (2008). *A guideline developer's handbook*. NHS Quality Improvement Scotland. Retrieved 25-9-2011, from www.sign.ac.uk/pdf/sign50.pdf .

Seenan P, Long M, & Langhorne P. Strokes in their natural habitat. *Stroke* 2007; 38: 1886-1892.

Segal M, Whyte J. Modeling case mix adjustment of stroke rehabilitation outcomes. *American Journal of Physical Medicine & Rehabilitation* 1997; 76(2): 154-161.

Speigelhalter D. Funnel plots for institutional comparison. *Quality and Safety in Healthcare* 2002; 11: 390-391.

Speigelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; 24 1185-1202.

StatSoft Inc (2011). *Classification and regression trees in Electronic Statistics Textbook*. Retrieved 17-10-2011, from www.statsoft.com/textbook/

Stroke Improvement Programme (2011). *Operational definitions and guidance for ASI collection*. Retrieved 24-09-2011, from system.improvement.nhs.uk/ImprovementSystem /ViewDocument.aspx?path=Cardiac%2fNational%2fStroke%20Improvement%20Programm e%2fAccelerating%20Stroke%20Improvement%202010-11%2fOperational%20Guidance%20for%20ASI%202011-12%20final%20version.pdf

Stroke Unit Trialists' Collaboration. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews* 2007; Issue 4 Art. No.: CD000197. DOI: 10.1002/14651858.CD000197.pub2.

Sudlow C, Warlow C. Getting the priorities right for stroke care. *British Medical Journal* 2009; 338: 1419-1422.

Teale EA, Young JB. A review of stroke outcome measures valid and reliable for administration by postal survey. *Reviews in Clinical Gerontology* 2010; 20: 338-353.

Teale EA, Forster A, Munyombwe T, Young JB "A systematic review of case-mix adjustment models for stroke" Clinical Rehabilitation published online ahead of print January 2012 available from http://cre.sagepub.com/content/early/2012/01/17/0269215511433068

The Care Quality Commission (2009). Retrieved 30-12-2009, from www.cqc.org.uk/aboutcqc/whoweare.cfm.

The EuroQoL Group. Euroqol - A New Facility for the Measurement of Health-Related Quality-Of-Life. *Health Policy* 1990; 16(3): 199-208.

The Information Centre (2011a). *HESonline*. Retrieved 25-09-2011a, from www.hesonline.org.uk/Ease/ContentServer?siteID=1937&categoryID=537

The Information Centre (2011b). *Patient Reported Outcome Measures*. Retrieved 25-09-2011b, from www.ic.nhs.uk/proms

The Information Centre (2011c). *Provisional PROMs data 2010-11*. Retrieved 21-09-2011c, from www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1582

The King's Fund (2011). *Patient-centred care*. Retrieved 25-09-2011, from www.kingsfund.org.uk/topics/patientcentred_care/#background

The National Institute of Neurological Disorders and Stroke rt-PA stroke study group. Tissue plasminogen activator for acute ischemic stroke. *The New England Journal of Medicine* 1995; 333(1581): 1587.

Thomassen L, Thorén M, Leys D, Roine R, & Anderson T (2006). On behalf of the 6th Karolinska Stroke Update, Stockholm Sweden. 'Organised acute stroke care'. Karolinska Stroke Update Consensus Statement. Retrieved 7-6-2011, from www.strokeupdate.org/Cons_organised_2006.aspx .

Tilling K, Sterne J, Rudd A, Glass T, Wityk R, & Wolfe C. A new method for predicting recovery after stroke. *Stroke* 2001a; 32(12): 2867-73.

Tilling K, Sterne JA, & Wolfe CD. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Statistics in Medicine* 2001b; 20(5): 3474-86.

Tirschwell DL, Longstreth WT, Jr., Becker KJ, Gammans RE, Sr., Sabounjian LA, Hamilton S et al. Shortening the NIH Stroke scale for use in the prehospital setting. *Stroke* 2002; 33(12): 2801-6.

Trigg R, Wood VA. The Subjective Index of Physical and Social Outcome (SIPSO): A new measure for use with stroke patients. *Clinical Rehabilitation* 2000; 14: 288-299.

Trigg R, Wood VA. The validation of the subjective index of physical and social outcome (SIPSO). *Clinical Rehabilitation* 2003; 17: 283-289.

Twisk J, Rijmen F. Longitudinal tobit regression: A new approach to analyze outcome variables with floor or ceiling effects. *Journal of Clinical Epidemiology* 2009; 62: 953-958.

UCLA: Academic Technology Services, SCG (2011). *Likelihood ratio test*. Retrieved 17-08-2011, from www.ats.ucla.edu/stat/stata/faq/nested_tests.htm

Wade DT, Skilbeck CE, & Langton Hewer R. Predicting Barthel ADL score at 6 months after an acute stroke. *Archives of Physical Medicine and Rehabilitation* 1983; 64: 24-28.

Walsh K, Gompertz PH, & Rudd AG. Stroke care: how do we measure quality? *Postgraduate Medical Journal* 2002; 78(920): 322-326.

Wang Y, Lim L, Heller R, Fisher J, & Levi C. A prediction model of 1-year mortality for acute ischemic stroke patients. *Archives of Physical Medicine and Rehabilitation* 2003; 84(7): 1006-11.

Wardlaw JM, Murray V, Berge E, & del Zoppo GJ. Thrombolysis for acute ischaemic stroke. *Cochrane Database of Systematic Reviews* 2009; Issue 4. Art. No.: CD000213. DOI: 10.1002/14651858.CD000213.pub2.

Waterman R (1999). *Leverage, residuals and influence*. Retrieved 15-10-2011, from www-stat.wharton.upenn.edu/~waterman/Teaching/701f99/Class04/class04.pdf

Weimar C, Konig IR, Kraywinkel K, Ziegler A, Diener HC, & German Stroke Study C. Age and National Institutes of Health Stroke Scale Score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models. *Stroke* 2004; 35(1): 158-62.

Weimar C, Roth M, Willig V, Kostopoulos P, Benemann J, & Diener HC. Development and validation of a prognostic model to predict recovery following intracerebral hemorrhage. *Journal of Neurology* 2006; 253(6): 788-93.

Weimar C, Ziegler A, Konig IR, & Diener HC. Predicting functional outcome and survival after acute ischemic stroke. *Journal of Neurology* 2002; 249(7): 888-95.

Weir N, Dennis MS. Towards a national system for monitoring the quality of hospital-based stroke services. *Stroke* 2001; (32): 1415-1421.

Weir N, Counsell C, McDowall M, Gunkel A, & Dennis M. Reliability of the variables in a new set of models that predict outcome after stroke. *Journal of Neurology, Neurosurgery and Psychiatry* 2003; 74(4): 447-451.

Whitley E, Ball J. Statistics review 4: Sample size calculations. *Critical care* 2002; 6 335-341.

Williams G, Jiang J. Development of an ischemic stroke survival score. *Stroke* 2000; 31(10): 2414-20.

Wyatt JC. Acquisition and use of clinical data for audit and research. *Journal of Evaluation in Clinical Practice* 1995; 1(1): 15-27.

Xian Y, Holloway RG, Chan PS, Noyes K, Shah MN, Ting HH et al. Association Between Stroke Center Hospitalization for Acute Ischemic Stroke and Mortality. *Journal of the American Medical Association* 2011; 305(4): 373-380.

Young J, Bogle S, & Forster A. Determinants of social outcome measured by the Frenchay Activities Index at one year after stroke onset. *Cerebrovascular diseases* 2001; 12(2): 114-20.

List of Abbreviations

ANOVA	Analysis of variance
ARR	Absolute Risk Reduction
ASI	Accelerating Stroke Improvement
ASU	Acute Stroke Unit
AUC	Area under the (ROC) curve
BCOS	Bakas Carer Outcomes score
BI	Barthel Index
BPT	Best Practice Tariff
CBS	Carer Burden Score
CCU	Coronary Care Unit
CI	Confidence interval
CIMSS	Clinical Information and Management System for Stroke
CLAHRC	Collaboration for Leadership in Applied Health Research and Care
CQC	Care Quality Commission
CQUIN	Commissioning for Quality in Innovation
CRAG	Consumer Research Advisory Group
CRF	Case report forms
CSI	Carer Strain Index
DH	Department of Health
d.p	Decimal place
ED	Emergency Department
EPV	Events per variable
EQ5D	EuroQoL
ESD	Early Supported Discharge
FAI	Frenchay Activities Index
FCE	Finished consultant episode
GCP	Good Clinical Practice
GCS	Glasgow Coma Score
GHQ_12	General Health Questionnaire-12
GP	General Practitioner
HDU	High Dependency Unit
HES	Hospital Episode Statistics
HQIP	Healthcare Quality Improvement Partnership
HRG	Healthcare Resource Group
IC	Information Centre
ICD-10	International Classification of Diseases v 10
ICH	Intracerebral haemorrhage
ICU	Intensive Care Unit
IHD	Ischaemic heart disease
IPM	Integrated performance Measures
ISWP	Intercollegiate Stroke Working Party
IT	Information Technology
IV	Instrumental Variable
LACS	Lacunar stroke
LHS	London Handicap Score
LOS	Length of Stay
LSSS	London Stroke Satisfaction Questionnaire
LYBRA	Leeds, York, Bradford Research Alliance
MAU	Medical Admissions Unit
MCMC	Markov Chain MonteCarlo
MDT	Multidisciplinary Team

MI	Myocardial Infarction
mRS	Modified Rankin Score
NAO	National Audit Office
NCGS	National Clinical Guideline for Stroke
NEADL	Nottingham Extended Activities of Daily Living
NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NIHR	National Institute for Health Research
NIHSS	National Institute of Health Stroke Score
NINDS	National Institute of Neurological Disorders and Stroke
NNT	Number needed to treat
NQB	The National Quality Board
NSF	National Service Framework
NSS	The National Stroke Strategy
NSSA	National Sentinel Stroke Audit
OCSP	Oxford Community Stroke Project
OCSP-4	Classification of Interventions and Procedures version 4
OHS	Oxford Handicap Scale
OT	Occupational Therapy
PACS	Partial anterior circulation stroke
PAS	Patient Administration System
PbR	Payment by Results
PCT	Primary Care Trust
PEG	Percutaneous gastroenterostomy
POCS	Posterior circulation stroke
PROMs	Patient Reported Outcome Measures
PT	Physiotherapy
QOF	Quality Outcomes Framework
R&D	Research and Development
RCP	Royal College of Physicians
RCT	Randomised Controlled Trial
REC	Regional Ethics Committee
ROC	Receiver Operating Curve
rtPA	Recombinant tissue plasminogen activator (a thrombolytic agent)
SAH	Subarachnoid haemorrhage
SD	Standard deviation
SIGN	Scottish Intercollegiate Guidelines Network
SINAP	Stroke Improvement National Audit Project
SIP	The Stroke Improvement Programme
SIPSO	Subjective Index of Physical and Social Outcome
SLT	Speech and Language therapy
SMR	Standardised Mortality Rate
SSNAP	Sentinel Stroke National Audit Project
SSV	Six simple variables case-mix adjustment model
SU	Stroke Unit
SUT	Stroke Unit Trialists
SW	Social Worker
TACS	Total anterior circulation stroke
TIA	Transient Ischaemic Attack
TTO	Time trade off
UIC	Urinary Incontinence
VIF	Variance Inflation Factor
VS	Vital Sign
WHO	World Health Organisation
YSRN	Yorkshire Stroke Research Network

Appendix A Case-mix adjuster systematic review

A-1 MEDLINE Search Strategy

Devised by Deirdre Andre at the University of Leeds Healthcare Library

1. cerebrovascular disorders/
2. exp basal ganglia cerebrovascular disease/
3. exp brain ischemia/
4. exp carotid artery diseases/
5. stroke/
6. exp brain infarction/
7. exp cerebrovascular trauma/
8. hypoxia-ischemia, brain/
9. exp intracranial arterial diseases/
10. exp intracranial arteriovenous malformations/
11. exp "intracranial embolism and thrombosis"/
12. exp intracranial hemorrhages/
13. vasospasm, intracranial/
14. vertebral artery dissection/
15. aneurysm, ruptured/ and exp brain/
16. brain injuries/
17. brain injury, chronic/
18. exp carotid arteries/
19. endarterectomy, carotid/
20. *heart septal defects, atrial/ or foramen ovale, patent/
21. *atrial fibrillation/
22. (stroke or poststroke or post-stroke or cerebrovasc$ or brain vasc$ or cerebral vasc$ or cva$ or apoplex$ or isch?emi$ attack$ or tia$1 or neurologic$ deficit$ or SAH or AVM).tw.
23. ((brain$ or cerebr$ or cerebell$ or cortical or vertebrobasilar or hemispher$ or intracran$ or intracerebral or infratentorial or supratentorial or MCA or anterior circulation or posterior circulation or basal ganglia) adj5 (isch?emi$ or infarct$ or thrombo$ or emboli$ or occlus$ or hypox$ or vasospasm or obstruction or vasculopathy)).tw.
24. ((lacunar or cortical) adj5 infarct$).tw.
25. ((brain$ or cerebr$ or cerebell$ or intracerebral or intracran$ or parenchymal or intraventricular or infratentorial or supratentorial or basal gangli$ or subarachnoid or putaminal or putamen or posterior fossa) adj5 (haemorrhage$ or hemorrhage$ or haematoma$ or hematoma$ or bleed$)).tw.
26. ((brain or cerebral or intracranial or communicating or giant or basilar or vertebral artery or berry or saccular or ruptured) adj5 aneurysm$).tw.
27. (vertebral artery dissection or cerebral art$ disease$).tw.
28. ((brain or intracranial or basal ganglia or lenticulostriate) adj5 (vascular adj5 (disease$ or disorder or accident or injur$ or trauma$ or insult or event))).tw.
29. ((isch?emic or apoplectic) adj5 (event or events or insult or attack$)).tw.
30. ((cerebral vein or cerebral venous or sinus or sagittal) adj5 thrombo$).tw.
31. (CVDST or CVT).tw.
32. ((intracranial or cerebral art$ or basilar art$ or vertebral art$ or vertebrobasilar or vertebral basilar) adj5 (stenosis or isch?emia or insufficiency or arterioscler$ or atheroscler$ or occlus$)).tw.
33. ((venous or arteriovenous or brain vasc$) adj5 malformation$).tw.
34. ((brain or cerebral) adj5 (angioma$ or hemangioma$ or haemangioma$)).tw.
35. carotid$.tw.
36. (patent foramen ovale or PFO).tw.
37. ((atrial or atrium or auricular) adj fibrillation).tw.
38. asymptomatic cervical bruit.tw.
39. exp aphasia/ or anomia/ or hemiplegia/ or hemianopsia/ or exp paresis/ or deglutition disorders/ or dysarthria/ or pseudobulbar palsy/ or muscle spasticity/

40. (aphasi$ or apraxi$ or dysphasi$ or dysphagi$ or deglutition disorder$ or swallow$ disorder$ or dysarthri$ or hemipleg$ or hemipar$ or paresis or paretic or hemianop$ or hemineglect or spasticity or anomi$ or dysnomi$ or acquired brain injur$ or hemiball$).tw.
41. ((unilateral or visual or hemispatial or attentional or spatial) adj5 neglect).tw.
42. or/1-41
43. Risk Adjustment/
44. (case mix$ adj3 adjust$).tw.
45. exp "Severity of Illness Index"/
46. Diagnosis-Related Groups/
47. DRG$1.tw. or diagnosis related group*.mp. or diagnostic related group*.mp.
48. Prognosis/
49. exp "Outcome and Process Assessment (Health Care)"/
50. Comorbidity/
51. morbidity/
52. mortality/
53. survival rate/
54. or/43-53
55. exp models, statistical/
56. ROC Curve/
57. roc curve.tw.
58. exp Survival Analysis/
59. Data Interpretation, Statistical/
60. multivariate analysis/
61. or/55-60
62. 42 and 61 and 54

A-2 Data extraction tables for studies describing models included in the review

Belfast	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Fullerton (1988)	Albert's test score, leg function, conscious level, arm power, weighted mental score, non-specific ECG changes	Within 48 hours of stroke	Yes	Yes	4 level measure of dependency	206	Linear logistic regression analysis (canonical discriminant analysis)	35 predictor variables and >40 dummy variables EPV<10	No	No		No	No

Belfast	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Gladman (1992)	Unselected consecutive patients admitted with stroke over 3 years	'On admission'	Prospective cohort study		Death at 3 months	102	Sensitivity 94%, specificity 29% Likelihood ratio 1.3 (1.1-1.6)

Bristol	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Wade et al (1983)	Age, hemianopia, arm motor deficit, sitting balance, urinary incontinence	On admission to rehabilitation unit	Yes	48/162 with insufficient data =30% No significant difference between 48 patients with insufficient data and whole sample	BI at six months post stroke (measured on 0-100 scale)	162	Multiple linear regression	Yes	Not stated	Yes	Correct prediction of 6 month BI (within 5 points) in 55% of cases	Attendees at a non-residential rehabilitation facility – time from stroke to recruitment not uniform	Yes

Bristol	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Gladman et al (1992)	Unselected consecutive patients admitted with stroke over 3 years	7-10 days following acute admission to hospital	Prospective cohort study		Barthel Index at 3 months	102	Sensitivity 100%, specificity 0%

Edinburgh	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Prescott et al (1982)	Upper limb motor function, proprioception, postural stability	?	Retrospective use of RCT data (patients randomised to treatment on a stroke unit)	Yes, however 30/100 surviving patients untestable on at least one test at 4 weeks and ascribed worst outcome score	7 level dependency scale	155	Linear regression	No	No	No	75% correct prediction of independence at week 4	Minor and very severe strokes excluded	Yes

Edinburgh	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Gladman et al (1992)	Unselected consecutive patients admitted with stroke over 3 years	At four weeks from admission to acute hospital	Prospective cohort study	2/102	Death or prolonged hospital stay	102	Sensitivity 55%, specificity 65%

G score	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Gompertz et al (1994)	Limb paralysis, higher cerebral dysfunction+ hemiparesis+ hemianopia, drowsy, age, unconscious at onset, uncomplicated hemiparesis	Within 24 hours of stroke	Prospective cohort study	No 12% loss to follow-up, a further 5 had incomplete data. Characteristics of non-responders not examined	BI at six months	361 recruited (314 with complete data)	None-adaptation of Guys score to simplify weights	Yes			Prediction of BI<13 at 6 months Sens: 47% Spec:73% LR 1.74	No	Yes
	Validation studies: NONE												

Guys	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Allen (1984)	Limb paralysis, higher cerebral dysfunction+ hemiparesis+ hemianopia, drowsy, age, unconscious at onset, uncomplicated hemiparesis	Within 2 weeks	Yes	7% at 2 months, 14% at six months (excluded from analysis)	Four point scale of dependency (dichotomised) at two months and six months	148	Stepwise logistic regression	50 patients with poor outcome, 10 variables entered into model EPV<10	No	Stepwise variable selection	89% correct allocation	Patients over 76 excluded	Yes

Guys	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Gompertz (1994)	Consecutive patients admitted with stroke 1990-91 (UK)	Within 24 hours of stroke	Prospective cohort study	42 (12%)	BI (postal) at six months	361	Sensitivity 0.72, specificity 0.63 for prediction of poor outcome (Likelihood ratio 1.97)
Gladman (1992)	Unselected consecutive patients admitted with stroke over 3 years	'On admission'	Prospective cohort study		Death at 3 months	102	Sensitivity 58%, specificity 83% Likelihood ratio 3.3 (1.8-6.0)
Muir et al (1996)	All patients with Ischaemic and haemorrhagic stroke admitted to a single stroke unit. No restriction on age or stroke subtype	Within 72 hours of admission	Prospective data collection	Less than 10%	Alive at home versus in care or dead at 3 months	408	Prediction of poor outcome *when added to model with NIHSS (i.e. not an assessment of performance of Guys score in isolation)* Sens 70% Spec 89% Predictive accuracy 82%

Johnston	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Johnston et al (2000)	Age, NIHSS score, small vessel stroke, previous stroke, diabetes, pre-stroke disability, infarct volume	6 hours from stroke onset	Retrospective use of RCT data (RANTTAS) – intervention and control groups	No NIHSS 35/256 BI 27/256 GOS 27/256 (excluded from analysis)	Excellent or poor outcome based on dichotomised NIHSS score, BI and Glasgow Outcome Score (GOS) at 3 months	256	Logistic regression models. All seven variables used.	EPV<10 (NIHSS), EPV≥10 (BI and GOS) Six models specified	Yes	No	C statistics >0.8 for all models except prediction of devastating outcome with NIHSS (0.79)	Unclear	No

Johnston	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Johnston et al (2003)	Ischaemic stroke population eligible for thrombolytic therapy	Within 3 hours of symptom onset	Retrospective use of placebo arm of RCT (NINDS trial)		Excellent or very poor outcome as above	299 EPV >10 for all models	Five out of six models have excellent discrimination (c statistic >0.8) C statistic for prediction of devastating outcome with NIHSS 0.75 Calibration: Over optimistic predictions of excellent recovery with NIHSS for patients in middle band of stroke severity
Johnson et al (2004)	Ischaemic stroke population eligible for thrombolytic therapy	Within 3 hours of symptom onset	Retrospective use of intervention and control arm of RCT (NINDS trial)		Excellent or very poor outcome as above	615	Study used model to calculate differences in unadjusted (univariate) and adjusted (using pre-specified models) odds ratios for prediction of excellent or very poor outcome. Model performance not measured.

Lincoln	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Lincoln et al (1989)	Age, sex, marital status, side of stroke, weeks post stroke, tests of motor function, ADL, perception, language, memory, cognition, incontinence	No – one week post admission to rehab facility (1-13 weeks post stroke)	Yes	No 16/70 lost to follow up at 9 months	Rivermead gross function score, ADL status at discharge and nine months, discharge destination	70	Stepwise regression	No	No	Stepwise variable selection	81% of cases correctly classified	Post-acute patients admitted to rehabilitation unit	No

Lincoln	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Lincoln et al (1990)	Prospective observational cohort admitted to rehab stroke unit	One week post transfer to stroke unit	Data capture for purposes of validation study		Discharge destination	57 EPV<10	Sensitivity 98%, specificity 25%

mNIHSS	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Lyden et al (2001)	Items 1B, 1C, 2,3,5 a&b, 6 a&b, 8, 9, 11 from the NIHSS Conscious level, gaze, visual fields, upper and lower limb power, sensory function , language and neglect	Within 24 hours of stroke onset	Retrospective use of data from 2 placebo-controlled trials of rt-PA in acute ischaemic stroke (NINDS-rtPA (REF)).	Not specified	Good/poor outcome BI >95, mRS <1 and GOS=1 mNIHSS<1 at 90 days	291	Developed through factor analysis of the NIHSS_15 , redundant items dropped to produce the mNIHSS (Lyden et al 1999)				Performs identically to the NIHSS when substituted into a model to predict outcome after ICH	Patients eligible for thrombolysis mNIHSS performed by specialists certified in administration of NIHSS	

	Validation studies (psychometric testing)						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Meyer et al (2002)	Ischaemic and haemorrhagic stroke Inpatients and outpatients	No specification for time since event	Prospective cohort		BI, mRS	27 for validity assessments	Examines reliability / validity of mNIHSS rather than model performance. Good inter-rater reliability and concurrent validity. Valid predictor of NIHSS.

NIHSS + age	Development											
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Weimar (2004)	Age NIHSS	Within 6 hours of onset	Data extracted from prospective stroke database	Yes	BI≤95 (model 1) and mortality 100 days mortality (model 2)	1079	Backwards and forwards selection logistic regression analysis	Model 1 EPV>10 Model 2 EPV<10	Yes	Yes, and interaction terms	Prediction of BI≤95 Sensitivity 63% Specificity 83% Prediction of mortality Sens 59% Spec 92%		No

NIHSS+age	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Weimar (2004)	Same sample as used to validate Weimar models 1&2. 13 acute hospitals in Germany 2001-2002. Pre-stroke mRS>2.	Within 6 hours of onset	Prospective data collection	Centres with>10% loss to follow up not included. Patients with incomplete data excluded 275/1582 (17%) but did not differ significantly from those included	BI<95 and death at 120 days	1307	120 day BI<95 Sens 63% Spec 83% 120 day mortality Sens 58% Spec 92%
König et al (2008)	Combined data from 11 randomised stroke trials. Inclusion/exclusion criteria for individual trials not specified	Within 6 hours of onset	Retrospective analysis of VISTA data (Virtual International Stroke Trials Archive).	BI 795/5843=14% Mortality <10%	BI or mortality at 90 days	5843	BI <95 at 90 days c statistic = 0.808 90 day mortality c statistic = 0.706

NIHSS_8	Development											
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Tirschwell et al (2002)	NIHSS_15 items 1a, 2,3,4 6a&b 9, 10 conscious level, gaze visual fields, facial paresis and lower limb motor scores, language and dysarthria	Within 24 hours	Secondary use of data from placebo arm of three RCTs	Patients with complete data selected	Dichotomised 'Global outcome score' (good/poor) derived from NIHSS_15 (≤1), mRS (≤1 and BI ≥95)	223	Forward and backward stepwise logistic regression	No	No	Stepwise regression	C statistic for model to predict good outcome = 0.87	NIHSS>5 at onset	Yes

NIHSS_8	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Tirschwell et al (2002)	Acute ischaemic stroke	3-5 hours post onset	Treatment and control arms of RCT of rt-PA in acute ischaemic stroke	Only pts with complete data included	Dichotomised (good/poor) NIHSS_15≤1, mRS≤1, GOS=1, BI ≥95 at three months	531	C statistic for prediction of good outcome = 0.77

Orpington	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Kalra & Crome (1993)	Arm power, proprioception, balance, cognition	Within 72 hours of admission	Yes	Yes	BI at discharge or 16 weeks	96	Linear regression	Yes	No (BI treated as interval variable)	No	Strong correlation between Orpington score and discharge or 16 week BI (r2 = 0.89, p<0.001)	Stroke patients >75 years.	

Orpington	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Lai (1998)	Patients with severe strokes, coma, dependent or from nursing home prior to stroke excluded	Within 14 days of stroke	Prospective cohort study		BI SF-36 At 1,3 and 6 months post stroke	184	Linear regression modelling, BI treated as interval data. R2 = 0.62 to predict BI at one month, less than 0.5 at 3 and six months
Studenski (2001)	Patients with coma, hepatic, renal or heart failure excluded, patients admitted from nursing care or dependent prior to stroke excluded	Within 2 weeks of stroke	Retrospective use of data from a prospective cohort study	11%	Five markers of functional independence at 3 and 6 months: (in)dependence in personal care, independent in meal preparation, medication and community mobility	413	Area under ROC (equivalent to c statistic for dichotomous outcomes) greater than 0.8 for all outcomes at 3 months, and 0.74-0.80 at six months
Kalra et al (1994)	Patients over 75 admitted to hospital with acute stroke, excluding patients with pre-stroke dependency, cognitive impairment or those admitted from institutional care	At two weeks from stroke	Prospective cohort study		BI, discharge destination, level of dependence (3 level score) at discharge from hospital	217	OPS measured at two weeks to predict independent living at discharge Sens 96% Spec 36%

Six Simple Variables	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability**	**Feasible to collect at ward level?**
Counsell et al 2002	Age Living alone Independent pre stroke Normal GCS verbal score Able to lift both arms Able to walk	Assessments performed up to 30 days after stroke, proportion of assessments after 14 days small, median delay 4 days	Retrospective use of data collected prospectively (Oxford Community Stroke Study data)	No loss to follow up	Survival at 30 days, 6 month independent survival	530	Forward stepwise logistic regression (independent survival) and Cox proportional hazards (30 day survival)	18 variables entered. 30 day survival EPV=3.8 6 month independent survival: EPV = 15	Yes	Stepwise variable selection	C statistic 0.88 30 day survival 0.84 6 month independent survival	Developed on community stroke data, 45% of patients were not admitted to hospital.	Yes

SSV	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Counsell et al 2002	Two cohorts one community and one hospital inpatients	Within 30 days of onset	Retrospective use of prospective cohort study data		Survival at 30 days, 6 month independent survival	538 community 1330 Hospital based	30 day survival: Community cohort c statistic 0.88 Hospital cohort c statistic 0.86 6 month independent survival: Community cohort c statistic 0.84 Hospital cohort 0.84
FOOD trial (2003) Dennis (2006) Dennis (2003)	Patients hospitalised with acute stroke	Within 7 days	Prospective RCT trial data (FOOD trial)		6 month independent survival	2955 EPV >10	Independent survival c statistic 0.79 Calibration: tends to predict over optimistic outcomes in patients with milder strokes, pessimistic predictions for more severe strokes
Lewis et al (2008)	Patients with ischaemic stroke eligible for thrombolysis	Within 6 hours of acute stroke	Prospective RCT trial data (IST-3)		Independent survival 30 day survival	537 EPV>10	6 month independent survival: c statistic = 0.82 30 day survival, c statistic = 0.73 Calibration for 6 month independent survival was good 30 day survival, higher number of observed than predicted outcomes (i.e. Over pessimistic prediction)

	Population	Inception cohort	Data source	Loss to follow-up	Outcome assessed	Sample size	Model performance
SSV continued							
Reid et al (2007)	Acute and hyperacute ischaemic and haemorrhagic stroke	At first assessment, 273/538 (51%) within 6h	Prospective cohort study (Stroke Outcomes Study)		6 month mRS≤2	538 EPV>10	mRS≤2 at 6 months c statistic 0.79 Good calibration
Weir et al (2001)	Five Scottish hospitals 1995-97. Two teaching hospitals, 3 district hospitals	Within 30 days of admission	Retrospective data extraction from case-notes		6 month mortality	2724	C statistic 0.84 Hosmer-Lemeshow goodness of fit χ^2 14.2, df 10, p=0.164 (good calibration)
Weir et al (2003)	Acute stroke	On admission	Prospective cohort			92	Aimed to establish inter-rater reliability of variable measurement for the SSV model. Kappa statistics for prospective and retrospective study were > 0.6 for all or all variables except ability to walk obtained from retrospective case-note review (κ 0.55)
	Five Scottish hospitals 1995-97. Two teaching hospitals, 3 district hospitals	Records from the day of admission	Retrospective case-note data as part of an observational study (Weir et al 2001)			200	

Tilling	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Tilling (2001)	Sex, age, ethnicity, prestroke handicap, limb weakness, dysphasia, dysarthria, incontinence, conscious, swallowing deficit, stroke subtype	Within 2 weeks of stroke	Retrospective use of randomised controlled trial data	Includes patients assessed at least once	Barthel Index at 2, 4, 6 and 12 months post randomisation	299 patient	Multilevel modelling	Yes	Yes (BI treated as continuous variable)	No	Not specified for development study	Hospital based cohort able to transfer independently	Yes

Tilling	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Tilling et al (2001)	Unselected observational cohort or first strokes		South London Stroke Register 1995-1998		Barthel Index	710	Average difference between predicted and observed BI -0.4 (limits of agreement -7 to +6)

Uppsala	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Frithz et al (1976)	Adaptation of Mathew's score (0-100) Conscious level, orientation, dysphasia, conjugate gaze palsy, facial weakness, arm power, Performance Disability scale, reflexes, sensation	On admission to hospital	Data extracted from case-notes.		Mortality at one month	344	Logistic regression	Yes	Unclear	Yes	Not reported	Patients over 70 excluded	

Uppsala	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Gladman et al (1992)	Unselected consecutive patients admitted with stroke over 3 years	'On admission' to acute hospital	Prospective cohort study		Death at 3 months	102	Sensitivity 30%, specificity 96% Likelihood ratio 7 (2.1-24)

Weimar models	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Weimar et al (2002)	Model 1 Neurological complications, fever, lacunar infarct, diabetes, previous stroke, sex, age, mRS, NIHSS score on admission	Within 72 hours of admission	Data extracted from prospective stroke database (the German Stroke Database)	No, 53/260 (20.4%) lost to follow up. Their characteristics were specifically examined and did not differ significantly from sample	BI ≤95 at 100 days	1754	Backwards stepwise logistic regression modelling	41 variables EPV>10	Yes	Yes	Sensitivity 77% Specificity 84% R^2 0.55		No
	Model 2 Fever, age, NIHSS score on admission				Death at 100 days			EPV<10	Yes	Yes	Sensitivity 49% Specificity 95% R^2 0.41		No

Weimar models	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
German Stroke Study Collaboration (2004)	13 acute hospitals in Germany 2001-2002. Pre-stroke mRS>2.	Within 24 hours of admission	Prospective data collection	Centres with>10% loss to follow up not included	BI<95 and death at 120 days	1470	120 day <BI95 Sens 68% Spec 86% 120 day mortality Sens 47% Spec 96%

Weimar_ICH	Development												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Weimar (2006)	NIHSS Age	Within 6 h of intracerebral haemorrhage	Yes from hospitals with an acute stroke unit	No 53/260=20% (did not differ significantly from those with complete data)	BI at 100 days	260	Forwards and backwards logistic regression analysis	>10		Stepwise selection	C statistic BI>95 = 0.861	Pre-stroke mRS of ≥3 Only includes ICH and excludes comatose patients	No

	Validation studies						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Weimar (2006)	Consecutive admissions with ICH 1998-1999 in 30 hospitals, with prestroke mRS≥3	Within 6 hours	Retrospective cohort		BI at 100 days	173	C statistic 0.876

Young	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Young et al (2001)	Gait speed, prestroke FAI score, AMT, AMT missing, sensory neglect, COPD, side of hemiplegia	Recruitment on discharge from hospital or within 6 weeks of stroke if not admitted	Yes	Yes. Only complete data used	FAI at 12 months	207	Forwards and backwards stepwise logistic regression	No 17 predictors, 100 patients with poor outcome	Yes	Yes	Poor FAI at 1 year Sens:75% Spec 80%	Post-acute patients admitted to rehab unit, some not patients not admitted to hospital	Yes

Young	**Validation studies**						
	Population	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Young	Community post-acute cohort		Community based stroke trial		FAI at six months	108	Correct assignment to good/poor outcome in 76% of cases

A-2.1 Studies using existing impairment or severity scales to predict outcome

Canadian Neurological Score (CNS)								
	Validation studies							
Citation	**Variables included in model**	**Population**	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Muir et al (1996)		All patients with Ischaemic and haemorrhagic stroke admitted to a single stroke unit. No restriction on age or stroke subtype	Within 72 hours of admission	Prospective data collection	Less than 10%	Alive at home versus in care or dead at 3 months	408	Prediction of poor outcome *when added to model with NIHSS (i.e. not an assessment of performance of CNS in isolation)* Sens 71% Spec 89% Predictive accuracy 82%
Middle Cerebral Artery Neurological Score (MCANS) or Orgogozo score								
	Validation studies							
Citation	**Variables included in model**	**Population**	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Muir et al (1996)	Conscious level, communications, gaze, facial movement, arm raise, hand movement, upper and lower limb tone, leg raise, foot dorsiflexion	All patients with Ischaemic and haemorrhagic stroke admitted to a single stroke unit. No restriction on age or stroke subtype	Within 72 hours of admission	Prospective data collection	Less than 10%	Alive at home versus in care or dead at 3 months	408	Prediction of poor outcome Sens 71% Spec 89% Predictive accuracy 82%
NIHSS								
	Validation studies							
Citation	**Variables included in model**	**Population**	**Inception cohort**	**Data source**	**Loss to follow-up**	**Outcome assessed**	**Sample size**	**Model performance**
Muir et al (1996)	NIHSS_15	All patients with Ischaemic and haemorrhagic stroke admitted to a single stroke unit. No restriction on age or stroke subtype	Within 72 hours of admission	Prospective data collection	Less than 10%	Alive at home versus in care or dead at 3 months	408	Prediction of poor outcome Sens 71% Spec 90% Predictive accuracy 83%
Lai (1998)	NIHSS_15	Patients with severe strokes, coma, dependent or from nursing home prior to stroke excluded	Within 14 days of stroke	Prospective cohort study		BI SF-36 At 1,3 and 6 months post stroke	184	Linear regression modelling, BI treated as interval data. R^2 =0.56 for at 1 month. R^2 below 0.5 at 3 and six months

A-2.2 Studies using split sample (internal) validation

Anderson (Development)													
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Anderson et al (1994)	Coma, cardiac failure, urinary incontinence, severe paresis, atrial fibrillation	Within weeks of stroke onset (median 5 days)	Retrospective use data from patients registered in a population based study of acute stroke	Yes	Mortality at 1 year	492	Stepwise Cox proportional hazards	14 variables		Stepwise variable selection	Sens 90% Spec 83%	19% of patients not admitted to hospital,	Yes
	No external validation studies (authors used split-sample internal validation)												

Ischaemic Stroke Survival Score (ISSS) (Development)													
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Williams (2000)	Age Scandinavian Stroke Score, Rapid Disability Rating Scale score, previous stroke	Within 3 hours of stroke onset	Retrospective use both arms of placebo-controlled trial data (Stroke Treatment with Ancrod Trial STAT)	Yes	1 year survival	453	Logistic regression modelling	EPV>10		Yes	ISSS model in training data set R^2 =0.3 C statistic in validation set (split sample) 0.86	Ischaemic strokes, exclusion of minor or very severe strokes	No
	No external validation studies (authors used split-sample internal validation)												

Masiero	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Masiero et al (2007)	Age TCT FIM	Within 24 hours of admission to rehab unit, and within 8 weeks of stroke	Yes		No Dichotomised Functional Ambulation Classification at discharge from rehabilitation facility	100	Logistic regression modelling	No, 12 variables, 48 with poor outcome	Yes	No	To predict good/poor outcome C statistic = 0.94 Sens 87% Spec 96%	Up to 8 weeks post first stroke, patients with hemiplegia, admitted for inpatient rehab, no additional preclusion to gait or exercise training, significant dysphasia or cognitive impairment	No
	No external validation studies (authors used split-sample internal validation)												
Wang	**Development**												
	Internal validity					Statistical validity						Feasibility	
Citation	**Variables included in model**	**Adequate inception cohort**	**Prospective data collection**	**Less than 10% loss to follow-up**	**Assessment of a reliable outcome and at a fixed time point**	**Sample size**	**Modelling method and method of variable selection**	**An EPV (events per variable) of 10 or more**	**Linearity assumptions tested and met?**	**Collinearity addressed**	**Model performance**	**Inclusion or exclusion criteria that may limit generalisability?**	**Feasible to collect at ward level?**
Wang et al (2003)	Conscious level, dysphagia, UIC, both sides affected, hyperthermia, IHD, peripheral vascular disease, diabetes		Retrospective cohort study (data extraction from case-notes) for ICD stroke coded admissions 1995-1997	Yes	1 year mortality	Split sample 223 training set 217 validation set	Cox-proportional hazards. Variables selected through uni-variate analysis	EPV<10 (48 deaths in training set)			Validation set to predict 1 year mortality Sens 56% Spec 91% PPV 60%		Yes
	No external validation studies (authors used split-sample internal validation)												

A-3 Standardised and studentised residuals

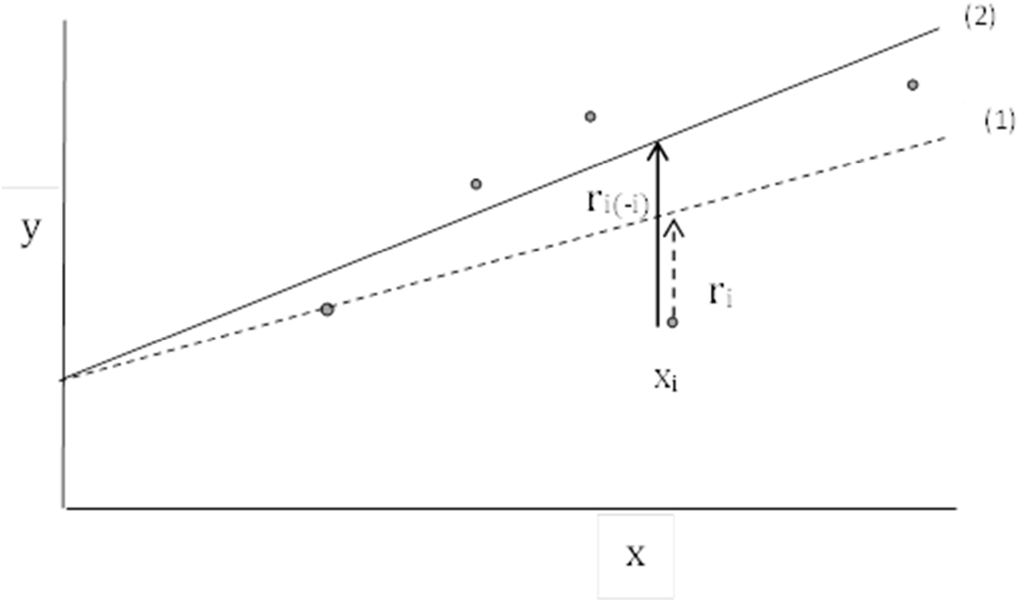
Examination of 'raw' residuals retains the units for the Y variable and these must therefore be interpreted in this context (Waterman R 1999). More useful for identification of outliers are residuals that have been standardised according to their standard deviation from the expected sample mean.

However, as the calculation of the standard deviation for an individual point and its estimate are not independent, particularly influential points will alter the regression line thus affecting the size of the residual (Fox J, 1997 p 272; Waterman R 1999).

This problem may be overcome through calculation of studentised residuals, where an individual point xi is omitted from the estimation of the standard deviation, such that the standardisation becomes independent of the observed value of xi (Fox J, 1997 p 272; Waterman R 1999).

For example, if point xi in Figure 94 is exerting undue influence (leverage) on a regression line (1), calculating the standardised residual from estimates including xi will falsely lower the magnitude of |ri| to give ri'. However, if xi is excluded from the calculation of model estimates, the studentised residual is created (ri(-i)) based on the unbiased regression line (2) (Fox J, 1997 p 272; Waterman R, 1999).

Figure 94 Demonstration of studentised residuals



Appendix B Study variables

Demographics
Age at stroke onset
Type of residence (pre-stroke)
Has main carer
Lived alone pre-stroke
Ethnicity
Gender
Process indicators
Admitted directly to stroke unit
Proportion of stay spent on stroke unit
Planned follow up by ESD
Post-hospital spell in NHS facility (e.g. intermediate care)
Discharge to the same address
Imaging within 24 hours
Thrombolysis given (date and time)
Swallowing screen within 24 hours of admission to hospital
Commenced antiplatelet within 48 hours of stroke
Physiotherapy assessment within 72 hours of admission to hospital)
Weighed at least once during admission
Evidence of mood assessment before discharge
Evidence of MDT rehabilitation goal setting
Occupational therapy assessment within 4 working days of admission
Visual field testing (RCP)
Sensory assessment (RCP)
Formal swallowing assessment within 72 hours of admission (RCP)
Formal communication assessment within 7 days (RCP)
Evidence within MDT notes of SW assessment within 7 days of referral (RCP)
Evidence of cognitive status assessment (RCP)
Malnutrition screening (RCP)
Continence promotion plan (RCP))
Receipt of fluids within 24 hours of stroke (RCP)
Receipt of nutrition within 72 hours of admission (RCP)
Case-mix data
Classification of stroke
Radiological classification of stroke
Pathological classification of stroke (OCSP)
Side of weakness
Six-simple variable case-mix adjustment variables
(Age)
Lived alone prior to stroke
Independent in ADL prior to stroke
Able to lift both arms above head (MRC power score>=3)
Able to walk unaided
Normal verbal GCS score
Univariate predictors
Drowsy since onset of stroke
Speech or language problems
Confusion at presentation
New urinary incontinence or newly catheterised since stroke onset
Previous disabling stroke

B-1 Instructions to delegates at the group decision making workshop to refine study outcome instruments

Part 1

We would like you (on the post it notes provided and working individually) to generate a list of the properties of a stroke outcomes instrument which you consider to be important. Write one idea on each piece of paper. Be inclusive and generic at this stage (e.g. does it measure relevant constructs, depth of questions, breadth of questions, length etc.)

[Similar constructs are grouped together on a flip-chart and numbered]

On an index card, please choose the five ideas that you feel are most important, and write the numbers, vertically down the side of the card

Please then rank the ideas from 1 (most important) to 5 (least important)

Part 2 – Paired weighting

Please consider each scale in relation to the criteria we have collectively identified to be most important in the instruments.

For each pair of scales please circle the one which you feel fulfils these criteria the best.

At the end of each row, please add up the number of times you have circled each instrument. This gives your ranking as to which you feel is the most useful instrument according to the criteria we have established.

				Total	
NEADL FAI	NEADL SIPSO	NEADL LHS	NEADL EQ5D	NEADL	=
	FAI SIPSO	FAI LHS	FAI EQ5D	FAI	=
		SIPSO LHS	SIPSO EQ5D	SIPSO	=
			LHS EQ5D	LHS	=
				EQ5D	=

B-2 The Oxford Handicap Scale and modified Rankin Scale

Table 63 Oxford Handicap Scale (postal version) from Dennis et al 2006

Grade	Description
0	I have no symptoms at all
1	I have a few symptoms but these do not interfere with m everyday life
2	I have symptoms which have caused some changes in my life but I am still able to look after myself
3	I have symptoms which have significantly changed my life and I need some help in looking after myself
4	I have quite severe symptoms which mean I need to have help from other people but I am not so bad as to need attention day and night
5	I have major symptoms which severely handicap me and I need constant attention day and night

Table 64 Modified Rankin Scale from van Swieten et al 1988

Grade	Description
0	No symptoms at all
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 assistance
4	Moderately severe disability: unable to walk without assistance, and unable to attend to own bodily needs without assiatance
5	Severe disability: bedridden, incontinent, and requiring constant nursing care and attention

B-3 Subjective Index of Physical and Social Outcome (from Trigg et al 2000)

Please answer all questions

Physical Subscore

1. Since your stroke, how much difficulty do you have dressing yourself fully?
(Circle One Number)

No difficulty at all	4
Slight difficulty	3
Some difficulty	2
A lot of difficulty	1
I cannot dress myself fully	0

2. Since your stroke, how much difficulty do you have moving around all areas of the home?
(Circle One Number)

No difficulty at all	4
Slight difficulty	3
Some difficulty	2
A lot of difficulty	1
I cannot move around all areas of the home	0

3. Since your stroke, how satisfied are you with your overall ability to perform daily activities in and around the home?
(Circle One Number)

Completely satisfied	4
Mostly satisfied	3
Fairly satisfied	2
Not very satisfied	1
Completely dissatisfied	0

4. Since your stroke, how much difficulty do you have shopping for and carrying a few items (1 bag of shopping or less) when at the shops?
(Circle One Number)

No difficulty at all	4
Slight difficulty	3
Some difficulty	2
A lot of difficulty	1
I cannot shop for and carry a few items	0

5. Since your stroke, how independent are you in your ability to move around your local neighbourhood?
(Circle One Number)

I am completely independent	4
I prefer to have someone else with me	3
I need occasional assistance from someone	2
I need assistance much of the time	1
I am completely dependent on others	0

Social subscore

6. Since your stroke, how often do you feel bored with your free time at home?
(Circle One Number)

I am never bored with my free time	4
A little of my free time	3
Some of my free time	2
Most of my free time	1
All of my free time	0

7. Since your stroke, how would you describe the amount of communication between you and your friends/associates?
(Circle One Number)

A great deal	4
Quite a lot	3
Some	2
A little bit	1
None	0

.

8. Since your stroke, how satisfied are you with the level of interests and activities you share with your friends/associates?
(Circle One Number)

Completely satisfied	4
Mostly satisfied	3
Fairly satisfied	2
Not very satisfied	1
Completely dissatisfied	0

9. Since your stroke, how often do you visit friends/others?
(Circle One Number)

Most days	4
At least once a week	3
At least once a fortnight	2
Once a month or less	1
Never	0

10. Since your stroke, how do you feel about your appearance when out in public?
(Circle One Number)

Perfectly happy	4
Slightly self-conscious	3
Fairly self-conscious	2
Very self-conscious	1
I try to avoid going out in public	0

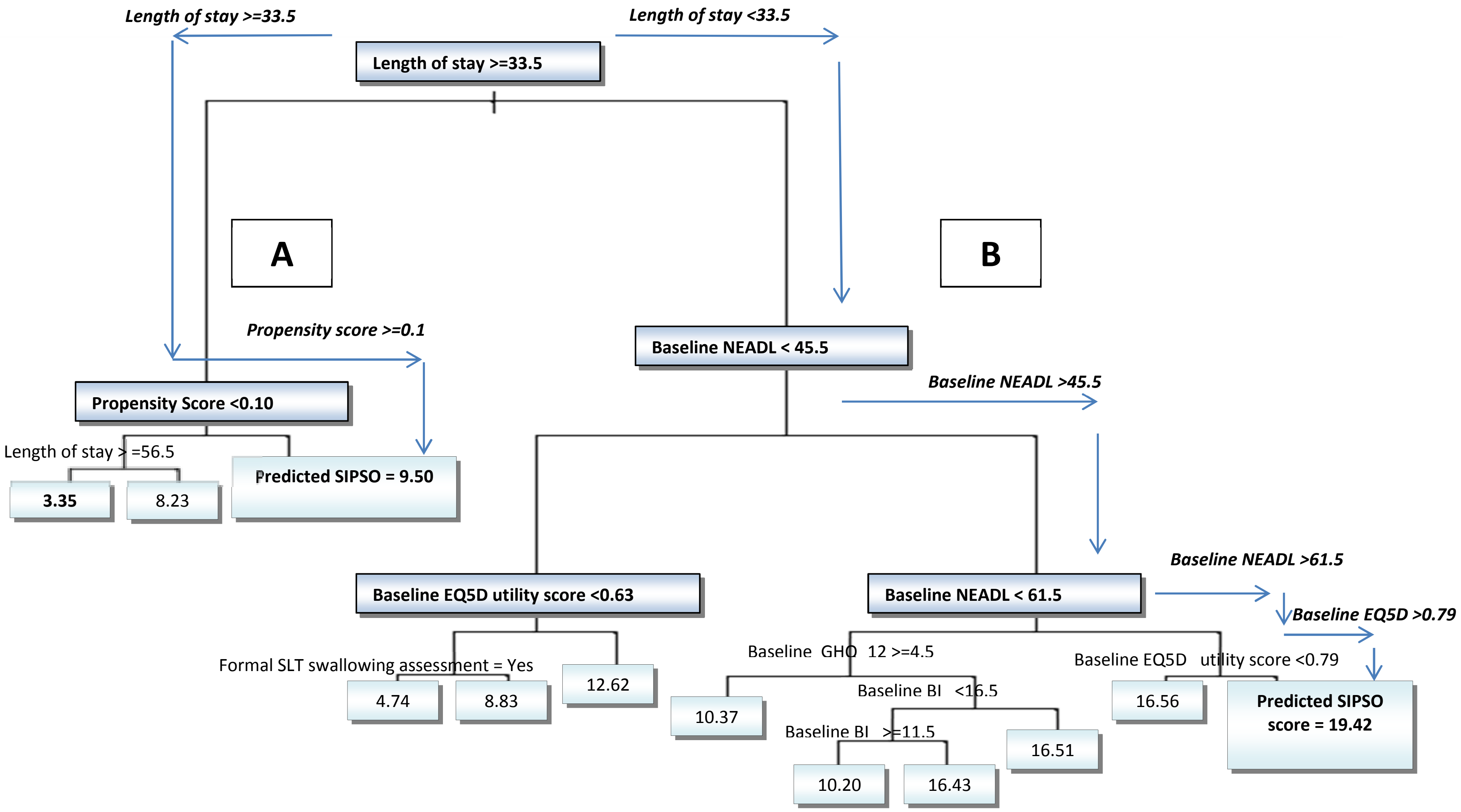
B-4 Statistical plan

- Data cleaning and missing data pattern analysis (baseline data)
- Outliers and tests of normality of continuous variables
- Examination of return rates and missing data analysis for outcomes questionnaire packs
- Descriptive statistics including:
- Floor and ceiling effects of baseline and six month patient completed questionnaires
- Examination of representativeness of study sample
    - Exploration of process-outcome linkages in the study population
    - Univariate (unadjusted) analyses
- Construction of decision trees to predict CIMSS study outcomes to identify important predictors
- Identification and testing of potential interaction terms
- Stratification of the sample using the SSV model (e.g. using propensity score as a continuous variable in models, or stratification according to matched propensity score)
- Construction of regression models to predict study outcomes using important clinical variables and predictors identified in decision trees
- Performance of the SSV case-mix adjuster in terms of:
    - Model discrimination (measured with c statistics)
    - Calibration of the SSV in the CIMSS study population (calibration plots)
- Exploration of potential univariate predictors of outcome that could be used in addition to, or instead of the SSV case-mix adjuster
- Replication of models in MLWin software and with Markov Chain MonteCarlo (MCMC) iterations to explore stability and convergence of the beta coefficients
- Identification of key process and case-mix variables that are important in determining outcome to be included in core dataset for further testing

Appendix C Regression and Classification Trees

Figure 95 shows an example of a regression tree to predict the (continuous) outcome of the physical subscore of the SIPSO and includes all the predictors, case-mix variable and baseline assessments used in the CIMSS study population. At the top of the tree, the condition length of stay >=33.5 is stipulated. For the purposes of interpretation of the tree, this has been interpreted as <=33 or >=34 (as length of stay has been recorded in whole days). If the length of stay was longer than 33 days, the left hand branch is followed; otherwise the right hand branch is selected. The length of the 'legs' for each predictor denotes its relative importance. Thus it can be seen from the example that, in this regression tree, the length of stay is the main determinant of physical SIPSO subscore. Other predictors and their relative importance are presented until, at the bottom of each terminal branch, a value for the predicted physical SIPSO score is given if all the preceding conditions are met. Thus, using this tree in this dataset, a patient with a length of stay of greater than 34 days and a probability of poor outcome as predicted with the SSV case-mix adjuster (propensity score) of greater than 0.1 has a predicted SIPSO physical score of 9.5 (path A highlighted on Figure 95), whilst a patient with a length of 33 days or fewer, a baseline NEADL of greater than 62 and a baseline EuroQoL utility score of greater than 0.79 has a predicted physical SIPSO subscore of 19.42 (path B on Figure 95).

Figure 95 Interpretation of regression trees using example of prediction of physical SIPSO subscore



Appendix D Descriptive statistics

D-1 Equivalence of proportions for sex between screened and recruited populations

The working for statistical tests performed during data cleaning and descriptive statistics are shown here

D-1.1 Equivalence of proportions for sex between screened and recruited populations

Screened 337
Recruited 312

	Mean	Standard error	95% confidence interval	
Screened	0.56	0.27	0.51	0.61
Recruited	0.51	0.28	0.45	0.56
Difference	0.054	0.040		
Probability difference ≠0: 0.16				

D-1.2 Significant difference in age by gender (recruited patients)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test age by gender

```
   Gender  |     obs   rank sum    expected
-------------+---------------------------------
     Male    |    154    20100.5      24101
   Female   |    158    28727.5      24727
-------------+---------------------------------
  Combined |    312      48828       48828


Ho: age at stroke(males) = age at stroke(females)
         z =  -5.024
  Prob > |z| =   <0.001
```

D-1.3 Age by recruitment to study

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

```
patient_recruited |     obs    rank sum    expected
-------------+---------------------------------
not recruited      |    343     126016      112504
  recruited          |    312      88824     102336
-------------+---------------------------------
  Combined         |    655     214840      214840

Ho: age(patient not recruited) = age(patient recruited)
         z =   5.589
  Prob > |z| =   <0.001
```

D-2 Baseline stroke severity data

D-2.1 Significant difference in Baseline Barthel Index between patients recruited and not recruited into study

```
patient recruited |     obs    rank sum    expected
-------------+---------------------------------
      false           |    319      79578     97454.5
       true           |    291     106777     88900.5
-------------+---------------------------------
   combined         |    610     186355      186355


Ho: total BI(not recruited) = total BI(recruited )
             z =  -8.303
    Prob > |z| =  <0.001
```

D-2.2 Kruskal-Wallis equivalence of medians test Barthel Index by site in patients not recruited into study

```
Kruskal-Wallis equality-of-populations rank test

  +---------------------------+
  |     site              | Obs      | Rank Sum          |
  |----------+-----+----------     |
  | Bradford          |  94      | 14721.00          |
  |    Leeds            |  56      |  8103.50          |
  |     York             | 169     | 28215.50          |
  +---------------------------+

chi-squared =     2.628 with 2 d.f.
probability =      0.2687
```

D-2.3 Two way Mann-Whitney U tests to identify significant differences in baseline Barthel Index between sites for patients recruited into the study

```
Two-sample Wilcoxon rank-sum (Mann-Whitney) tests

    site        |    obs   rank sum   expected
-------------+---------------------------------
  Bradford      |     63     5746.5     5638.5
    Leeds       |    115    10184.5    10292.5
-------------+---------------------------------
  combined      |    178      15931      15931

Ho: total BI(Bradford) = total BI(Leeds)
         z =   0.330
  Prob > |z| =   0.7417

    site        |    obs   rank sum   expected
-------------+---------------------------------
  Bradford      |     63     4808.5     5575.5
    York        |    113    10767.5    10000.5
-------------+---------------------------------
  combined      |    176      15576      15576

Ho: total BI(Bradford) = total BI(York)
         z =  -2.392
  Prob > |z| =   0.0168

    site        |    obs   rank sum   expected
-------------+---------------------------------
    Leeds       |    115      11576    13167.5
    York        |    113      14530    12938.5
-------------+---------------------------------
  combined      |    228      26106      26106

Ho: total BI (Leeds) = total BI (York)
         z =  -3.215
  Prob > |z| =   0.0013 (reject Ho)
```

D-2.4 Difference in median baseline Barthel Index between categories of response

D-2.4.1 Kruskal Wallis test (BI by response category)

```
 +------------------------------+
 |   response     | Obs   | Rank Sum      |
 |-------------+-----+---------- |
 | no response  |  67   | 10855.50      |
 |   response   | 187   | 32814.00      |
 |     dead     |  44   |  3296.50      |
 |   withdrew   |  13   |  1550.00      |
 +------------------------------+

chi-squared with ties =    47.670 with 3 d.f.
probability =    <0.001
```

D-2.4.2 Pairwise comparisons of BI between levels of response (Mann-Whitney U tests)

```
  Response     |     obs    rank sum    expected
-------------+---------------------------------
 no response   |      67     7975.5      8542.5
  response     |     187    24409.5     23842.5
-------------+---------------------------------
  combined     |     254     32385       32385
Ho: Baseline BI (non-responders) = Baseline BI (responders)
        z =  -1.111    Prob > |z| =   0.2666
```

```
  response     |     obs    rank sum    expected
-------------+---------------------------------
 no response   |      67     4601.5      3752
     dead      |      44     1614.5      2464
-------------+---------------------------------
  combined     |     111      6216       6216
Ho: Baseline Barthel (non response) = Baseline BI (dead)
        z =   5.155    Prob > |z| =   <0.001
```

```
  response     |     obs    rank sum    expected
-------------+---------------------------------
 no response   |      67     2834.5      2713.5
  withdrew     |      13      405.5       526.5
-------------+---------------------------------
  combined     |      80      3240       3240
Ho: Baseline BI (no response) = Baseline BI (withdrew)
        z =   1.591    Prob > |z| =   0.1117
```

```
  response     |     obs    rank sum    expected
-------------+---------------------------------
  response     |     187    24291.5      21692
     dead      |      44     2504.5      5104
-------------+---------------------------------
  combined     |     231     26796       26796
Ho: Baseline BI (responders)= Baseline BI (dead)
        z =   6.563    Prob > |z| =  <0.001 (reject Ho)
```

```
  response     |     obs    rank sum    expected
-------------+---------------------------------
  response     |     187     19269     18793.5
  withdrew     |      13      831      1306.5
-------------+---------------------------------
  combined     |     200     20100      20100
Ho: Baseline BI (response) = Baseline BI (withdrew)
        z =   2.377    Prob > |z| =   0.0174
```

```
  response     |     obs    rank sum    expected
-------------+---------------------------------
     dead      |      44     1157.5      1276
  withdrew     |      13      495.5       377
-------------+---------------------------------
  combined     |      57      1653       1653
Ho: Baseline BI (dead) = Baseline BI (withdrew)
        z =  -2.293    Prob > |z| =   0.0219
```

Appendix E Model construction and checks of assumptions

E-1 Association between specific impairments and assessments

E-1.1 Chi squared tests of association between specific impairments and corresponding assessments

```
                |   SLT communication assessment
                |     No      Yes    No but              |    Total
-----------+--------------------------------+----------
  No dysphasia  |          4        9         104        |     117
                |          3.42     7.69      88.89      |   100.00
-----------+--------------------------------+----------
    Dysphasia   |          53       82        59         |     194
                |          27.32    42.27     30.41      |   100.00
-----------+--------------------------------+----------
    Total       |          57       91        163        |     311
                |          18.33    29.26     52.41      |   100.00

     Pearson chi2(2) = 100.1835   Pr =< 0.001
```

```
New urinary |    Urinary continence care plan
Incontinence                |        No        Yes        No but   |          Total
-----------+--------------------------------+----------
  No new incontinence    |          19       20         197        |          236
                         |          8.05     8.47       83.47      |          100.00
-----------+--------------------------------+----------
   New incontinence      |          14       46         9          |          69
                         |          20.29    66.67      13.04      |          100.00
-----------+--------------------------------+----------
  Total                  |          33       66         206        |          305
                         |          10.82    21.64      67.54      |          100.00

     Pearson chi2(2) = 130.1537   Pr = <0.001
```

E-1.2 Likelihood ratio test for transformed length of stay in the prediction of physical subscore of the SIPSO

Model 10   Likelihood ratio test to determine if linearity is improved through categorising the log transformed length of stay variable

Regression model of SIPSO physical subscore on categorised length of stay (log transformed) cut into 5 equally sized groups (Model A)

```
---------------------------------------------------------------------------
SIPSO physical
subscore                   Coef.      Std. Err . t        P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
_LOS_cut_1     |           0.22       1.27      0.17      0.87   -  2.29      2.73
_LOS_cut_2     |           1.27       1.41      0.90      0.37      4.05      1.50
_LOS_cut_3     |           3.63       1.38      2.62      0.01      6.36      0.90
_LOS_cut_4     |           4.92       1.51      3.26      0.001     7.90      1.94
_LOS_cut_5     |           10.17      1.42      7.17      <0.001    12.96     7.37
     _cons     |           15.76      1.03      15.27     <0.001    13.72     17.79
```

Linear regression model of SIPSO physical subscore on log transformed length of stay (continuous variable) (Model B)

```
-----------------------------------------------------------------------------

SIPSO Physical
Subscore        |       Coef    Std. Err  t       P>|t|    [95% Conf. Interval]
-------------+---------------------------------------------------------------
log_(LOS+1)     |       3.14    0.34      9.12    <0.001   3.82     2.46
     _cons      |       20.39   0.91      22.37   <0.001   18.59    22.19
Estimates store B
```

Likelihood ratio test that model A is significantly different from model B

```
Likelihood-ratio test               LR chi2(4)  =    4.76
(Assumption: B nested in A)         Prob > chi2 =   0.3124
```

Therefore cannot reject the assumption that model A deviates from the linear model (model B is nested in model A)

E-1.3 Variance inflation factors (VIF) for examination of potential collinearity (Model 2)

```
Variable                            VIF      1/VIF

Baseline EQ5D                       1.85     0.541835
log_(LOS+1)                         1.66     0.601100
Baseline NEADL                      1.21     0.826535
Age at stroke                       1.09     0.917088
Discharged to same address          1.07     0.938279

Mean VIF                            1.38
```

E-1.4 Sample size required to detect significant difference in SIPSO social subscore dependent on receipt of SLT communication assessment

Calculation of sample size of two equal sized groups is given by

$$(7)\quad N = \frac{2}{\left(\frac{E}{s}\right)^2} * P$$

where N=total sample size (equal groups), E=expected difference in outcome score between groups, s = standard deviation and P = 7.9, a constant based on the α-significance level (set here at 0.05) and the power (set here at 80%) (Whitley E et al 2002). The standard deviation of a sample can be calculated from the standard error of the mean for two groups given by the formula:

$$(8)\quad se = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$=> s = se\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where n = number of patients in each of “no” (n_1) and “yes” (n_2) groups for receipt of SLT communication assessment in the sample used to calculate the standard error.

Using Model 6 (p 177), se = 1.311, $n_1 = 29$, $n_2 = 108$, such that s = 4.65, E = -2

Therefore, N =[$2/(-2/4.65)^2$] * 7.9 = 85 for each group such that the total sample size =170.

Using formulae from (Whitley E et al 2002), the adjustment for calculation of the sum total of two unequal groups is given by:

$$(9)\quad N' = \frac{N(1+k)^2}{4k} = \frac{85 * [1+(\frac{108}{29})]^2}{4(\frac{108}{29})} = \mathbf{255}$$

Where k = n_2/n_1

And each individual group is given by

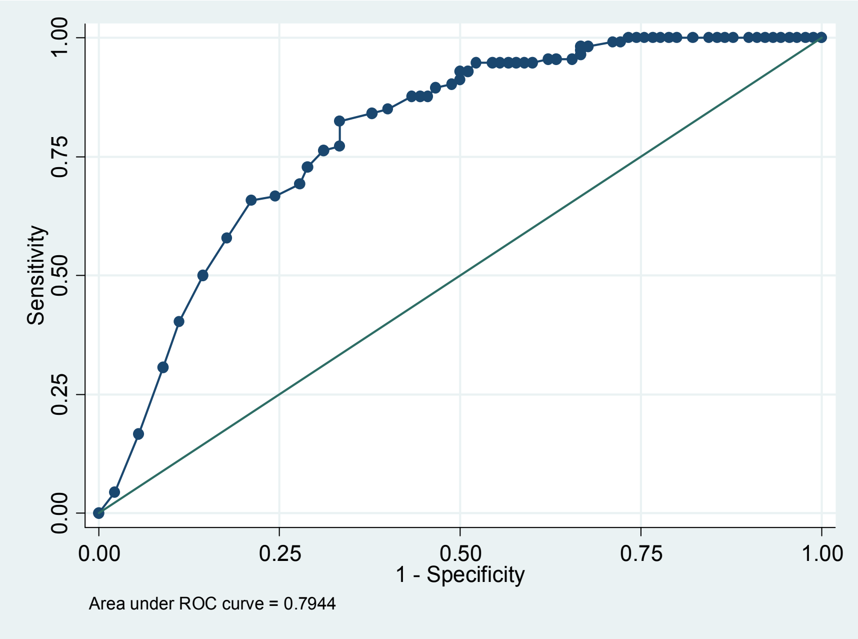
$$(10)\quad \frac{N\prime}{(1+k)} \text{ and } \frac{kN\prime}{(1+k)} \text{ such that } \mathrm{n}_1 = \mathbf{54}, \mathrm{n}_2 = \mathbf{201}$$

Accounting for the 55% of patients in whom SLT assessments are not indicated, 255/0.45≈567 patients with complete data would be required to detect the difference with power of 80%.

E-2 Length of stay as a univariate case-mix adjuster

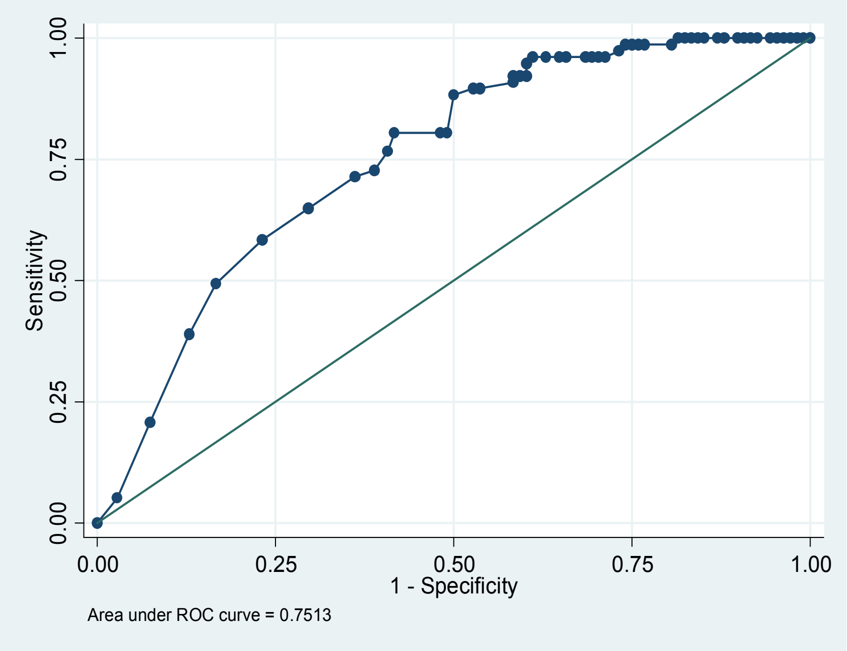
E-2.1 ROC curves to calculate c statistic for length of stay to predict dichotomised study outcomes

E-2.1.1 Receiver Operating Curve (ROC) and c-statistic for length of stay to predict dichotomised OHS,



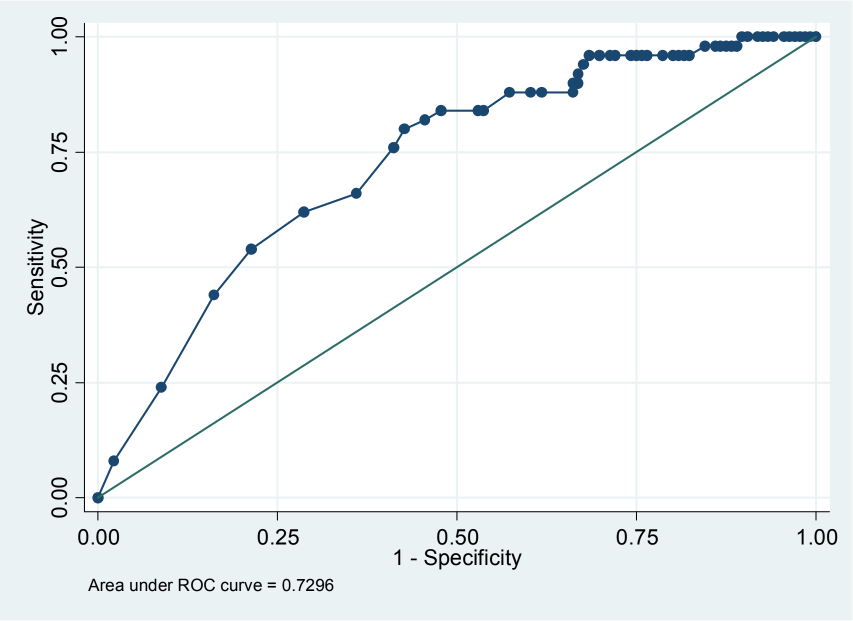
C statistic (AUC) = 0.794 [95% CI 0.73-0.85]

E-2.1.2 ROC and c-statistics for length of stay to predict physical SIPSO subscore (dichotomised at 15 and excluding dead patients)



C statistic 0.75 95% confidence interval 0.68-0.81

E-2.1.3 Receiver Operating Curves (and c-statistics) for length of stay to predict social SIPSO subscore (dichotomised at 15 and excluding dead patients)



C-statistic 0.73 95% confidence intervals 0.66-0.79

Appendix F CIMSS dataset fields

The requisite fields from which the important predictors identified in the study may be derived are outlined in Table 65.

Table 65 Fields required to derive important predictors of SIPSO physical and social subscores in the study

Field	Definition
Baseline NEADL	NEADL completed by patient, or proxy within 7 days of admission with respect to activities performed in the few weeks leading up to stroke
Baseline EQ5D	EQ5D completed by patient, or proxy within 7 days of admission, with respect to the current day
Date of birth	
Independent in ADL prior to admission	Record as yes / no – pre-stroke BI of 19 or 20. No report of requirements for assistnace in ADL from pt or carer
Lived alone prior to admission	No other person registered as living at address
GCS (verbal score)	Five point verbal component of the Glasgow Coma Score
Able to walk without assistance at presentation	Able to walk without support of another person. Does not include use of walking aids
MRC power score (both arms)	MRC power grade (scored 0 to 5) in both upper limbs
Date / time SLT therapy session commenced (communication)	Date and time SLT start therapy session
Reason SLT intervention not required/indicated	Unconscious, no speech, language or communication deficit, receiving palliative care
Date/time admission to hospital/trust	Date and time patient first arrived at hospital (A&E or assessment unit)
Date discharged from acute hospital/death	Date patient discharged from the acute trust (or rehabilitation unit within the trust if inpatient rehabilitation has been provided) to home, community rehabilitation facility or care home.
Date/time admission to ward/bed	Date / time patient arrives at allocated bed.
Ward type	Acute stroke unit, MAU, CCU, HDU, ITU, general medical ward etc.
Previous disabling stroke	Any previous stroke resulting in limitations to ADL, a pre-stroke OHS >=3 or a pre-stroke BI < 19
Address on admission	
Address on discharge	