The harmonisation of stroke datasets: A case study of four UK datasets

Theresa Munyombwe

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

The University of Leeds

The faculty of Medicine and Health

March 2016

Intellectual Property Statement

The candidate confirms that the work submitted is his/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.

Publications

Chapter 4 contains work based on the following publication:

1. Munyombwe, T., Hill, K.M., West, R.M. (2015). Testing measurement invariance of the GHQ-28 in stroke patients, *Quality of Life Research*, 24(8), pp. 1823-1827

   As the first author Theresa Munyombwe carried out all the statistical analyses and prepared the first draft of the manuscript. The other authors provided feedback on the statistical analyses and proof read drafts of the manuscript.

Chapter 7 contains work based on the following publication:

2. Munyombwe T., Hill. K.M., Knapp. P., West.R.M. (2014). Mixture modelling analysis of one-month disability after stroke: stroke outcomes study (SOS1). *Quality of Life Research*, 23(8), pp. 2267-2275.

   As the first author Theresa Munyombwe carried out all the statistical analyses and prepared the first draft of the manuscript. The other authors provided feedback on the statistical analyses and proof read drafts of the manuscript.

Dedication

To my late mom, dad, and my late brothers John, Tinashe, and Edmore

Acknowledgements

Thanks to my supervisors Professor Robert West, Dr Kate Hill, and Dr George Ellison for their unwavering support and confidence that I would complete the work. Your patience and encouragement was invaluable.

Thanks to all members of the Epidemiology & Biostatistics Division, and School of healthcare for your help, support and advice during the time I have been working on this thesis.

Thanks to the Leeds stroke database team, SOS1, SOS2, and CIMSS for providing the datasets that were used in this thesis. I am extremely grateful to Dr Elizabeth Teale (Tizzy) for the support and advice on stroke.

This PhD work would not have been possible without funding from the CLARCH and the Division of Biostatistics, University of Leeds. I am extremely grateful.

Thanks to my family, especially my husband, Raymond and my children Bernadette, Shelton, and Shamilla for coping with the mood swings. Shelly Belly my beloved son, I am sorry for not being there for you when you needed me most.

Thanks to all my Zimbabwean Catholic Community friends and members of the Leeds St Anne Guild for providing the spiritual guidance during this long journey.

Abstract

Introduction

Longitudinal studies of stroke patients play a critical part in developing stroke prognostic models. Stroke longitudinal studies are often limited by small sample sizes, poor recruitment, and high attrition levels. Some of these limitations can be addressed by harmonising and pooling data from existing studies. Thus this thesis evaluated the feasibility of harmonising and pooling secondary stroke datasets to investigate the factors associated with disability after stroke.

Methods

Data from the Clinical Information Management System for Stroke study (n=312), Stroke Outcome Study 1(n=448), Stroke Outcome Study 2 (n=585), and the Leeds Sentinel Stroke National Audit (n=350) were used in this research. The research conducted in this thesis consisted of four stages. The first stage used the Data Schema and Harmonisation Platform for Epidemiological Research (DataSHaPER) approach to evaluate the feasibility of harmonising and pooling the four datasets that were used in this case study. The second stage evaluated the utility of using multi-group-confirmatory-factor analysis for testing measurement invariance of the GHQ-28 measure prior to pooling the datasets. The third stage evaluated the utility of using Item Response Theory (IRT) models and regression-based methods for linking disability outcome measures. The last stage synthesised the harmonised datasets using multi-group latent class analysis and multi-level Poisson models to investigate the factors associated with disability post-stroke.

Results

The main barrier encountered in pooling the four datasets was the heterogeneity in outcome measures. Pooling datasets was beneficial but there was a trade-off between increasing the sample size and losing important covariates. The findings from this present study suggested that the GHQ-28 measure was invariant across the SOS1 and SOS2 stroke cohorts, thus an integrative data analysis of the two SOS datasets was conducted. Harmonising measurement scales using IRT models and regression-based methods was effective for predicting group averages and not individual patient predictions. The analyses of harmonised datasets suggested an association of female gender with anxiety and depressive symptoms post-stroke.

Conclusions

This research concludes that harmonising and pooling data from multiple stroke studies was beneficial but there were challenges in measurement comparability. Continued efforts should be made to develop a Data Schema for stroke to facilitate data sharing in stroke rehabilitation research.

Table of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

ADAMS	Aging, Demographics, and Memory Study
ADL	Activities of Daily Living
AHA SOC	American Heart Association Classification of stroke outcome
AIC	Akaike Information criteria
ART	Anti- Retroviral Therapy
ANOVA	Analysis of Variance
A-QOL	Assessment of Quality of Life
ASD	Autism Symptom Disorder
BADL	Basic Activities of Daily Living
BI	Barthel Index
BIC	Bayesian Information criteria
BioSHaRE	Bio bank standardisation and harmonisation for research excellence
BLRT	Bootstrap Likelihood Ratio Test
CAPE	Clifton Assessment Procedures for the Elderly
CART	Classification Regression Trees
CERERRA	Collaborative Registries for the Evaluation of Rituximab in rheumatoid arthritis
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CHUD	Child health utility
CLAD	Censored least absolute deviations
CIMSS	Clinical Information management System for stroke
CLARHC	Collaboration for Leadership in Applied Health Research and Care
CLESA	Comparison of Longitudinal European Studies on Aging
CPT	Canadian Partnership for Tomorrow Project
DATASHaPER	Data Schema and Harmonisation Platform for Epidemiological Research
DIF	Differential Item Function

EORTC	European Organisation for Research and Treatment of Cancer
EURALIM	EURope ALIMentation
EU-ADR	European Adverse drug reactions
EQ-5D	European QoL-5 Dimensions
EFA	Exploratory Factor Analysis
FAI	Frenchay Activities Index
FIM	Functional Independence measure
GAD	General Anxiety Disorder
GHQ	General Health Questionnaire
GLM	Generalised Linear model
GWAS	Genome Wide Association Studies
HAART	Highly Active Anti -retroviral Therapy
HICDEP	HIV Cohort Data Exchange Protocol
HRQoL	Health Related Quality of Life
HOP	Health Obesity Project
HTA	Health Technology Assessments
IADL	Instrumental Activities of daily living
IALSA	Integrative analysis of longitudinal studies of aging
ICC	Intra cluster correlation
ICF	International Classification of function
IDA	Integrative Data Analysis
IOAB	Idiopathic Overactive Bladder
I-QOL	Incontinence Quality of Life
LMR	Lo-Mendel-Rubin
LPA	Latent Profile Analysis
LREC	Local Research Ethics committee
MAE	Mean Absolute Error
MDD	Major Depressive Disorder
MDS	Minimum Data Set

MFIS	Modified Fatigue Impact Scale
MGCFA	Multi Group Confirmatory Factor Analysis
MGLCA	Multi Group Latent Class Analysis
MFIS	Modified Fatigue Impact Scale
MI	Multiple Imputation
MICE	Multiple imputation chained Equations
MMSE	Mini Mental State Evaluation
MLE	Maximum Likelihood Estimation
MS	Multiple Sclerosis
NDAR	National Database for Autism research
NEADL	Nottingham Extended Activities of Daily Living
NICE	National institute of Clinical excellence
NIHR	National Institute Health Research
NIHSS	National Institute Health Stroke Services
NSAID	Non-steroidal anti-inflammatory drug
NLSAA	Nottingham Longitudinal Study on Activity and Ageing
PANAS	Positive and Negative Affect Schedule
PCA	Principal Component Analysis
PCM	Partial Credit Model
PHOEBE	Promoting Harmonisation of Epidemiological Bio banks in Europe
PG	Public Population Project in Genomics
PROMs	Patient Reported Outcome Measures
PROMIS	Patient Reported Outcome Measures Information Systems
PSAT	Patient Satisfaction
PSE	Present State Examination
QLQ-MY	Quality of life questionnaire
R&D	Research and Development
RMI	River mead Mobility Index
RMSE	Root Mean Square Error

RMSEA	Root Mean Square Error of Approximation
RMSD	Root Mean Square Deviation
RS	Rankin Score
SEM	Structural Equation Modelling
SF	Short Form
SIPSO	Subjective Index of Physical and Social Outcome
SNPs	Stroke National Programmes
SSNAP	Sentinel Stroke National Association Programme
TIA	Transient Ischemic Attack
TICS	Telephone Interview for Cognitive Status
TLI	Tucker-Lewis Index
OLS	Ordinary Least squares
UK	United Kingdom
UGIB	Upper gastrointestinal bleeding
WLSMV	Weighted least squares means and variance

Chapter 1

1 INTRODUCTION

The World Health Organisation (WHO) defines stroke as "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, 1976) . Stroke is the leading cause of disability worldwide (Adamson et al., 2004) and the second commonest cause of death worldwide after ischaemic heart disease (Donnan et al., 2008). Every year in the UK, it is estimated that 152,000 people have a stroke (Townsend et al., 2012), with those above 65 years of age being the most susceptible. In 2010 Morse and General (2010) estimated that stroke costs the National Health Service (NHS) £3 billion a year. Hence preventing stroke is of great importance to reduce mortality, disability, morbidity, and the financial costs of associated treatment and care.

A stroke occurs when the blood supply to the brain is interrupted, depriving the brain of oxygen. There are two main types of stroke: ischaemic stroke (which accounts for 85% of strokes in the UK); and haemorrhagic stroke (which accounts for the remaining 15%) (Morse and General, 2010). Ischaemic stroke occurs when a blood clot blocks the flow of blood to the brain, and haemorrhagic stroke occurs when blood vessels that supply blood to the brain rupture. Sometimes individuals will experience stroke-like symptoms for less than 24 hours. This is known as a 'Transient Ischemic Attack' (TIA), often referred to as a 'mini stroke'.

The main risk factor for stroke is high blood pressure (Donnan et al., 2008), while other risk factors include: obesity; high cholesterol levels; atrial fibrillation; and diabetes (Stroke Association, 2015). Approximately 25% of stroke patients die within the first month post-stroke, 33% within six months, and 50% within one year (Morse and General, 2010; Donnan et al., 2008).

1.1 Stroke disability

The World Health Organisation (WHO, 2001 p.3) defines disability and functioning as "an umbrella term, covering impairments, activity limitations and participation restrictions" (Organization, 2001). 'Impairments' are problems relating

to body functions or structure (e.g. organs or limbs); 'activity limitations' are difficulties in carrying out tasks (e.g. walking or feeding), and 'participation restrictions' are problems relating to involvement in life situations. About 50% of stroke survivors are left with disability (Adamson et al., 2004) hence mortality is not the only important outcome after stroke. The type and severity of post-stroke disability depends on the part of the brain that is affected and the health of the affected person prior to their stroke. As a result the stroke population is heterogeneous (Flick, 1999). For example one stroke might result in loss of power in the right or left hand, while another in the loss of speech or a loss of bladder control, or an inability to swallow. Major strokes may also affect mobility and can lead to dependency in activities of daily living. Similarly, depending on the type and severity of stroke, a stroke can leave individuals with residual impairment of psychological, social and cognitive functions (Tobin et al., 2008). Post-stroke depression is common in stroke patients, with about 33% of stroke survivors suffering from post-stroke depression (Hackett et al., 2005). The impact of stroke is often devastating and some stroke survivors may require major lifestyle adjustments (Pan et al., 2008).

1.1.1 Patient reported outcome measures in stroke rehabilitation

Within health care systems, the various forms of post-stroke disability are assessed using Patient Reported Outcome Measures (PROMs). These types of measures are considered to be particularly important because they define health in terms of how individuals feel, and how they themselves evaluate their health and their prospects for the future (Swenson and Clinch, 2000). In stroke rehabilitation studies, PROMs can be used as single assessments, or as repeated assessments, of disability to assess the progress of a patient over time. However, due to the heterogeneity of disability post-stroke, a multitude of validated outcome measures have been developed to assess the different forms of disability, and there can be considerable inconsistency in the selection of measures or the frequency and timing of patient assessment (Duncan et al., 2000). For example, some researchers have assessed disability following recovery from stroke 3 months post-stroke; others at 6 or 12 months post-stroke. The heterogeneity in outcome measurement scales and patient assessment frequencies/intervals makes comparisons of disability patterns across stroke rehabilitation studies difficult. There is therefore substantial scope for data harmonisation.

In this section, details of some of the commonly used measures that are used to assess cognitive function, physical function, and psychological function post-stroke are provided. Some of these measures were used in the datasets that were harmonised in this present study.

The Mini Mental State Examination (MMSE) questionnaire (Folstein et al., 1975) is one of the most commonly used measure of cognitive impairment post-stroke. The MMSE has 11 questions that tests five areas of cognitive function: attention and calculation, orientation, registration, recall, and language. The summed scores range from 0 to 30 and a higher score is indicative of better cognitive function. A score of 23 or lower is indicative of cognitive impairment (Barnes and Good, 2013). The MMSE is simple to administer but has been criticised for assessing too many functions. Its greatest limitation is its low sensitivity in people with mild cognitive impairment(de Koning et al., 1998) and in stroke patients with right-sided lesions (Dick et al., 1984). MMSE scores have also been shown to be affected by age, education level, and socio-cultural background (Tombaugh and McIntyre, 1992). These variables introduced bias resulting in misclassification of patients.

Neurological deficits in stroke patients at patient admission are commonly assessed using the National Institute Health Stroke Scale (NIHSS; (Dunning, 2011)). The NIHSS has 15 items that measure levels of consciousness, language, neglect, visual-field loss, extra ocular movement, motor strength, ataxia, dysarthria, and sensory loss. Items are scored on a 3 or 4-point scale depending on the item. The summed scores range from 0 (no impairment) to 42 and a higher score is indicative of greater impairment. NIHSS Scores >25 indicate very severe impairment. The NIHSS questionnaire is quick and simple to administer but there are concerns that the scale favours assessment of left hemisphere strokes compared to right hemisphere strokes (Meyer et al., 2002). The scale demonstrated differential function in stroke patients with left and right hemisphere lesions (Millis et al., 2007).

Physical functioning in basic Activities of Daily Living (ADL) post-stroke is commonly assessed using the Barthel Index (BI; (Mahoney, 1965)). The BI has 10 items that assess the performance of patients in such basic ADL as hygiene, continence, dressing, and mobility. The summed scores of the original BI ranged from 0 (totally dependent) to 100 (completely independent. The scores of the modified version range from 0 (totally dependent) to 20 (completely independent). The BI is a

reliable and validated measure but it is considered to lack responsiveness in stroke survivors with mild (as opposed to severe) impairment and is prone to 'ceiling effects' (Duncan et al., 2000) meaning that recovery may continue even when a person has achieved a BI score of 20, but that this improvement will not be captured by the index. It seems likely that the BI is prone to ceiling effects because it only captures basic ADL functions and not more complex/higher level ADL functions. It is recommended to use the BI in the sub-acute phase of stroke and the measures of extended ADL in the longer terms (Schepers et al., 2006).

In response to the limitations of BI, a more comprehensive measure, called the 'Functional Independence Measure' (FIM) was developed by Van der Putten et al. (1999). The FIM has 18 items, 13 of which are based on the BI and measure physical function; the remaining 5 measuring cognition. The FIM's summed scores range from 18 to 126 with a higher score indicating a greater level of independency. Measures of basic ADL are often supplemented with measures of 'extended activities of daily living' (EADL), also known as 'Instrumental activities of daily living' (IADL). These EADL or IADL measures aim to assess a higher level of activities of daily living such as: walking outside; cooking; light and heavy household work; and participation in social activities (Gladman et al., 1993). The commonly used measures of IADL in stroke rehabilitation research include: the Nottingham Extended Activities of Daily Living (NEADL; (Nouri and Lincoln, 1987)); and the Frenchay Activities Index (FAI; (Holbrook and Skilbeck, 1983)). The NEADL and FAI were developed specifically for use in stroke patients to assess post-stroke physical function.

The NEADL questionnaire was developed by compiling a list of 22 items that were thought to be important for daily living in stroke patients. The 22 items were grouped into four categories: mobility (6 items), kitchen (5 items), domestic (5 items), and leisure (6 items). These four subscales were shown to be unidimensional and hierarchical based on the Guttman scaling coefficients (coefficient of reproducibility, scalability) except the mobility subscale (Nouri et al., 1987). Scalability refers to the selection of items that can show response patterns which can be ordered from highest to lowest. The acceptable thresholds for the coefficient of reproducibility is >0.9 and scalability > 0.6. Lincoln and Gladman (1992) confirmed in stroke patients that the four NEADL subscales form hierarchical scales using Guttman scaling coefficients, but the total scale was not unidimensional. The kitchen and domestic subscales were combined and also showed acceptable coefficient of reproducibility >0.9 and scalability >0.6 indicating that these two subscales can be combined to produce a

household score (Lincoln and Gladman, 2009). Nicholl et al. (2002), using factor analysis in multiple sclerosis patients also reported four NEADL subscales and these were labelled: mobility, kitchen, domestic, and communication. The communication subscale contained items for leisure activities. A Rasch analysis of the NEADL by Nair et al. (2011) supported the use of the four NEADL subscales (mobility, kitchen, domestic and leisure) and not the total scores as it was not unidimensional. In stroke rehabilitation research, these four NEADL subscales are used.

Other studies use 21 NEADL items instead of 12 because there is evidence that item 12 (Do you manage your own money when you are out) is problematic as it did not fit well with the other items. The NEADL items have a 4-item response scale: 'not at all'; 'with help'; 'on my own with difficulty'; 'on my own'. The scoring system is (0, 0, 1, 1) or a 4-point Likert type response (0, 1, 2, 3). The summed scores of the NEADL-22 range from 0 to 63 and a higher score is indicative of better function in IADL. A NEADL threshold of 18 or more has been used to determine better function in extended activities of living in elderly patients with chronic airflow limitation (Yohannes et al., 1998). A comparison of the NEADL BI, and FAI by Sarker et al. (2012) showed that in stroke patients, the NEADL was a more sensitive measure of extended activities of daily living without ceiling or floor effects compared to the other measures.

The FAI is a validated measurement scale with 15 items measuring extended activities of daily living. The originators of the FAI scale using factor analysis showed that the scale measures three domains: domestic, leisure/work, and outdoor activities. The three factor solution was consistent with findings from Schuling et al. (1993) using principal component analysis but two items were recommended for deletion from the FAI scale ("Gainfully work" and "reading books"). These two items were found to be of no value in stroke patients since most patients are elderly retired patients. The FAI has a 4-point Likert scoring system (0, 1, 2, 3) and this scoring system produces scores of 0 – 45 for the total scale. A higher score indicates better function in IADL. The FAI is criticised for having a large Smallest Real Difference (SRD) in chronic stroke patients. The SRD indicates real improvement or deterioration for an individual beyond measurement error (Schreuders et al., 2003). The SRD for the FAI scale in stroke patients is 6.7 (Lu et al., 2012) but this is considered to be too large. The FAI measure is also influenced by gender with females having higher scores on domestic activities, and males having higher scores on outdoor activities (Hoolbrook and Skillbeck, 1983). There is also evidence that the FAI measure is age biased, with younger age being associated with higher scores (Appelros, 2007).

Furthermore, other studies have found a floor effect for the FAI scale (Sarker et al., 2012).

The Rivermead Mobility Index (RMI; (Collen et al., 1991)) is also commonly used in stroke patients to assess mobility. It is a validated measure of functional mobility developed for patients with neurological impairment, and has 15 mobility items. The scale is scored on a binary scoring system (0/1), with a score of one indicating ability to carry out the task and zero inability to carry out a task. The RMI summed scores range from 0 to 15, with a higher score indicating better patient mobility. The RMI is quick to administer but there is evidence that the RMI may have ceiling effects in high function patients (Ashford et al., 2015).

There are various validated outcome measures that are used to assess depression or identify depression in stroke survivors. These measures include: the Geriatric Depression Scale (GDS; (Yesavage et al., 1983)); Beck Depression Inventory (BDI; (Beck et al., 1996)); Hamilton Rating Scale for Depression (HRSD; (Hamilton, 1960)), and the Hospital Anxiety and Depression Scale (HADS; (Hamilton, 1960)). These scales were not designed specifically for stroke patients but validation studies of the BDI and HRSD have nonetheless shown that these are acceptable screening instruments for stroke patients (Aben et al., 2002).

The Geriatric Depression Scale (GDS) has 30 items used to identify depression in the elderly. The scores range from 0 to 30 and a higher GDS score is indicative of more depressive symptoms. The GDS-SF is a 'short form' (i.e. a shorter version) of the GDS with 15 items and scores range from 0 to 15. GDS-30 scores that are greater than 10 indicate the presence of depression, and the threshold is > 4 for the shorter version, the GDS-15 (Barnes and Good, 2013). The GDS requires less time to administer compared to some long interview based assessments but tend to have more false negatives for men compared to women (Stiles and McGarrahan, 1998).

The original Beck Depression Inventory (BDI) has 21 items used to measure the severity of depression. The BDI items are scored from 0 to 3. The scores range from 0 to 63 and a higher total score indicate severe depressive symptoms. BDI scores that are greater than 10 are indicative of presence of depression (Barnes and Good, 2013). The BDI scale was recommended as the most suitable scale for assessing depression in stroke patients as it does not rely heavily on somatic components of depression (Aben et al., 2002). However, the threshold of >10 for indicating depressive

symptoms produces a high rate (31%) of misdiagnosis in the stroke population, especially for women (Aben et al., 2002).

The Hamilton Rating Scale for Depression (HRSD) is used to rate the severity of depression in patients who have been diagnosed as depressed. The original HRSD has 17 items, and some are scored from 0(absence) to 4 (extreme presence) and others 0 to 2. A higher score is indicative of more depressive symptoms. HRSD summed scores of 0 to 7 are considered to be normal (Barnes and Good, 2013).

The Hospital Anxiety and Depression Scale (HADS) is commonly used to identify cases of depression and anxiety disorders in physically ill patients , and the total score is considered as a global measure of psychological distress (Barnes and Good, 2013). The scale has 14 items that measure two domains: Anxiety subscale (HADS-A, 7 items) and depression subscale (HADS-D, 7 items). The HADS item are scored from 0 (absence) to 3(extreme presence). The total scale score ranges from 0 to 42 and a higher score is indicative of greater levels of anxiety or depression. HADS subscale scores of greater than 7 are indicative of presence of depression or anxiety (Barnes and Good, 2013). The HADS scale is quick and easy to use but one of the items (*I feel as if I am slowed down*) was shown to be problematic in elderly patients and did not fit with any of the two subscales((Helvik et al., 2011).

The General Health Questionnaires (GHQ; (Goldberg 1992)) are commonly used for measuring psychological distress post-stroke. The GHQ was originally developed as a 60 item measure but shorter versions were developed and these include the GHQ-30, GHQ-28, GHQ-20, and GHQ-12. The factor analysis of the original GHQ-60 measure supported a four-factor structure of: somatic, anxiety/insomnia, social dysfunction, and severe depression (Goldberg and Hillier, 1978). Shorter versions (GHQ-30, GHQ-20, and GHQ-12) were developed by excluding the items relating to physical illness that are often endorsed by physically ill people (Goldberg and Williams, 1988). Unlike the other shorter GHQ versions, the GHQ-28 was derived from the GHQ-60 using factor analysis (Goldberg and Hillier, 1979). Factor analysis of the GHQ-28 supported four factors that measure: somatic (items 1-7); anxiety (items 8-14); social (items15-21); and depression (items 21-28). There are 14 items in the GHQ-30 that do not appear in the GHQ-28. Factor analysis of the GHQ-20 measure in stroke population by Sveen et al. (2004) yielded three dimensions :anxiety, coping, and satisfaction.

The shorter version GHQ-12 is commonly used to screen for psychiatric disorders in stroke survivors because of its ease of administration. The GHQ-12 items are embedded in the GHQ-30 but not in the GHQ-28. Although the GHQ-12 was originally developed as a unidimensional scale, alternative factor structures have been proposed as more appropriate. Only a few studies were found in literature that supported a single factor (Gao et al., 2004). Two factor structures (Werneke et al., 2000) and three factor structures (Graetz, 1991; Bun Cheung, 2002) have been proposed as more appropriate for the GHQ-12 measure. The two GHQ-12 factors reported by Werneke et al. (2000) were labelled: anxiety/depression and social dysfunction. In the GHQ-12 two factor structure, the positively worded items form a factor and the negatively worded items make the other factor. The three factors in Graetz's (1991) model are: anxiety (4 items), social dysfunction (6 items), and loss of confidence (2 items). In Graetz's (1991) model, the positively worded GHQ-12 items loaded on one factor and the negatively worded items on the other two factors.

The validity of the multidimensional GHQ-12 factor structures has been questioned (Gao et al., 2004; Hankins, 2008; Molina et al., 2014). Molina et al. (2014) have argued that the GHQ-12 multidimensional models could be a result of artefacts due to "methods or wording" effects. Methods effects occur when art factual groupings due to positively worded and negatively worded items are formed. The factors produced from factor analysis could be a result of measurement bias introduced by methods or wording effects. Hankins (2008) accounted for the wording effects of the negative items during factor analysis and found that the unidimensional model fitted the data better than the two and three factor models. The high correlations of the dimensions in the multidimensional GHQ-12 models suggest the existence of a higher order factor or unidimensionality of the GHQ-12. Studies by Hankins (2008), Smith et al. (2013) and Molina et al. (2014), also adjusted for the wording effects of the GHQ-12 and concluded that the unidimensional GHQ-12 model fitted the data better compared to the two factor (Werneke et al.,2000) and three factor (Graetz, 1991) models.

The four GHQ measures can be scored using a Likert scoring system (0, 1, 2, 3), 0 denoting (absence) and 3 (presence). Alternatively the items can be scored using a binary scoring method (0, 0, 1, and 1). Based on the binary scoring system (0, 1), the maximum GHQ-60 score is 60, GHQ-30 is 30, GHQ-28 is 28, GHQ-20 is 20 and, 12

for the GHQ-12. Based on the Likert scoring system (0, 3), the maximum score for the GHQ-60 is 180, GHQ-30:90, GHQ-28:84, GHQ-20:60, and GHQ-12:36.

The GHQ cut-off points vary in different studies and the originators of the scale have suggested that the mean GHQ score of a sample can be used to determine the best cut-off point for the sample (Goldberg et al., 1998). Sterling (2011) suggested that a GHQ-28 score of 4 out of 28 on the binary scoring system (0, 0, 1, and 1) indicates the presence of psychological distress, and this is 23/24 on the 0 to 3 scoring system. Similarly, using the binary scoring system, GHQ-30 scores of 4 or more indicate the presence of psychological distress. The threshold for the GHQ-20 for the binary scoring system is 10/11 and Likert scoring is 23/24. The threshold for the GHQ-12 is 3 for a binary scoring system and 11/12 for a Likert scoring system. There are no "case-ness" thresholds for GHQ subscales (House et al., 2001). The GHQ measures have been tested in many different populations but have not been validated adequately in the stroke population where they are frequently used (Salter et al., 2007).

The PROMS described in this section are merely examples of the many PROMS used in stroke research and are not intended to be an exhaustive list. A comprehensive discussion of the other outcome measures that are commonly used in stroke rehabilitation is provided by Barnes and Good (2013) and Salter et al. (2007) .

1.2 Stroke care

Stroke care is complex as it requires delivering highly individualised, complex treatments to a large number of patients (Young and Forster, 2007). The most substantial advance in stroke care has been the creation of specialist stroke care units. Meta-analyses of randomised controlled trials have shown that stroke patients who received stroke unit care have better outcomes compared to patients managed in general wards (Langhorne and Duncan, 2001). Specialist stroke care units reduce mortality rates, dependency and institutionalisation (Langhorne and Dennis, 2004; Toschke et al., 2010; Trialists'Collaboration, 2001; Candelise et al., 2007; Trialists'Collaboration, 1997) and improve a range of different aspects of long term quality of life (Indredavik et al., 1999). There is also evidence that specialist stroke care units can improve functional outcomes by about 20% (Donnan et al., 2008) and there is no evidence to restrict access by age, sex or stroke severity (Langhorne and Dennis, 2004).

Whilst there is evidence regarding the effectiveness of specialist stroke care units on patient outcomes, the key care components associated with good patient outcomes remain unclear. The lack of evidence on such components has been attributed primarily to the use of small studies that lacked statistical power (Candelise et al., 2007) and inadequate case-mix adjustment (Bravata et al., 2010; Mohammed et al., 2005; McNaughton et al., 2003). For example, Davenport et al. (1996) study failed to demonstrate any beneficial effects of specialist stroke care units on mortality rates and the failure was attributed to limited statistical power of their study, which had n= 468 patients identified over 27 months. With the low prevalence and level of variation in mortality rates, this number could have been insufficient.

As highlighted before, stroke patients are a heterogeneous group and there are many potential confounders influencing the relationship between stroke care and patient outcomes. As a result, large samples are required for adequate case-mix adjustment (Flick, 1999) to enable robust investigation of which (if any) key components of care are associated with good patient outcomes post stroke.

1.3 Stroke rehabilitation

Depending on the configuration of local services, some stroke patients in the UK are discharged home to be followed up by their General Practitioners after discharge from acute care. Others, often the more elderly, will instead be discharged to step-down care or to longer-term, institutional care homes. In between these two groups are those patients with potential for rehabilitation but who are not suitable for home discharge; and it is this group that are most likely to be admitted to a specialist stroke care and/or rehabilitation unit. The purpose of rehabilitation post-stroke is to limit the longer-term impact of stroke using a mixture of therapeutic and problem solving approaches (Young and Forster, 2007). While there is evidence that more intensive exercise therapy is beneficial (Veerbeek et al., 2014), there is still need for more research to identify which patients benefit most from specific interventions and the optimal timing, dosage and frequencies of interventions (Verbeek et al., 2014). Thus person-centred approaches which identify patients with specific needs for particular therapies are needed.

1.4 Stroke recovery prognostic factors

The majority of stroke prognostic studies focus on physical function post-stroke. There is a paucity of prognostic studies of psychological function post-stroke despite the fact that a third of stroke survivors suffer from post-stroke depression and about 24 % have anxiety symptoms at six months (Campbell Burton et al., 2013). A systematic review by Bartoli et al. (2013) reported that post-stroke depression is associated with high risk of mortality (House et al., 2001), poor long term outcomes (West et al., 2010; Hadidi et al., 2009; Pohjasvaara et al., 2001), poor rehabilitation outcomes (Ahn et al., 2015), and poor quality of life (Žikić et al., 2014) .

There is evidence that factors such as a previous stroke, pre-stroke disability, baseline stroke severity (Toschke et al., 2010), urinary incontinence, sex (Tilling et al., 2001), and depression (Hackett et al., 2005; West et al., 2010) are associated with long-term physical function outcomes post-stroke. An updated systematic review by Kutlubaev and Hackett (2014) showed that physical disability, stroke severity, cognitive impairment were the factors that were consistently associated with the development of depression. This systematic review reported inconsistent evidence of the relationship between age and depression post-stroke. Some studies showed that older age was associated with post-stroke depression whilst findings from the other studies had no evidence to support this association. An earlier systematic review by Hackett et al. (2009) found that younger patients were more likely to experience post-stroke depression, while Pohjasvaara et al. (2001) found no significant age effect. Another study by Buber and Engelhardt (2011) found that the relationship between age and depressive symptoms was mediated by health and living conditions of older people and age on its own had no explanatory power.

A lot of studies on the clinical association of lesion location and post-stroke depression have been reported but there is inconsistent evidence of the relationship between the site of the brain lesion and post-stroke depression. Studies at John Hopkins University by Robinson et al. (1975) claimed that post-stroke depression was frequent in stroke patients with left lesions rather than the right hemisphere, while other studies suggested the opposite (MacHale et al., 1998). Meta-analyses have failed to establish a definitive relationship between the site of the brain lesion and depression (Bhogal et al., 2004; Carson et al., 2000). The inconsistencies in study findings have been attributed to the heterogeneity in definitions and measurement scales for depression, sampling and study settings (Wei et al., 2015).

Although anxiety symptoms are common in stroke patients, there is limited research on prognostic factors for post-stroke anxiety (Menlove et al., 2015) and there is no consensus regarding some of the prognostic factors (Campbell et al., 2013). A systematic review by Menlove et al. (2015) showed that the predictors of anxiety that were consistent across the 18 studies in their review were pre-stroke depression, stroke severity, early anxiety, and cognitive impairment. Older age was not associated with post-stroke anxiety. Anxiety disorders are less common in older age (Menlove et al., 2015) and a large proportion of stroke patients are older greater than 65 years. There is conflicting evidence of the effect of gender female, previous stroke, physical function on post-stroke anxiety (Menlove et al., 2015) thus more research on the effects of these factors in large and representative stoke studies are needed.

1.5 Stroke Prognostic models

Stroke prognostic models are useful both in research and clinical practice. In observational stroke research, prognostic models are used for case-mix adjustment to correct for differences in patient characteristics when comparing individuals or groups in order to make meaningful group comparisons. In clinical practice, stroke prognostic models are used to predict patient outcomes and help in guiding patient treatment management (Counsell and Dennis, 2001). Stroke prognostic models have not gained much acceptance in clinical practice due to poor prediction accuracy and methodological flaws in their development (Counsell and Dennis, 2001; Teale et al., 2012). For example a systematic review by Counsell and Dennis (2001) found that the majority of the 83 prognostic models that were identified by their review, none were fit for purpose for case-mix adjustment in routine clinical care. The majority of the stroke prognostic models were developed using small samples, with a median sample size of n=209 patients (Counsell and Dennis, 2001). Due to the heterogeneity of stroke outcomes, small effect sizes, and the many confounding factors, complex case-mix adjustment is needed for meaningful group comparisons (Flick, 1999). For this reason, large sample sizes are needed to adjust for case-mix by, for example, propensity score matching or statistical modelling.

Other methodological issues that were found in studies of stroke prognostic models include: poor generalisability, poor handling of non-linear relationships, exclusion of patients with missing data, no follow-up of patients after hospital discharge, no external validation, omission of important baseline clinical variables,

and stroke severity data (Teale, 2011; Hackett et al., 2005; Veerbeek et al., 2011; Counsell and Dennis, 2001). The majority of the stroke prognostic models were developed using cross-sectional designs, there is a paucity of longitudinal stroke disability studies that follow patients over time (Kollen et al., 2006; Veerbeek et al., 2011; Tilling et al., 2001). More research based on large longitudinal studies is therefore urgently needed to better identify the prognostic factors, critical time periods for interventions, and normal recovery patterns of stroke survivors; and to identify how best to monitor patients during recovery (Tilling et al., 2001; Kollen et al., 2006).

However longitudinal stroke studies are expensive because recovery from stroke can take many months or years depending on the severity of the stroke, and thereby require collecting repeated measurements over long periods of time (which can be costly). Furthermore it is also difficult to recruit and retain participants in longitudinal studies (not least patients at increased risk of disability, cognitive deficit and/or psychological dysfunction). This too explains why the majority of longitudinal stroke disability outcome studies are characterised by small samples. For example the (Stroke Outcomes Study 2 (SOS2; (Hill et al., 2009)), from which data were used in the present thesis, initially aimed to recruit a sample of 900 patients into the main cohort during a three year period of active recruitment; and although the study identified 3108 patients of which 1070 (34%) were eligible, the cohort recruited 592 patients – much less than the estimated sample size required to achieve good statistical power (the shortfall in recruitment being partly due to high refusal rates). Other stroke longitudinal outcome studies in the literature are also characterised with small sample sizes (i.e. of less than 400) and these include studies by Tilling et al. (2001), Toschke et al. (2010), and Pan et al. (2008). Additional consequences of using small studies in stroke rehabilitation research are that they lack the statistical power necessary not only for the analysis of primary outcomes and exposures, but also for subgroup analyses; while the small samples used are more likely to be unrepresentative of the target population.

It therefore bears repeating that, due to the heterogeneity of the stroke population, stroke prognostic studies require large samples, with adequate power for subgroup analyses and sampling frames that generate samples that are representative of the stroke population. Larger samples are also necessary for the use of the more complex statistical models required for longitudinal data analysis (such as growth

curve models or multilevel models), which require large samples to model the correlations of the repeated measurements over time.

1.6 Study motivation

This PhD research work was motivated by the availability of secondary longitudinal stroke datasets and the potential of harmonising and combining these to provide the larger longitudinal samples necessary to support the analyses required to inform a better understanding of stroke outcomes. The small samples achieved by previous/existing longitudinal studies (caused by factors such as poor recruitment rates, high attrition rates, and the high costs of longitudinal studies) might be minimised by combining individual patient data from existing datasets to: enhance the statistical power of studies (Thompson, 2009; Allen et al., 2013); provide more precise estimates (Kjær and Ledergerber, 2004); and improve the generalisability of research findings. These benefits have encouraged researchers elsewhere to explore the possibility of pooling existing datasets to address issues inherent in the smaller numbers of participants in each of the constituent studies (Thompson, 2009; Fortier et al., 2010); and of course it is this that has made meta-analysis a popular technique for enhancing the statistical power and analytical value of separate, smaller studies.

The utility of combing multiple individual patient data to increase statistical power has also been demonstrated in studies of rare diseases, rare exposures and rare outcomes (Yoshida et al., 2013). Most notably, many Genome Wide Association Studies (GWAS) combine multiple datasets to generate the huge samples required to understand the role and interaction of genetic, lifestyle, and environmental factors in modulating the risk and progression of disease (Fortier et al., 2010). Although stroke is not a rare disease, and its determinants are not rare, combining datasets is an approach that has great appeal for stroke outcomes research where the nature of the condition has limited opportunities for recruiting large numbers of patients into individual studies.

However, while pooling individual patient data from multiple longitudinal stroke studies appears an attractive solution to the small samples achieved in most stroke studies, careful attention has to be paid to differences in: sampling, study designs, and measurement instruments used (Hofer and Piccinin, 2009). Pooling data from separate studies demands methodological rigor (Fortier et al., 2011) since heterogeneity

between studies poses major challenges that may limit the potential for data pooling (Curran and Hussong, 2009). Systematic differences in sampling and study designs across studies may confound inferences drawn from pooled data analyses (Allen et al., 2013), and for this reason the heterogeneity existing between studies needs to be addressed to permit valid inferences to be drawn from the analysis of pooled data. There is therefore a need for harmonising data from each of the different studies before pooling to allow for valid data integration. This 'data harmonisation' comprises procedures that place variables on the same scale in order to permit pooling of data from multiple independent studies (Hussong et al., 2013; Griffith et al., 2013) and this was the main idea of this thesis.

The present study set out to evaluate the feasibility of harmonising and pooling four different stroke datasets in order to generate large(r), high quality databases that could be analysed to inform a better understanding of stroke outcomes (in particular, the complex interplay of patient characteristics, stroke clinical factors, stroke severity, socio-economic factors, treatments, and patient disability outcomes). Challenges associated with pooling existing individual stroke patient datasets were identified and an attempt was made to address some of the challenges using novel statistical methods. It is important to note that this research was about the innovative application of statistical methods to address a substantive problem in stroke research, and was not about developing new statistical methods.

The availability of four stroke datasets: the Stroke Outcome Study 1 (SOS1; (House et al., 2001)); Stroke Outcome Study 2 (SOS2; (Hill et al., 2009)); the Clinical Information Management System for Stroke (CIMSS; (Teale, 2011)); and the Leeds Sentinel Stroke National Audit (SSNAP; (Rudd et al., 1998)) provided the opportunity to evaluate the feasibility of harmonising such studies and to identify barriers that might prevent data harmonisation and pooling (both here and elsewhere). The choice of datasets examined was influenced by the funding sponsors of this PhD: the NIHR-funded Leeds, York, Bradford CLARHC (Collaboration for Leadership in Applied Health Research and Care) project; stroke rehabilitation being one of the key themes of the CLAHRC project. Furthermore, the primary researchers of the SOS1, SOS2, and CIMSS studies were available to act as research collaborators, so that any necessary clarifications of uncertainties encountered in the datasets were readily available.

Somewhat fortuitously, the four studies presented similarities (for example, all four were based in Yorkshire, UK) and differences that offered an ideal opportunity to examine a wide range of challenges and to explore various statistical methods with which to address these. The SOS1 and SOS2 studies generated research datasets for which the research focus was similar. They were both longitudinal and were designed to generate the data necessary to investigate potential factors influencing depressive symptoms post-stroke. The CIMSS was also longitudinal but had a different objective; focussing on the impact of care processes on patient disability outcomes. And while the Leeds SSNAP data had also been collected to examine care processes following stroke, these data were generated for audit purposes, not for research and the data were not longitudinal. More details on each of the datasets are provided in Chapter 3.

1.7 Aims and objectives

The aim of this thesis was to evaluate the feasibility of harmonising and pooling secondary stroke datasets in order to create large(r), high quality datasets capable of determining the factors associated with patient disability outcomes post-stroke. Insight into factors associated with poor outcomes following stroke have the potential to help target services to those most in need and to assess the effectiveness of existing and future interventions. The challenges that hinder data pooling were therefore identified, and various techniques for harmonising these data were then explored to address these challenges.

Hypothesis: The present study hypothesised that harmonising and pooling secondary longitudinal stroke datasets to create large(r), high quality datasets capable of better analysing and understanding disability outcomes post-stroke was feasible.

More specifically, the objectives were to:

- Harmonise variables with comparable content from the multiple datasets
- Identify the challenges and requirements of harmonisation across independent stroke datasets.
- Evaluate the accuracy of using regression analysis and Item Response Theory (IRT) methods for harmonising patient-reported outcome measures.

- Conduct an illustrative, pooled, longitudinal data analysis to investigate factors associated with anxiety symptoms post-stroke.
- Conduct an illustrative multi-group analysis to identify: (any) subgroups of early disability post-stroke; and (any) factors associated with these subgroups.
- Develop recommendations for data harmonisation for future harmonisation studies.

1.8 Study structure

The study began with a literature review of data harmonisation studies and statistical methods that have been used to harmonise patient reported outcome measures. This was followed by a succession of studies conducted in four research strands.

The first strand of research was the comparison of the study characteristics and variables from the four participating studies, and also determining the potential for harmonising the datasets using the DataSHaPER approach.

Study 1: Qualitative harmonisation of four UK stroke datasets: Application of the data SHaPER approach.

The second strand of research, investigated the measurement invariance properties of the GHQ-28 measure before conducting an integrative data analysis of the GHQ-28. Measurement invariance is a pre-requisite for integrative data analyses of data from multiple sources.

Study 2: Measurement invariance of the GHQ-28 measure in the SOS1 and SOS2 datasets: Application of the Multi-Group Factor Analysis.

The third strand of research investigated the utility of using regression-based methods, item response theory models, and use of common items for harmonising patient reported outcome measures.

Study 3a: Harmonisation of the Frenchay Activities Index and Nottingham Extended Activities Index: Application of regression-based methods and Item response theory models

Study 3b: Harmonisation of the GHQ-12 and GHQ-28 measurement scales.

The fourth strand of research was an illustrative pooled data analysis using the harmonised datasets to determine factors associated with disability after stroke using multi-group latent class analysis and multi-level modelling approaches

Study 4a: Patterns of early disability after stroke: Application of multi-group latent class analysis

Study 4b: Multi-level modelling of anxiety outcomes after stroke: Integrative analysis of Stroke Outcomes Study 1 and Stroke Outcomes Study 2.

1.9 Structure of the thesis

The thesis is presented in nine chapters as follows:

Chapter 1 introduces the work that was undertaken and the motivation for this research. It also provides background information about stroke and challenges in stroke rehabilitation research; and discusses the potential benefits of combining individual longitudinal stroke patient data.

Chapter 2 is a literature review providing an overview of approaches commonly used for harmonising and pooling individual-level data from multiple sources. A critical review of statistical methods for harmonising Patient Reported Outcome Measures (PROMs) and establishing measurement invariance of PROMs is provided. Exemplars of previous data harmonisation studies are presented. The chapter ends with a consideration of the various options available for the methods used in this research work and a justification for the selection of these methods.

The first strand of research is reported in Chapter 3. Chapter 3 reports the Qualitative harmonisation of the four datasets that was conducted in Study 1 using the DataSHaPER approach

The second strand of research is reported in Chapter 4. Chapter 4 describes the measurement invariance analyses of the GHQ-28 measure that were conducted in Study 2 in order to facilitate the pooling of GHQ-28 scores from the SOS1 and SOS2 datasets. Measurement invariance of PROMs across studies is a pre-requisite of pooled data analyses.

The third strand of research is reported in Chapters 5 and 6. Chapter 5 reports the methods and findings from Study 3a which compared the effectiveness of using regression-based models and Item response theory models for harmonising PROMs. Chapter 6 reports the harmonisation of GHQ-12 and GHQ-28 that was conducted in Study 3b in order to facilitate the pooling of data from these scales generated by the SOS and CIMSS studies.

The fourth strand of research is reported in Chapters 7 and 8. Chapter 7 reports the analyses that were conducted in study 4a, which was an illustrative multi-group analysis of harmonised datasets using a latent class analysis framework to compare patterns of disability post-stroke across different stroke cohorts and the factors associated with these subgroups. Chapter 8 reports the analyses that were conducted in Study 4b to demonstrate the benefits of pooling existing datasets; in this instance to investigate factors associated with post-stroke anxiety.

The discussion in Chapter 9 integrates each of the distinct work strands that make up the preceding chapters in this thesis. The main findings emanating from each of these work strands are discussed, focusing on the following questions:

- Was it possible to harmonise and pool data from the four stroke studies; and what were the challenges/barriers to successful data pooling?
- Was harmonisation of multiple studies beneficial?
- What were the statistical issues raised by harmonisation and how effective were the statistical methods used, to address these issues?
- What recommendations can be made for future stroke data harmonisation studies?
- How has the data harmonisation conducted in this research contributed to our knowledge and understanding of recovery after stroke?
- How might the work from this thesis be further developed?

Appendix: The appendix contains: details of the literature search strategies; additional tables and figures; results that could not be included in the main body of the thesis; the R code for the IRT models fitted in Chapter 5; details of measurement model selection for Chapter 7; Mplus software syntax for the multi-group latent class analysis conducted in Chapter 7; Mplus syntax for the multi-group factor analysis for models fitted in chapter 4.

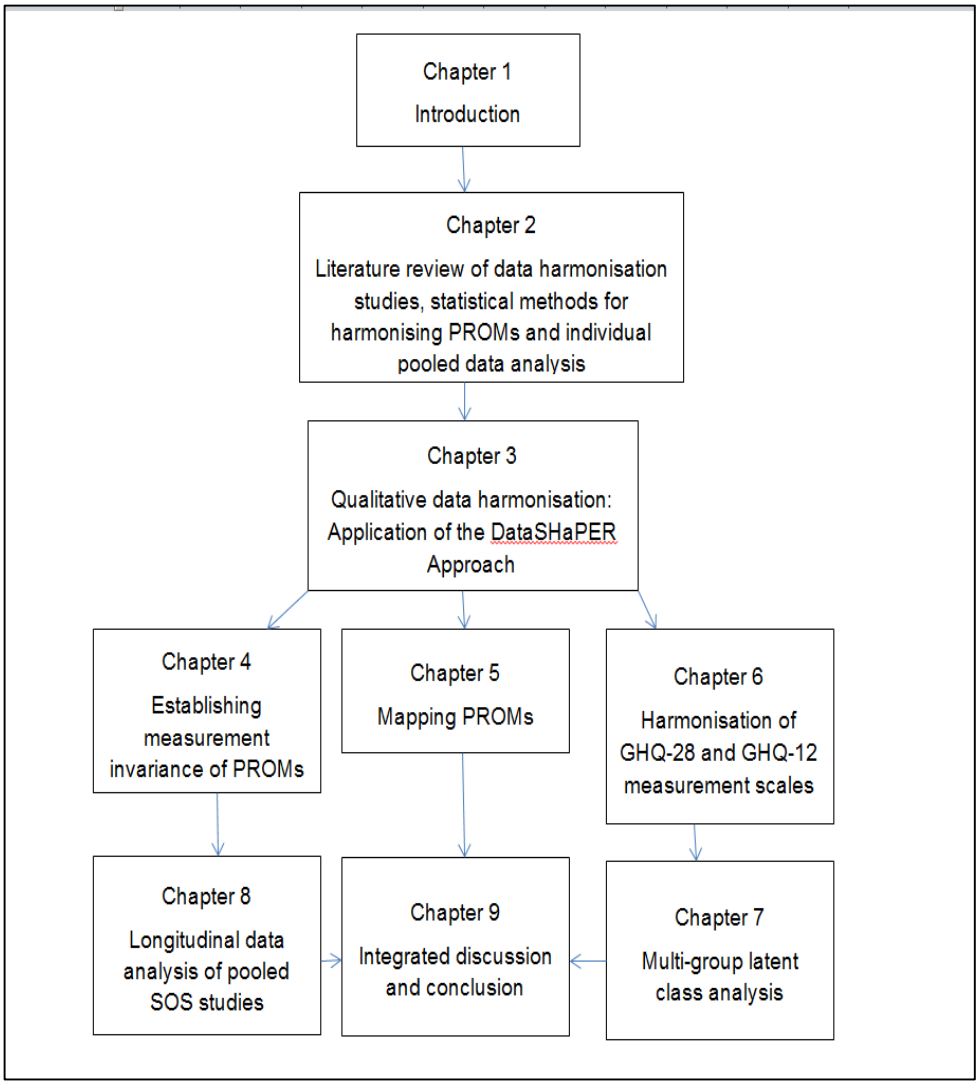

Figure 1.1 Thesis map: outline of the study structure

Chapter 2

2 LITERATURE REVIEW

2.1 Introduction

Chapter 1 provided the background and rationale for this thesis. This Chapter provides a literature review of three different areas: definitions of data harmonisation and approaches for qualitative harmonisation; statistical methods commonly used for harmonising patient reported outcome measures, and pooled individual person data analysis. Chapter 2 begins by providing the definition of data harmonisation that was used in this thesis. In section 2.3, a description of some examples of data harmonisation studies that were found in literature is provided and the challenges commonly encountered during data harmonisation are discussed. In section 2.4, a description of the DataSHaPER approach that is commonly used by data harmonisation studies is provided. Section 2.5 describes methods for establishing measurement invariance of PROMs. Measurement invariance is a pre-requisite of Integrative Data Analysis (IDA) of data from multiple sources. Section 2.6 describes statistical methods that are used for harmonising PROMs that measure similar constructs. Section 2.7 describes the statistical methods for analysing pooled individual patient data from multiple sources. Chapter 2 concludes by providing a summary of the methods that are commonly used for harmonising data from multiple sources and also justifying the methods that were explored in this thesis.

2.2 Data Harmonisation

Data harmonisation is making data from different sources compatible and comparable so that the data can be combined and used in research (Griffith et al., 2015). Researchers are increasingly harmonising and combining individual person data from multiple sources in order to generate large datasets that can be used to address research questions with more statistical power and precision (Kjær and Ledergerber, 2004). Lack of compatibility between studies makes the harmonisation process difficult or impossible (Curran and Hussong, 2009a). Like any research, data harmonisation starts by defining the research question, and this is followed by selecting the studies or collaborators (Griffith et al., 2015). In prospective

harmonisation, researchers will agree on a “core” or important set of variables to be collected, the measurement scales to be used, and the standard operating procedures for data collection (Fortier et al., 2011). When collaborative studies collect similar data this is known as “stringent harmonisation” (Fortier et al., 2011).

Retrospective harmonisation uses information collected by the participating studies and is flexible, it does not require studies to collect identical data, but requires the use of sound methodology to harmonise data (Fortier et al., 2011). The five main sources of heterogeneity in retrospective harmonisation that need to be addressed are: study aims, sampling, study designs, geographical location, and measurement (Curran et al., 2009). The heterogeneity in these five sources is a threat to statistical inference, thus it is important to account for the heterogeneity in study characteristics when analysing pooled data from multiple sources.

2.3 Motivating examples of harmonised individual person data analysis

In this present study, a literature review of data harmonisation studies was conducted to gain an understanding of data harmonisation and synthesis in medical research. The literature review was limited to medical research studies that harmonised individual person data from multiple studies. Searches were performed for English articles only within MEDLINE and dating back to 1996 and google scholar. The initially search in Medline produced 241 results. After removing 28 duplicates and 47 irrelevant articles, 166 articles remained. More articles were also identified from the references of the selected articles. The details of the Medline search strategy and search terms that were used are given in appendix A.

The literature search showed that pooling individual person data has become increasingly popular. Data are harmonised and pooled either for increased statistical power, increased generalisability or for comparative research. Combining multiple individual person datasets for increased statistical power is common in rare diseases, rare exposures, and rare outcomes (Yoshida et al., 2013). In comparative research, data from several cohorts are harmonised and used to test whether results are reproducible or consistent across studies or countries. For example in Genome Wide Association Studies (GWAS) data from multiple sources are pooled to generate large datasets for investigating the interactions of genetic, lifestyle, and environmental factors in chronic diseases (Fortier et al., 2010). A single research group cannot attain the large samples needed in GWAS (Ripke et al., 2013) hence mega analyses of

multiple existing studies are conducted to increase statistical power for subgroup analyses.

Similarly in public health, the HIV Cohort Data Exchange Protocol (HICDEP; Kjaer and Ledergerber, 2004)) was developed to facilitate data merging for joint analysis of observational HIV databases. The HICDEP made substantial contributions to the knowledge of HIV epidemiology and management. Another collaboration of HIV datasets was the Anti- Retroviral Therapy (ART) cohort collaboration (Collaboration, 2007) that was established in 2000 to monitor disease progression among HIV patients starting HAART (Highly Active Anti- Retroviral Therapy). The ART cohort collaboration was an international collaboration of data from 12 cohorts in Europe and North America with n=20 379 adults who started HAART between 1995 and 2003. The data from the collaborative analysis was used to develop 5 year prognostic models with high discriminatory power for patients starting HAART (Collaboration, 2007). ART Collaboration (2007) produced a risk calculator that calculates estimates for progression rates at years 1 to 5 after starting HAART.

In rheumatoid arthritis , a rare outcome, the European Collaborative Registries for the Evaluation of Rituximab in Rheumatoid Arthritis (CERERRA; (Chatzidionysiou et al., 2011)) investigated the effectiveness of rituximab using harmonised and pooled data from 10 European cohorts. Data were pooled to generate a larger sample (n=2019) with increased statistical power for subgroup comparisons. Another registry collaboration study was the EU-ADR Project (Coloma et al., 2011), which combined data from eight electronic healthcare records from four countries (Denmark, Italy, Netherlands, and UK). Coloma et al. (2011) used the harmonised datasets to confirm the increased risk of upper gastrointestinal bleeding (UGIB) in Non-Steroidal Anti-Inflammatory Drug (NSAID) users.

In autism research another rare outcome, large samples that can be used in Autism research are difficult to get. Similar to stroke, Autism is a heterogeneous condition, with severity and symptoms varying widely across those affected thus large samples are required in Autism prognostic research to ensure representativeness. In order to generate large samples that could be used in Autism research, the National Database for Autism Research (NDAR project; (Hall et al., 2012)) was developed. The NDAR project focuses on ways of aggregating existing data from multiple laboratories so as to speed up research through data sharing. The NDAR project

aggregated harmonised data are used to investigate the causes and treatments of Autism Symptom Disorder (ASD; Hall et al., 2012).

In the social sciences, data are mostly harmonised for comparative research. For example the Comparison of Longitudinal European Studies on Aging (CLESA (Minicuci et al., 2003)), was a collaborative study of six European longitudinal studies of aging to investigate the determinants of quality of life. A common database was developed for the CLESA project in five European and one Israeli population. The database provided opportunities to identify common risk factors for mortality and functional decline across the six countries (Minicuci et al., 2003). Measurement comparability was a problem where countries used different outcomes to assess physical function. These measures were harmonised by dividing each score by its maximum and converting scores to a 0 to 1 scale. The limitations of this approach will be discussed in section 2.6 of this chapter.

Another example from social sciences was the Integrative Data Analysis of Longitudinal Studies of Aging (IALSA;(Hofer and Piccinin, 2009)) which was a co-ordinated analysis of 35 longitudinal studies from 35 countries. The harmonised dataset was used to investigate the association between physical, cognitive and social activity and cognitive function in later years.

An Australasian collaborative study by Horwood et al. (2012) conducted an integrative data analysis of four cohorts to investigate the association between frequency of cannabis use and severity of depressive symptoms. The benefits of combining datasets in this study were to increase the sample size and representativeness of the sample. The combined dataset comprised repeated observations on over 6900 individuals assessed on between 3 and 7 occasions. In addition to increased sample size, the combined dataset had a wider range of adolescence age: ranging from 13 to 15 years. The availability of four datasets gathered by independent investigators in different centres offered the advantages of testing for robust and general associations between the use of cannabis and the development of depressive symptoms. Individual studies had produced inconsistent results on the direction of causation with two studies favouring a path from cannabis use to depression, while the other two studies favoured a path from depression to cannabis use. The analyses based on the harmonised integrated dataset favoured a model in which cannabis use led to depression and not depression leading to cannabis.

The four studies that were pooled in the Australasian collaborative study used different measures to assess depressive symptoms and the assessment intervals were also different. A common depression scale was obtained by rescaling depression scores from the different measure to a common mean of 100 and standard deviation of 10. The limitations of the approaches that were used in this study to harmonise the depression scales will be discussed in section 2.6 of this chapter.

The examples of data harmonisation studies described in this section showed that the benefits of harmonising and pooling datasets include: increased sample size, providing more assessment intervals, wider age range of participants, reproducing results in different cohorts, providing robust evidence of associations. As highlighted earlier stroke rehabilitation studies are characterised by small samples due to poor recruitment and attrition. Pooling existing stroke datasets might be an attractive solution to the small samples achieved in most stroke studies. Stroke is heterogeneous in terms of outcomes and the rehabilitation effects are small (Counsell and Dennis, 2001) thus large samples are required for increased statistical power of subgroups analyses.

The next sections of this chapter provide an overview of commonly used approaches for data harmonisation.

2.4 Data SHaPER approach

The data harmonisation process involves a systematic comparison of similarities and differences across studies/datasets to evaluate the potential for harmonisation. The literature search conducted in this present study identified a systematic approach for data harmonisation. The approach is called the "Data Schema and Harmonisation Platform for Epidemiological Research" (DataSHaPER; (Fortier et al., 2010; Fortier et al., 2011)). The DataSHaPER is a systematic approach which was developed to provide a flexible and structured approach for retrospective or prospective data harmonisation. It is co-ordinated by: the Public Population Project in Genomics (PG; (Knoppers et al., 2008)); PHOEBE (Promoting Harmonisation of Epidemiological Bio-banks in Europe, cited in Fortier et al. (2010)); CPT (Canadian Partnership for Tomorrow Project, (Borugian et al., 2010)); and Generation Scotland (Smith et al., 2006). The development of the DataSHaPER approach was motivated by the lack of statistical power in most biosciences research studies where large samples are required

to understand the complex relationships of the genetic, lifestyle, environmental, and social factors in chronic diseases (Fortier et al., 2010). Details of the origins, purpose and scientific foundations of the DataSHaPER approach were provided by Fortier et al. (2010).

The DataSHaPER approach starts by identifying a research question and selecting eligible studies. In retrospective harmonisation, after selecting the eligible studies, the following four steps are conducted:

- Identify the set of "core" or important variables to be shared across studies and create a Data Schema
- Formally assess the potential to share each variable across participating studies
- Define appropriate data processing algorithms to generate the required variables in each study
- Synthesis of harmonised data.

2.4.1 Identifying set of core variables

When pooling data from multiple studies, the core variables needed in the data schema are selected guided by the research question of interest. A data schema is a list of the core variables that are required for the research and their definitions. In retrospective harmonisation, this can be a list of variables that were collected by the original studies. After developing a data schema, the next step is to assess the potential to share each variable across the participating studies.

2.4.2 Evaluating the potential for harmonisation

After developing the data schema, the potential to share each variable in the data schema is assessed. This is called the harmonisation platform. The DataSHaPER approach assesses variables in the data schema on a three-level scale of matching to determine the potential of sharing the data. A process called "pairing" is conducted where variables are classified as "complete matching", "partial matching", and "impossible". "Complete" matching is where the meaning and format of a variable is the same across the studies; "partial matching" is where the variables required can be constructed from the existing variables; and "impossible" is when the variable cannot be constructed from variables available in the studies. Allen et al. (2013) referred to "complete matching" as "ideal circumstances", partial matching as "less than ideal

circumstances" and the third option is "circumstances requiring statistical and design solutions".

2.4.3 Defining appropriate data processing algorithms for harmonising variables

After identifying the "partial matching" variables or variables that need harmonisation, the next stage is the development of processing algorithms for putting the data onto the same metric. These algorithms can be qualitative or quantitative and their accuracy need to be evaluated because choosing the wrong method to harmonise the data may lead to biased results (Griffith et al., 2015). The quantitative harmonisation algorithms use statistical methods to put variables on to the same metric.

2.4.3.1 Synthesis of harmonised data

The last step of the DataSHaPER approach is data synthesis. Depending on the aims of the study, data synthesis of harmonised datasets can be for comparative research or statistical inference. If the data is to be used for statistical inference, it is important to account for the heterogeneity across the studies for valid statistical inference.

The utility of using the DataSHaPER approach for harmonising data from multiple sources has been demonstrated by various studies, these include harmonisation studies by: Fortier et al. (2011); Griffith et al. (2013), and Doiron et al. (2013). For example, the Bio-SHaRE's Healthy Obese Project (HOP, (Doiron et al., 2013)) piloted retrospective data harmonisation using the DataSHaPER approach to harmonise, integrate, and synthesis data collected by eight population-based cohorts across Europe. The harmonised HOP dataset was used to investigate the lifestyle and behavioural risk factors associated with obesity. Eligible studies that fitted the required inclusion criteria were identified. A set of core variables (data schema) that were relevant for answering obesity-related research questions were identified from each participating study. A data schema was generated for the HOP project with 96 variables that included anthropometric, biochemical measures, history of obesity-related disease outcomes, socio-demographic status, and lifestyle and risk factors. After developing the data schema the potential for each study to generate the required variable in the appropriate form was assessed. Where data were not in the required format, processing algorithms were used to transform study specific data into

harmonised formats were possible. In some instances, harmonisation was achieved by simply recoding data, while in other instances harmonisation involved the development of complex data processing algorithms. After generating the harmonised variables in the different study servers, the last step was to co-analyse harmonised datasets while addressing ethical and legal restrictions associated with pooling individual-level data. Special software was used to allow researchers to jointly analyse harmonised data while retaining individual-level data within their respective host institutions.

2.5 Measurement invariance

When conducting an integrative data analysis of PROMs data from multiple sources, it is important to establish measurement invariance (Meredith, 1993) of the measures before pooling the data (Hussong et al., 2013). An outcome measure is measurement invariant across studies if the items reliably and validly assess the same construct across studies (Curran and Hussong, 2009). Observed means of PROMs scores between groups are not directly comparable if measurement invariance is not met (Meredith, 1993; Curran and Hussong, 2009; Chen, 2007). In most multi-studies of PROMs , measurement invariance is often not investigated (King-Kallimanis et al., 2012).

Mellenbergh (1989) provided the mathematical definition of measurement invariance as shown in Equation 2.1:

$$f(X|W,G) = f(X|W) \qquad \text{Equation 2.1}$$

If the conditional distribution of the observed scores ‘X’, given the latent construct ‘W’ is independent of group membership ‘G’ then the measurement invariance assumption holds. If measurement invariance with respect to group membership holds, individuals with identical latent construct scores have the same probability of endorsing scores on the measurement scales regardless of group membership.

2.5.1 Statistical methods for establishing measurement invariance

The literature review conducted in this present study identified methods that are commonly used for establishing measurement invariance in multi-study analyses. The commonly used methods include: Multi-Group Confirmatory Factor Analysis (MG-CFA; (Vandenberg and Lance, 2000)), and item response theory models (Van Der

Linden and Hambleton, 1997). MG-CFA is an extension of confirmatory factor analysis (CFA; (Suhr, 2006)) to accommodate multiple groups. Confirmatory Factor Analysis (CFA) is a statistical method that is used to verify the latent structure of a set of observed variables (Suhr, 2006). Figure 2.1 shows a diagrammatic representation of a CFA model. The unobserved latent variable is shown by oval with an 'F' inside in Figure 2.1(A), the boxes with the 'U's represent the observed indicators (or questionnaire items) of the latent variable. The arrows pointing to the boxes indicate the relationship between the indicators and the underlying continuous latent variable. In a CFA model, observed variables can be questionnaire items and these are linked to the latent variable(s) or factor(s) through a linear function. Factor loadings are the coefficients that link the indicators to the underlying variable; they show the strength of the linear relation between the underlying latent variable and its associated items (Bollen, 1998). A cut-off point of 0.30 is often considered an acceptable magnitude of standardised factor loadings (Kim and Mueller, 1978). In a CFA model, the observed item response is a linear combination of the latent variables 'F', factor loadings, intercept, and the error value for that item (Suhr, 2006).

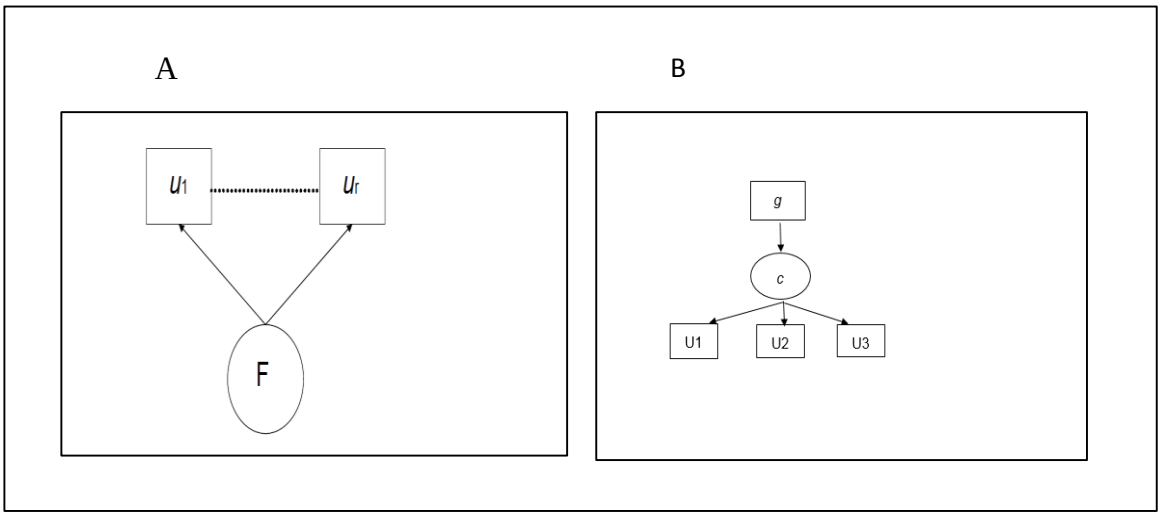

Figure 2.1 (A) Diagrammatic representation of a confirmatory factor analysis model, (B) Multi-group latent variable model

Figure 2.1(B) shows a diagrammatic representation of a MG-CFA model. The Boxes with Us inside represent the observed items/indicator variables. The square with a '*g*' inside represent the "known" grouping variable. The oval shape represents the underlying latent variable which is assumed to be continuous in MG-CFA and

categorical in Multi-Group Latent Class Analysis (MG-LCA). Details of MG-LCA will be provided in Chapter 7.

2.5.1.1 Testing for measurement invariance using MG-CFA

When using MG-CFA, assessing measurement invariance is achieved by conducting simultaneous CFA in multiple groups. MG-CFA is a powerful approach for testing measurement invariance across groups (Steenkamp and Baumgartner, 1998) because the framework allows the comparison of a set of hierarchical measurement models: configural invariance, weak invariance, and strong invariance (Meredith, 1993). A diagrammatic representation of these invariance tests are shown in Figure 2.2. A detailed description of these tests is provided by Vandenberg and Lance (2000). In this Chapter the measurement invariances tests (configural invariance, weak invariance, and strong invariance) that were relevant to this present study are described in the next section.

2.5.1.2 Configural invariance

The most basic level of measurement invariance is configural invariance. It requires the same number of factors across groups and that the same items load on to the same latent factors but factor loadings can vary across groups. If configural invariance exists, this implies that data collected from each group decompose into the same number of factors, with the same items associated with each factor (Meredith, 1993). The configural invariance model is the initial model for testing measurement invariance. To assess configural invariance, an unrestricted baseline model is fitted in each group, with the same number of latent factors but allowing parameter estimates to vary across groups. The parameters that are allowed to vary across groups could be the factor loadings, intercept, and the error value for the indicator items.

2.5.1.3 Factor loading invariance or metric invariance

When configural invariance is indicated additional constraints are imposed on to the configural model to test for factor loading invariance also known as metric invariance or weak invariance (Meredith, 1993). Factor loading invariance requires that all factor loadings are the same across groups and this is achieved by constraining factor loadings to be equal across groups. Figure 2.2 taken from Newsom (2015) shows a diagrammatic representation of the factorial loading invariance model, the constrained factor loadings in groups A and B are represented by greyed elements (λ). If factor loading invariance is indicated, this means that the strength of the association

between each item and the corresponding latent factor is equal across groups. To establish factor loading invariance, the fit of the configural invariance model is compared with the fit of the nested factor loading invariance model using a Chi-square difference test (Bollen, 1989).

Equation 2.2 shows the formula for the likelihood ratio Chi-square difference test.

$$G^2(df) = -2(L_{MO} - L_{MI}) \qquad \text{Equation 2.2}$$

Where *df* is the difference of degrees of freedom between the two nested models;

L_{M0} is the log-likelihood for the baseline model; L_{MI} is the log-likelihood for the more or less constrained model. A significant change in the log-likelihoods indicates that the less constrained model better fits the data than the more constrained model. A non-significant Chi-square difference test is indicative of measurement invariance. There are issues with using the Chi-squared difference test for testing for measurement invariance because it is sample size dependent (Cheung and Rensvold, 2002), large samples lead to inflated Type I error rate for rejecting a true model. Hence other goodness of fit indices such as the Comparative Fit Index (CFI) (Bentler, 1980) are also used to assess measurement invariance. A change in Comparative Fit Index (CFI) of 0.01 or lower suggest evidence of Measurement invariance (Cheung and Rensvold, 2002).

2.5.1.4 Scalar invariance or intercept invariance

When factor loading invariance is indicated, more restrictions are posed onto the factor loading invariance model to test for scalar invariance (Steenkamp and Baumgartner, 1998) also known as intercepts invariance. Scalar invariance is tested by constraining the item intercepts to be equal across groups. Intercepts correspond to the zero value of the underlying latent construct. If a scale achieves scalar invariance it means that scale scores from different groups have the same unit of measurement (factor loading) and the same origin (intercept) hence the factor means can be compared across groups (Chen et al., 2005). Achieving configural, metric, and scalar invariance across groups indicates strong invariance (Meredith, 1993). Figure 2.2 shows a diagrammatic representation of the strong invariance model, the constrained factor loadings (λ) and intercepts (ν) in groups A and B are represented by greyed elements. The Chi-square difference test and the change in CFI are also used to test for scalar invariance. Similar to testing for factor loading invariance, a non-significant

Chi-square difference test or a change of 0.01 or lower for CFI and Tucker-Lewis Index (TLI) is indicative of scalar invariance across groups.

The other invariance tests include strict factorial invariance and structural invariance (Meredith et al., 1993). The strict factorial invariance constraints the factor loadings (λ), intercepts(ν), and the measurement error (θ) to be the same across the groups as shown in Figure 2.2. The structural invariance further constraints the latent factor variance and mean to be the same across the groups (Figure 2.2). Meredith et al. (1993) have argued that valid comparisons of group mean scores can be conducted if configural, factor loading, and scalar invariance hold across groups. In this present study, interest was in assessing whether valid group comparisons of PROMs could be conducted hence no other restrictive measurement invariance models were tested. Details of the other restrictive measurement models are provided by Meredith (1993).

Other methods of testing for measurement invariance that were found in literature include the Item Response Theory (IRT) models and standardised mean difference. Details of measurement invariance using IRT models were provided by Van Der Linden and Hambleton (1997) and details of the use of standardised mean difference for measuring invariance across groups are provided by Dorans (2004). The advantage of using MG-CFA to test for measurement invariance is that the framework can be used to test various levels of invariance: configural; metric; scalar; and other restrictive measurement models. The pooled data analysis conducted in this thesis raised methodological problems of measurement invariance where studies used the same PROMs hence measurement invariance analyses was conducted, the details of are provided in Chapter 4 this thesis.

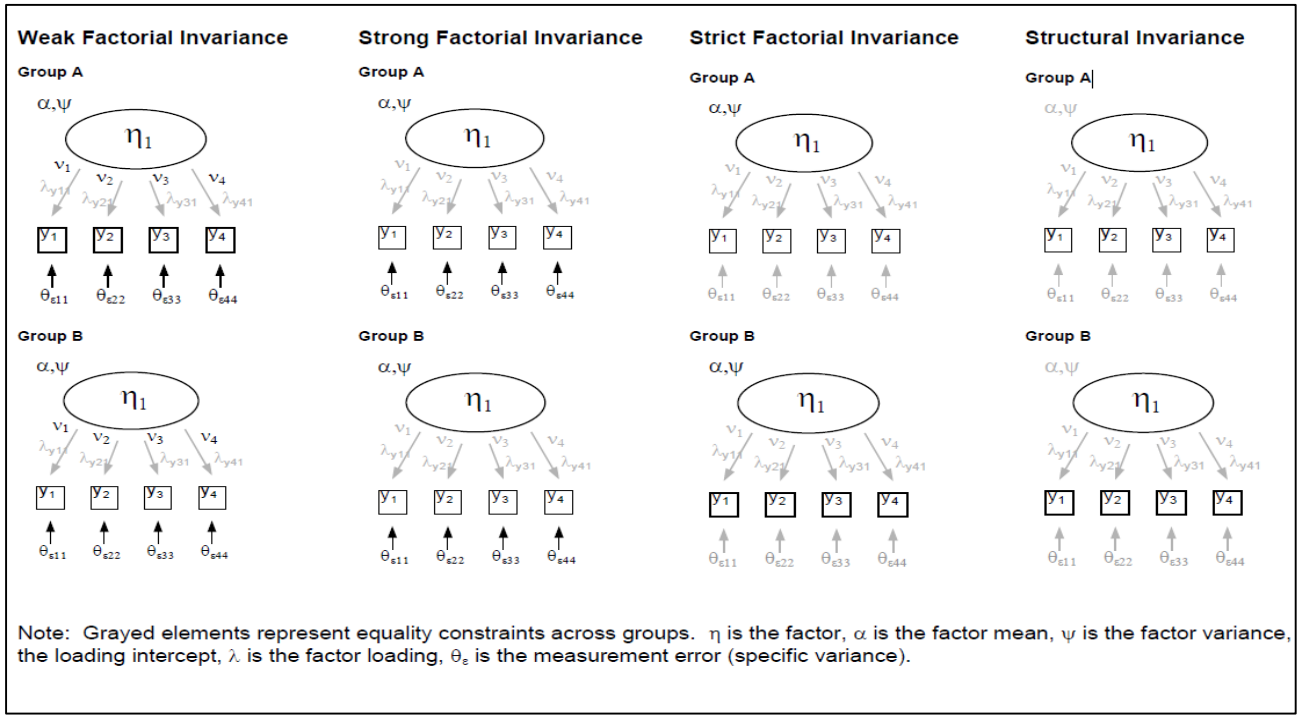

Figure 2.2 Diagrammatic representation of measurement invariance models in multi-groups (taken from Newsom (2015))

2.6 Measurement comparability and approaches for data harmonisation

In retrospective harmonisation, the main barrier of data harmonisation is measurement comparability. Measurement comparability is when studies use different PROMs to assess the same underlying latent construct (Curran and Hussong, 2009). When studies use different outcome measures, there is need to harmonise the data before integration (Bauer and Hussong, 2009). Harmonisation of PROMs describes the procedure of placing variables on the same scale or common metric (Hussong et al., 2013; Griffith et al., 2013). The DataSHaPER approach described in section 2.4 recommends the use of appropriate data processing algorithms to harmonise data from different studies, especially those that will have been recorded as partial matching during the matching exercise. There are various statistical methods that can be used to harmonise PROMS that assess similar constructs and these will be discussed in this section.

In this present study, a second literature search was conducted to identify statistical methods that are commonly used for harmonising PROMs. Searches were performed for English articles only within MEDLINE dating back to 1996 and google scholar. The initially search in MEDLINE produced 229 results. After removing 12 duplicates and 155 irrelevant articles, 62 articles remained. The references of relevant articles were checked to identify more articles. The details of the Medline search strategy and search terms are given in appendix A. The next section provides an overview of the statistical methods that are commonly used for harmonising found in literature.

2.6.1 Algorithmic harmonisation and standardisation

A systematic review by Griffith et al. (2015) identified four general classes of harmonising PROMs and these were: algorithmic harmonisation, standardisation, calibration using e.g. regression models and latent variable modelling. Algorithmic harmonisation is where data are put to same metric by e.g. categorisation using cut-off points to convert scores into categories. For example if studies use different measures of physical activity, these can be harmonised by categorising the data into low, medium, high or dependent and not dependent using some cut-off thresholds. Similarly psychological distress measured by GHQ-28 can be categorised into

presence of psychological distress (yes/no) using a score of 4 on the binary scoring system (Sterling, 2011).

Standardisation puts outcome measures onto the same scale by standardising to a fixed mean and standard deviation. For example standardising by using z scores creates harmonised data by subtracting the mean of the population from each score and dividing by the standard deviation of the population to create normalised z scores. Other methods of standardisation divide the summed scores by the maximum scores to have a score range of zero to one or use T Scores. A detailed description of T scores is provided by Tuokko and Woodward (1996).

Equi-percentile (Lord, 1982) is a traditional method of equating scores. In percentile equating, raw score frequency distributions and their corresponding percentile ranks are obtained for each scale. The ogives corresponding to the data from each scale are plotted and smoothed. Scores having identical percentile ranks in the smoothed cumulative distribution of both scaling are considered equivalent.

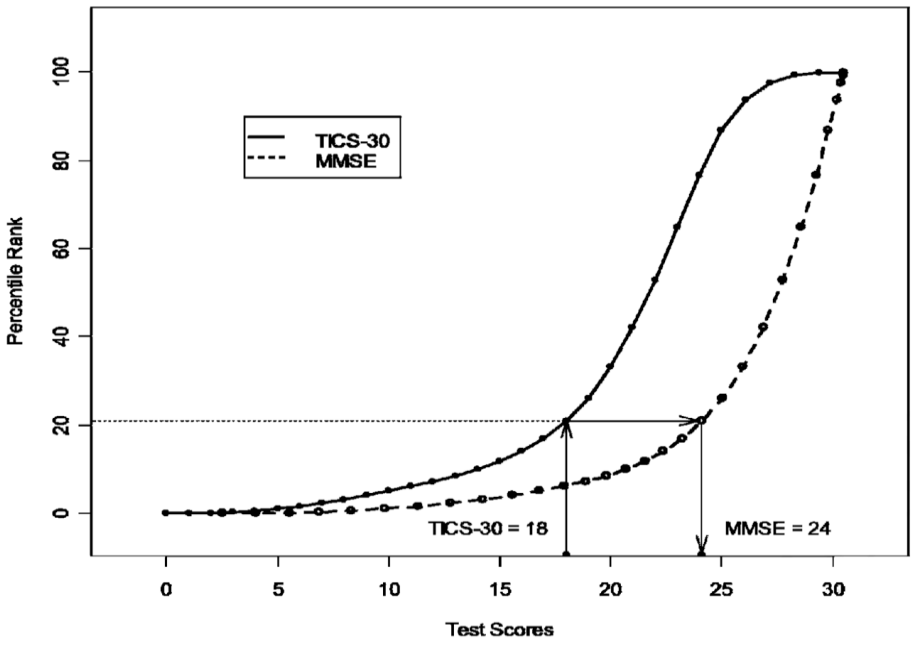

Figure 2.3 Corresponding scores and percentile ranks for TICS-30 and MMSE scores (taken from Fong et al. (2009))

For example a study by Fong et al. (2009) used the Equi-percentile method to harmonise the MMSE and the Telephone Interview for Cognitive Status (TICS-30) measures. The ogives corresponding to the data from the MMSE and TIC-30 were plotted and smoothed as shown in Figure 2.3. Reading off from the graph in Figure 2.3, a TIC score of 18 is equivalent to a MMSE score of 24.

Examples of some of the studies that were found in the literature that used algorithmic harmonisation, standardisation, and Equip-percentile equating are shown in Table 2.1.

Table 2.1 Summary of articles that used categorisation, standardisation, Equi-percentile linking for harmonising PROMs

Study & Title	**Study aim**	**Harmonisation method**
Minicuci et al. (2003) Cross-national determinants of quality of life from six longitudinal studies on aging: The CLESA project.	To investigate the factors contributing to the maintenance of health and function in older persons in different countries and to identify the determinants of morbidity, disability and mortality	**Categorisation** of social and psychological measures **Continuous scores were divided by their maximum** score to have a 0-1 common scale
Bath et al. (2010) The harmonisation of longitudinal data: A case study using data from cohort studies in the Netherlands and the United Kingdom	To develop harmonised data from two independent cohort studies of older people in Netherlands and UK.	Cognitive impairment, anxiety and depression measurement scales were harmonised by **dividing by their maximum scores** to have a 0-1 common scale
Horwood et al. (2012) Cannabis and depression: An integrative data analysis of four Australasian cohorts	This study was an integrative data analysis to investigate the association between frequency of cannabis use and severity of depressive symptoms using data from four Australasian cohort studies.	A common measurement scale of depression was established by **rescaling** all scores from the different depression scores to a **common mean of 100 and standard deviation of 10** within waves for each study.
Fong et al. (2009) The Telephone Interview for Cognitive Status: Creating a crosswalk with the Mini-Mental State Exam	The aim of the study was to develop a metric that allows the linkage of scores on permutations of the TICS and TICS-M to the MMSE. using data from , the Aging, Demographics, and Memory Study (ADAMS) study	**Equi-percentile equating** was used to directly link the Mini-Mental State Examination (MMSE and Telephone Interview for Cognitive Status (TICS)
Noonan et al. (2012) Measuring fatigue in persons with multiple sclerosis	The aim of the study was to create cross-walk tables to associate scores for the Modified Fatigue Impact Scale (MFIS) with scores for the Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) in persons with Multiple Sclerosis (MS)	Cross-walk tables were created using **Equi-percentile linking** to link the MFIS, PROMIS, and SF

For example a study by Bath et al. (2010) summarised in Table 2.1 harmonised cognitive data from two studies the (Longitudinal Aging Study Amsterdam (LASA) and the Nottingham Longitudinal Study on Activity and Ageing (NLSAA)). The LASA study used the Mini Mental State Exam (MMSE-30 point scale) to assess cognitive function and the NLSAA used the Clifton Assessment Procedures for the Elderly (CAPE-12 point scale). The two PROMs were harmonised by simply dividing the MMSE-30 by 30 and CAPE-12 by 12 to obtain a common 0-1 scale. Similarly the Comparison of Longitudinal European studies on Aging (CLESA) project (Minicuci et al., 2003) previously described in section 2.3 and also summarised in Table 2.1, harmonised different social and psychological measurement scales from six longitudinal studies on aging by dividing the scores by their maximum scores to obtain a 0 to 1 scale.

A collaborative study by Horwood et al. (2012) previously described in section 2.3 harmonised different measures of depressive symptoms by standardising all depression scores to a common mean of 100 and standard deviation of 10 within waves for each study. Equi-percentile was used by Fong et al. (2009) described in Table 2.1, to equate the Mini-Mental State Examination (MMSE) and the Telephone Interview for Cognitive Status (TICS) in 746 community dwelling elders who were participants in the Aging, Demographics, and Memory Study (ADAMS, Fong et al., 2009).

Data harmonisation using algorithmic methods or standardisation using z scores or dividing by the maximum score has strengths and weaknesses. The advantage of harmonising PROMs using: categorisation, converting to a common 0-1 scale by dividing scores by their maximum score; or standardising using z scores is that these methods are easy to perform and do not require common items across studies or special software. The disadvantage of categorising is that information on intermediate states may be lost. Griffith et al. (2015) , Curran and Hussong (2009) have argued that standardising using z scores may not be appropriate because this approach assumes that the underlying variable follows a normal distribution and the distribution of the standardised scale is mean and variance invariant. The normality assumption may not be valid for some of the outcome measures because of the ceiling or floor effects in these measures hence failing to take into account the difference in the distributions across groups may bias the data harmonisation (Curran and Hussong, 2009; Griffith et al., 2013).

2.6.2 Harmonising outcome measures by statistical linking

Data harmonisation by calibration requires linking outcome measures by using for example statistical models to transform scores from one outcome measure to another measure. Dorans (2007) describes linking outcome measures as general classes of transformations between one measures to the other. There are three commonly used methods of linking outcome measures: predicting, scale aligning, and equating (Dorans, 2004). The definitions of these linking methods are shown in Table 2.2. In prediction, scores from one outcome measure can be linked to scores on another instrument by a prediction model. Dorans (2004) considered prediction as the weakest form of linking because it does not make any of the assumptions shown in Table 2.2. Furthermore prediction using regression models can only be used if the measures were administered to the same sample.

Equating is considered the strongest form of linking; it makes all the assumptions shown in Table 2.2, and establishes equivalence between the measures being linked (Dorans, 2004). Scale alignment is a lesser form of linking compared to equating but still makes assumptions 1-3 shown in Table 2.2. The details of the assumptions of the various linking types provided by Dorans, 2004 are shown in Table 2.2.

Table 2.2 Types of linking (taken from Dorans (2004))

Type of Linking	**Definition**	**Assumptions**
Equating	Establishes an effective equivalence between scores on two measures to allow scores from both to be used interchangeably	1.Equal(same) constructs measured in both measures 2.Equal reliability (measurement errors) in both measures 3.Symmetrical (function for linking scores of Y to the scores of X should be the inverse of the linking function for equating the scores of X to those of Y) 4.Equity(should not matter if a person is assessed by either one of the two measures that have been equated) 5. Population invariance (linking function used to link measures X and Y should be population invariant.
Scale alignment	Transforms scores from two different measures onto the same metric	All approaches to scale alignment meet assumptions 1-3) as described above. Two approaches relevant to health outcomes are concordance and calibration Concordance: linking two measures developed according to different test specifications that measure similar constructs and have similar reliability estimates (e.g. linking two different fatigue measures) Calibration: Linking two measures developed using the same test specifications that measure the same constructs and have dissimilar reliability
Prediction	Estimates a score from a measure using information from the respondent	Does not require meeting any of the assumptions 1 to 5.

2.6.2.1 Regression-based mapping methods

Harmonisation by prediction commonly uses regression-based methods to link PROMs. For example in economic evaluation studies regression-based methods are commonly used in mapping generic instruments to the EuroQol-5 dimension (EQ-5D) (Torrance, 1986). Longworth and Rowen (2011) defined mapping as "the development of an algorithm that can be used to predict health state utility values using data on other indicators or measures of health" (p.4). Mapping is recognised by the National Institute for Health and Care Excellence (NICE) for use in economic evaluation studies for predicting group averages and individual values of EQ-5D when the EQ-5D data is not available (Chuang and Whitehead, 2011). NICE considers mapping as the second best solution for generating ED-5D data, and a quarter of economic evaluations submitted to NICE technological appraisals rely on mapping to generate the EQ-5D utility scores (Kearns et al., 2012). The literature review conducted in this present study showed that regression analysis using ordinary least squares (OLS; (Dismuke and Lindrooth, 2006)) was the most widely used method used for mapping in economic evaluation studies (Brazier et a., 2010).

The mathematical representation of the classical linear regression model can be written in scalar notation as shown below:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad \text{Equation 2.4}$$

$$\varepsilon_i \sim (0, \sigma^2), \quad i = 1 \ldots n;$$

Where y_i is the outcome variable on the i^{th} observation, βs are the parameter estimates, X_i is the value of the predictor variable on the i^{th} observation, ε_i are the normally distributed random errors. Errors are the differences between the observed and predicted values. For example the outcome variable y_i could be a summed score from a measurement scale and the predictor variables X_i can be scores from the starting measure. The assumptions of the linear regression model are as follows:

- Linearity in the parameters
- Independence: error terms are uncorrelated
- Normality: the residuals are independent and normally distributed with mean 0, variance σ^2
- Homoscedasticity. The variances of the error terms are constant for every $i = 1 \ldots n$.

Longworth and Rowen (2011) produced NICE guidelines that should be followed by Health Technological Assessment mapping studies and these guidelines, are as follows:

- Mapping should be based on statistical associations rather than opinion.
- Characteristics of the estimation sample should be similar to the target sample.
- Standard statistical techniques such as frequency tables, correlations and graphical plots showing the distributions of the measures should be used to examine the data prior to mapping.
- The range of observed outcome measure values from the source sample, and predicted values used in the mapping model, should be fully described to provide information on whether the predicted values have involved extrapolation (which should be avoided).
- An appropriate model that suits the data type should be selected, and prior knowledge of the clinical relationship between variables should inform model selection and application. A justification should be provided explaining why a particular regression model was selected.
- The statistical properties of the mapping algorithms should be clearly described. The root means squared error or mean squared error should be reported, and a plot of observed and predicted values should be used.
- The model should be validated in an external sample similar to the target sample. If an external sample is not available, and the sample size is large, it is recommended that the sample is randomly split to provide an estimation subsample and a validation sample.

Longworth and Rowen (2011) recommended that mapping algorithms can be used to predict values in an independent study provided the study has the predictor variables that were used to develop the algorithm. To reduce prediction errors, a mapping algorithms needs to be applied to a similar population in which the algorithm was developed and validated. There are recent guidelines for reporting mapping studies which were produced by Petrou et al. (2015).

2.6.2.2 Examples of studies that used regression analysis to map patient reported outcome measures

A summary of key studies that were found in literature that used prediction models for mapping PROMs is presented in Table 2.3. The commonly used regression models for harmonising PROMs by prediction include: linear regression models, regression trees, and multiple imputation models. For example studies by Proskorovsky et al. (2014), Chen et al. (2014), Ghatnekar et al. (2013), and Fryback et al. (1997) summarised in Table 2.3 used the Ordinary Least Squares (OLS) estimator to harmonise measurement scales. The disadvantage of using OLS regression models in mapping outcome measures is that it assumes conditional normality of errors and may produce biased estimates if the errors are not normally distributed (Wailoo et al., 2014). The OLS is also known to produce inconsistent regression coefficients in the presence of ceiling effects (Brazier et al., 2010). This bias has been demonstrated in many other disease areas (Alava et al., 2013). Ceiling effects occur when a large proportion of individuals are at the upper end of the scale and floor effects occur when a large proportion is at the lower end. OLS models under predict the upper end of the scale if a measure has ceiling effects and over predicts the lower end of the scale when a measure has floor effects (Brazier et al., 2010). Furthermore, the majority of measurement scales have ordinal data and may require models for ordinal data.

To overcome problems of ceiling or floor effects , other estimators such as the Tobit Model (Tobin, 1958) or the Censored Least Absolute Deviation Model (CLAD) (Cameron and Trivedi, 2005) are used in mapping studies. The Tobit model takes into account the bounded nature of outcome measures but assumes the error terms to be homoscedastic (constant variance)(Sullivan and Ghushchyan, 2006) . The CLAD model relaxes the normality and homoscedastic assumptions for the error terms. Robust estimators such as the MM estimators (Yohai, 1987) have also been used to account for the skewed distributions of the outcome measures. Other models that allow for non-normal errors include the generalised linear models (GLM)(Fox, 2015) and multi-nominal logit regression models (Brazier et al., 2010).

Studies by Ghatnekar et al. (2013) and Rowen et a. (2009) summarised in Table 2.3 explored the utility of using the Tobit and CLAD models to overcome the issues with OLS estimators. Rowen et al. (2009) developed a mapping algorithm for mapping SF-36 onto the EQ-5D index using both random effects Tobit model and

censored least absolute deviations (CLAD) model. Both the Tobit and CLAD models suffered from over- prediction of more severe EQ-5D health states.

Chen et al. (2014) summarised in Table 2.3, compared the prediction performance of an MM-estimator, OLS and GLM to map the Incontinence Quality of Life (I-QOL) scores to the Assessment of Quality of Life 8D (AQoL-8D) utilities in patients with idiopathic overactive bladder. The best model was obtained using the I-QOL total score as the outcome and AQoL-8D, age and gender as predictors using the a robust GLM estimator with a Gaussian family and log link function.

The majority of studies that used regression-based mapping found in literature used summed scores as predictors in the mapping models. Brazier et al. (2010) indicated that using the summed scores of the starting measures as predictors assumes that all items carry equal weight, and the response choices to each item lie on an interval scale for example (the intervals between 'all of the time', 'most of the time', 'some of the time', 'a little of the time' and 'none of the time') will be considered to be equal. Variants of this model relaxes this assumption by modelling subscale summed scores or item responses as predictors. Mapping models that include items as predictors treat items as discrete dummy variables. The only disadvantage of using items as predictors is that this can result in a large number of predictor variables if the measure has many items. To overcome the problem of many predictors, Brazier et al. (2010) recommended the use of significant items only and exclude non-significant items or items with counter-intuitive regression coefficient signs. Some mapping studies include squared terms, interactions, and demographic characteristics such as age and gender to improve the performance of the mapping function.

A few studies that were found in literature used multiple imputation (Little and Rubin, 2014) for harmonising data. Multiple imputation can be used to harmonise different measures even if there is no overlap in outcome measures across the studies (Siddique et al., 2015). The multiple imputation methods treat the unobserved measures as missing data and use imputation models to generate the missing data (Resche-Rigon et al., 2013; Gelman et al., 1998). In multiple imputations, data are assumed to be missing by design or missing at random. An imputer model is developed using combined data from all the participating studies. Missing values are replaced with imputed values to create multiple completed datasets and analysis are conducted in the imputed datasets separately and estimates from the separate imputed

datasets are combined using rules that account for within – imputation and between imputation variability (Siddique et al., 2015). Gelman et al. (1998) recommended the imputer models should use hierarchical models (Raudenbush and Bryk, 2002) to account for the heterogeneity in the multiple studies. The hierarchical modelling will also allow for the inclusion of individual and study specific covariates in the imputation model. A few studies were found in literature that used multiple imputation models to harmonise data from multiple studies. For example Gelman et al. (1998) used hierarchical multiple imputation models to harmonise data from several cross sectional surveys in which some questions were not asked in the other surveys. A recent study by Siddique et al. (2015) extended Gelman, (1998) approach to longitudinal data and harmonised depression data across multiple trials using multiple imputation methods. A harmonisation project of nine Australian longitudinal studies of aging by Burns et al. (2011) summarised in Table 2.3 used multiple imputation to derive estimates of MMSE total scores for participants who reported missing data on any MMSE item. An imputer model was developed using MMSE items, age, gender, years of education, study, and study interactions. Multiple imputation with chained equations (MICE) (Royston, 2009; White et al., 2011) was used to impute missing MMSE item scores. The authors concluded that multiple imputation was an effective method for imputing missing item-level data for the Mini-Mental State Examination (MMSE) and was preferred to other methods of dealing with missing data such as list wise deletion or case wise deletion, and replacement with mean item substitution.

The advantages of using multiple imputation over single imputation methods such as stochastic regression or imputation from a conditional distribution is the ability to account for uncertainties in the missing values by using multiple imputed datasets (Little and Rubin, 2014). Multiple imputation accounts for uncertainty in the missing value by using multiple imputed datasets. The limitations of using multiple imputation approach to impute total scores with partial data have been identified, so the imputation of item-level data is recommended (Graham, 2009). Unlike other methods of harmonisation, in multiple imputation data analysis of the individual person data is based on the measurement scales of interest and not a z score or a latent variable (Siddique et al., 2015). Putting the measures onto the metric that is familiar to researchers makes the interpretation of results easier Harmonisation using multiple imputation also generates imputated datasets that can be shared with other researchers.

Other mapping studies found in literature used regression-based methods together with mixture models to overcome problems of poor predictions in some disease severity states. Wailoo et al. (2014) summarised in Table 2.3, used mixture modelling to develop regression models for predicting EQ-5D from the WOMAC osteoarthritis index. A 2-stage approach was used to develop the mapping algorithms. The first stage classified patients into homogeneous groups using mixture modelling and the second stage developed mapping models within each homogeneous group. Wailoo et al. (2014) overcame the problems of ceiling effects, bi and tri-modal distribution of EQ-5D by using a mixture modelling approach. Standard regression models were not suitable to model the EQ-5D because of the bimodal distributions of the data. Standard regression models that assume a unimodal distribution for the data may produces biased estimates in non-unimodal distributions hence mixture modelling was used to account for this heterogeneity. Mixture modelling is described in detail in chapter 7 of this thesis.

Table 2.3 Summary of articles describing statistical methods for predicting PROMs

Study	Study Aim	Statistical method for harmonisation
Chen et al. (2014) From KIDSCREEN-10 to CHU9D: creating a unique mapping algorithm for application in economic evaluation	To develop an algorithm for generating CHU9D utility scores from KIDSCREEN-10 index summary scores,	KIDSCREEN-10 to CHUD9D Several econometric models were fitted using ordinary least squares estimator, censored least absolute deviations estimator, robust MM-estimator and generalised linear model
Parmigiani et al. (2003) Cross calibration of disability measures: Bayesian analysis of longitudinal categorical ordinal data using negative dependence	To provide a tool for translating between Barthel Index (BI) and Rankin Stroke outcome Scale (RS) aims to get the conditional distribution of RS given BI and BI given RS.	2x2 tables of cross classification of patients by BI and modified RS Estimated the conditional distribution of one measure given the other using Bayesian methods
Proskorovsky et al. (2014) Mapping EORTC QLQ30 and QLQ-MY20 to EQ-5D in patients with Multiple Myeloma	Developed a mapping algorithm for Multiple Myeloma that relates HRQoL scores from the European Organisation for Research and Treatment of Cancer (EORTC) questionnaires QLQ-C30 and QLQ-MY20 to a utility value from the European QoL-5 Dimensions (EQ-5D) questionnaire.	Mapping from EORTC QLQ-C30 and QLQMY20 to EQ-5D scores Multiple linear regression analysis was used to develop the prediction model
Ghatnekar et al. (2013) Mapping health outcome measures from a stroke register to EQ5D weights.	To developed an algorithm for translating variables used for stroke health care quality assessment into EQ-5D	Three regression techniques, ordinary least squares, Tobit, Censored least absolute deviation (CLAD) were used for mapping the Rankin scale to EQ-5D.

Study	Study Aim	Statistical method for harmonisation
Brazier et al. (2010) Review of mapping studies (cross walking studies) from non-preference based measures of health to generic preference based measures.	To evaluate the validity of the mapping approaches and to report lessons learnt for future mapping studies.	The review reported that the most widely used method of mapping PROMs were additive regression models regressing target score e.g. EQ-5D to total index, item scores, dimension scores and the most common method was OLS Other estimators used include the Tobit regression, Censored least absolute deviation model(CLAD) -The commonly used models for categorical outcomes were the ordinal logit and multinomial logit regression models. -A few studies used generalised linear models with random effects.
Burns et al. (2011) Multiple imputation was an efficient method for harmonising the Mini-Mental State Examination with missing item-level data	A harmonisation project of nine Australian longitudinal studies of aging.	Multiple imputation of MMSE items Imputer model was developed and the predictors were gender, years of education, study and study interactions.
Siddique et al. (2015) Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis	A harmonisation project of different measures of depression	Uncollected depression measures were considered as missing data and an imputation model was developed to generate the missing data

Study	**Study Aim**	**Statistical method for harmonisation**
Rowen et al. (2009) Mapping SF-36 onto the EQ-5D index: how reliable is the relationship?	The mapping relationship between the EQ-5D index and the SF-36, a generic non-preference-based health status measure commonly used in clinical trials	Mapping algorithm of SF-36 to EQ-5D Explored three regression models with predictors: (1) all dimensions; (2) all dimensions and squared terms; (3) all dimensions, squared terms and interactions. -Random effects Tobit model suitable for censored data -Censored least absolute deviations (CLAD produces consistent estimates in the presence of heteroscedasticity and non-normality.
Gray et al. (2006) Estimating the association between SF- 12 and EQ-5D utility values by response mapping.	Explored the merits of different methods for mapping between generic quality-of-life instruments.	Mapping algorithm for SF-12 to EQ-5D Multinomial logit regression models Monte Carlo simulation was then employed to place respondents on response levels and called this approach response mapping.
Fryback et al. (1997) Predicting quality of wellbeing scores from the SF-36	The study aimed to develop an empirical equation for predicting Quality of Well Being (QWB) index from SF-36.	Mapping algorithm for SF-36 to QWB scores Multiple linear regression
Wailoo et al. (2014) Modelling the relationship between the WOMAC osteoarthritis index and EQ-5D	This study compared linear regression and mixture modelling for predicting EQ-5D from WOMAC scale	WOMAC to ED5D First used mixture models to identify disease severity classes within the data. A five class mixture model was preferred and a linear regression model was developed to predict EQ-5D within each latent class

Study	Study Aim	Statistical method for harmonisation
Chen et al. (2014) Mapping of Incontinence Quality of Life (I-QOL) Scores to Assessment of Quality of Life 8D (AQoL-8D) Utilities in Patients with Idiopathic Overactive Bladder	The aim of this study was to develop an algorithm to map I-QOL to the Assessment of Quality of Life (AQoL) 8D utility instrument in patients with idiopathic overactive bladder (IOAB).	I-QOL to Quality of Life (AQoL) 8D Compared the performance of OLS, GLM and MM estimator

Advantages and disadvantages of mapping using regression based methods

The advantage of using regression analysis to link outcome measures is that it does not require any of the equating assumptions described in Table 2.2 (same constructs, equal reliability, symmetrical, population invariance) hence it can be used even if measures do not measure similar constructs. The disadvantage of using regression based linking is that the mapping algorithms may lead to increased uncertainty and error around the estimates (Brazier et al., 2010). Furthermore methods such as OLS regression does not restrict the range of predicted values therefore may lead to predicted values that are outside of the required range (Longworth and Rowen, 2011). The predictive performance of regression models may also vary across a range of disease severity (Grootendorst et al., 2007) and may lead to biased predicted values in some disease severities.

2.6.3 Linking outcome measures using latent variable approaches

A sophisticated method of harmonising PROMs reported in a systematic review by Griffiths et al. (2013) was latent variable modelling. The word "latent" means that the true value of the variable cannot be observed but can be measured through the observed variables. The latent variable linking approach posits that a latent factor(s) underlies a set of items, and the items from the different measurement scales measure part of the underlying construct (Griffith et al., 2015). In the context of data harmonisation, latent variable approaches places items from the different measures along the same underlying latent construct hence providing a basis for comparing studies or samples directly (Kern et al., 2014). The first step in linking PROMs using latent variable approaches is to develop a "conversion key" using statistical models such as: factor analysis, Item response theory models, and nonlinear factor analysis (Van Buuren et al., 2005). The "conversion key" models the relationship between the underlying latent construct and the items. The second step uses the "conversion key" to put the data onto a common scale. The advantage of using latent variable approaches for harmonising PROMs is that the equating requirements summarised in Table 2.2 can be checked thus producing strong equating conversion tables that can be used to equate the scores from one measure to another.

2.6.3.1 Latent variable linking designs

There are various linking designs that are used in latent variable linking and these include: single group design, anchor test design also known as the common item non-equivalent groups design (Kolen and Brennan, 2004). Figure 2.4 shows a diagrammatic presentation of the anchor test or non-equivalent designs taken from Ryan and Brockmann (2009). As shown in Figure 2.4, in non-equivalent group designs, a subset of common items from each questionnaire referred to as 'anchors' are used for linking the measurement scales. For example the common items from the GHQ-28 and GHQ-12 measures of psychological distress can be used as anchors when putting the scores from the two measures onto a common scale. The advantage of the anchor test design is that measurement scales can be linked or equated even if the scales were taken by different groups, what is required are common items. The disadvantage is that there may be contextual effects that may bias the linking (Dorans, 2007). Unlike anchor tests designs, single group designs require that individuals complete both measurement scales. The advantage of using the single group design is that differences in abilities across groups are controlled, but there may be order effects that may affect the relationship between the measurement scales (Dorans, 2007). Having respondents completing multiple measurement scales also puts greater burden on respondents. A detailed description of the other designs used in IRT linking is provided by Ryan and Brockmann (2009).

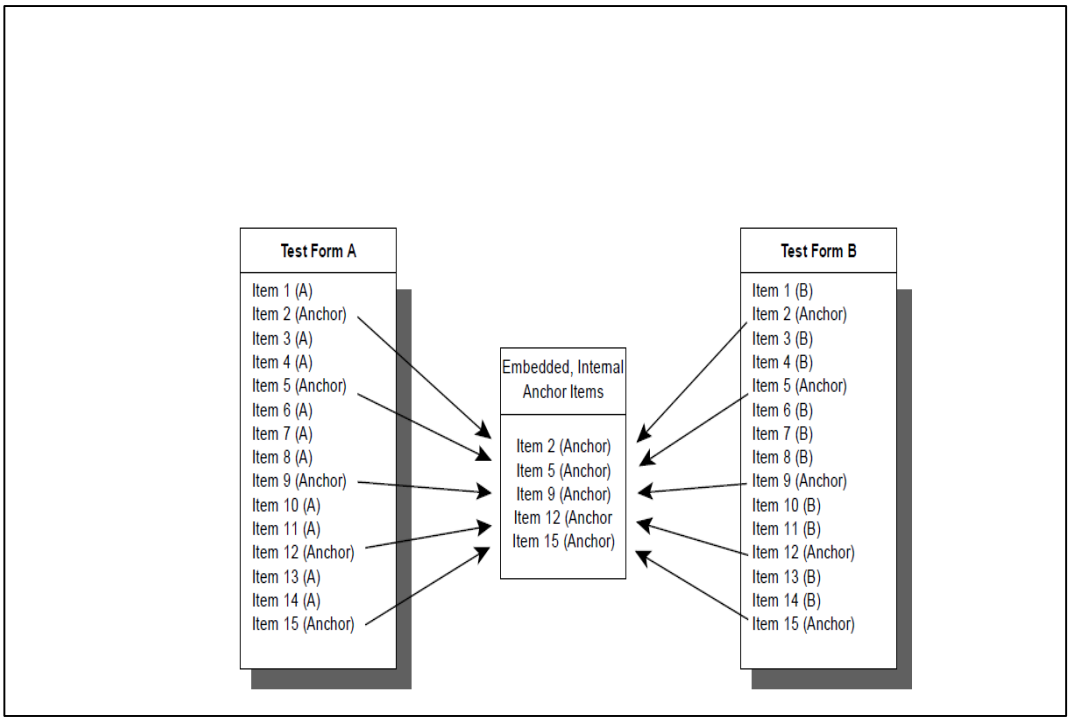

Figure 2.4 Anchor test design or common item non-equivalent group design (taken from Ryan and Brockmann (2009))

The literature review conducted in this present study showed that the majority of studies that used latent variable linking, used item response theory models (Van Der Linden and Hambleton, 1997) to calibrate measurement scales. In this thesis linking using Item response theory models was explored in Chapter 5 and details of linking using IRT models are provided in the next section.

2.6.4 Linking outcome measures using item response theory models

Item Response Theory (IRT) models are used to link PROMs when there is an overlap of items across the studies (Chen et al., 2009). These models are commonly used in educational and psychological research (Dorans, 2007; Velozo et al., 2007). IRT linking is achieved by putting item parameters from different instruments onto the same metric (Chen et al., 2009) using a suitable IRT model. A necessary condition for IRT linking is that the scales to be linked should measure the same or highly similar constructs.

2.6.5 Stages in developing cross walks using IRT methods

Dorans (2007) described the stages for linking similar measures using IRT methods as follows: comparison of the items in the two scales to determine the level

of overlap; establishing the dimensionality of the measurement scales; IRT calibration and scoring; checking for measurement invariance or DIF of the linking algorithm.

2.6.5.1 Dimensionality

In order to use IRT linking there is need to evaluate the dimension(s) underlying a set of observed items. The dimensionality of a measurement scale refers to the number of factors the items fit. The main assumption of traditional IRT models is that a single dimension underlies a set of observed items. There are various methods that can be used to determine the dimensionality of measurement scales and these include: factor analysis(Gorsuch, 1983); and Mokken analysis (Mokken, 1971). Details of factor analysis are provided in this thesis because this is the method that was used in some of the data harmonisation that was conducted in this present study.

The notion behind factor analysis is that a set of observed items can be reduced to fewer unobservable latent variables that share a common variance (Bartholomew et al., 2011). The mathematical model for the classical factor analytical model is shown in equation 2.5

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + e_j \qquad \text{Equation 2.5}$$

Where X_j is the variable represented in the latent factor, j = 1, 2… p, (F_1, F_2, Fm) are the underlying factors, a_{j1}, a_{j2}…a_{jm} are the factor loading, e_j represents the unique factor.

Prior to extracting factors using factor analysis, the suitability of conducting a factor analysis on the data is checked. There are various methods that can be used for this and the commonly used are the Kaiser-Meyer-Olkin (KMO:(Kaiser, 1974)) a measure of sampling adequacy and Bartlett's test of sphericity (Bartlett, 1954) and examination of the item correlation matrix. KMO values range from 0 to 1, and 0.6 is considered suitable for factor analysis (Kaiser, 1974). Tabachnick and Fidell (2007) recommended item correlations of 0.3 or above as appropriate for conducting factor analysis.

There are varying opinions on sample sizes for factor analysis and several rules of thumb have been used by researchers. The commonly used ratio of respondents to variables is 10:1, but other ratios that are also used for factor analysis include are 3:1, 6:1, 15:1, 20, 30:1 (Tabachnick and Fidell, 2007; Hair, 2010). Tabachnick and Fidell (2007) recommended large samples of at least 300 participants for conducting EFA.

The use of large samples reduce the error in the data (Yong and Pearce, 2013). Other researchers have argued that the sample size needed for factor analysis is conditional upon the strength of the factors and the items (Guadagnoli and Velicer, 1988). According to Guadagnoli and Velicer (1988), if the factors have four or more items with loadings of 0.60 or higher, then the size of the sample is not relevant. If the factors have 10 to 12 items that load moderately (0.40 or higher), then a sample size of 150 or more is needed to be confident in the results. If factors are defined with few variables and have moderate to low loadings, a sample size of at least 300 is needed. The use of rules of thumb for determining sample sizes suitable for factor analysis has been criticised as these are misleading and often do not take into account many of the complex dynamics of a factor analysis (MacCallum et al., 1999).

Exploratory factor analysis (EFA) aims to find the smallest number of the common factors that accounts for the correlations of the observed items. EFA identifies enough factors that adequately represent the data. The first factor extracted accounts for the largest percentage of the variance in the data and the second factor accounts for the greatest percentage of the remaining variance not included in the first factor (Suhr, 2006) and factor extraction continues until all the variance in the data has been explained. The commonly used extraction methods for factor analysis are Principal Component Analysis (PCA), Principal Axis Factoring (PAF) and Maximum Likelihood Estimation (ML)(Beavers et al., 2013), unweighted least squares and generalised least squares. The ML requires multivariate normality of the data and PAF makes no distributional assumptions. The default method in most statistical software is PCA and is recommended to use when no prior theory model exists (Gorsuch, 1983).

The extracted factors are rotated in an attempt to achieve simple structure (Bryant and Yarnold, 1995) that makes interpretation easier. A simple structure is achieved when each factor is represented by several items that load strongly onto that factor only (Pett et al., 2003). The common rotation methods are orthogonal (Thompson, 2004) or oblique. Orthogonal produces factor structures that are uncorrelated and oblique produces factor structures that are correlated. Gorsuch (1983) reported four different orthogonal rotation methods and these are varimax, equamax, orthomax and quartimax and 15 different Oblique methods and these include olbimin and promax. Rotation is conducted depending on whether the factors are believed to be correlated (oblique) or uncorrelated (orthogonal). Orthogonal Varimax produces factor structures that are uncorrelated and oblique rotation produce

factors that are correlated. Oblique rotation methods produce more accurate results in research involving human behaviours (Williams et al., 2012).

There are various methods that can be used to determine the optimum number of factors and these include the Kaiser's criteria (eigenvalue > 1 rule) (Kaiser, 1960), the scree test (Cattell, 1966) and the cumulative percent of variance extracted. The eigenvalue describes the amount of variance in the items that can be explained by the associated factor (Pett et al., 2003). There are no fixed thresholds that exist for percent of variance extracted. The Kaiser rule is to drop all factors or components with eigenvalues under 1.0. A scree plot is a plot of eigenvalues against components (Figure 2.5). The inspection of the scree plot indicates the number of factors that should be considered for a given set of indicator variables. The position on the plot where the curve levels off ('elbow') determines the number of the factors. For example in Figure 2.5 there is a pronounced elbow at factor 3, thus a three factor solution might be appropriate for the data. The use of the scree plot is criticised for being subjective as the plot can have multiple elbows or no clear breaks.

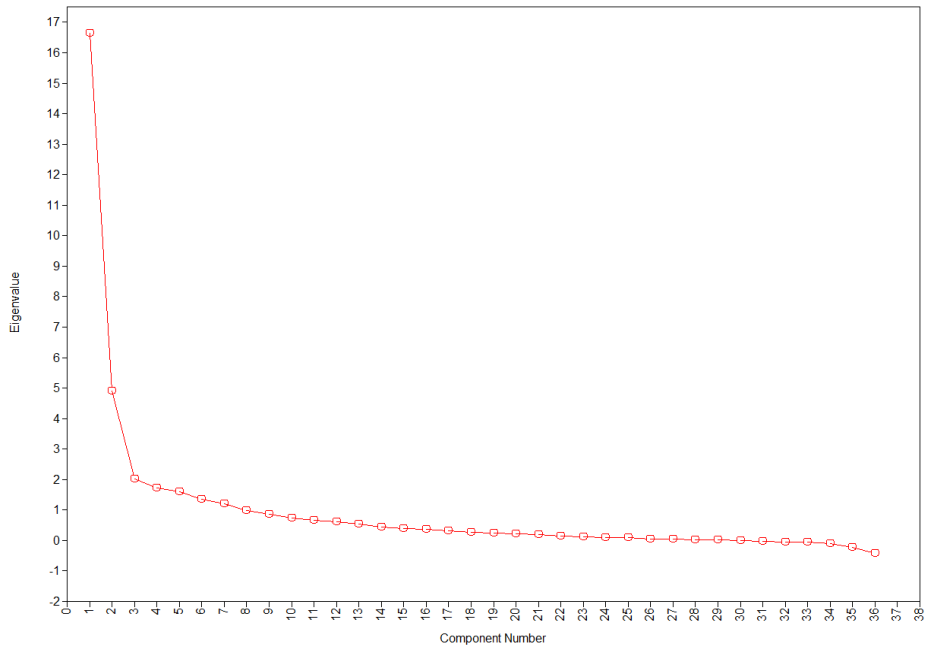

Figure 2.5 Scree plot of eigenvalue from factor analysis

Due to the subjective nature of some of the criteria for determining the number of factors to extract, researchers are encouraged to use multiple criteria to determine the number of factors explained by the observed variables or indicators.

After identifying the number of factors the researcher then examines items that load onto the factor(s) and those that are redundant and decides whether the items should be discarded. Items with factor loading close to 1 are important in interpretation of the factor and those that load close to 0 are not important. Osborne and Costello (2009) recommend that stable factors contain at least 3 to 5 items with significant factor loadings. The factors are then given a name depending of the items that load to the factor. For example a factor may have items that all relate to anxiety loading onto one factor, therefore the researcher can name the factor "anxiety".

2.6.5.2 IRT Calibration

When linking outcome measures using IRT methods, after establishing unidimensionality of the items, the items are calibrated using a suitable IRT model. In IRT linking, calibration refers to the process of estimating item parameters using an appropriate IRT model (Dorans, 2007). Items can be calibrated by using common subjects, common items or both. IRT calibration can be achieved by using suitable IRT models.

The item response theory models use a logistic regression model to describe the relationship between observed item responses and the underlying latent variable. Logistic regression models are used to model the probability of a person choosing a particular response category on an item given the person's latent variable score. For a binary item with responses, such as yes/no, the item response function gives the conditional probability (p_i) of endorsing a "yes" given the person's latent variable score. Figure 2.6 shows the diagrammatic representation for an item response function for a binary item. The item response function shown in Figure 2.6 is sigmoid or 'S' shaped and shows that higher scores of the latent variable (theta) are associated with higher conditional probability of endorsing the item.

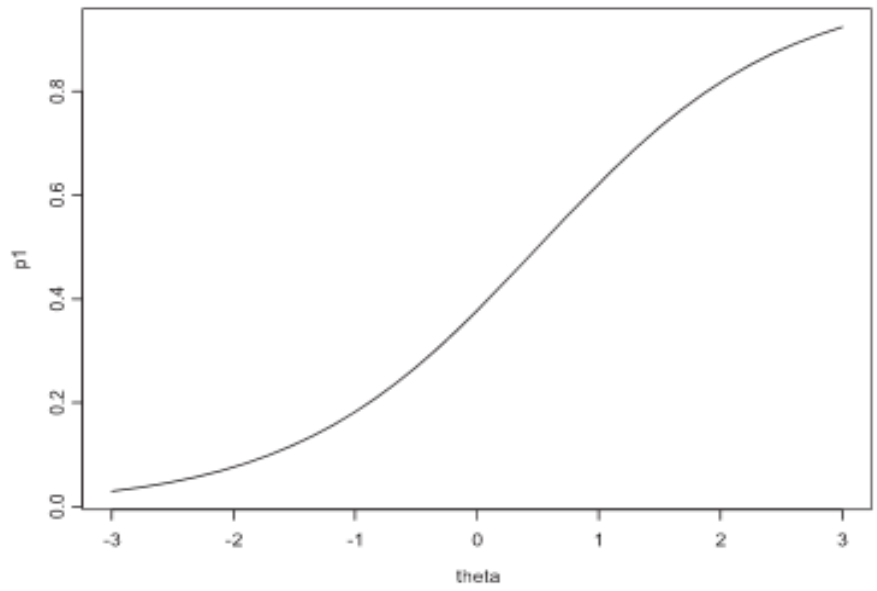

Figure 2.6 Item response function for a binary item (taken from Millsap (2010))

There are three commonly used IRT models that can be used to analyse binary items and these are the one, two and three IRT parameter models. The one-parameter Rasch model (Rasch, 1993) is the simplest IRT model and has one parameter, the item difficulty. The Rasch model assumes the item discrimination parameter to be constant. Assuming the discriminatory parameters to be constant implies that all items are equally reliable. Equation 2.2 shows the mathematical presentation of a one-Parameter Logistic (1PL) model.

$$P(X_{ij} = 1|\theta_i) = \frac{1}{1 + \exp[-D_{aj}\,(\theta_i - b_j)} \qquad \text{Equation} \quad 2.2$$

Where X_{ij} is the observed response to item j, θ_i is the latent variable, b_j is the difficult parameter, D is the scaling constant and for the logistic function D is 1.7.

The Rasch model assumes that the total sum of the scores is a sufficient statistics, i.e. it contains all the information in the data that is needed to estimate a person's latent variable score Millsap (2010). The assumption of equally reliable parameters made by the Rasch model may not be realistic hence other models that do not make this assumption may be required. The two Parameter Logistic (2 PL) model (Millsap, 2010) is an extension of the one-parameter Logistic model and unlike the one-parameter model, it allows the discrimination parameter to vary across items. The item response function for a two parameter model is described by two parameters, the item difficulty and item discrimination. The item discrimination is the slope of the item response function and determines the steepness of the Item response function at

each difficult value. The three-Parameter Logistic (3PL) model (Millsap, 2010) extends the two parameter model by adding a guessing parameter to the 2PL model. The model takes into account that respondent may just guess for example in multiple choice questions.

The IRT models for binary items described above can also be used for polytomous items. Polytomous items are items with more than two responses. For example response items on the NEADL "not at all'; 'with help'; 'on my own with difficulty'; 'on my own'. The commonly used IRT models for polytomous items are the Partial Credit Model (PCM: (Masters, 1982)) and the Graded Response Model (GRM: (Samejima, 1997)). The partial credit model extends the binary one parameter Rasch model to a polytomous model. In a polytomous model, the categorical response function is the conditional probability of choosing a particular response category given the individual's latent variable score. An example of a category response function for a four item is given in Figure 2.7. The category response function shows that Individuals with higher theta values are more likely to endorse category four and individuals with low theta values more likely to endorse category one.

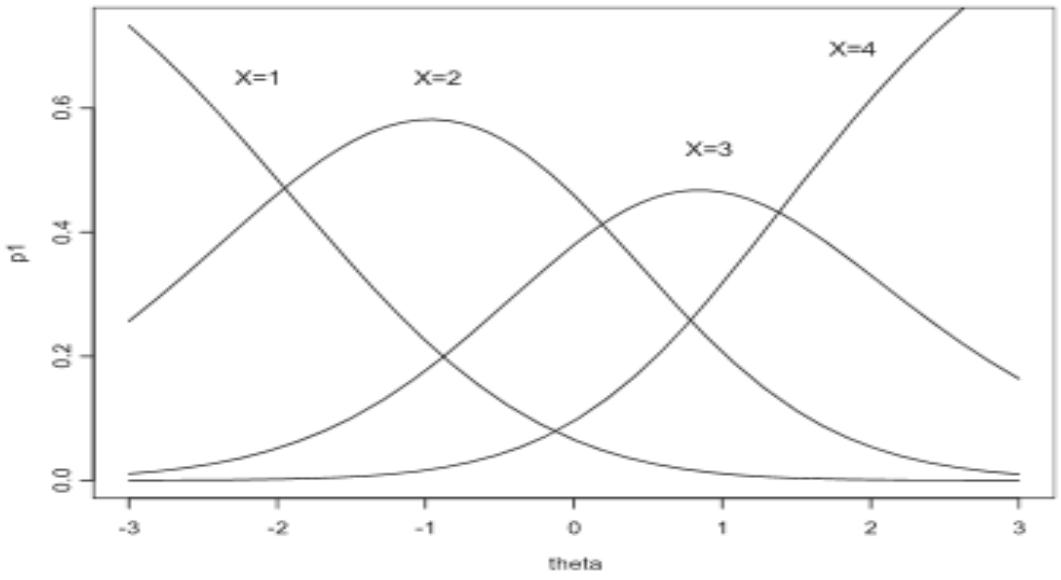

Figure 2.7 Item response function for the Partial Credit Model for an item with four response categories (taken from Millsap (2010))

The polytomous Rasch model allows the different items to have different category response functions but assumes that the discrimination parameter is constant

implying that all items are equally reliable (Millsap, 2010). The polytomous GRM allows the categorical response functions to vary showing variations in the discrimination parameters across items.

When linking measurement scales using IRT models, items from the different measures can be calibrated concurrently/simultaneously or separately (McHorney and Cohen, 2000). In concurrent or simultaneous IRT calibration, the item parameters are estimated simultaneously in both the base (starting) measurement scale (e.g. FAI) and target (e.g. NEADL). In separate IRT calibration, the parameters of the common items from the base (starting) test are estimated first and then held fixed and used as anchors when estimating the parameters of remaining items from the target test (Chen et al., 2009). Other separate IRT calibrations calibrate each sample separately and then link the scales using some scale transformation. There is evidence that the use of marginal maximum likelihood with concurrent calibration is slightly more accurate than separate calibration and linking (Kim and Cohen, 1998).

2.6.5.3 Checking subgroup measurement invariance of the linking algorithm

The IRT methodology for linking outcome measures require checking for subgroup invariance (measurement bias) of the linking algorithm. Subgroup variants of the linking algorithm occurs when it performs differently in different patient subgroups e.g. in male and females. Dorans and Holland (2000) suggested the use of the standardised root mean square deviations (RMSD) to determine measurement invariance of the linking algorithm across patient characteristics such as age group and gender. The RMSD compares the differences between the standardised difference of subgroups such as age group (>65 years, < 65 years, gender (male, females) or study group (SOS1, SOS2).

2.6.5.4 Test Equating

When items from different measurement scales are simultaneously calibrated using a suitable IRT model, all item parameters are automatically on the same metric scale Kim and Cohen (1998). The individual person IRT scores from simultaneous IRT calibration can be used instead of the original scores. In separate calibrations of measurement scales with common items, items parameter estimates are not automatically on the same metric as in concurrent/simultaneous calibration. To link the item calibrations from the separate calibrations, a scale transformation is conducted (Kim and Cohen, 1998). The scale transformation establishes a mathematical relationship that places the item parameters onto the same metric. Scale

"transformation constants" or equating coefficients are calculated and then used to place item parameters from the separate calibrations on a common mathematical metric (Kolen and Brennan, 2004; McHorney and Cohen, 2000). Lord (1980) showed that, under IRT the relationship between the metric of any two calibrations is linear and can be expressed as:

$$\theta^* = A\theta + B \qquad \text{Equation 2.5}$$

Where A is the slope and B is the intercept of the linear transformation and θ^* is θ_i expressed in the target metric. The new item parameters can be transformed to the target metric using the same coefficients as shown in equation 2.6 and equation 2.7. Linking the two metrics requires finding constants A and B.

$$a_j^* = \frac{a_j}{A} \qquad \text{Equation 2.6}$$

$$b_j^* = Ab_j + B \qquad \text{Equation 2.7}$$

Where * indicates a transformed value, aj and aj* are the slope parameter and bj and bj* are the location or threshold parameters.

The commonly used methods for transforming scores from one measure to another in IRT scoring are the mean/mean, mean/sigma and test characteristic curve methods (Chen et al., 2009). Sophisticated approaches such as the Stocking and Lord method (Stocking and Lord, 1983) obtain the slopes and intercepts coefficients by minimising a quadratic loss function based on the difference between characteristic curves estimated in each sample (Kim and Cohen , 1998). For an IRT Samejima graded response model described in section 2.6.5.2, Baker (1992) extended the Stocking and Lord (1983) method to obtain the two constants A and B by minimising the quadratic loss function shown in Equation 2.8.

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_{i1} - T_{i2}^*) \qquad \text{Equation 2.8}$$

Where N is an arbitrary number of points along the theta metric, T_{i1} and T_{i2} are the expected number of correct scores for groups1 and 2 respectively. The algorithms that minimises the quadratic response functions under a GRM model are implemented in special software such as the EQUATE software version 2 (Baker, 1993). Details of IRT calibration and scale transformations for the Graded Response Model are provided by Cohen and Kim (1998). The IRT approaches for linking measurement

scales are useful if the data does not violate the IRT assumptions of unidimensionality and DIF (Dorans, 2007) therefore it is important to check the IRT assumptions for effective linking. The next section describes some examples of studies inform the literature that used IRT linking to harmonise patient reported outcome measures (PROMs).

2.6.6 Examples of studies that used IRT to link measurement scales

The utility of using IRT co-calibration to harmonise outcome measures has been demonstrated by several studies and some key examples found in the literature are shown in Table 2.4.

Table 2.4 Summary of articles describing IRT methods for linking PROMs

Study	Study Aim	Statistical method for harmonisation
Curran & Hussong (2009) Integrative data analysis: The simultaneous Analysis of multiple data sets	The primary aim was to identify developmental pathways that lead to substance use and disorder	IRT method was used to put Anxiety & depression scales from multiple studies on to the same metric (IRT scores)
Hussong et al. (2008) Disaggregating the distal, proximal and time varying effects of parent alcoholism on children's internalizing symptoms	Integrative data analysis of two prospective studies of children to investigate effects of parent alcoholism on children	Categorised summed items as present and absent and then used the 2 parameter logistic IRT model to create commensurate measures. Pooled data was analysed using a Random effects modelling approach
Bauer and Hussong (2009) A review of Psychometric approaches for developing commensurate measures across independent studies: Traditional and New models	A review of methods for producing commensurate predictors and outcomes	The review identified commonly used approaches for linking measurement scales such as : Latent factor models, 2-parameter logistic model
Byers (2004) Testing the accuracy of linking healthcare data across the continuum of care	This PhD study was an external validation of a cross walk or conversion table designed to transform a score on the physical ability component of the Functional Independence Measure (FIM) to its corresponding score on the Minimum Data Set (MDS) and vice versa	A FIM-MDS conversion table was developed using Rasch analysis

Study	Study design	Statistical method for harmonisation
Hawthorne et al. (2008) Deriving utility scores from the SF-36 using the Rasch analysis	To derive EQ 5D utility scores from the SF-36 using Rasch analysis	SF-36 mapped to ED-5D using Rasch analysis
Holzner et al. (2006) Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research	To derive direct Conversion tables for the European Organization for Research and Treatment of Cancer Core Questionnaire (EORTC QLQ-C30) and the Functional Assessment for Cancer Therapy – General (FACT-G).	Confirmatory factor analysis was used to confirm the unidimensionality of the EORTC QLQ-C30 and FACT-G scores The pooled set of items in each pair of corresponding EORTC QLQ-C30 and FACT-G subscales was calibrated using the Rasch model
Velozo et al. (2007) Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set	This study demonstrated the utility of using Rasch analysis for the creation of a crosswalk between the Functional Independence Measure (FIM), which is used in inpatient rehabilitation, and the Minimum Data Set (MDS), which is used in skilled nursing facilities.	A crosswalk between the Functional Independence Measure (FIM) and MDS was created using Rasch models
Askew et al. (2013) Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form	To develop a cross walk for two pain interference measures.	Items were calibrated by combining the data from two pain measurement scales and used a 2-parameter logistic graded response model to estimate the discrimination and difficult parameters for the items. The calibration was anchored on the established parameters for one of the measures

Study	Study design	Statistical method for harmonisation
Edelen et al. (2014) Correspondence between the RAND-Negative Impact of Asthma on Quality of Life item bank and the Marks Asthma Quality of Life Questionnaire	Provided an example of how to transform scores across disparate measures (the Marks Asthma Quality of Life Questionnaire (AQLQ-Marks) and the newly developed RAND-Negative Impact of Asthma on Quality of Life item bank (RAND-IAQL-Bank)) using an IRT-based linking method	AQLQ-Marks to crosswalk to the RAND-IAQL toolkit was created using IRT models

For example a study by Curran et al. (2008) summarised in Table 2.4 combined data from three cohorts and used IRT methodology to put three measurement scales of depression and anxiety onto a common scale. IRT scores were then used in subsequent analysis of the pooled datasets. Another study by Velozo et al. (2007) (Table 2.4) developed a cross walk between the Functional Independence Measure (FIM) and the Minimum Data Set (MDS)-Post Acute Care using a one parameter IRT Rasch model. A cross-walk table was created that can be used by other researchers to convert between these two measures. In stroke research a study by Hsueh et al. (2004) used the one parameter Rasch model to co-calibrate the combined 23 items from BI and FAI and produced conversion tables that can be used by prospective users to derive the Rasch-transformed scores from the raw scores. A multiple sclerosis study by (Askew et al., 2013) summarised in Table 2.4, used IRT methodology and developed and tested a cross-walk table to transform the Brief Pain Inventory pain interference scale (BPI-PI) scores to PROMIS-PI short form (PROMIS-PI SF) scores. After establishing sufficient overlap between the two measures, item data from the BPI-PI and PROMIS-PI SF items were combined. The two measures were administered to the same respondents therefore a single design was used for calibration. A two-parameter logistic GRM response model was fitted to the data. Calibration was anchored onto the established parameters of the PROMIS-PI SF and parameters of the BPI-PI were freely estimated. The accuracy of the conversion table was assessed by comparing observed and predicted scores using the standardised root mean square difference

(RMSD) and Bland Altman plots. The cross-walk was validated in independent samples.

A study by Schalet et al. (2014) used IRT methodology and a single group design (same individuals completed all measurement scales) to produce a cross-walk linking three anxiety scores in order to compare anxiety scores across three studies. The researchers reported that IRT method was preferred over other harmonisation methods such as percentile rank scores and standardised scores because these were problematic as they produce scores that are highly sensitive to sample characteristics such as restricted range (Baguley, 2009). Concurrent or simultaneous calibration of three measurement scales: the Mood and Anxiety Symptom Questionnaire (MASQ) a 90 item scale, Generalized Anxiety Disorder Scale (GAD-7), Positive and Negative Affect Schedule (PANAS) a 20 item scale for general dimensions of mood to the PROMIS anxiety measure was conducted. A simultaneous calibration was conducted with the parameters of the PROMIS items fixed on their published results. Fixing the parameters of the PROMIS items on to their published results placed the other three measures onto the PROMIS metric. The simultaneous or fixed parameter calibration was compared with separate calibration whereby the different measures were calibrated separately and then linked by determining a transformation constant and equi-percentile linking. The results from separate calibration and fixed calibration were similar. The accuracy of the linking algorithms was assessed by comparing the actual and predicted scores. The IRT (fixed-calibration) linking method was slightly more accurate than equi-percentile methods.

Edelen et al. (2014) developed a cross-walk that can be used by researchers and clinicians to convert between the Marks Asthma Quality of Life Questionnaire (AQLQ-Marks) and the RAND-Negative Impact of Asthma on Quality of Life item bank (RAND-IAQL) toolkit using a single group design. The researchers first determined whether the two measures were measuring the same constructs by determining their correlation and this was found to be > 0.8. The items from the two measures were combined and EFA was used to determine the dimensionality of the pooled items. After confirming the unidimensionality of the combined measure, an IRT calibration of the entire 85-item set was conducted using IRTPRO software (Cai et al., 2011) and conversion tables were produced that can be used to convert between the scales were produced. The cross-walked pain interference scores adequately

approximated observed PROMIS-PI SF scores in both the calibration and validation samples.

IRT harmonisation can also be used in longitudinal datasets. For example Curran and Hussong (2009) demonstrated the utility of using IRT calibration in an integrative analysis of three longitudinal studies. Twelve items were drawn from the anxiety and depression subscale of the BSI and 15 items were drawn from the anxiety and depression subscale of the CBCL questionnaire. A total of 27 items were drawn from the anxiety and depression subscales of the two different scales. Six items were common across the two scales as they had same wording across the two scales. The four steps of IRT linking dimensionality, DIF, calibration and scoring were used to derive individual IRT scores that were used for secondary analysis to develop trajectories of internalising symptomatology in the pooled data analysis of n=1827 individuals.

2.6.7 Harmonising patient reported outcome measures using common items

In other data harmonisation studies, different measurement scales that assess similar constructs were harmonised by selecting common items across the measures and these were used to construct the desired measure. For example the CLESA project (Minicuci et al., 2003) previously described in section 2.3 of this Chapter developed common databases across six countries and harmonised ADL measurement scales by selecting four ADL items that were common across the different measurement scales that were used by the different countries. The four items that were selected are: "bathing, dressing, transferring and toilet ability items". The items were dichotomised into needed help (yes/no) and the scores of the four items were summed. Similarly three items on Instrumental Activities of Daily Living (IADL): "preparing meals", "shopping", and "doing light housework", were harmonised and dichotomised into "able to perform the activity without assistance" vs "inability/need of help". Harmonising outcome measures by selecting common items is easier but the disadvantage is that important items may be lost resulting in a measure with poor psychometric properties. When this approach is used studies need to establish the psychometric properties of the selected items to determine the performance of the reduced scale.

The next section discusses approaches that are used to analyse harmonised data from multiple sources.

2.7 Statistical approaches for analysing combined data from multiple sources

Integrative methods of combining data from multiple studies offer multiple advantages and these include: replication of findings across studies, increased statistical power for statistical tests, broader psychometric assessments of constructs, more follow-up periods (Hofer and Piccinin, 2010; Curran and Hussong, 2009). The main issue for pooled data analysis in epidemiologic studies is whether differences in the populations and methods used in original studies influence the results obtained from the pooled data analysis (Friedenreich, 1993). Hence pooled data analysis should be mindful of misidentifying effects as theoretically meaningful when they are artefacts resulting from differences in sampling composition across studies (Hussong et al., 2013). It is therefore critical to identify important sources of between-study heterogeneity when conducting meta-analysis or integrative data analysis (Hussong et al., 2003). If heterogeneity is present there is need to account for it in the pooled data analysis using statistical methods that account for heterogeneity across studies.

2.7.1 Meta-analysis: fixed effects and random effects models

In epidemiological studies, data from multiple studies can be combined using meta-analytical approaches. The traditional meta-analysis combine summarises from multiple studies by putting the data on a common metric such as standardised mean differences or log odds ratios. In traditional meta-analytic methods, a two stage approach is used to synthesise the data. In the first stage, summary statistics are first calculated for each study separately. In the second stage, the study estimates are combined using meta-analytical models that produce a pooled common estimate. The pooled estimate is a weighted estimate of summaries from different studies. Pooled estimates are calculated using fixed effects or random affects(Hedges and Vevea, 1998) meta-analytical models depending on the heterogeneity of the studies. The fixed effect meta-analytical model assumes that studies are measuring the same underlying true treatment effect and any variation is due to chance alone (Sutton et al., 2003). In fixed effects meta-analytic models, different methods are used to calculate the pooled estimate and these include inverse variance, Mantel-Haenzel and Peto for combining odds ratios (Sutton et al., 2003). Details of these methods were provided by Sutton et al., (2003).

The random effect model assumes that the treatment effect is not the same across studies but varies from one study to another (Sutton et al., 2003) and is commonly used to account for heterogeneity across studies. The limitations of traditional meta-analysis that uses summary statistics from published articles is that the studies may use different research methods or different modelling approaches creating difficulties in combining the results from such studies (Thompson, 2009). Furthermore using summary statistics restricts the analysis to the aims that were focused by the primary studies and no sophisticated statistical analysis such as mediation moderation, growth modelling and subgroup analysis can be conducted due to lack of individual data (Siddique et al., 2015).

2.7.2 Integrative data analysis: fixed and random effects models

In response to the limitations of traditional meta-analysis, researchers are increasingly using individual person data meta-analysis (Riley et al., 2010) also known as integrative data analysis(Curran and Hussong, 2009). The advantages of synthesising individual person data are that the researcher can adjust for the same patient-level covariates, account for differential in follow up times, missing data and also conduct sophisticated analyses which were impossible from summaries from individual studies (Thompson, 2009; Siddique et al., 2015). While individual person data analyses offer many benefits, data from individual studies should not be pooled and analysed as if it's from a single study, without accounting for the heterogeneity in the studies. Ignoring the heterogeneity in the datasets might lead to biased estimates (Verma et al., 2009; Riley et al., 2010) hence it is important to check for heterogeneity across studies and account for heterogeneity during pooled data analysis.

In individual person data from multiple studies can be analysed using integrative data analyses (IDA) approaches suggested by Curran and Hussong (2009). IDA fits models directly to the combined pooled individual dataset. Fixed effects IDA models takes into account between-study heterogeneity by modelling the effects of study membership directly into the model. The main advantage of modelling study membership directly into the model is that one can also estimate multiplicative interactions between individual characteristics such as gender and study membership (Hussong et al., 2013). Having interaction terms will allow the testing of differential impact of individual characteristics on outcomes across the set of studies. A significant interaction between study and a covariate indicate that the relationship vary by study. If the study membership indicator is statistically significant this will help to

identify differences in the outcomes between the two studies. Random effects IDA assumes hierarchy in the data for example patients nested within studies. There are two sources of variability in random effects IDA; variability due to sampling of studies and variability due to sampling of individuals within studies (Raudenbush and Bryk, 2002). Random effects statistical models such as multi-level models (Raudenbush and Bryk, 2002) can be used to analyse pooled data from multiple sources. The advantages of conducting pooled individual patient data analysis using the random effects modelling framework are that researchers can disaggregate patient-level effects, study-level effects, and patient by study cross-level interactions (Curran and Hussong, 2009). Study-level predictors could be type of sampling used by the study, mode of data collection (e.g. postal or face to face interviews), and geographical location of study.

Multilevel models are commonly estimated using maximum likelihood methods that require large number of groups/studies for estimating group/study level variance parameters. There is debate on the minimum acceptable number of units at group/study level. Some researchers have argued that the multi-level approach requires about 20 to 30 studies(Van der Leeden and Busing, 1994; Kreft et al., 1998), while Busing (1993) suggested more than 100 groups/studies and Gelman and Hill (2006), argued that the number of groups does not matter. The researchers in favour of having a large number of groups indicated that a large number of units at the group/study level is needed in order to have better estimates of the group-level variance parameter. In stroke rehabilitation research, finding 20 to 30 studies would be difficult. In the absence of a large number of contributing studies, a fixed effect IDA in which study membership is treated as a fixed factor (rather than a random effect) is preferable (Hussong et al., 2013).

2.8 Summary of literature review

In this literature review four topics were discussed and these are: (1) definitions of data harmonisation and approaches for qualitative harmonisation; (2) statistical methods commonly used for harmonising patient reported outcome measures in medical research; (3) examples of data harmonisation studies in medical research and challenges in quantitative data harmonisation; and (4) statistical models for pooled individual person data analysis. The literature review showed that there are many data harmonisation studies in medical research. The main reasons for pooling individual

person data varied from increased statistical power for subgroup analysis and precision of estimates, comparative research for descriptive and inferential purposes, and wider variation in patient characteristics. Following guidelines from (Fortier et al., 2010) there are some studies that were found in literature that used the DataSHaPER approach to conduct a systematic harmonisation of data from multiple sources. The advantage of using this approach is its flexibility as it does not require studies to use identical data collection tools and procedures but require the use of sound methodology to ensure inferential equivalence of harmonised information. The data harmonisation platform of this approach recommends the development of harmonisation algorithms that can be used to derive harmonised variables using exiting variables where studies collected similar but not the same variables.

In this present study, the DataSHaPER approach for data harmonisation was preferred because of its flexibility. The approach does not require studies to use identical data collection tools and procedures but requires the use of sound methodology to ensure inferential equivalence of harmonised information. Because of the heterogeneity in PROMs used in stroke rehabilitation research, the DataSHaPER approach may be useful for data harmonisation and pooling or synthesis of existing stroke patient reported outcome surveys. This thesis explored the utility of using the DataSHaPER approach to harmonise and pool data from four UK studies. The approach is not a new method but has not been utilised in stroke research. Fortier et al. (2010) suggested future work to develop future DataSHaPERs on particular conditions (e.g. stroke, type 2 diabetes) and this current study has initiated that process by testing the utility of this approach in stroke studies.

The main methodological complexities encountered by data harmonisation studies found in this current literature review are measurement invariance and measurement comparability of PROMs. Even if studies have used the same outcome measures, there is need to establish measurement invariance across studies before pooling for comparative research for valid comparisons. However the majority of comparative research studies found during the literature review did not check for measurement invariance before comparing averages of PROMs across studies. The most popular method for establishing measurement invariance found in literature was multi-group confirmatory factor analysis (MG-CFA). The second most popular method for establishing measurement invariance was IRT methodology. In this present study, the utility of using MG-CFA to establish measurement invariance of

PROMs was explored. The results are reported in Chapter 4 of this thesis. The MG-CFA was chosen because it provided a framework for testing many measurement invariance assumptions such as configural, metric, and scalar invariance and these analyses do not require special software.

The use of different PROMs to assess the same constructs complicates data pooling or comparisons across studies. The literature review showed that various methods are used to deal with measurement comparability. Some researchers only pooled common data across studies and excluded variable that were not common across studies. Using common variables across studies is advantageous but may result in losing many important variables that are needed to answer the research question of interest. In response to measurement comparability, other researchers have used linking methodologies to create crosswalk tables or mapping algorithms that associate scores from one measure to the corresponding score on the other measure in order to compare studies or pool data from multiple sources. The commonly methods of data harmonisation of different outcomes that measure the same or similar constructs are linear transformation, z transformations, latent variable methods and multiple imputations. Some linear transformation methods use prediction or mapping using regression based models to estimate the relationship between measures and creating algorithms that predict the missing measure. Mapping has been recommended by the National Institute for Health and Care Excellence (NICE) for use in economic evaluation studies for predicting group averages and individual values of EQ-5D when the EQ-5D data is not available. Calibration of PROMs using latent variable approaches use approaches such as IRT methodology, Latent factor analysis or nonlinear factor analysis. IRT scores from simultaneous calibration automatically places the measurement scales on to the same metric. The disadvantage of using regression based approaches is that it requires the two questionnaires to be administered to the same sample, whereas IRT models can be used even if the two questionnaires were not administered to the same sample as long as there are common items that can be used to link or anchor the two questionnaires. Multiple imputation methods treat the unobserved measures as missing data and use imputation models to generate the missing data.

In this present study the utility of using both regressions-based models and IRT methods to link the FAI and NEADL measures was explored in Study 3a reported in Chapter 5. These two methods were preferred because as highlighted in the literature

review, they are widely used to link PROMs in psychological research and economic evaluation studies and their utility has been confirmed by the various studies that were discussed in this literature review. The SOS1 provided the ideal opportunity to investigate the effectives of linking measures using a single group design since it collected data on both measures of IADL.

Key messages from chapter two

- Data harmonisation and synthesis is increasingly being used in medical research
- The main problem in data harmonisation is the heterogeneity in variables collected by participating studies.
- The commonly used statistical methods for harmonising different PROMs are z score transformation, regression-based methods and latent variables approaches
- The most common statistical method for establishing measurement invariance is multi-group confirmatory factor analysis.

The next chapter describes the first strand of research that was conducted in study 1 to evaluate the feasibility of harmonising and pooling the four studies that were used in this thesis.

Chapter 3

3 HARMONISATION OF FOUR UK STROKE DATASETS: AN APPLICATION OF THE DATASHAPER APPROACH

Study 1 Qualitative harmonisation of four UK stroke datasets: Application of the DataSHaPER approach

3.1 Introduction

As we have seen, Chapter 2 comprised a literature review of examples of data harmonisation studies and of the methods that are commonly used to harmonise and pool data from multiple studies. In the present Chapter, the first strand of research that was carried out in study 1 of this thesis is described. The Data Schema and Harmonisation Platform for Epidemiological Research (DataSHaPER) approach (which was described in some detail in Chapter 2) was applied to each of the four datasets examined for the present thesis, to determine their comparability and potential for harmonisation. It is important to determine the comparability of the studies on all possible dimensions to allow for valid pooled data analysis (Curran and Hussong, 2009). In this Chapter, the between dataset heterogeneity and challenges that could affect pooled data analyses were identified. The statistical methods that could be used to address these challenges were discussed.

This Chapter begins by stating the aims and objectives of Study 1. The methods section describes the datasets that were used and how the first two steps of the DataSHaPER approach were applied to highlight the similarities and differences across studies. Section 3.3 presents the results from the application of the DataSHaPER approach and the descriptive analyses of variables that were common across the datasets. The Chapter concludes by discussing the feasibility of harmonising and pooling the four datasets that were used in the present study.

Aims and objectives

The aims of Study 1 were to apply the DataSHaPER approach to evaluate the potential of harmonising and pooling the four stroke datasets that were used in the present study. It was desirable to harmonise and pool similar datasets to create larger

datasets that could be used to understand the factors associated with disability outcomes after stroke.

The specific objectives of study 1 were:

- To compare the similarities and differences across the four constituent datasets in terms of their sampling, study designs, patient characteristics, and outcome measures
- To create a data schema of relevant variables that were contained in the datasets
- To assess the potential of harmonising relevant variables across datasets
- To conduct a descriptive analyses of the common variables across the four datasets

To identify the barriers that could prevent the pooling of the four datasets

3.2 Method

3.2.1 Data sources

In order to address the aims of the present study, eligible stroke datasets that could be used were identified. The choice of the datasets was influenced by the funding sponsors of this PhD: the NIHR-funded Leeds, York, Bradford (Collaboration for Leadership in Applied Health Research and Care) CLARHC project. Stroke rehabilitation was one of the themes of the CLAHRC project which provided the majority of the funding for this PhD research. The present study had access to data from four UK stroke datasets, the SOS1, SOS2, CIMSS, and Leeds SSNAP. The data from these four datasets provided the opportunity to evaluate the feasibility of harmonising and pooling independent yet similar datasets to create a high(er) quality large(r) database. The four datasets had a similar geographical basis (Yorkshire) but none were identical in terms of their aims. The similarities and differences of the studies provided an ideal opportunity to examine a wide range of challenges to data harmonisation and to explore various statistical methods to address these challenges.

3.2.2 Stroke Outcome Study 1 dataset

The Stroke Outcome Study 1 (SOS1 (House et al., 2001)) was a prospective, population-based observational study of n=448 stroke patients which was conducted between 1995-1999. The aim of the SOS1 study was to evaluate the effect of a problem-solving therapy on depression post-stroke in patients admitted to general

medical and neurology wards. The SOS1 study recruited patients one month after stroke from a stroke population of n=1387 consecutive admissions at Leeds and Bradford hospitals. Informed consent was sought from participants. Patients with severe cognitive impairment or language disorders, and those who were too ill to participate in the study, were excluded. A detailed description of the SOS1 study is provided elsewhere by House et al. (2000) and Dempster et al. (1998).

3.2.3 Stroke Outcome Study 2 dataset

The Stroke Outcomes Study 2 (SOS2 (Hill et al., 2009)) was a prospective cohort of n = 585 stroke patients and was conducted between 2002- 2005. The aim of the SOS2 study was to determine the trajectories of psychological symptoms after stroke and their impact on physical recovery. Patients were recruited in the first few weeks after stroke from three acute and four rehabilitation units in two NHS Acute Hospital Trusts in West Yorkshire. Life-time first or recurrent stroke survivors, aged 18 years or older who were fit to be seen at 2–4 weeks were included in the study. Informed consent was sought from participants. Non-English-speaking patients and those with subarachnoid hemorrhage or transient ischemic attack or severe cognitive impairment, concurrent major illness were excluded. The SOS2 study had full ethical approval from the relevant Local Research Ethics Committees (LREC) in the areas in which the study was conducted. Project reference numbers were as follows: Leeds Teaching Hospitals NHS Trust: St. James's University Hospital LREC Ref No: 01/182; Leeds General Infirmary LREC Ref No: CA02/131; Bradford Hospitals NHS Trust LREC Ref No: 02/06/222. A detailed description of the SOS2 study is provided elsewhere by Hill et al. (2009).

3.2.4 CIMSS dataset

The CIMSS study (Teale, 2011) was a prospective cohort of n=312 patients which was conducted in 2011. The aim of the CIMSS study was to identify key stroke indicators that could be included in a stroke minimum dataset generated from routine care. Participants were recruited from three NHS acute Hospital Trusts in the Yorkshire & Humber area. Patients were eligible for inclusion if they had a primary diagnosis of stroke, but were excluded if participation in the study was clinically inappropriate, for example: patients receiving palliative care or those with subarachnoid haemorrhage. Ethical approval for the CIMSS study was obtained from the Bradford Regional Ethics Committee. Research and Development (R&D)

approvals were obtained individually from each of the three study sites. A detailed description of the CIMSS study is provided elsewhere by Teale (2011).

3.2.5 Leeds Sentinel Stroke National Audit programme (SSNAP) dataset

The Leeds Sentinel Stroke National Audit Programme (SSNAP) is a prospective, longitudinal audit that measures the quality of care provided to stroke patients throughout the whole care pathway up to 6 months post-stroke. It is administered by the Royal College of Physicians Stroke working party. The SSNAP has been collecting data since January 2013 and aims to improve the quality of stroke care by auditing stroke services against evidence-based standards. The SSNAP collects data on stroke care delivered within the first 72 hours following acute stroke. The present study was provided with anonymised data of n=350 patients that had been collected in 2013.

3.2.6 Study Ethics

Ethical approval for the present study was provided by the NRES Committee North East - Sunderland in May 2013, reference number 13\NE\0157. NHS R&D approval for accessing the Leeds SSNAP data was issued by Leeds Teaching Hospital NHS trust, reference number LTHT R&D EP13/10784.

3.2.7 Comparability of dataset characteristics: application of the DataSHaPER approach

Identifying sources of between study-heterogeneity is a critical step in individual patient data analysis (Curran and Hussong, 2009). In the present study, the initial assessment of the feasibility of harmonising and pooling the four constituent datasets involved assessing the comparability of these datasets in terms of: their aims, study designs, sampling, patient characteristics, and the data that were collected. The comparability of the datasets was conducted using the DataSHaPER approach. As described in Chapter 2, the DataSHaPER approach is a systematic and structured approach of harmonising data from different sources. Details of the DataSHaPER approach have already been provided in Chapter 2. This approach was chosen in this present study because it provides a systematic and structured approach to data harmonisation. It is flexible, does not require studies to have collected identical data, but requires the use of sound methodology to ensure the inferential equivalence of information that needs to be harmonised (Fortier et al., 2010). The utility of using the DataSHaPER approach has been demonstrated in epidemiological studies that

combine data from multiple independent sources (Hofer and Piccinin, 2009; Fortier et al., 2011).

3.2.8 Identifying and documenting the set of core variables

In retrospective data harmonisation, the DataSHaPER approach identifies and documents the set of variables that were collected by each of the constituent datasets. To this end, the variables that were collected for each of the different datasets were first documented so that a data schema could be developed. This data schema comprised a list of all the 'core' variables that are required in the pooled database, together with their respective definitions. The choice of what constituted a 'core' variable in the data schema was influenced by the overall aim of the present study, which was to create a large, high quality dataset that could be used to understand factors associated with disability outcomes after stroke. However, because the present study was a retrospective harmonisation study, the development of the data schema also depended on the variables that were collected by the primary studies, unlike in prospective harmonisation where the data schema is developed before data is collected.

3.2.9 Assessing the potential to share each variable between participating datasets

After identifying the variables required for the data schema, the next step was to determine the data that might be validly combined across the datasets. The potential for sharing the variables in the data schema across datasets was evaluated using the DataSHaPER matching approach called "pairing". The "pairing" approach classifies each variable on a three level matching scale: complete matching, partial matching, and impossible. Complete matching is where the meaning of a variable, and the format in which this has been measured, is in the (same) required format across all the datasets. Partial matching is where the meaning and format of the variable would allow the construction of the required variable but with unavoidable information loss. Impossible matching is when there is insufficient information to allow construction of the 'core' variable in the combined database. These definitions were used in the present study to assess the potential of sharing variables across studies.

Table 3.1 illustrates, using selected variables how the pairing process was conducted in the present Chapter. In Table 3.1, age is recorded as complete matching meaning that the format and definition of the variable was provided, as required, in all

the four constituent datasets. NEADL was recorded as impossible matching in SSNAP indicating that it was not possible to generate this variable from the information that was available in the SSNAP dataset. Occupation was recorded as partial matching in the SOS1 dataset indicating that the required variable and format could be generated from existing data therein.

Table 3.1 A demonstration of pairing variables

Variable	SOS1	SOS2	CIMSS	SSNAP
Occupation (Yes/no)	partial	complete	complete	complete
Age	complete	complete	complete	complete
Extended activities of daily living(NEADL)	complete	partial	complete	impossible

3.2.10 Defining data processing algorithms for harmonising variables

The "pairing" exercise identified the common variables that did not require harmonisation and the variables that needed harmonisation. The DataSHaPER approach calls this process the "data harmonisation platform". As described in Chapter 1, data harmonisation seeks to make data from different sources compatible and comparable.

Unlike standardisation, the data harmonisation platform does not impose a single method but seeks to find ways of making the data comparable (Fortier et al., 2010). In the present study, flexible harmonisation was adopted, which does not require studies to have collected the same data as in stringent harmonisation. Stringent harmonisation in this instance would have resulted in restricted data sharing across the datasets because the participating/constituent datasets had different aims and objectives and it was unlikely that the datasets would have collected similar data.

3.2.11 Descriptive analysis of patient characteristics

In order to assess whether the samples from the four different datasets were comparable, patient characteristics and distributions of common variables were compared using descriptive statistics. Categorical data were summarised using

frequencies and percentages. Quantitative data were summarised using means and standard deviations if normally distributed, and using medians and Inter-Quartile Ranges (IQRs) otherwise. Comparisons of categorical data were conducted using Chi-square tests, and continuous data were compared using independent t-tests, ANOVA and Mann Whitney U-tests (or Kruskall Wallis tests where the independent t-test and ANOVA assumptions were violated). A *p* value of 0.05 was considered statistically significant.

3.3 Results

3.3.1 Comparability of dataset characteristics

The characteristics of the datasets that were used in the present study are shown in Table 3.2. The focus of SOS1 and SOS2 were similar; both investigated depressive symptoms after stroke. CIMSS had a different research focus: this study investigated the key stroke indicators that could be included in a stroke minimum dataset in routine care. The SSNAP was different from the other three studies as it was not a research project but collected audit data to evaluate stroke care during the acute phase. The four datasets also comprised data that had been collected at different times: SOS1 between 1995-1999; SOS2 between 2002-2005; CIMMS in 2011; and SSNAP (provided data that had been collected) in 2013. The heterogeneity in the times the studies were conducted was a potential threat to pooled data analysis and was therefore addressed during data analysis using statistical models that account for heterogeneity across datasets/data sources.

Table 3.2 Dataset characteristics: Aims, Sampling, Study design, Inclusion and Exclusion criteria

List of variables	SOS1 n=448	SOS2 n=585	CIMSS data n=312	SSNAP Audit n=350
Aims of study.	To evaluate the effect of a problem-solving therapy on post stroke depression in patients admitted to general medical and neurology wards	To determine the trajectory of psychological symptoms and their impact on physical recovery.	To identify key stroke indicators that should be included in a stroke minimum dataset in routine care.	Audit data aiming to improve the quality of stroke care by auditing stroke services against evidence based standards, and national and local benchmarks.
Study design and recruitment.	Observational study Patients were recruited from the Leeds stroke database.	Observational study Patients were recruited from the Leeds stroke database.	Observational study Patients were recruited from Leeds stroke database.	Audit
Year(s) of study	1995-1999	2002-2005	2011	2013
Type of study	Research	Research	Research	Audit
Centres recruited	448 patients recruited from Leeds and Bradford.	Leeds: 135 (23.1%) Bradford: 138 (23.6%) SJUH: 165 (28.2%) CAH: 96 (16.4%) Sea croft: 49 (8.4%) WGH: 2(0.3%)	Leeds: 125 (40.1%) Bradford: 71 (22.8%) York: 116 (37.2%)	Yorkshire

List of variables	**SOS1** **n=448**	**SOS2** **n=585**	**CIMSS data** **n=312**	**SSNAP Audit** **n=350**
Exclusion criteria	-Subarachnoid haemorrhage -Severe cognitive impairment -Non English speakers	-Subarachnoid haemorrhage -TIA -Severe cognitive impairment -Major illness -Non English speakers	-Subarachnoid haemorrhage -TIA -Severe cognitive impairment -Patients receiving palliative care	No inclusion and exclusion criteria
Method of data collection	Interviews	Interviews	Postal survey	Stroke register audit data captured by the Blues pier system.
Baseline and Follow-up times	Baseline (within 4 weeks), 12 months and 24 months	Baseline (within 2-4 weeks) 9 weeks, 13 weeks, 26 weeks, 52 weeks	Baseline (within 4 weeks) 6 months follow-up.	No follow-up data available at the time data was requested

The designs of the three research studies (SOS1, SOS2, and CIMSS) were comparable. All were prospective observational studies which recruited patients from Yorkshire, and their inclusion and exclusion criteria were also similar: all included patients with a definite diagnosis of stroke and excluded patients with subarachnoid haemorrhage, TIA and severe cognitive impairment. Having similar inclusion and exclusion criteria was advantageous for the harmonisation analyses conducted in the present study since the patient samples in each of these three research datasets were directly comparable. Between-sample heterogeneity due to sampling is a threat to the internal validity of analyses based on pooled individual patient data (Hussong et al., 2013). Hussong et al. (2013) argue that ignoring between-sample heterogeneity during pooled data analysis may result in misidentifying effects as theoretically meaningful when these are actually artefacts caused by the heterogeneity in sampling. Thus pooled data analysis needs to address between-sample heterogeneity.

The modes of data collection were similar for both SOS1 and SOS2; data was collected by face-to-face interviews. In contrast, CIMSS collected data using a postal survey, and the Leeds SSNAP data were captured at patient admission. The response rate for SOS1 and SOS2 were high (>90%), while CIMSS had a much lower response rate (60%; though relatively good for a postal survey). The response rate for the SOS studies was high because the data was collected by interviews.

The baseline assessments of the SOS1, SOS2, and CIMSS were all conducted within a month after stroke, but there was some heterogeneity in patient follow-up. SOS1 collected follow-up data at 12 and 24 months post-stroke, while SOS2 collected follow-up data at 9, 13, 26 (6months) and 52 weeks (12months), and CIMSS made only one follow-up at 6 months post-stroke. And while the intended follow-up time points were recorded as 9, 13, 26 and 52 weeks in the SOS2 the exact follow-up timing at each time point varied between patients. This heterogeneity in follow-up times across studies was a challenge for pooled data analysis. It necessitated the use of statistical methods such as multi-level modelling (see Chapter 8 of this thesis) which accommodate heterogeneity and thereby allow valid statistical inferences to be drawn.

3.3.2 Identifying and documenting the set of core variables collected within datasets

The core variables in the data schema that was produced in the present study are shown in Table 3.3. The data schema included data on: patient demographic and

socio-economic characteristics; stroke severity; treatments; and Patient Reported Outcome Measures (PROMs). It covered some of the dimensions in the International Classification of Function, disability and Health (ICF) core set for stroke (Geyh et al., 2004), which was developed using the World Health Organisation ICF framework (WHO, 2007 (Organization, 2007)). The WHO ICF model is a multi-dimensional concept that includes the physical function, psychological function, personal life situation and social role (Geyh et al., 2004). Stroke outcomes research has been strongly influenced by the WHO ICF model (Mayo et al., 2013).

Table 3.3 The core set of variables in the data schema

Personal factors and socio-economic factors	**Disability domains**	**Stroke Care processes**
Age, gender, marital status, residential status before stroke, education, ethnicity, employment status, independent before stroke, living at home before stroke, length of stay, co-morbidities, occupation and smoking **Environmental factors** Area of residence (Postcode), carer-wellbeing, hospital , and centre	-Physical function: Activities of Daily Living(ADL), Instrumental Activities of Daily Living (IADL) -Social function -Mental function: Anxiety, depression, emotion -Quality of life, patient satisfaction -Cognition -Carer :carer well-being, carer satisfaction **Clinical assessments** Variables collected for clinical assessment of the patient such as stroke type, stroke side, previous stroke, stroke clinical classification, stroke severity, hemianopia, aphasia, urinary incontinence, and aphasia	-Screened for swallowing disorders -Brain scanning within 24 hours of stroke -Commenced aspirin by 48 hours -Assessed by physiotherapy assessment within first 72hours -Assessment by an occupational therapist within 7 days -Assessed by speech therapist -Weighed during admission -Mood assessed by discharge -On anti-thrombotic therapy by discharge -Rehabilitation goals agreed by multi-disciplinary team **Medical interventions** -Home visit performed before discharge, treated in stroke unit -Visuals assessed -Admitted in a stroke unit

3.3.3 Assessing the potential to share each variable between participating datasets

Comparability of demographic and socio-economic factors

Table 3.4 shows the results from the "pairing" of variables across the four datasets. All four datasets contained patient data on age, gender, hospital of admission, previous stroke, stroke type, and functional independence. The definitions of these variables were similar across the four datasets and needed no harmonisation; hence these variables were recorded in Table 3.4 as "complete" matching in all datasets.

There was substantial heterogeneity in the socio-economic data which were collected by the four datasets. The SOS2 dataset had data on: postcode, occupation, education level, and house ownership. These variables were recorded in Table 3.4 as "complete" for the SOS2 dataset following the guidelines of the DataSHaPER pairing approach. Although the SOS1 dataset collected occupation data it used different response categories from those in the SOS2 dataset; thus the occupation variable was recorded in Table 3.4 as "partial" for the SOS1 dataset. The SSNAP and the CIMSS datasets had no socio-economic data. Therefore socio-economic data such as education and employment were recorded as impossible for the SSNAP and CIMSS datasets (Table 3.4).

All four datasets collected data on ethnicity but there was heterogeneity in the response categories used by the different datasets. The SOS1 dataset had three ethnicity response categories, while the SOS2 and CIMMS datasets had more than three. The variable ethnicity was therefore recorded as "partial" for the SOS1 dataset and complete for the other three datasets. Data on patient's residential status before stroke were collected by all datasets. The SOS2, SOS1, and SSNAP used the same response categories for the variable "residential status", but the SOS1 used fewer response categories (though with some overlaps). Hence this variable was recorded "complete" for the SOS1, SOS2 and CIMSS datasets and "partial" for the SOS1 dataset.

Table 3.4 Pairing of variables in the four datasets

Variable	**SOS1 n=448**	**SOS2 n=592**	**CIMSS n=312**	**SSNAP n=350**	**Harmonised Variable**
Age in years	complete	complete	complete	complete	Age
Gender	complete	complete	complete	complete	Sex(male/female)
Marital status	impossible	complete	complete	impossible	-
Residential	partial	complete	complete	complete	-
Education	impossible	complete	impossible	impossible	-
Ethnicity	partial	complete	complete	complete	Ethnicity(white/other)
Previous stroke	complete	complete	complete	complete	Previous stroke(yes/no)
Employment	complete	complete	impossible	impossible	-
Independent	complete	complete	complete	complete	Independent before stroke
Living Alone	complete	complete	complete	impossible	-
Length of stay	impossible	partial	complete	complete	-
Smoker	impossible	complete	impossible	impossible	-
Clinical variables					
Stroke type	complete	complete	complete	complete	Stroke type
Stroke side	impossible	complete	complete	complete	-
Clinical class	impossible	impossible	complete	complete	-
Aphasia	impossible	complete	complete	complete	-
Urine	complete	complete	complete	impossible	-
Hemianopia	impossible	complete	complete	complete	-

Variable	**SOS1 n=448**	**SOS2 n=592**	**CIMSS n=312**	**SSNAP n=350**	**Harmonised Variable**
Environment					
Hospital	complete	complete	complete	complete	Hospital
Centre	complete	complete	complete	complete	centre
Carer	complete	impossible	complete	impossible	-
Carer strain	complete	impossible	complete	impossible	-
Carer satisfaction	complete	impossible	impossible	impossible	-
Housing tenure	impossible	complete	impossible	impossible	-
Anti-depressant therapies	complete	complete	impossible	impossible	-
Co-morbidities					
Diabetes	impossible	impossible	impossible	complete	-
Myocardial Infarction	impossible	impossible	impossible		-
Hypertension	impossible	impossible	impossible	complete	-
Dukes Severity illness scale	impossible	complete	impossible	complete	-
Atrial Fibrillation	impossible	impossible	impossible	complete	-
Congestive heart failure	impossible	impossible	impossible	complete	-

Variable	SOS1 n=448	SOS2 n=592	CIMSS n=312	SSNAP n=350	Harmonised Variable
PROMS					
Cognitive	complete	complete	impossible	impossible	-
ADL	complete	complete	complete	partial	-
IADL	complete	partial	complete	impossible	-
Psychological distress	complete	complete	partial	impossible	-
Patient satisfaction	complete	impossible	impossible	impossible	-
Carer data	complete	impossible	complete	impossible	-
Quality of life	impossible	complete	complete	impossible	-

Key

Complete: meaning and format of the question in studies is as require;
Partial: can construct the required variable but with unavoidable loss of Information;
Impossible: No information or insufficient information in the data to allow Construction of the variable;
-: Denote missing data

Comparability of Clinical variables and care processes

The SOS2, CIMSS, and Leeds SSNAP datasets all had data on: stroke side, stroke clinical classification, aphasia, and hemianopia but these data were not available in the SOS1 dataset and were therefore recorded as "impossible" for the SOS1 dataset (Table 3.4). The SSNAP and CIMSS datasets had similar information on care processes and these data were missing in the SOS2 and SOS1 datasets. The SOS studies collected data on anti-depressant use but these data were missing in the CIMSS and SSNAP datasets. Co-morbidities data were only available in the SSNAP audit data.

Comparability of Patient reported Outcome measures

The pairing exercise showed that there were similarities and differences in the outcome measures that were collected by the different studies (Table 3.4). Table 3.5

summaries the Patient Reported Outcome Measures (PROMs) data in the four datasets.

Table 3.5 Patient Reported Outcome Measures collected by the four datasets

	SOS1	SOS2	CIMMS	SSNAP
Cognitive impairment	MMSE	MMSE	-	NIHSS
ADL **IADL**	BI NEADL FAI	BI FAI, FIM, RMI	BI NEADL SIPSO	pre Rankin score -
Quality of life	-	SF-36	EQ-5D	-
Psychological function	GHQ-28 PSE	GHQ-28 PSE	GHQ-12 -	- -
Patient satisfaction	PSAT	-	-	-
Carer General well being	GHQ-28	-	GHQ-12	-

Cognitive impairment

There was heterogeneity in the cognitive impairment measurement scales which were used across datasets. The two SOS datasets assessed cognitive impairment using the Mini Mental State Examination (MMSE; Folstein et al. (1975)), while the Leeds SSNAP audit assessed neurological deficits at patient admission using the National Institute Health Stroke Scale (NIHSS; (Dunning, 2011)), and there were no baseline cognitive assessments in the CIMSS dataset. (Details of the MMSE, NIHSS and BI have already been provided in Chapter 1).

Activities of daily living (ADL) and Instrumental Activities of Daily Living (IADL)

There were similarities across datasets in the measurement scales used to assess basic ADL. The CIMSS, SOS1, and SOS2 datasets all collected data on independence in basic ADL using the Barthel index (BI), while these data were missing in the SSNAP dataset. Meanwhile, there was considerable heterogeneity across datasets in the measurement scales that were used to assess IADL. The SOS1 and CIMSS studies assessed IADL using the Nottingham Extended Activities of Daily Living (NEADL) tool while the SOS2 used the Functional Independence Measure (FIM), the Frenchay Activities Index (FAI), and the Subjective Index for Physical and Social Outcomes (SIPSO; (Kersten et al., 2010)). Though SOS1 and CIMSS both used the NEADL to

assess extended activities of daily living, the two studies used different versions of the measurement scale. The SOS1 study used the 21-item version of the NEADL, while the CIMMS study used the original NEADL with all 22 items. A description of the NEADL, FAI, BI and FIM has already been provided in Chapter 1. The Leeds SSNAP data that were available for analysis in the present study had no follow-up outcome data. When these were requested for the SSNAP data, it was discovered that the Leeds teaching hospital trust had just started using an electronic system (the Bluespier) to capture the SSNAP data; hence outcomes data were not available at that time. Due to time restrictions on the present study, only incomplete SSNAP data were available at the time the thesis was finalised.

Psychological function

The two SOS studies assessed psychological distress using the GHQ-28 questionnaire (Goldberg and Hillier, 1979), while the CIMSS study used the GHQ-12 and the SSNAP did not collect any patient psychological data.

Quality of life and patient satisfaction

There was heterogeneity in the outcome measures which were used to assess patient quality of life across the datasets. The CIMSS study assessed quality of life using EuroQol-5 (EQ-5D; (Group, 1990)), while the SOS2 used the Short Form 36 (SF-36) questionnaire (Ware Jr and Sherbourne, 1992), and the SOS1 did not collect data on patient quality of life, but instead collected data on patient satisfaction. The SF-36 questionnaire has 36 items which measures eight 8-scale profiles of functional health and well-being. The Leeds SSNAP did not collect any quality of life data.

Carer well being

The SOS1 and CIMSS datasets collected data on carer well-being using the GHQ-12 measure but these data were missing in the SOS2 and Leeds SSNAP datasets. Hence carer well-being data were not considered in the present study.

3.3.4 Descriptive analysis

Descriptive statistics were produced to assess the comparability of the patient characteristics cross datasets. As highlighted earlier on, differences in patient characteristics are a threat to pooled data analysis hence the need for comparable patient samples across datasets to enable valid data pooling. Table 3.6 shows the descriptive statistics that were produced for these data in the present study. The

majority of patients (>80%) in all four datasets were of 'White British' ethnicity. The patients in the SOS1 and SOS2 datasets were comparable in terms of age, there being no statistically significant differences between the mean age for the SOS1 and SOS2 datasets (Table 3.6). The mean age of the patients in the CIMSS and Leeds SSNAP showed that the patients in these two datasets were slightly older compared to patients in the SOS1 and SOS2 datasets ($p<0.001$). However these mean age differences may not be clinically important as the age distributions are clinically similar. The proportion of females was also comparable across the SOS datasets and SSNAP, while the CIMSS dataset had a slightly higher proportion of females (50.6%) than the other datasets. The proportion of patients who had a previous stroke was comparable in SOS1, SOS2, and SSNAP but lower in the CIMSS study ($p<0.001$; Table 3.6). The SOS2 and CIMSS datasets had higher proportions of patients with urinary incontinence compared to patients in the SOS1 dataset ($p<0.001$; Table 3.6). The SOS2 dataset also had a higher proportion of missing data on stroke type, while the Leeds SSNAP had a higher proportion of missing data on previous stroke. Comparison of baseline BI across the datasets showed that SOS2 had fitter patients compared to SOS1 and CIMSS ($p<0.001$; Table 3.6). In the present study, the Leeds SSNAP data were considered to be more representative of the stroke population as these were generated as registry data with unselected stroke patients unlike the other data generated in dedicated research studies.

Table 3.6 Characteristics of the samples

	SOS1 n=448	SOS2 n=585	CIMSS n=312	SSNAP n=341	*p* -value
Mean age (SD)	70.75	70.34	72.63	74.04	<0.001
(Age range)	(11.61)	(11.89)	(12.59)	(13.75)	
	(18-94)	(22-97)	(31-95)	(19-100)	
Female n (%)	207 (46.2%)	253 (43.2%)	158 (50.6%)	170 (48.7%)	0.145
Ethnicity					
White n (%)	426 (95.1%)	579 (99%)	298 (95.5%)	300 (89.2%)	
Previous stroke	94 (21%)	129 (22.1%)	48 (15.4%)	82 (24%)	< 0.001
Living Status					
Alone	175 (39.1%)	193 (33%)	115 (38.8%)	-	0.112
Co-habits		366 (62.6%)	182 (60.3%)		
Nurs/resid/shel		23 (3.9%)	5 (1.7)		
Urinary incontinence	30 (6.7%)	107 (18.6%)	64 (20.6%)	-	<0.001
Baseline BI (median, range)	15 (0-20)	18 (1-20)	14 (0-20)	-	<0.001
MMSE					
Median(range)	26 (18-30)	27 (14-30)	-	-	<0.001

Continuous data were expressed as mean and SD in parentheses; categorical data were expressed as number of patients with % in parentheses; '–'denote missing data; Nur/resid/shel relates to: Nursing home/residence/shelter, respectively.

3.3.5 Rationale for pooling data across datasets

The aim of pooling data in the present study was to create a larger, high quality dataset that could be used to better quantify the factors associated with patient disability outcomes after stroke. It was desirable to include, in the pooled datasets, as many of the disability domains recommended in the WHO ICF model (WHO, 2001) as possible. The choice of which datasets to pool was guided primarily by the intended aims of the subsequent analyses and by the availability of the variables required across each of the participating datasets. Pooling all four datasets (SOS1, SOS2, CIMSS, and SSNAP) would have resulted in a dataset of n = 1686 patients with 8 variables: age, gender, ethnicity, stroke type, previous stroke and independence before stroke, hospital, and center (see Table 3.4). While pooling all four datasets would have resulted in a suitably large(r) sample, any variables that could not be harmonised

would have been lost if only common or harmonised data were considered. Thus including variables that were considered 'impossible' in the "pairing" exercise would lead to missing data in the harmonised/pooled dataset. For example, the Leeds SSNAP dataset had no patient follow-up data, thus it was dropped from the present study in order to avoid losing a significant number of important variables from the pooled dataset. In order to optimise both the sample size and number of variables in the pooled dataset, only datasets with comparable data were harmonised and pooled.

3.3.6 Harmonising the SOS datasets

Hussong et al. (2013) describes data harmonisation as the procedure of placing variables on the same scale in order to permit pooling of data from multiple independent sources. In this section the data harmonisation procedures that were conducted for pooling the SOS datasets are described. The pairing exercise conducted in the present study showed that the two SOS datasets were broadly comparable in terms of their study characteristics and the data they collected. The descriptive analyses also showed that the two datasets had similar samples in terms of patient age, gender, ethnicity, the proportion of patients with previous stroke, and the proportion of patients living alone before stroke. These similarities in study and patient characteristics provided the rationale for pooling the two SOS datasets and conducting a pooled data analysis in Chapters 7 and 8 on these once pooled.

A harmonised, pooled dataset containing n=1,033 patients and n=10 variables was produced by combining the two SOS datasets. The variables in the harmonised dataset were: age (continuous), gender (male/female), ethnicity (white/other), previous stroke (yes/no), living status before stroke (alone, not alone), stroke type (ischemic/hemorrhagic), urinary incontinence (yes/no), GHQ-28, BI, and MMSE. Variables such as age, gender, stroke type, urinary incontinence, GHQ-28, BI, and MMSE that were common across the two datasets did not require any harmonisation, and were directly pooled. While the common data on outcome measures did not require any harmonisation, pooling common outcome measures data from different sources raised statistical issues of measurement invariance; hence the measurement invariance properties of the GHQ-28 were investigated in Chapter 4 of this thesis. Item data were missing for the BI and MMSE questionnaires thus it was impossible to determine measurement invariance for these measures.

Variables such as ethnicity, residential place before stroke, and living status before stroke, though common across the two datasets, required harmonisation before combining the datasets as a result of the different measurement categories used. For

example, the response categories for the “ethnicity” variable were different across the SOS datasets. The SOS1 dataset had three response categories, while the SOS2 dataset had more than three such categories. The category with more than 80% of patients in both datasets was the ethnic group “White British”. For this reason a new, harmonised ethnic group variable was created by recoding data into two response categories (White British vs. other). However this resulted in unavoidable information loss for other (non-White British) ethnic groups.

Similarly the two SOS datasets used different response categories for the variable “residential status”. Harmonisation of the residential status variable was achieved by creating a new variable called “Living alone before stroke (yes/no)” but, again, information on other response categories were lost from the SOS2 dataset. Other data that were lost due to combining data from the two SOS datasets included data on: marital status, stroke classification, aphasia, hemianopia, and socio-economic data that were not collected by the SOS1 study.

At the same time, in the SOS1 follow-up time was recorded in months while the SOS2 dataset recorded follow-up time in weeks. This difference was harmonised by generating follow up-time in months for both datasets. Furthermore, the heterogeneity in follow-up times was problematic but was addressed by using multi-level modelling. After harmonising the SOS datasets, the pooled SOS database was used to investigate factors associated with anxiety post-stroke in Chapter 8.

3.3.7 Harmonising the SOS1, SOS2 and CIMSS datasets

Harmonising and pooling the three datasets from each of the research studies (SOS1, SOS2, and CIMSS) would have resulted in a dataset of n =1345 patients with a total of 11 covariates (age, gender, ethnicity, hospital, center, living at home before stroke , living alone before stroke , urinary incontinence, independent before stroke, previous stroke, and baseline ADL). Variables that were common across theses datasets (such as urinary incontinence, age, and gender) would not have required harmonisation. Pooling IADL data across all three datasets raised statistical issues of measurement comparability because each of these datasets used different measurement instruments to assess this construct. As highlighted earlier, the SOS1 assessed IADL using the FAI and NEADL measures, while CIMSS used the NEADL and SIPSO, and SOS2 used the FAI and FIM. Similarly the two SOS datasets assessed

psychological distress using the GHQ-28, while the CIMSS dataset used the GHQ-12 questionnaire.

In the present study, harmonisation of the measurement scales could have been achieved by using simple methods described in Chapter 2 such as categorising the data or by standardising using z-scores. For example, thresholds of GHQ-28 total scores >4 and GHQ-12 total scores > 3 indicate presence of psychological distress (for the binary scoring system) and these could have been used to harmonise the GHQ-12 and GHQ-28 measures. However categorisation of data was not preferred because it would have resulted in substantive data loss on intermediate states. Likewise, Curran and Hussong (2009) have argued that the use of z scores is an even weaker way of harmonising patient reported outcome measures because it does not take into consideration the differences in the distributions of the data from the different cohorts, hence this approach was not preferred in the present study. For this reason, an attempt was made in the present study to develop mapping algorithms that might be used to harmonise measures of IADL used in stroke rehabilitation research. Details of this harmonisation work are presented in Chapters 5 and 6 of this thesis.

3.4 Chapter summary

In Study 1, an initial evaluation of the feasibility of pooling the SOS1, SOS2, CIMSS, and Leeds SSNAP datasets was conducted using the DataSHaPER Approach. Comparison of the four datasets showed that pooling all of the datasets would have resulted in a trade-off between increasing sample size and the loss of important variables and/or intermediary response categories. In this regard, the SSNAP dataset did not fit with the efforts to harmonise its data with those from the research studies. But even so, the SSNAP offered a comparison of age and sex distribution of the research datasets in the later Chapters of this thesis.

The next Chapter presents the second strand of research that was conducted in Study 2 of this thesis to investigate the measurement invariance analysis of the GHQ-28 questionnaire that was conducted prior to combining the GHQ-28 scores from the SOS datasets.

Chapter 4

4 MEASUREMENT INVARIANCE ANALYSIS OF GHQ-28 DATA FROM DIFFERENT DATASETS: APPLICATION OF MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS

Study 2: Measurement invariance of the GHQ-28 measure in the SOS1 and SOS2 datasets: Application of the Multi-Group Factor Analysis

The following publication has arisen from the analysis and results in this Chapter:

3. Munyombwe, T., Hill, K.M., West, R.M. (2015). Testing measurement invariance of the GHQ-28 in stroke patients, *Quality of Life Research*, 24(8), pp. 1823-1827

As the first author Theresa Munyombwe carried out all the statistical analyses and prepared the first draft of the manuscript. The other authors provided feedback on the statistical analyses and proof read drafts of the manuscript.

4.1 Introduction

Chapter 3 reported the investigation that was conducted in Study 1 to determine the feasibility of harmonising data from the four datasets that were used in this thesis using the DataSHaPER approach. The qualitative harmonisation conducted in Study 1 which was reported in Chapter 3 showed that there was some measurement scales which were common across the datasets to which harmonisation is being applied in this thesis. Although these measurement scales were common across datasets, it was nonetheless important to establish measurement invariance of all such measures before combining the data so as to ensure valid group mean comparisons (Hussong et al., 2013; Curran and Hussong, 2009). This is because an outcome measure is

‘measurement invariant’ across groups if its constituent items/criteria assess the same construct across the groups (Curran and Hussong, 2009).

The second strand of the research that was conducted in this thesis was study 2, which investigated the measurement invariance properties of the GHQ-28, prior to pooling the SOS datasets in Chapter 8 of the thesis. The measurement invariance properties of the GHQ-28 measure with respect to SOS1 and SOS2 form the focus of the present Chapter. Multi-Group Confirmatory Factor Analysis (MG-CFA) was used in Study 2 to investigate the measurement invariance of the GHQ-28 measure along with some testing methodologies that are applied for the very first time in this particular disease area. The GHQ-28 measure has been tested in many different populations but have not been validated comprehensively in stroke populations (Salter et al., 2007). Thus the measurement invariance analysis of the GHQ-28 questionnaire conducted in Study 2 makes a novel contribution to the literature on the psychometric properties of a scale (the GHQ-28 measurement scale) that is widely used in stroke rehabilitation research. The methods that were used to investigate measurement invariance are described in section 4.2, the results of these analyses are reported in section 4.3, and the Chapter concludes by discussing their findings.

4.2 Methods

4.2.1 Datasets

In Study 2, the baseline data from SOS1 (n=448) and SOS2 (n=585) datasets were used to investigate the measurement invariance properties of the GHQ-28. The details of these datasets have already been described in Chapter 3.

4.2.2 Measure

The details of the GHQ-28 measure have already been described in Chapter 1 of the thesis. The SOS1 and SOS2 used different scoring methods for the GHQ-28 measure. The SOS2 used the Likert type response (0, 1, 2, and 3), while the SOS1 used the bimodal symptom present scoring (0, 0, 1, 1) for the following item responses: ‘0-Not at all’, ‘0-No more than usual’, ‘1-rather more than usual’, and ‘1-Much more than usual’. Study 2 harmonised the GHQ-28 item responses in the SOS2 by re-scoring these as follows: ‘0-not at all’, ‘0-no more than usual’, ‘1-rather more than usual’, and ‘1-much more than usual’. Hence the measurement invariance analysis of the GHQ-28 that was conducted in Study 2 was based on the 0,0,1,1 scoring system. Finally, because

there were no data for the GHQ-28 item 1 in the SOS1 dataset, therefore the analyses conducted in the present Chapter excluded the GHQ-28 item 1 data.

4.2.3 Measurement invariance analyses

Multi-group Confirmatory Factor Analysis (MG-CFA) was used to investigate measurement invariance of the GHQ-28 measure in the SOS1 and SOS2 datasets. MG-CFA is a powerful approach for testing measurement invariance across groups (Steenkamp and Baumgartner, 1998), because many specific aspects of measurement invariance are readily testable within a MG-CFA framework (Vandenberg and Lance, 2000). It provides a framework for comparing a set of hierarchical measurement models: configural invariance, weak invariance, and strong invariance (Meredith, 1993). Details of the MG-CFA have already been described in Chapter 2 of this thesis. As highlighted in Chapter 2, MG-CFA is an extension of the factor analytic model that accommodates multiple groups. The MG-CFA framework was chosen in the present study because it provided an elegant approach for examining the different levels of invariance within a single procedure rather than using many separate procedures (Cheung and Rensvold, 2002). Guidelines for testing measurement invariance were provided by Vandenberg and Lance (2000), and these guidelines were followed in the present Chapter. Details of these guidelines have already been described in Chapter 2.

The measurement invariance analyses conducted in the present study, started by verifying the four factor structure of the GHQ-28 questionnaire (as proposed by the originators of the scale; (Kihç et al., 1997; Gibbons et al., 2004)). The verification of the four factor structure of the GHQ-28 was conducted using single group Confirmatory Factor Analysis (CFA) in each of the constituent datasets (a description of single group CFA has already been provided in Chapter 2). The goodness-of-fit of the four factor structure was assessed using the Root Mean Square Error of Approximation (RMSEA; (Gelman et al., 1998)), Tucker-Lewis Index (TLI; Bentler, 1980)), and Comparative Fit Index (CFI; (Bentler, 1980)). Acceptable fit was indicated by CFI >0.95, TLI >0.95 (Tucker and Lewis, 1973; Bentler, 1980), and RMSEA < 0.06 (Browne et al., 1993).

After verifying the four factor structure of the GHQ-28 in each of the constituent datasets, three levels of measurement invariance of the GHQ-28 measure were tested with respect to ‘dataset’: configural, factor loading invariance, and scalar invariance.

4.2.4 Configural invariance

The first step of measurement invariance analysis involved testing for configural invariance (also known as 'equal factorial form'). Configural invariance requires the same number of factors and the same pattern of item factor loadings across groups. To test for configural invariance, an unrestricted baseline four factor model was fitted across the studies. The configural invariance model served as the baseline model (i.e. the starting point) for subsequent more rigorous tests of measurement invariance.

4.2.5 Factor loading invariance or metric invariance

When configural invariance of the GHQ-28 measure was evident, factor loading invariance (also known as 'metric invariance', (Horn et al., 1983)) or weak invariance, (Meredith, 1993) was tested for. Factor loadings are the strengths of the association between each item and the corresponding underlying latent factor; and factor loading invariance tests the hypothesis that there is equal factor loading across groups. Factor loading invariance of the GHQ-28 measure was tested by constraining factor loadings to be equal across the SOS1 and SOS2 datasets whilst allowing intercepts to vary across datasets, where 'intercept' corresponds to the zero value of the underlying latent construct. The fit of the configural invariance model was compared with the fit of the nested factor loading invariance model using a Chi-square difference test (Bollen, 1989). Details of the Chi-square difference test have already been described in Chapter 2 of the thesis). A non-significant Chi-square difference test is indicative of measurement invariance. However, the Chi-squared difference test is sample size depended and produces significant results when the sample size is large (Cheung and Rensvold, 2002). Therefore, in Study 2, the change in Comparative Fit Index (CFI) was also used to evaluate measurement invariance; Cheung and Rensvold (2002) recommending that a change in CFI of 0.01 or less is suggestive of measurement invariance.

4.2.6 Scalar invariance or intercept invariance

Scalar invariance (Steenkamp and Baumgartner, 1998; Meredith, 1993)also known as 'intercepts invariance') was tested after establishing both configural and factor loading invariance. This form of invariance is only tested on items that show both configural and factor invariance. It is a stricter form of invariance compared to factor loading invariance and suggests strong measurement invariance (Meredith, 1993). Evidence of scalar invariance implies that the scale scores from different

groups have the same unit of measurement (factor loading) and the same origin (intercept), so that the factor means of the underlying construct can be compared across groups (Chen et al., 2005).

In Study 2, scalar invariance was tested by further constraining the factor invariance model to have equal item thresholds across each of the datasets. The fit of the metric invariance model was then compared with that of the scalar invariance model using the Chi-square difference test. A non-significant Chi-square difference test together with a change in CFI of 0.01 or lower provided evidence that scalar invariance was present. The other stricter forms of invariance, such as strict factorial invariance and structural invariance described in Chapter 2 were not tested for in the present Chapter because establishing scalar invariance was deemed sufficient for making meaningful group comparisons. Indeed, Wang and Wang (2012) suggested that testing for the other stricter forms of invariance is of no interest if, in practice, the aim is to make group mean comparisons.

4.2.7 Model Estimation

In Study 2, the CFA and MG-CFA analytic models were fitted using Mplus version 7 (Muthén and Muthén, 2012). These models were estimated using a robust Weighted Least Square Mean and Variance (WLSMV; (Muthén, 2004)). The WLSMV is used to model categorical outcome measures or a combination of binary, ordered, and continuous measures (Wang and Wang, 2012). In Study 2, the GHQ-28 item data were considered categorical hence the models were estimated using WLSMV. Maximum Likelihood Estimation (MLE) was not used because it assumes that the observed data follows a multivariate normal distribution and uses Pearson correlations for the relationships amongst items. However, the Pearson correlations underestimate the true relationships when the variables are categorical thus the WLSMV is preferred. Simulation studies have shown that MLE underestimates the size of factor loadings for variables with two or three response categories, and that WLSMV performs better than MLE under these circumstances (Beauducel and Herzberg, 2006). The Mplus syntax for the measurement invariance analyses that was conducted in this Chapter is shown in Appendix C.

4.3 Results

4.3.1 Confirmatory factor analysis of the GHQ-28 measure

The results of the CFA of the GHQ-28 are shown in Table 4.1. There was evidence of good fit for the four factor structure of data from both the SOS1 and SOS2 datasets based on a CFI >0.95, a TLI >0.95, and a RMSEA <0.06. The item factor loadings on each of the four subscales of the GHQ-28 were similar across data from the SOS datasets (Tables 4.2).

Table 4.1 Goodness of fit indices from the confirmatory factor analysis of GHQ-28

Items	SOS2 (n=585)	SOS1 (n=448)
χ^2	740.79, p <0.001	490.71, p<0.001
CFI	0.958	0.969
TLI	0.954	0.965
RMSEA	(0.048(0.04-0.05)	0.035(0.03-0.04)

Table 4.2 Standardised factor loadings from the confirmatory factor analysis of GHQ-28

Domains	SOS2	SOS1	Domains	SOS2	SOS1
Somatic (A)			**Anxiety(B)**		
Item1	-	-	Item8	0.750	0.851
Item2	0.811	0.838	Item9	0.593	0.767
Item3	0.895	0.890	Item10	0.811	0.847
Item4	0.859	0.710	Item11	0.684	0.731
Item5	0.746	0.651	Item12	0.743	0.743
Item6	0.672	0.735	Item13	0.865	0.872
Item7	0.713	0.650	Item14	0.895	0.864
Social function (C)			**Depression(D)**		
Item15	0.875	0.833	Item22	0.720	0.705
Item16	0.944	0.926	Item23	0.699	0.787
Item17	0.926	0.865	Item24	0.922	0.848
Item18	0.929	0.813	Item25	0.870	0.881
Item19	0.903	0.886	Item26	0.838	0.806
Item20	0.914	0.895	Item27	0.781	0.730
Item21	0.916	0.815	Item28	0.760	0.801

4.3.2 Measurement Invariance of the four factor model for the GHQ-28

The results of the measurement invariance analyses of the GHQ-28 measure are shown in Table 4.3. There was evidence for configural invariance of the GHQ-28 measure with respect to dataset based on the model fit indices, TFI and CFI > 0.95, and RMSEA < 0.06. Comparisons of the Configural invariance model with the factor loading invariance model produced a non-significant Chi-square difference test ($\Delta \chi^2$ =14.9, Δdf=23, p=0.898), and a change of CFI (ΔCFI) of 0.008 which was lower than 0.01; thereby supporting factor loading invariance for the GHQ-28 measure with respect to data from the two SOS datasets. Comparisons of the factor loading invariance model with the scalar invariance model also produced a non-significant Chi-square difference test ($\Delta \chi^2$ = 39.4, Δdf= 27, p=0.06), and a change of CFI of zero, supporting scalar invariance of the GHQ-28 with respect to the two SOS datasets.

Table 4.3 Results of testing measurement invariance of the GHQ-28 across the SOS1 and SOS2 using MG-CFA, overall fit indices

Measurement invariance model	χ^2, df	CFI	TLI	RMSEA
Configural	1270.42 df = 640	0.960	0.956	0.044(0.04-0.047)
Metric	1158.24 df =663	0.968	0.967	0.038(0.034-0.042)
Scalar	1189.398 df=690	0.968	0.968	0.038(0.034-0.041)

CFI: Comparative Fit Index, TLI: Tucker–Lewis Index, RMSEA: Root Mean Square Error of Approximation.

4.4 Discussion

In Study 2, the measurement invariance properties of the GHQ-28 measurement scale with respect to the SOS1 and SOS2 datasets were investigated, as a preliminary step towards pooling the GHQ-28 scores for an integrative data analysis in Chapter 8 of the thesis. The CFA conducted in both SOS1 and SOS2 datasets supported the four factor structure of the GHQ-28 measure that was proposed by the originators of the scale (Goldberg and Hillier, 1979). The four factor structure of the GHQ-28 found in

this present study is also consistent with previous studies by Kihç et al. (1997); (Gibbons et al., 2004; Werneke et al., 2000). The SOS1 and SOS2 recruited patients at different times, and obtaining the same factor structure in each of their two cohorts suggested that the factor structure of the GHQ-28 was stable over time and context.

The Multi-Group Confirmatory Factor Analyses conducted in Study 2 of this thesis, supported configural, metric, and scalar invariance of the GHQ-28 questionnaire with respect to the SOS1 and SOS2 datasets. Establishing configural invariance indicated that the GHQ-28 questionnaire measured the same constructs across both SOS1 and SOS2. Establishing both configural and metric invariance of the GHQ-28 implied that: the scale measured the same factor structure across both SOS datasets; and the structural relationships of the items and latent constructs were also similar across the two studies. Establishing scalar invariance with respect to SOS1 and SOS2 implied that the item thresholds were the same across studies. Gregorich (2006) have argued that establishing scalar invariance implies that group differences in observed means will be directly related to group differences in the factor means, and these group differences in observed means are considered to be unbiased estimates of group differences in corresponding factor means. The findings of Study 2, supported scalar invariance of the GHQ-28 measure with respect to the SOS datasets. Thus, following Gregorich (2006)'s argument, comparisons of the GHQ-28 summed scores can be validly conducted across the SOS datasets. In other words, the observed GHQ-28 score mean differences across the studies will be unbiased estimates of the underlying factor means.

Only one previous study was found in the literature that had investigated the measurement invariance properties of the GHQ-28 measure. This study by Prady et al. (2013) based on the Born in Bradford cohort, concluded that the GHQ-28 measurement scale worked differently in women from different ethnic groups. However, Prady et al.'s (2013) findings might not be consistent with those from the present study because the majority of patients used in the SOS1 and SOS2 were of the same, 'White British' ethnicity. The violation of the measurement invariance assumption for the GHQ-28 in the Born in Bradford cohort might have therefore been due solely to cultural differences of the participants whose data were analysed.

4.4.1 Limitations

The analyses conducted in Study 2 have a number of potential limitations that warrant discussion. The measurement invariance analyses was conducted in a selected cohort that excluded people with severe strokes, language impairments and other ethnic groups were not adequately represented hence the findings may not be generalisable if the scale has differential function in these patient groups. People with severe stroke are often self-selected out of research studies because they are/feel too unwell to give consent, or they may be unable to complete questionnaires without the help of a carer or researcher. The selection bias associated with the use of fitter patients was not considered a major limitation for Study 2 as the main aim was to demonstrate the statistical validity of combining existing stroke datasets. The majority of existing stroke datasets will have high proportions of severe stroke patients excluded. The self-selection of fitter/healthier/more competent patients in each of the studies examined here (i.e. SOS1 and SOS2) may, of course, explain why measurement invariance was successful in this case.

Another limitation was that only one method was used to evaluate measurement invariance in Study 2. There are several methods that can be used to test for measurement invariance (such as items response theory models), and these might yield different results. More research is needed to clarify the impact of using alternative statistical methods to assess measurement invariance. Furthermore data on GHQ-28 item 1 were missing from the SOS1 dataset so that the analyses conducted in Study 2 excluded GHQ-28 item 1. Hence there is need to test the measurement invariance properties of GHQ-28 item 1 data in future research.

4.5 Conclusion

In conclusion, this Chapter involved the statistical testing of whether or not there was a good statistical basis for combining GHQ-28 scores from two seemingly similar stroke datasets. The measurement invariance (configural, metric, and scalar) for the GHQ-28 questionnaire was established in data from two stroke cohorts. These findings contribute new knowledge about the psychometric properties of the GHQ-28 measure in stroke survivors. The analyses conducted in Study 2 provided support for conducting integrative data analyses of the GHQ-28 scores from the SOS1 and SOS2 pooled database. This integrative data analyses will be presented in Chapter 8.

Key message from this Chapter

- The GHQ-28 measurement scale showed measurement invariant properties across two stroke cohorts.
- The utility of using multi-group confirmatory factor analysis to establish measurement invariance was demonstrated.

The next Chapter describes the third strand of research that was conducted in Study 3a of this thesis to develop mapping algorithms that can be used to harmonise the FAI and NEADL measures.

Chapter 5

5 STATISTICAL HARMONISATION OF FRENCHAY ACTIVITIES INDEX AND NEADL: APPLICATION OF ITEM RESPONSE THEORY MODELS AND REGRESSION BASED MODELS

Study 3a: Statistical harmonisation of Frenchay Activities Index and Nottingham Activities of Daily Living: Application of item response theory models and regression analysis

5.1 Introduction

Chapter 4 reported the measurement invariance analyses of the GHQ-28 questionnaire that was conducted in Study 2, as a pre-requisite for pooling the GHQ-28 scores from the SOS1 and SOS2 datasets. The utility of using MG-CFA for establishing measurement invariance was demonstrated in Study 2 of thesis. The third strand of research that was conducted in this thesis investigated in Studies 3a and 3b, the potential of harmonising different measurement scales that assess similar constructs. The pairing exercise conducted in Study 1, reported in Chapter 3, showed that measurement comparability was a threat to pooling Patient Reported Outcome Measures (PROMs) from all three datasets (SOS1, SOS2, and CIMSS). The problem of measurement comparability occurs when studies use different PROMs to assess similar constructs. For example the SOS1 assessed Instrumental Activities of Daily Living (IADL) using the Nottingham Extended Activities of Daily Living (NEADL) and Frenchay Activities Index (FAI); SOS2 used FAI and the Functional Independence Measure (FIM), while the CIMSS study used the NEADL. Due to the multiple measurement scales used in stroke rehabilitation research, measurement comparability is a common problem in stroke rehabilitation research.

When studies use different measurement scales to assess similar constructs, these measures can be harmonised in order to produce comparable assessments. Data harmonisation places variables or measurement scales on the same scale in order to permit pooling of data from different sources (Hussong et al., 2013). There are various

statistical methods that can be used to harmonise measurement scales and these include: algorithm harmonisation, standardisation, regression analysis, and item response theory models (Griffith et al., 2015). Details of these methods were provided in Chapter 2.

Study 3a of this thesis explored the utility of using regression-based mapping and Item Response Theory (IRT) linking to harmonise the FAI and NEADL measures. Mapping involves the development of algorithms that can be used to predict one measure from the other. As highlighted earlier in Chapter 2, the harmonisation platform of the DataSHaPER approach recommends the construction and application of processing algorithms to generate the required variable in appropriate forms. The harmonisation analyses conducted in Study 3a started by investigating whether the FAI and NEADL scales measure the same constructs using exploratory factor analysis. The exploratory factor analysis of the combined items from the FAI and NEADL measures conducted in Study 3a, contributed to the literature on psychometric properties of these two measurement scales that are commonly used in stroke rehabilitation research. Furthermore the mapping algorithms that were developed in Study 3a could be used in future data harmonisation studies or by other stroke researchers or clinicians to follow patients longitudinally across the continuum of care.

In this Chapter the analyses that were conducted in Study 3a are reported. Chapter 5 begins by stating the aims of study 3a. The methods that were used to harmonise the measurement scales are presented in sections 5.3. The results of the factor analysis of the combined FAI and NEADL items, the mapping analysis of the FAI and NEADL using regression-based methods, and linking using IRT methods are reported in section 5.4. The chapter concludes by discussing the findings of the analyses that were conducted in Study 3a.

5.2 Aims

The aims of Study 3a were twofold:

(1) To develop mapping algorithms for linking the NEADL and FAI outcome measures to facilitate data pooling across the SOS and CIMSS datasets,

(2) To explore the utility of using regression-based and IRT methods for harmonising the NEADL and FAI measures.

5.3 Methods

5.3.1 Data sources

The SOS1 dataset had both the FAI and NEADL assessments therefore the SOS1 dataset was used in Study 3a to develop mapping algorithms for converting between the FAI and NEADL measures. Details of the methods of recruitment, data collection, and patient demographic characteristics of the SOS1 have already been reported in Chapter 3.

5.3.2 Measures

The NEADL and FAI assess higher level of activities of daily living such as: walking outside, cooking, light and heavy household work, and participation in social activities. Details of the NEADL and FAI measures have already been given in Chapter 1. The SOS1 study used the 21 item version of the NEADL and the (0, 0, 1, 1) scoring system. As described in Chapter 1, the NEADL-22 scale yields total scores that range from 0 to 63 when using the (0, 0, 1, 1) scoring system. A higher NEADL total score is indicative of greater independence in extended activities of daily living. The SOS1 study used the 0-3 scoring system for the FAI measure, this scoring system yields a total score of 0 – 45. A higher FAI score is indicative of better independence in extended activities of daily living. The harmonisation of the FAI and NEADL measures conducted in Study 3a developed mapping algorithms for converting FAI scores (starting measure) onto the NEADL scores (target measure) and vice versa.

5.3.3 Descriptive analyses

The initial analysis in Study 3a was an examination of the NEADL and FAI items to evaluate the extent of item overlaps between the two measures. Poor content overlap between the starting and the target measure reduces the performance of mapping functions (Brazier et al., 2010). The distribution of NEADL and FAI total scores were displayed using the kernel density estimator in STATA version 13 software (StataCorp, 2013) and were not normally distributed. Thus the correlation between the two measures was investigated using the Spearman correlation coefficient. The Spearman's correlations range between -1 and 1 where values closer to one signify a strong positive correlation and values closer to -1 signify a strong

negative correlation. Poor correlation between the measures could indicate that the questionnaires measure different constructs.

5.3.4 Dimensionality of combined NEADL and FAI measures

Further examination of the relationship between the NEADL and FAI measure was conducted using Exploratory Factor Analysis (EFA). Factor analysis was chosen because it is a widely accepted method for deriving the internal structure of measurement scales (Williams et al., 2012) and its utility has been demonstrated elsewhere to evaluate the relationship between measures (Fairhurst et al., 2014). Details of EFA have already been given in Chapter 2. The suitability of using the data for factor analysis was determined using , the Kaiser-Meyer, Olkin (KMO) measure of sampling Adequacy (Kaiser, 1974) and Bartlett's test (Bartlett, 1954) of sphericity. A KMO value of 0.06 or more, and a significant Bartlett test were indicative of suitability.

In Study 3a, EFA was performed using a sample size of n=448. Based on a minimum of 10 observations per variable recommended by (Hair, 2010), this sample size was considered adequate for the 36 items from the FAI and NEADL measures. A sample size of n=448 for 36 items provided acceptable ratio of 12 cases per variable. Tabachnick and Fidell (2007) recommended samples sizes of n=300 for EFA. The sample size of n=448 used in Study 3a was greater than the recommended sample size of 300.

Item data from the FAI and NEADL measures were treated as ordinal therefore factor analysis was conducted using the Mean and Variance-adjusted Weighted Least Squares Algorithm (WLSMV) in Mplus software version 7 (Muthén and Muthén, 2012). In order to aid the interpretation of the factors, the EFA solutions were rotated using oblique geomin rotation. The items were assumed to represent more than one latent factor and these factors were also assumed to be correlated hence geomin rotation (Gorsuch, 1983 cited in Williams et al. (2012)) which produces correlated factors was used. Oblique rotation methods produce more accurate results in research involving human behaviours (William et al., 2012).

The best factor solution was selected based on Root Mean Square Error of Approximation (RMSEA) ≤ 0.06; Tucker-Lewis index (TLI) ≥ 0.95 (Tucker and Lewis, 1973); Comparative Fit Index (CFI) ≥ 0.95 (Gelman et al., 1998), and the extent of interpretability of the emerging factors. A cut-off of 0.3 as recommended by

Tabachnick and Fedell (2007) was set for factor loadings. Exploratory Factor analysis was conducted using Mplus software version 7.11 (Muthén and Muthén, 2012) and the analysis was repeated in SPSS version 22 software to validate the findings.

5.3.5 Mapping NEADL and FAI using regression based methods

After confirming that the NEADL and FAI scales were measuring similar constructs using EFA. Mapping algorithms for converting between the FAI (starting measure) onto the NEADL (target measure) and vice versa were first developed using regression-based methods. Mapping using regression-based methods was chosen because its utility has been demonstrated in economic evaluation studies (Brazier et al., 2012; Longworth and Rowen, 2011; Chen et al., 2014). For example regression-based models have been widely used for mapping from health status measures onto generic preference-based measures when health state utility values are not directly available (Rowen et al., 2009). Mapping using regression-based models is considered the second best solution for the National Institute of Clinical Experience (NICE) Health Technological Appraisals (HTA) submissions when EQ-5D data is not available (Kearns et al., 2012). As highlighted in Chapter 2, the advantage of regression-based mapping is that it makes few assumptions to linking outcome measures compared to IRT methods. However, regression-based approaches require that the two measurement questionnaires be administered to the same sample, while IRT linking can be conducted even if the measures were not administered to the same sample, but requires common items across the measures.

The regression based-mapping which was conducted in Study 3a followed the guidelines for mapping studies recommended by Longworth and Rowen (2011) and these are as follows: defining the source and target measures, selecting the estimation and validation samples, model specification, model estimation, and model validation.

5.3.5.1 Model specification

In Study 3a, various regression model specifications for mapping the FAI and NEADL measures were explored to identify the best predictive model. The outcome variable was the total score for the target measure in all the model specifications. Two model specifications for predictor variables were used. The first used total scores of the starting measure as the predictor and the second used the items of the starting measure as the predictor variable(s). Using totals as predictors is parsimonious but gives all items equal weighting. Using the items as predictors has the advantage of not

assuming that all items carry equal weights (Chen et al., 2014), but can be cumbersome if the measure has many items.

Scatters plots of NEADL and FAI total scores were produced to determine whether the relationship between the two measures was linear. The locally weighted scatterplot smoothing (lowess) (Cleveland, 1981) function in STATA software was used to produce the smoothed scatter plots. The plots showed that the relationship between the FAI and NEADL was non-linear. The relationship appeared to be quadratic. The non-linear relationship between the FAI and NEADL measures was accounted for in the predictive models using squared terms and fractional polynomials (FP) (Royston and Sauerbrei, 2004). Fractional polynomials were selected because, unlike standard polynomials, they are not limited to positive integers, but can also include negative fractional powers (Royston and Sauerbrei, 2004). Royston and Sauerbrei (2004) have argued that, compared to other approaches of modelling non-linear functions such as splines, fractional polynomials are easier to implement and simulation studies have shown their favorable performance. The mfp command in STATA version 13 (StataCorp, 2013) was used to fit the fractional polynomial models. In STATA software fractional polynomial power terms are selected by an automated process and in Study 3a, power terms were selected from the default provided in STATA software: (-2,-1,-0.5, 0, 0.5, 1, 2, 3) and 0 denotes log transformation. The final polynomial degree was decided based on the deviance tests.

5.3.5.2 Model estimation

In study 3a, various model estimators were considered for mapping the NEADL and FAI measures and these included: the Ordinary Least Squares (OLS), quantile regression (Koenker and Bassett Jr, 1982), and robust regression (Yohai, 1987). Ordinary Least squares assume normally distributed errors (residuals) and may produce biased predictions if data is skewed or has celling or floor effects (Brazier et al., 2010). Errors also known as residuals are the difference between the observed and predicted values. Robust estimators are considered efficient for non-normal errors or if the data has outliers (Susanti and Pratiwi, 2014; Yohai, 1987) . A detailed explanation of robust estimators such as MM estimators is provided by Susanti and Pratiwi (2014). Quantile regression is also used to model skewed data and it is more robust to statistical outliers compared to OLS. Details of quantile regression are provided by Koenker and Bassett Jr (1982). In Study 3a, the kernel density plots of the FAI and NEADL total scores showed that the distribution of the scores were left-skewed hence

quantile and robust regression were used and the results from the various estimators were compared to identify the best predictive model. Other popular estimators that are widely used in mapping studies include the Tobit estimator (Tobin, 1958), the Censored Least Absolute Deviation (CLAD; (Jolliffe et al., 2001)) regressions and two part models (Leung and Yu, 1996). These other models are used to account for the floor or ceiling effects (high proportion of individuals at the maximum or minimum). The methods that account for floor and ceiling effects methods were not used in this present study because the distributions of the NEADL or FAI total scores did not show any ceiling or floor issues.

The quantile regression and robust regression were fitted using the qreg and rreg commands in STATA version 13 (StataCorp, 2013). To improve the performance of the mapping algorithms, demographic factors such as age and gender were included in the models. Including demographic factors to improve the prediction performance of the mapping algorithms is a common approach in economic evaluation studies (Brazier et al., 2010).

5.3.5.3 Missing data

The mapping analyses conducted in Study 3a, used baseline data hence no data was missing at baseline.

5.3.5.4 Estimation of predicted scores

The coefficients of the regression models can be used to map or predict scores of one measure to the other. In Study 3a predicted scores from the mapping functions (regression equation) were estimated using the "predict post-estimation" command in STATA version 13.

5.3.5.5 Measures of model performance

Model performance was evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Willmott and Matsuura, 2005). The MAE and RMSE show the average prediction errors at individual level. The MAE is the mean of the absolute differences between the observed and predicted value and the MSE is the mean of the squared differences between the observed and predicted values. The formulae for MAE and Root MSE are shown below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad \text{Equation 5.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} e_i^2} \qquad \text{Equation 5.2}$$

Where n is the sample size and e_i is the residual from individual i.

Low RMSE or MAE values are indicative of good prediction. In Study 3a, the best mapping function or predictive model was identified as the model with the lowest combination of MAE and RMSE values.

Some mapping studies use the acceptable minimum clinical important change for the outcome measure as the threshold for determining predictive performance. Predictive models are considered accurate if prediction errors are less than the acceptable minimal clinical important change of the measures. For example Byers (2004) used a minimum clinically important change of 5 points as acceptable prediction error for the Functional Independence Measure (FIM) scores to Minimum Data Set (MDS) scores conversion table. Wu et al. (2011) suggested that patient's change score has to reach 4.9 points on the total NEADL to indicate a true change, and the mean change score of a stroke group on the total NEADL scale should achieve 6.1 points to be regarded as clinically important. A study by Lu et al. (2012) found the Small Real Difference (SRD) for FAI to be 6.7, but this is considered too large. The SRD indicates real improvement or deterioration for an individual beyond measurement error (Schreuders et al., 2003). In Study 3a, absolute prediction errors of 4.9 points for the NEADL and 6.7 points for the FAI measure were considered too large hence these thresholds were not used to evaluate the prediction accuracy of the mapping algorithms. Acceptable prediction errors close to zero were considered reasonable.

Other criteria for evaluating the prediction performance of mapping algorithms include the use of the four moments (mean, standard deviation, skewness and kurtosis) of the distribution of predicted and observed values. Kolen and Brennan (2013) suggested that all four moments of the distribution should be similar, if the mapping or equating function is accurate. In this present study, the four moments (mean, standard deviation, skewness and kurtosis), minimum, and maximum scores of the observed and predicted scores were also compared to determine the accuracy of the mapping functions. Bland Altman plots of observed and predicted values and the limits of agreement and their corresponding 95% limits of agreement were also produced.

5.3.5.6 Internal validation

Internal cross validation (Efron, 1983) was conducted using the five-fold cross-validation command in STATA version 13 (StataCorp, 2013). Internal validation assesses the performance of the mapping algorithm in the same sample that was used to develop it. In five-fold cross-validation the original sample is randomly partitioned into 5 equal sized subsamples. Of the five subsamples, a single subsample is retained as the validation data for testing the model, while the remaining 4 subsamples are used as training datasets. The cross-validation process is then repeated 5 times, with each of the five subsamples used once as the validation data. The commonly used cross validation is the 10-fold validation. In this present study, the results of the 10-fold validation were similar to five-fold validation hence the five-fold validation was chosen. More details of *k*-fold validation using STATA software can be obtained from (Daniels, 2012).

5.3.5.7 External validation sample

In mapping analyses the performance of the preferred model should be externally validated in an independent dataset to assess how the mapping algorithms would generalise to independent data (Longworth and Rowen, 2011). In Study 3a, there was no independent data for external validation of the developed mapping algorithms, hence the SOS1, wave 2 data was used. However, the use of the wave 2(one year data) for external validation could produce biased results because data in wave 2 were of the same people and the patients will have improved in disability outcomes at one year after stroke.

5.3.5.8 Model diagnostics

The OLS regression models make the assumption of normal errors (residuals) and homoscedasticity (constant variance of residuals). The normal errors assumption assumes that the errors/residuals follow a normal distribution. This assumption was tested using normal probability plots (Ryan, 1974). The plot should be straight if the errors are normally distributed. Formal statistical tests for normality were also conducted using the Smirnov–Kolmogorov test (Lilliefors, 1967). Non-normal errors are indicated by a statistically significant Smirnov-Kolmogorov. In Study 3a *p*-value of ≤ 0.05 was considered significant. The homoscedasticity assumption assumes that the variance of the error term in the regression model is constant. In this present study,

the homoscedasticity assumption was tested using plots of standardised residuals against observed values and the Cook–Weisberg Test (Cook and Weisberg, 1982). Standardised residuals are obtained by dividing a residual by its standard deviation. When testing for the constant variance assumption using a plot of standardised residuals against the observed values, a pattern less plot with residuals equally distributed around zero indicates constant variance. A significant Cook–Weisberg Test indicates heteroscedasticity.

Extreme negative or positive values (outliers) were detected using plots of standardised residuals against fitted values. Observations with standardised residuals in excess of 3.5 and -3.5 were considered to be outliers (Hawkins, 1980). Outliers that are also influential might distort the mean predictions from regression models. Influential and outlier observations were checked using Cooks DFITs (Cook, 1977) and graphical tools such as Added-Variable Plots (Avplots) and Leverage versus squared residual plots (Chatterjee and Hadi, 1986). Leverage measures the deviation of an observation from the mean of the variable. Points with high leverage can also bias the estimated regression coefficients. All regression models and model diagnostic analysis was conducted using STATA version 13(StataCorp, 2013).

5.3.6 Linking FAI and NEADL using IRT methods

Since the mapping functions developed using regression based methods were good at predicting mean scores and not individual patient scores, the utility of using Item Response Theory (IRT) linking to map the FAI and NEADL measures was also explored in study 3a. The utility of using IRT methods for linking outcome measures has been demonstrated elsewhere (McHorney and Cohen, 2000; Velozo et al., 2007) (Chen et al., 2014; Hsueh et al., 2004). Details of IRT linking have already been provided in Chapter 2 of this thesis.

Item response theory linking approach is used to link measurement scales that assess similar underlying construct(s) (Velozo et al., 2007) and the scores from the two measures should be highly correlated. (Dorans, 2007) recommended a correlation coefficient of $r > 0.86$ for successful IRT linking. In this present study the examination of the FAI and NEADL items showed item overlap across the two measures. The IRT linking conducted in Study 3a followed guidelines provided by Dorans (2007) and Holland et al. (2006) which are: establishing dimensionality of the combined measures, calibration, checking for subgroup invariance, and scoring.

Group invariance of the linking algorithm with respect to study was not assessed because linking was conducted using an equivalent or single group design (FAI and NEADL measures were administered to the same sample) hence group invariance by study was not investigated. Group invariance of the linking algorithm with respect to age-group was not investigated because the majority of stroke patients were elderly.

The dimensionality of the combined FAI and NEADL measures was determined using exploratory factor analysis. The items from the two measures loaded on to the same factors hence the foundation of the IRT linking conducted in Study 3a was that the items of the FAI and NEADL measures were measuring similar latent disability constructs. Item calibration was conducted using the two parameter Samejima's 1969, 1997 graded IRT model (Samejima, 1997). In IRT linking, item calibration refers to the process of estimating item parameters using an appropriate IRT model (Dorans, 2007). The Samejima IRT model was chosen because of the graded nature of the responses from the NEADL and FAI items. The details of the Samejima graded IRT model was provided in Chapter 2. Two Samejima graded IRT models were fitted for the "mobility" and the "household/domestic" subscales. Since the FAI and NEADL questionnaires were administered to the same respondents in the SOS1, the IRT calibration conducted in Study 3a used a single sample designs(where both the starting and target measures are administered to the same people) and the items from the two measures were simultaneously calibrated. Simultaneous IRT calibration places all item parameters on the same metric scale (Kim and Cohen, 1998). An R package for latent variable modelling (ltm version 0.9)(Rizopoulos, 2006) and grm library in R software (Mair et al., 2009) were used for the simultaneous IRT calibration of the items that loaded on the same factors.

5.3.6.1 IRT score to summed score conversion

After co-calibrating the set of items from the two subscales ("Mobility" subscale and "Household/Domestic") using an IRT two parameter model, the item parameters from the co-calibrations were used to produce the summed scores corresponding to the different IRT scores. The SS_IRT software (Orlando et al., 2000) was used to estimate the average IRT scores that correspond to the summed score on each subscale. The SS_IRT software uses the expected a posteriori (EAP) summed scoring discussed in Chapter 2. The item parameters (item difficulty and discrimination parameter) from the two parameter graded IRT model obtained from the R software were used as separate inputs into the SS_IRT software. The item parameters were first

saved as a text file that can be read by the SS_IRT software. The resultant summed score to IRT score conversion tables were used to crosswalk between the FAI and NEADL subscales. The SS_IRT software that was used in this thesis was provided by Maria Orlando, North Carolina University. The mathematical details of the IRT score to summed score translation are not provided in this thesis but can be found in Orlando et al. (2000). The accuracy of IRT equating depends on sample size. Fitzpatrick and Yen (2001) recommended sample sizes of at least 200 cases for 20 items. In Study 3a, a sample size of n=448 was used for the IRT calibration for a subset of 15 items or less for the two subscales. This sample size was approximately 30 cases per item, more than the recommended 10 cases per item.

5.3.6.2 Testing the accuracy of the IRT conversion tables

The accuracy of the IRT conversion tables were tested using FAI and NEADL data from wave 2 of the SOS 1. These conversion tables were used to convert FAI subscale scores to NEADL subscale scores and vice versa. The observed scores were compared with their converted scores to determine the accuracy of the conversion tables. RMSE and MAE were calculated to determine the prediction accuracy of the conversion tables. The conversion tables developed using IRT methods were considered accurate if the differences between the converted and the actual were close to zero.

5.4 Results

The next section reports the results of mapping analyses that was conducted in Study 3a. The results are presented following the order of analyses. The reporting of the results from the mapping analyses conducted using regression based methods followed the guidelines for reporting mapping studies recommended by Petrou et al. (2015), described in the "Mapping onto Preference-based measures reporting Standards" (MAPS) statement. The MAPS statement is a checklist of essential items that should be considered by mapping studies for complete and transparent reporting.

5.4.1 Demographic characteristics and final sample sizes

The total number of individuals used in the estimation sample was n=448 and internal validation was conducted using wave 2 (one year data) with n=386 individuals. The average age of the estimation sample was 70.75 years (SD=11.61) and 207 (46.2%) were females. The average age of the validation sample (SOS1 wave

2 data) was 70.52 years (SD=11.43), and 173(44.8%) were females. The wave 2 sample size was smaller compare to wave 1 data due to missing data at wave 2 because of attrition.

5.4.2 Item content overlap

Table 5.1 shows a comparison of the NEADL and the FAI items. The two measures have similar items and these are shown in bold. For example mobility items (e.g. "walking outside"), kitchen/domestic items (e.g. " preparing meals", "washing up", "shopping"), and leisure items ("reading books" and "driving a car"), were common across the two measures. Differences in the scales are that the NEADL has more mobility items, compared to the FAI measure, which focuses mostly on domestic activities.

Table 5.1 Comparison of NEADL with FAI Items

NEADL	FAI
Mobility **Walk around outside** Climb stairs Get in and out of car **Walk over uneven ground** Cross roads Travel on public transport	**Walking outside more than 15 minutes**
Kitchen Manage to feed yourself **Manage to make yourself hot drink** **Make yourself a hot snack** Take hot drinks from one room to another **Do the washing up**	**Preparing main meals** **Washing up after meals(dishes)**
Domestic Manage your own money **Do your own shopping** **Do a full clothes was** **Wash small items of clothing**	Light housework Heavy housework Household / car maintenance (D-I-Y) **Local shopping's** **Washing clothes** Gainful work
Leisure **Read newspapers and books** Use the telephone Write letter **Manage your own garden** **Drive car** **Go out socially**	**Reading books** **Gardening** **Drive/ going on bus** Travelling outings/car rides Actively pursuing hobby **Social outings**

5.4.3 Dimensionality of the combined FAI and NEADL measures

Testing assumptions of factor analysis

The Kaiser-Meyer-Olkin test produced a value of 0.90 which exceeded the recommended value of 0.6 (Kaiser, 1974) for factor analysis, supporting strong partial correlations between the combined items. The Bartlett's test of sphericity was statistically significant [Chi-square =9101, degrees of freedom (df) = 630, $p < 0.001$], suggesting that the 36 combined items were sufficiently correlated. The data met the assumption of factor analysis thus it was suitable to use factor analysis for the combined items.

Table 5.2 shows the model fit indices of the various factor solutions for the combined FAI and NEADL items. The results of the EFA supported solutions with more than one factor as indicated by the, CFI $\geq$ 0.95, TLI $\geq$ 0.95, and RMSEA $\leq$ 0.08. The scree plot supported a three factor structure (Figure 5.1), with a pronounced 'elbow' at three factors.

Table 5.2 The model fit indices from the EFA of the combined FAI and NEADL measures

	CFI	TLI	RMSEA
1 factor	0.86	0.86	0.116 (0.113-0.120)
2 factors	0.95	0.95	0.072 (0.068-0.075)
3 factors	0.97	0.96	0.059 (0.059-0.063)
4 factors	0.98	0.97	0.053 (0.049-0.058)

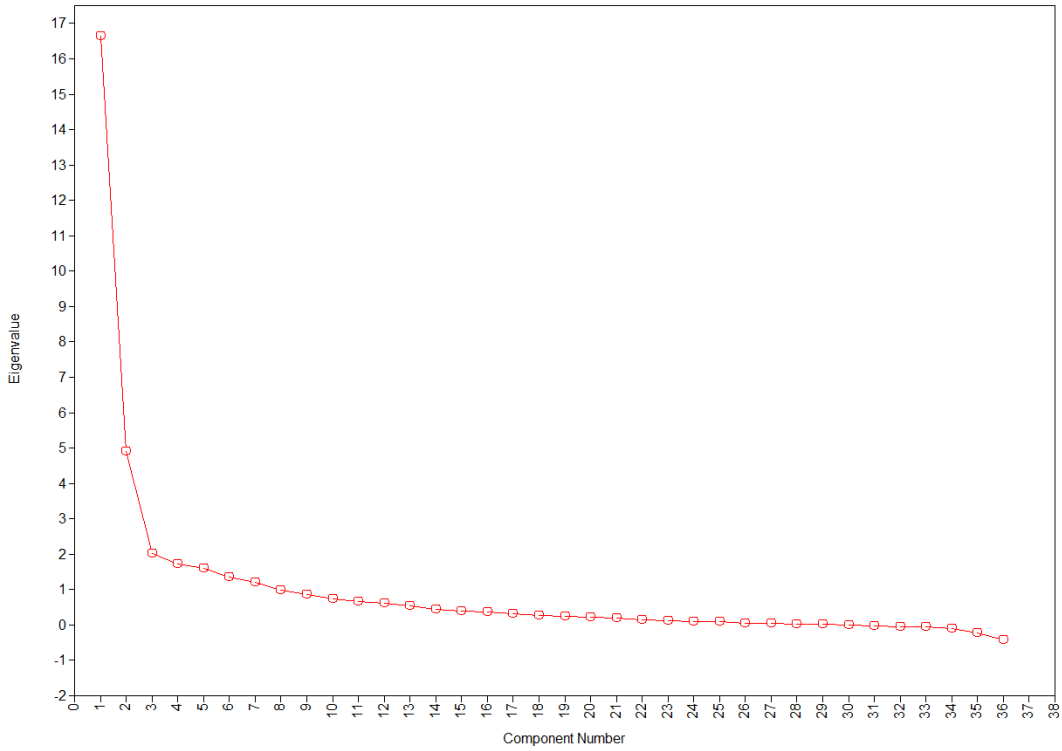

Figure 5.1 Scree plot from the EFA of the combined NEADL and FAI items

The Geomin rotated factor loadings of the 2, 3 and 4 factor solutions are shown in Table 5.3.

Two factor solution

The 2 factor structure showed that 18 NEADL items (1- 9, 11, 12, 14, and 16-21) and 9 FAI items (6-13, and 15) were correlated with factor 1 with loadings of at least 0.38 (Table 5.3). This factor could be labelled "Mobility/Domestic/Leisure". Three NEADL Items (10, 13, and 15) and 5 FAI items (1 to 5) correlated with factor 2 with loadings of at least 0.52; this factor could be labelled "Household work".

Three factor solution

The 3 factor structure showed that 18 NEADL items (1-9, 11, 12, 14, and 16-21) and 9 FAI items (6-13, and 15) were correlated with factor 1. This factor could be labelled "Mobility/Domestic/Leisure". Six NEADL Items (8-11, 13, and 15) and 5 FAI items (1 to 5) were correlated with factor 2 with loadings of at least 0.49. This factor could be labelled "Household work". Two NEADL items (19, 20) and two FAI item (7, 12) were correlated with factor 3. The four items that loaded on factor 3 also cross loaded on factor 1.

Four factor solution

The 4 factor structure showed that 10 NEADL items (1-6, 12, 14, 19, 21) and 6 FAI items (6-8, 10, 13, 15) were correlated with factor 1. Eight NEADL Items (7-11, 13, 15, and 21) and 5 FAI items (1 to 5) were correlated with factor 2. Four NEADL items (10, 16, 18, and 19) and 2 FAI items (item 7, 14) were correlated with factor 3. NEADL item 20 and FAI items 12 and 13 were correlated with factor 4. The 4 factors could be labelled as follows: Factor 1: "Mobility", Factor 2:"Housework/Domestic", Factor 3: "Reading books & writing letters" and Factor 4: "Gardening". Some items cross loaded on more than one factor and these are shown in Table 5.3. Based on goodness-of-fit indices (CFI, TFI, and RMSEA) and the interpretability of the factor solutions, the preferred factor structure was the 4 factor structure for the combined FAI and NEADL measures. The 4 factor solution accounted for 50.5 % of the variance. The 4 factor solution was also reproduced using SPSS software version 22 (Santoso, 2014).

Table 5.3 Geomin rotated factor loadings of the two, three and four factor solutions for the `combined FAI and NEADL items

Items	F1	F2	F1	F2	F3	F1	F2	F3	F4
EXT1	**0.848***	0.102*	**0.832***	0.176*	-0.056	**0.809***	0.126*	-0.031	0.108
EXT2	**0.818***	0.034	**0.803***	0.104	0.001	**0.717***	0.060	0.008	0.174
EXT3	**0.826***	0.003	**0.822***	0.058	-0.040	**0.733***	0.010	0.079	0.119
EXT4	**0.893***	0.090*	**0.875***	0.169*	-0.018	**0.800***	0.120*	-0.011	0.171
EXT5	**0.921***	-0.001	**0.900***	0.081	0.016	**0.761***	0.032	0.068	0.208
EXT6	**0.689***	0.166*	**0.703***	0.193*	-0.157*	**0.748***	0.141*	0.036	-0.054
EXT7	**0.603***	0.340*	**0.554***	0.431*	0.168	0.355*	**0.419***	-0.030	0.371*
EXT8	**0.647***	0.497*	**0.561***	**0.612***	0.174*	0.180	**0.582***	0.335*	0.320*
EXT9	**0.704***	0.410*	**0.663***	**0.498***	0.072	0.397*	**0.466***	0.243*	0.234
EXT10	0.410*	**0.642***	0.371*	**0.706***	0.013	0.005	**0.664***	**0.559***	0.082
EXT11	**0.622***	0.523*	**0.567***	**0.609***	0.112	0.231*	**0.574***	0.341*	0.244*
EXT12	**0.777***	0.182*	**0.778***	0.235*	-0.185*	**0.759***	0.190*	0.137	-0.050
EXT13	-0.008	**0.945***	-0.038	**0.950***	-0.093	0.028	**0.947***	-0.072	-0.041
EXT14	**0.704***	0.292*	**0.713***	0.325*	-0.244*	**0.826***	0.270*	-0.031	-0.109
EXT15	0.074	**0.949***	0.047	**0.960***	-0.109	0.141*	**0.942***	-0.120	-0.029
EXT116	**0.429***	-0.051	**0.443***	-0.033	-0.037	0.003	-0.001	**0.603***	0.090

EXT117	**0.519***	-0.235*		**0.461***	-0.134	0.230*		0.096	-0.124	0.213	0.370*
EXT118	**0.465***	0.073		**0.445***	0.133	0.051		0.061	0.157*	**0.419***	0.192
EXT119	**0.772***	-0.020		**0.802***	-0.038	**-0.436***		**0.641***	-0.053	0.**565***	-0.270*
EXT120	**1.043***	-0.554*		**0.635***	-0.016	**0.700***		0.065	-0.026	-0.024	**0.951***
EXT121	**0.767***	-0.514*		**0.785***	-0.473*	0.123		**0.497***	**-0.505***	0.225*	0.288*
FREN1	0.021	**0.755***		-0.004	**0.769***	-0.088		-0.057	**0.792***	0.126	-0.063
FREN2	0.060	**0.730***		0.016	**0.766***	-0.007		-0.305*	**0.787***	0.492*	-0.011
FREN3	-0.017	**0.992***		-0.051	**0.998***	-0.078		-0.009	**0.998***	-0.071	-0.013
FREN4	0.161*	**0.772***		0.116	**0.816***	-0.002		-0.011	**0.835***	0.101	0.083
FREN5	0.439*	**0.523***		0.402*	**0.595***	0.006		0.309*	**0.584***	0.014	0.145
FREN6	**0.646***	0.344*		**0.649***	0.374*	-0.259*		**0.792***	0.320*	-0.055	-0.137
FREN7	**0.686***	0.010		**0.705***	-0.007	**-0.498***		**0.574***	-0.006	**0.585***	-0.361*
FREN8	**0.752***	0.104*		**0.742***	0.169*	-0.057		**0.704***	0.123*	0.010	0.095
FREN9	**0.568***	-0.087		**0.505***	0.035	0.228*		0.069	0.042	0.310*	0.399*
FREN10	**0.814***	-0.127*		**0.833***	-0.097	-0.089		**0.729***	-0.148*	0.188*	0.049
FREN11	**0.381***	-0.199*		**0.408***	-0.199*	-0.057		0.212*	-0.200*	0.297*	0.031
FREN12	**0.997***	-0.519*		**0.589***	0.008	**0.731***		-0.053	0.011	0.032	**0.998***
FREN13	**0.754***	-0.300*		**0.726***	-0.203*	0.240*		**0.504***	-0.249*	-0.012	**0.427***
FREN14	0.163*	0.075		0.181*	0.065	-0.067		-0.203*	0.123	**0.587***	-0.018
FREN15	**0.503***	-0.081		**0.530***	-0.073	-0.015		**0.602***	-0.126	-0.160	0.101

*Significant at 5% level–

Validation of the exploratory factor models

EFA was repeated in different software, SPSS Software version 22 (Santoso, 2014). The 4 factor solution was also reproduced using different software. Similar factor loadings were produced.

5.4.4 Descriptive statistics of measures

The distributional plots of the baseline NEADL and FAI total scores showed that the scores from the two measures were left-skewed (Figure 5.2). The mean NEADL total score in the estimation sample was 49.30(SD= 11.50) range (6 - 63) and mean FAI total score was 26.20 (SD=9.33) range (2- 45). In the validation sample (SOS1 wave 2 data) the average NEADL score was 30.82(SD=16.76) and average FAI score was 13.34 (SD=10.74). The correlation between the two measures at baseline was strong positive (r = 0.83, p<0.001). The relationship between the two measures was also non-linear, it appeared to be quadratic (Figure 5.3).

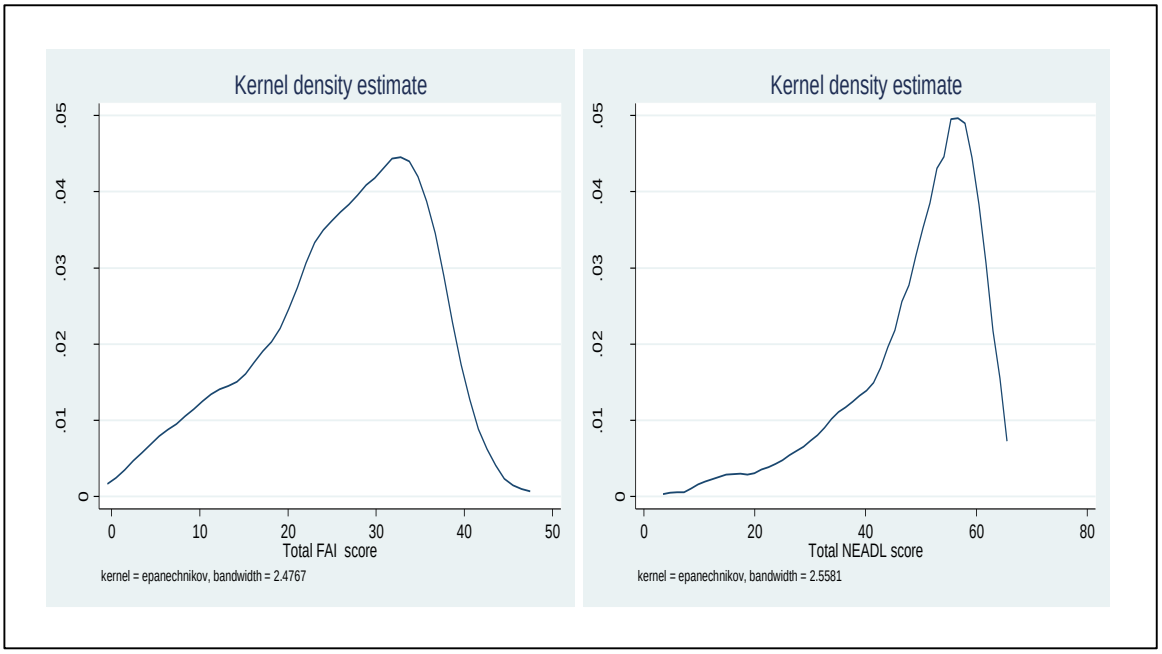

Figure 5.2 Kernel density estimate of one month NEADL and FAI totals

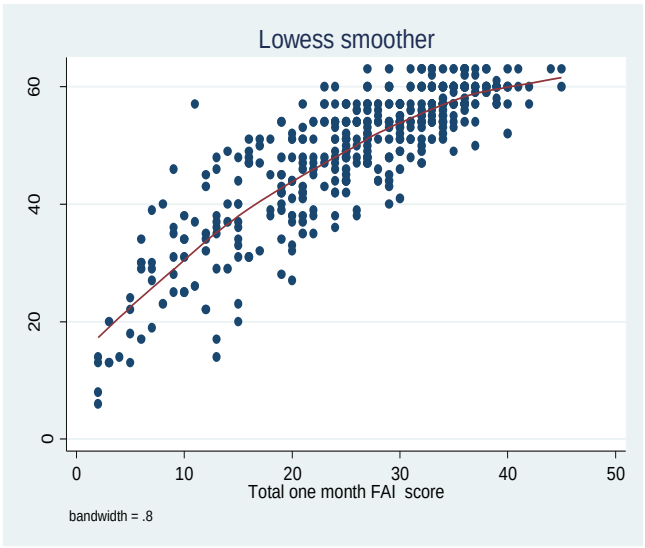

Figure 5.3 Scatter plot of baseline NEADL scores against FAI scores with lowess smoother

5.4.5 Results from mapping using regression based methods

5.4.5.1 Mapping FAI onto NEADL prediction accuracy

The prediction performance indices (RMSE, MAE) of the various models that were fitted for mapping the FAI (starting measure) onto the NEADL (target measure) in the estimation sample are shown in Table 5.4. The five-fold validation results are shown in Appendix D. The results in Table 5.4 showed that in the estimation sample all the models predicted very well the average baseline NEADL score. The observed NEADL was 49.3 and the predicted averages for the models with no fractional polynomials were 49.3, 49.5 and 49.51 for the ordinary least squares, quantile, and robust regressions models respectively. The lower limits of estimation were 15 for all the three estimators with no fractional polynomials and exceeded the observed average NEADL lower limit of 6 (Table 5.4). This result suggested that the mapping functions over predicted the lower end of the NEADL scale. The average upper limit of estimation was 62 for the ordinary least squares, quantile estimator and robust estimators and this estimate was very close to the observed average upper limit of 63 (Table 5.4). These results showed that the FAI-NEADL mapping algorithm predicted the average NEADL score and the upper limit of the NEADL scale very well compared to the lower limit of the scale.

The various estimators that were used to predict NEADL scores from the FAI scores showed large RMSE and MAE far from zero (Table 5.4). The RMSE values were larger than the MAE values. By definition the RMSE is never smaller than the MAE. Willmott and Matsuura (2009) argued that the RMSE was not a good indicator of average prediction errors and recommended the use of the MAE. However (Chai and Draxler, 2014) showed that the RMSE was superior than the MAE when the error distribution is normally distributed. In Study 3a the prediction errors were considered acceptable if both RMSE and MAE were close to zero. The prediction errors(RMSE, MAE) in the estimation sample for all the estimators that were used in study 3a to predict the NEADL from FAI were considered large as these were all far from zero (Table 5.4, Appendix D), suggesting poor individual level predictions in the estimation sample. Using fractional polynomials to account for the non-linear relationship between the measures was of no significant gain since the MAE and RMSE remained poor, and were similar to those of simpler models that accounted for the non-linear relationship by using a squared term in the model (Table 5.4). Using FAI items as the predictors of NEADL total slightly reduced the RMSE and MAE, but the values were still not close to zero.

Table 5.4 Model performance of various estimators for mapping FAI onto NEADL measure

Function	Predictors	Mean(min, max)	MAE Internal	RMSE Internal
Observed NEADL scores		49.3(6 - 63)		
OLS	FAI , FAI*FAI , age,	49.3(15 - 62)	4.19	5.55
Quantile	gender	49.5(15- 62)	4.16	5.55
Robust		49.5(15- 62)	4.16	5.30
Fractional Polynomials	FAI, age, gender			
OLS(0.5)		49.3(14 - 66)	4.22	5.53
Quantile(0.5, 2)		49.6(11.3 - 63)	4.11	5.52
Robust(-0.5, 0)		49.5(10 - 64)	4.10	5.49
Items as predictors				
OLS	FAI items + age+	49.3(17 - 65)	3.58	5.03
Quantile	gender	49.3(17 -66)	3.46	4.89
Robust regression		49.3(17 -65)	3.54	4.80

5.4.5.2 Model coefficients: FAI-NEADL Mapping function

The model coefficients for the ordinary least squares, quantile and robust regression models for the FAI-NEADL mapping functions are shown in Table 5.5. Age and gender were statistically significant in most of the models that were fitted and were

therefore included in the prediction models. The results of the model coefficients that used items as predictors are shown in Appendix D. The responses to the items were treated as categorical variables in the models that used items as predictors hence the models resulted in many predictor variables and some of the items had regression signs that were counter intuitive.

Table 5.5 Regression coefficients from OLS, Quantile and Robust estimators for mapping the FAI onto NEADL measure

	OLS β **(95% CI)**	**Quantile** β **(95% CI)**	**Robust** β **(95% CI)**
FAI	1.98(1.73-2.23)***	2.05(0.86,-1.03)***	2.06(1.83, 2.30)***
FAI*FAI	-0.02(-0.03,-0.01)***	-0.02(-0.03,-0.02)***	-0.02(-0.03, -0.02)***
Age	-0.07(-0.11, -0.02)***	-0.06(-0.12,-0.008)*	-0.07(-0.11, -0.03)***
Female	-1.77(-2.82, -0.73)***	-0.76(-1.99, 0.47)	-1.17(-2.17, -0.17)***
Constant	18.74(14.23, 23.25)	17.28(11.99, 22.57)	18.20(13.89, 22.51)
	Adj R = 0.77	Pseudo R squared = 0.51	

*P<0.05, ***p<0.001

5.4.5.3 Model Selection: FAI-NEADL mapping functions

Even though the FAI-NEADL mapping functions from the various estimators were not good for individual level predictions, group average predictions were accurate. For example the predicted NEADL averages were similar to the observed average (mean: 49.30, range: 15 - 62) suggesting good prediction for group averages (Table 5.4). A model for predicting group averages was then selected and further validated using SOS1 wave 2 data. The preferred FAI-NEADL mapping function was the ordinary regression model with a quadratic term. Using complex models such as robust estimators, quantile regression or including fractional polynomials added no substantial reduction in the RMSE and MAE values. As highlighted earlier, the models with FAI items as predictors of NEADL had slightly smaller RMSE and MAE compared to models that used totals, but the number of predictors was too much and using such a model for mapping will be cumbersome. The FAI questionnaire has 15 items and these were fitted as categorical variables hence the model ended up with 30 predictors. The number of predictor items could have been reduced by considering only the significant items as recommended by Brazier et al. (2010) but this approach was not adopted as only two items (4, 15) were not statistically significant. Since the MAE and RMSE of the various estimators (OLS, quantile regression and robust regression) were similar, for parsimony the OLS mapping function with FAI total, quadratic term of FAI, age and gender was preferred. Furthermore the model

diagnostics for the OLS did not show any serious violations of the assumptions of OLS (see Appendix D). The equations of the preferred ordinary least squares FAI-NEADL mapping function are shown below:

NEADL=1.98FAI-0.02xFAIsquared -0.07xAge -1.77 x female + 18.74: for females

NEADL=1.98FAI-0.02xFAIsquared -0.07xAge + 18.74: for males

Where age is in years, female is a dummy variable equal to 1 if they are females and zero is they are male.

The variance covariance matrix of the preferred ordinary least squares regression FAI-NEADL mapping function is presented in Appendix D.

5.4.5.4 Model External Validation: FAI-NEADL mapping function

Validation of the preferred model (OLS) FAI-NEADL mapping function was conducted using wave 2 (one year) SOS1 data. The regression coefficients of the preferred OLS regression mapping function shown in section 5.4.5.3 were used to predict the NEADL scores in the SOS2 wave 2 dataset ("external" validation sample). The predicted and observed NEADL scores were highly positively correlated with a Spearman correlation coefficient of 0.90, and was statistically significant (p<0.001). However, the RMSE and MAE were very high 7.85 and 6.22 respectively and were all far from zero, suggesting poor individual predictions in the validation sample.

Kolen and Brennan (2013) suggested that an equating function is successful if all four first moments of the distribution are statistically equivalent. In the validation sample, the four moments (mean, standard deviation, skewness and kurtosis) of the observed and predicted scores were calculated and compared; these are shown in Table 5.6. The results in Table 5.6 showed that in the validation sample, the mean of the predicted NEADL was close to the observed NEADL mean, approximately 2.7 points greater than that of the observed suggesting good group average predictions. The bias and the 95% limits of agreement from the Bland Altman plots was -2.7 (CI -3.476 to -1.998). About 47.9% for the predicted NEADL scores from the FAI-NEADL mapping algorithm lies within less than 5 points of one another. The standard deviation of the actual and predicted mean NEADL scores was within two points of each other, with the predicted values having lower variances compared to the actual observed values. The predicted NEADL scores showed negative kurtosis similar to the actual observed scores. The observed NEADL scores showed negative skewness, while the predicted values showed positive skewness. Based on these four moments

the mapping function was successful in predicting group level summary statistics as these were similar, except skewness.

Table 5.6 Four moments of the observed and predicted NEADL scores. OLS mapping function, SOS1 Wave 2(one year) data, n=386

	NEADL Observed	NEADL Predicted
Mean	30.82	33.57
SD	16.76	14.83
Skewness	-0.014	0.095
Kurtosis	-1.089	-1.291
Median	30.00	33.18

5.4.5.5 Mapping NEADL onto FAI prediction accuracy

Mapping functions that predicted the FAI (target measure) from the NEADL (starting measure) were also developed. The prediction performance indices (RMSE, MAE) of the various models that were fitted for mapping the NEADL (starting measure) onto the FAI (target measure) in the estimation sample are shown in Table 5.7. The RMSE and MAE from the cross-validation are shown in Appendix D. The prediction errors (RSME and MAE) in the estimation sample for all the estimators to predict the FAI from NEADL were considered large as these were all far from zero (Table 5.7, Appendix D), suggesting poor individual level predictions in the estimation sample. However, the results in Table 5.7 showed that in the estimation sample all the models predicted very well the average baseline FAI and the average lower limits. The NEADL-FAI mapping functioning under predicted the upper limits of the FAI scale as all the predicted values were lower than the observed FAI average score for the various estimators (Table 5.7). The average observed FAI score and range were 26.20(2-45) and the predicted average FAI scores and range were: 26.20 (1.69, 38.43), 26.30 (0.81, 39.17), and 26.35(1.36, 38.66) for the ordinary least squares, quantile and robust regression respectively. Using the NEADL items as predictors of FAI total resulted in no significant change in MAE and RMSE values (Table 5.7). The predictions from the models that used NEADL items as predictors of the FAI measure had negative average lower limits for the FAI measure (Table 5.7), probably because some items had some regression signs that were counter intuitive, not working as expected.

Table 5.7 Model performance of various estimators for mapping NEADL onto FAI measure

Function	Predictors	Mean(min, max)	MAE Internal	RMSE Internal
Observed FAI scores		26.20(2-45)		
Totals as predictors	NEADL+NEAD*NEADL+age+ gender			
OLS		26.20 (1.69, 38.43)	3.57	4.73
Quantile		26.32(0.81, 39.17)	3.57	4.37
Robust		26.35 (1.36, 38.66)	3.35	4.34
Items as predictors	NEADL items+ age + gender			
OLS		26.20(-1.56, 38.86)	5.28	4.52
Quantile		26.52(-0.009 , 39)	5.11	3.57
Robust regression		26.35(-1.30 , 38.72)	3.86	5.36

Table 5.8 Regression coefficients from various estimators for mapping the NEADL onto FAI measure

	OLS β (95%CI)	Quantile β (95%CI)	Robust β (95%CI)
NEADL	0.25(0.03, 0.47)*	0.31(0.03, 0.60)*	0.28(0.05, 0.50)**
NEADL*NEADL	0.005(0.003, 0.008)***	0.004(0.001, -0.008)**	0.005(0.002, 0.008)***
Age	-0.01(-0.05, -0.03)	-0.03(-0.08, 0.02)	-0.01(-0.05, 0.03)
Female	1.85(0.96, 2.73)***	1.80(0.65, 2.95)**	1.77(0.88, 2.69)***
Constant	0.14(-5.30, 5.58)	0.49(-7.57, 7.54)	-0.24(-5.79, 5.31)
	Adj R =0.74	Pseudo *R* squared = 0.51	

*P<0.05, ***p<0.001

5.4.5.6 Model selection NEADL-FAI mapping

Even though the FAI-NEADL mapping functions were not good for individual level predictions, group average predictions were good. For example the predicted NEADL averages were similar to the observed average (mean: 26.20, range: 2-45) suggesting good prediction for group averages (Table 5.7). A model for predicting FAI group averages was then selected and further validated using SOS1 wave 2 data. The ordinary least squares NEADL-FAI mapping function with predictors NEADL total, quadratic term of NEADL, age, and gender was preferred since the MAE and RMSE of this simple model was similar to that of the quantile and robust regression models. The ordinary least squares mapping functions for mapping NEADL total scores onto the FAI total scores s are shown in the equations below:

FAI total = 0.25xNEADL+0.005 x NEADLsquared+-0.01xAge+1.85 gender female+ 0.14: for females

FAI total = 0.25xNEADL+0.005 x NEADLsquared+-0.0x Age + 0.14: for males

The variance covariance matrix of the preferred OLS NEADL-FAI mapping model is presented in Appendix D.

5.4.5.7 Model External Validation: NEADL-FAI mapping function

The external validation of the NEADL-FAI mapping function was conducted using the SOS 1 wave 2 data. The regression coefficients from the NEADL-FAI mapping functions presented in section 5.4.5.6 were used to predict the FAI scores in the validation sample. The correlation between the observed and predicted FAI scores in the validation sample (wave 2 data) was high ($r=0.90$) and statistically significant ($p<0.001$). In the validation sample (SOS2 wave 2 data) the RMSE and MAE for the preferred NEADL-FAI mapping function were 4.65 and 3.50 respectively suggesting poor individual level predictions as these values were far from zero. However, the observed and predicted mean, standard deviation, skewness and kurtosis values were similar in the validation sample (Table 5.9) suggesting good group average predictions in wave 2 data. The Mean difference (bias) from the Bland Altman plot was: -0.781 (CI:-1.241 to -0.321), which was close to zero suggesting good group average predictions. Similar to the OLS FAI-NEADL mapping function, the standard deviation of the predicted FAI values were slightly less compared to the observed values

showing less variation in the predicted values. About 76% of the predicted FAI scores from the NEADL-FAI mapping function lies within less than 5 points of one another.

Table 5.9 Four moments of the observed and predicted FAI scores, OLS mapping, SOS1 Wave 2(one year) data, n=386

	FAI Observed	FAI Predicted
Mean	13.34	14.12
SD	10.74	9.12
Skewness	0.49	0.40
Kurtosis	-0.94	-0.92
Median	12.00	12.78

5.4.5.8 Summary of mapping using regression based methods

Based on RMSE and MAE, the NEADL-FAI and FAI-NEADL mapping functions developed in Study 3a using regression-based methods were poor at predicting individual level predictions. However, the mapping functions were good at predicting group averages. The variation of the predicted values was less than the variation of the actual observed scores.

The FAI-NEADL mapping functions developed using ordinary least squares, quantile regression, and robust regression over predicted the lower limit of the NEADL scale and under-estimated the upper limit. The NEADL-FAI mapping function was very accurate in predicting the lower limit of the FAI but the upper limit was slightly under predicted.

The bias from the Bland Altman plot was almost zero for the NEADL-FAI mapping function and 2.7 for the FAI-NEADL mapping function. The NEADL-FAI mapping algorithm could have performed better than the FAI-NEADL measure due to the wider coverage of extended activities of daily living of the NEADL compared to the FAI measure. The NEADL scale seemed to predict the FAI scale better than the FAI predicting the NEADL measure.

5.4.6 Results from linking the FAI and NEADL measures using item response theory models

5.4.6.1 Calibration using Item response theory modelling

The IRT calibration of the NEADL and FAI items was based on the four factor structure that was identified using factor analysis in section 5.4.3. The factor loadings from the four factor structure reported in section 5.4.3 showed that the NEADL items

(1-6, 12, 14, 19, 21) and FAI items (6, 7, 8, 10, 13, and 15) loaded onto the same factor which was labelled “Mobility”. The FAI items (1-5) and NEADL items (7-11, 13, 15 and 21) loaded onto the same factor which was labelled Household/Domestic”. Items that loaded onto the same factors were calibrated simultaneously using the two parameter IRT model. Items that cross loaded on multiple factors were included in the factor in which they had the highest factor loadings. The 15 items from NEADL (items 1-6, 12, 14, 21) and FAI (items 6, 7, 8, 10, 13, 15) that loaded onto the “Mobility factor” were co-calibrated simultaneously. Similarly the 12 items from FAI (items 1-5) and NEADL (items 7-11, 13 and 15) that loaded on to “Household/Domestic” factor were co-calibrated simultaneously.

The results of the co-calibrations using the two-parameter graded IRT models are presented in Table 5.10. The item parameters from the two-parameter graded IRT models for the items that loaded onto the “Mobility” factor and those that loaded onto the “Household/Domestic” factor are shown separately in Table 5.10. In Table 5.10, the extreme 1, extreme 2 and extreme 3 values are the estimates of the location parameters (ability) of the items and the discrimination parameters indicate how good an item is at differentiating among individuals with different abilities. The higher the discrimination value, the more discriminating the item. The discrimination parameters of the “Mobility” subscale suggested that the majority of items from the NEADL measure that loaded onto the “Mobility” factor were more discriminating compared to the items from the FAI scale since they had higher discriminating values (Table 5.10). The most discriminating items on the “Mobility” subscale were: NEADL items 1 (“Walking around outside”), NEADL item 4 (“Walk over uneven ground”), NEADL item 3 (“Get in and out of car”), NEADL item 2 (“Climb stairs”), NEADL item 12 (“manage your own money when out”), NEADL item 5, (“Cross roads”), NEADL item 6 (“Travel on public transport”), FAI item 6 (“local shopping”) and FAI item 8 (“ Walking outside for greater than 15 minutes”) (Table 5.10). The most discriminating items on the “Housework/Domestic” domain were: NEADL item 15 (“Do your own shopping”), NEADL item 13 (“Washing small items of clothes”), FAI item 3 (“Washing clothes”), FAI item 4 (“Light housework”), and FAI item 5(“Gainful work”). The majority of the FAI items that loaded onto the “Housework/Domestic” factor were also more discriminating compared to the items from the NEADL measure.

Table 5.10 Item parameter estimates from the simultaneous calibration of the pooled FAI/NEADL items for the "mobility" and "housework/domestic" subscales

		"Mobility Dimension"					"Housework/Domestic dimension"		
Item	**Extreme1**	**Extreme2**	**Extreme3**	**Discrimination**	**Item**	**Extreme1**	**Extreme2**	**Extreme3**	**Discrimination**
NEADL1	-2.329	-1.841	-0.899	**3.225**	FAI1	-1.025	-0.986	-0.923	**2.679**
NEADL2	-2.050	-1.564	-0.687	**2.448**	FAI2	-2.118	-1.935	-1.706	1.784
NEADL3	-2.734	-1.895	-1.115	**2.422**	FAI3	-0.347	-0.346	-0.346	**9.105**
NEADL4	-1.491	-1.200	-0.556	**3.897**	FAI4	-0.574	-0.572	-0.569	**6.367**
NEADL5	-1.901	-1.367	-0.819	**3.619**	FAI5	-0.294	-0.292	-0.285	**6.261**
NEADL6	-0.966	-0.843	-0.625	**1.856**	NEADL7	-4.483	-3.776	-2.926	1.494
NEADL12	-2.072	-1.974	-1.798	**2.715**	NEADL8	-2.984	-2.579	-1.976	1.795
NEADL14	-1.186	-0.702	-0.527	**2.080**	NEADL9	-2.120	-1.824	-1.385	2.073
NEADL21	1.159	1.173	1.216	0.942	NEADL10	-2.261	-2.177	-2.054	2.010
FAI6	-1.120	-1.086	-0.804	**1.707**	NEADL11	-1.583	-1.466	-1.241	2.776
FAI7	-1.944	-1.412	-0.574	1.022	NEADL13	-0.387	-0.387	-0.386	**8.701**
FAI8	-1.358	-1.104	-0.872	**2.088**	NEADL15	-0.315	-0.315	-0.315	**9.830**
FAI10	-1.330	-1.078	-0.841	1.665					
FAI13	0.239	0.830	1.304	1.266					
FAI15	2.176	2.242	2.311	0.862					

5.4.6.2 IRT score to summed score

The ability and discrimination parameters shown in Table 5.10 from the co-calibrated items were entered into the SS_IRT software to produce IRT score to summed scores conversion tables for the "Mobility" and "Housework/Domestic" subscales. Table 5.11 shows the IRT score to summed scores conversion table for the "Mobility" subscale produced using the SS_IRT software. The conversion table 5.11 shows that an FAI "Mobility" score of 5 was equivalent to a score of 14 on to the NEADL "Mobility" subscale since both had an IRT score associated with each score of -1.1. Similarly an FAI "Mobility" score of 1 was equivalent to a score of 8 on to the NEADL "Mobility" scale.

Table 5.11 IRT score to summed score conversion table for the 9 NEADL "Mobility" items and 6 FAI mobility items

NEADL "Mobility" items 1 to 6, 12, 14, 21			FAI "Mobility" items 6,7,8,10,13, 15		
Summed Score	IRT score	SD	Summed Score	IRT score	SD
0	-2.9	0.41	0	-2	0.59
1	-2.6	0.34	1	-1.7	0.53
2	-2.4	0.32	2	-1.6	0.53
3	-2.2	0.32	3	-1.4	0.55
4	-2.1	0.3	4	-1.2	0.5
5	-2	0.29	5	-1.1	0.5
6	-1.9	0.28	6	-0.88	0.53
7	-1.8	0.27	7	-0.72	0.52
8	-1.7	0.27	8	-0.59	0.52
9	-1.6	0.26	9	-0.32	0.56
10	-1.5	0.26	10	-0.18	0.57
11	-1.4	0.26	11	-0.051	0.57
12	-1.3	0.26	12	0.28	0.62
13	-1.2	0.27	13	0.47	0.63
14	-1.1	0.27	14	0.59	0.65
15	-1.0	0.28	15	0.92	0.7
16	-0.91	0.29	16	0.91	0.67
17	-0.8	0.31	17	1.0	0.67
18	-0.67	0.34	18	1.5	0.73
19	-0.57	0.34			
20	-0.43	0.36			
21	-0.14	0.48			
22	-0.14	0.45			
23	-0.0036	0.43			
24	0.62	0.61			
25	0.17	0.55			
26	0.33	0.55			
27	1.1	0.69			

SD: standard deviation

Table 5.12 shows the IRT score to summed score for the NEADL items (7-10, 11, 13 and 15) and FAI items (1-5) for the "Housework/Domestic" subscale produced using the SS_IRT software. The conversion table shows that an FAI "Housework/Domestic" score of 4 was equivalent to a score of 13 on to the NEADL "Housework/Domestic" factor since both have an IRT score of -1.0. Similarly an FAI "Housework/Domestic" score of 15 was approximately equivalent to a score of 21 on the NEADL "Housework/Domestic" subscale.

Table 5.12 IRT score to summed score conversion table for the 7 NEADL items and 5 FAI domestic subscales

NEADL "Housework/Domestic" items 7-10, 11,13,15			FAI "Housework/Domestic" items 1-5		
SS	**IRT score**	**SD**	**SS**	**IRT score**	**SD**
0	-3.0	0.5	0	-1.8	0.59
1	-2.9	0.49	1	-1.5	0.49
2	-2.7	0.48	2	-1.5	0.47
3	-2.5	0.47	3	-1.2	0.44
4	-2.3	0.45	4	-1	0.35
5	-2.2	0.44	5	-0.98	0.34
6	-2	0.44	6	-0.81	0.31
7	-1.9	0.42	7	-0.57	0.21
8	-1.7	0.42	8	-0.56	0.21
9	-1.5	0.43	9	-0.46	0.21
10	-1.4	0.41	10	-0.3	0.21
11	-1.3	0.4	11	-0.29	0.21
12	-1.1	0.38	12	-0.13	0.3
13	-1	0.36	13	0.14	0.39
14	-0.9	0.33	14	0.14	0.39
15	-0.72	0.29	15	0.72	0.65
16	-0.39	0.21			
17	-0.37	0.21			
18	-0.25	0.28			
19	0.12	0.39			
20	0.15	0.4			
21	0.68	0.66			

5.4.6.3 Testing the accuracy of the conversion tables produced using IRT methods

The accuracy of the IRT conversion tables shown in Tables 5.11 and 5.12 was conducted using SOS1 wave 2 data (1 year). The correlation of the observed and converted NEADL "Household /Domestic" subscale scores was r = 0.860, and 0.857 for the observed and converted NEADL "Mobility" subscale. Both correlation coefficients were statistically significant, $p<0.001$. In the validation sample, the

average individual level prediction errors for the NEAD -FAI Mobility conversions was MAE =3.89 and RMSE = 4.89 suggesting poor individual level predictions. Similarly the individual level predictions were high (MAE=3.06, RMSE=3.98) for the NEADL-FAI Housework/Domestic" subscale.

The distributions of the four moments (mean, standard deviation, skewness and kurtosis) for the actual and predicted scores from the conversion tables developed using IRT methods are shown in Table 5.13. The predicted average NEADL "Mobility" subscale score was within 2.49 of the observed, suggesting good group average predictions. The bias estimates between the predicted and observed from the Bland Altman plot was (-2.48, 95% CI: -2.91, -2.06). The predicted average NEADL "Household/Domestic" subscale score was within 1.25 of the observed, suggesting accurate group average predictions. The bias estimates between the predicted and observed from the Bland Altman plot was -1.25 (95% CI: -1.63, -0.87).

Table 5.13 Four moments of the NEADL distributions for the observed and predicted data, SOS1 wave 2 (one year) data

	Observed NEADL Mobility n=386	**Predicted NEADL Mobility n=386**	**Observed NEADL Housework/Domestic n=386**	**Predicted NEADL housework/ Domestic n=386**
Mean	11.65	14.14	11.73	12.98
SD	8.34	7.40	6.73	4.70
Skewness	0.23	0.20	-0.16	0.372
Kurtosis	-1.25	-1.32	-1.23	-1.264
Median	10.00	13.00	12.00	13.00

5.5 Discussion

The aims of Study 3a were twofold, to develop mapping algorithms for linking the NEADL and FAI outcome measures, and to explore the utility of using regression-based and IRT methods for harmonising the NEADL and FAI measures. The ability to harmonise PROMs by mapping or linking is important when pooling data from different PROMs that measure similar construct(s). In Study 3a, mapping functions and conversion tables were developed to relate scores from the FAI and NEADL measures. The strengths of the analyses conducted in Study 3a were that a single sample design was used for IRT linking and multiple methods of harmonisation (regression based and IRT linking) were explored and compared. The single sample

design is considered to produce more robust links as it controls for differences in abilities across groups (Dorans, 2007).

Exploratory data analysis showed that the NEADL and FAI measures were highly correlated (r =0.83). This finding is consistent with Sarker et al. (2012) who also found that these two measures were highly correlated, with a correlation coefficient of r >0.8. The exploratory factor analysis of the NEADL and FAI measures showed that the combined items from the two measures were measuring four latent factors, and were labelled: Factor 1: "Mobility", Factor 2:"Housework/Domestic", Factor 3: "Reading books & writing letters" and Factor 4: "Gardening".

The first mapping analyses conducted in study 3a used regression based methods to map the FAI onto the NEADL measures and vice versa. The findings from these analyses showed that regression-based mapping functions were effective in predicting the group means and not patient level predictions. The predicted group level moments (mean, standard deviation, skewness and kurtosis) were close to the actual observed group statistics. However the variation of the predicted values was lower than that of the observed. The findings from Study 3a recommends mapping for predicting group level estimates and not individual level predictions. Similarly, in economic evaluation studies, the purpose of mapping functions is to predict differences across groups of patients or differences between arms over time in clinical trials and not between individual level index values and accuracy of mapping functions focuses on predicting mean values for subgroup of patients and not on individual level predictions (Brazier et al., 2010).

An explanation for poor individual level predictions from regression based mapping was provided by Fayers and Hays (2014). They attributed poor individual level predictions by regression-based models to a statistical phenomenon known as "regression to the mean". Fayers and Hays (2014) explained that at individual level, "regression to the mean" will unfairly award patients with lower observed scores higher predicted scores closer to the mean and individuals with higher scores will be awarded lower predicted scores closer to the mean. Therefore when mapping using regression-based methods such as OLS estimators, lower scores or higher scores may become unfairly biased towards the mean.

In the estimation sample, the regression based FAI-NEADL mapping algorithms over predicted the lower end of the NEADL scale and slightly under-estimated the upper boundary of the scale. The NEADL-FAI mapping algorithm was accurate in predicting the lower end of the scale but slightly under-estimated the upper end of the scale. The issue of over predicting the lower boundary and under predicting the upper boundary of the scale is common in regression-based mapping analyses (Rowen et al., 2009). For example in mapping studies of EQ-5D, Rowen et al. (2009) found that the Tobit model and CLAD estimators that were used to map the SF-36 to EQ-5D suffered from over prediction of severe health states. In their review, Brazier et al. (2010) found that in most mapping studies the level of prediction error for EQ-5D was far greater at the lower (severer health) end of the scale. Grootendorst et al. (2007) also reported a similar finding, the standard regression models over predicted utility values for patients with relatively severe disease and under predicted values for those patients at higher levels of health. Brazier et al. (2010) explained that ceiling and floor effects produce heteroskedastic residuals, causing modes to under-estimate scores for patients at ceiling and over-estimating scores for patients at the floor. In study 3a, the quantile regression and robust regression estimators were used to account for the non-normal residuals and heteroskedastic residuals but the predictions at the lower end of the scales were still poor for the FAI-NEADL mapping function.

The other sources of poor predictions reported by Brazier et al. (2010) was the strength and the degree of conceptual overlap between the two measures, and the differences in the severity ranges covered by the measurement scales. Successful mapping is achieved when there is good conceptual overlaps between the target and start measures. The target measure should cover all important aspects of health of the start measure and if there are important dimensions of one instrument not covered by the other, the performance of the mapping algorithm may be undermined (Brazier et al., 2010). In Study 3a, the examination of the items from the NEADL and FAI measures showed that there were common items across the two measures, but the NEADL had a wider coverage of important aspects of HRQoL. Exploratory factor analysis conducted in Study 3a showed that the two measures were measuring similar constructs, but the NEADL had more items loading on to the factors that were identified. The NEADL captures a wider range of extended activities of daily living compared to the FAI and this could have led to large individual level prediction errors for the FAI-NEADL mapping function. The NEADL-FAI mapping function had

smaller prediction errors compared to the FAI-NEADL mapping function because the NEADL covers almost all the dimensions in the FAI measure.

In Study 3a, the predicted values had smaller variance compared to the observed values. A systematic review by Brazier et al. 2010 also highlighted this issue. Fayers and Hays (2014) attributed the less variation in predicted values to "regression to mean". Mapping studies using regression-based approaches have also shown that the cumulative distribution function of the predicted scores is shrunk at the tails in comparison with the observed values of the target distribution (Brazier et al., 2010; Rowen et al., 2009).

The second analyses conducted in Study 3a used IRT methods to link the FAI and NEADL measures. The use of latent variable approaches such as IRT linking is considered to be a strong form of linking compared to regression based linking (Dorans, 2007). Similar to the regression-based mapping, the IRT linking conducted in Study 3a also produced accurate group mean predictions but poor individual level predictions. These findings from Study 3a were consistent with other studies that have used IRT approaches to link outcome measures. For example Byers (2004) evaluated the accuracy of a FIM-MDS conversion table that was developed using IRT methodology and also concluded that linking was accurate for producing group level predictions and not individual level predictions. Poor individual level predictions using IRT methods have been attributed to: measures not assessing similar constructs, poor conceptual overlap between measures, group variance (linking function should be population invariant), use of unsuitable IRT models for calibration, and poor scoring algorithms (Dorans, 2007). In Study 3a, 36 items from both FAI and NEADL were factor analysed together to determine whether items were measuring similar constructs and calibration was conducted on items that loaded on similar constructs. Since the IRT calibration was based on items measuring similar domains, poor item overlap between measurement scales was not considered a possible source of poor individual level predictions for conversion tables developed using IRT methods.

Fitzpatrick and Yen (2001) have argued that greater accuracy in equating outcome measures using IRT methods is gained by using more items and too few items produce poor calibrations. In Study 3a, IRT calibration of items from the two measures had more than 5 items; these might not have been enough for accurate IRT equating. Simulation studies by Fitzpatrick and Yen (2001) suggested that to obtain acceptable reliabilities and accurate equated scores, tests should have at least eight 6-

point items or at least 12 items with 4 or more score points per item. In non-equivalent designs, Angoff cited in (Chen et al., 2009) suggested that for IRT linking to produce more precise and stable calibration the common items should constitute 20% of the total items but other researchers suggested 5 to 10 items (Wright and Bell, 1984). However, other studies have demonstrated that even less than 5 common items can be used for simultaneous calibration of items (Chen et al., 2009). In Study 3a, of this present thesis, IRT simultaneous calibration of items from the FAI and NEADL measures had more than 5 common items but still the conversion tables produced poor individual level predictions.

There is evidence that the accuracy of IRT linking also depends on sample size. Fitzpatrick and Yen (2001) reported that n= 200 cases were too few to obtain precise item parameter estimates and recommended a sample size of n= 500 cases for more precision and n= 1000 cases was considered to offer even more precision. In study 3a, a sample size of n=448 was used which was slightly less than the 500 recommended by Fitzpatrick and Yen (2001), therefore this could have affected the precision of the parameter estimates of the IRT model. More research based on advanced IRT multi-dimensional models is needed to improve the accuracy of IRT linking.

Discussion of statistical methods

Based on RMSE and MAE there was comparable predictive performance of ordinary least squares, quantile regression and the robust estimators despite the other models overcoming the limitations of OLS. Using complex models rather than the OLS for mapping was of no significant gain. Using polynomial functions to model the non-linear relationships between the FAI and NEADL measures produced similar results with models that accounted for non-linearity using a quadratic term in the model. A review of mapping studies by Brazier et al. (2010) also found that simple additive models with an index score as the dependent variable and main effects of either total or dimension scores as independent variables, performed nearly as well as those for more complex models. Another study by Ghatnekar et al. (2013) also found that Ordinary least square produced the best prediction model compared to the Tobit and Censored least absolute deviation (CLAD) estimators.

Implications for future research

The mapping conducted in Study 3a using the regression- based mapping and the IRT linking showed that statistics at the group level tended to support the accuracy of the conversions compared to individual level predictions. The implications of these finding are that mapping using regression-based methods or IRT methods are suitable for the purposes of predicting group mean scores and not for predicting individual patient scores. The mapping algorithms developed in Study 3a of this thesis can be used for predicting group level statistics such as means. In economic evaluation studies, the purpose of mapping functions is also for predicting differences across groups of patients or differences between arms over time in clinical trials, and not between individual level index values (Brazier et al., 2010).

5.5.1 Limitations

The analysis conducted in Study 3a has limitations that warrant discussion. The SOS1 dataset that was used to develop the mapping algorithms was a selected cohort as it excluded patients with severe strokes. Thus the SOS1 dataset might not be representative of the stroke population for which the conversion tables are intended and the mapping algorithms might lack generalisability. There is need to develop mapping algorithms in more representative samples with complex or severe strokes.

The validation of the mapping algorithms or cross walks developed in Study 3a was conducted using 1 year (wave 2) data of SOS1 study. The mapping algorithms were developed using baseline data (within four weeks after stroke). One year may be a long period such that the performance of the algorithms might have been affected by the long time difference. At one year most stroke patients are expected to have completed the major part of their recovery. Using the same sample for external validation might have biased the validation analysis. A more precise approach would be to externally validate the mapping algorithms and conversion tables developed in study 3a in an independent sample. A cross-validation would be ideal once a suitable external dataset becomes available.

In Study 3a, the 2-parameter IRT model was used to develop conversion tables for linking measures, due to time limitations, other complex IRT models that do not assume unidimensionality were not explored. More research is needed to investigate the utility of using such IRT models. Furthermore in Study 3a IRT linking was used to develop conversion tables for harmonising between measures. Harmonising outcome

measures using IRT models can also be achieved by putting scores on the same metric e.g., IRT scores and use these in further analysis. Future research may explore harmonisation of PROMs by putting the different measures on a common metric e.g. IRT scores and use these in secondary analysis.

The other limitations was that the regression based mappings and IRT conversion tables that were developed in Study 3a were not checked for group invariance by gender hence more research is needed to establish whether these mapping functions and conversion tables do not exhibit any differential function with respect to gender.

5.6 Conclusion

In conclusion, the analysis conducted in Study 3a showed that both the regression based and IRT methods of harmonising outcome measures seem to be promising methods for predicting group means and not individual patient predictions in mapping analysis. Good conceptual overlap between measures is required for accurate mappings or conversion tables.

The next Chapter describes the third strand of the research that was conducted in Study 3b of this thesis to harmonise the GHQ-12 and GHQ-28 measures.

Chapter 6

6 HARMONISATION OF GHQ-12 AND GHQ-28 MEASURES OF PSYCHOLOGICAL DISTRESS

Study 3b: Harmonisation of GHQ-12 and GHQ-28 measures of psychological distress

6.1 Introduction

In Chapter 5, regression-based and item response theory models were used to harmonise the NEADL and FAI questionnaires. The third strand of research that was conducted in this thesis investigated in Study 3b, another harmonisation approach which uses common items across measures. As highlighted before in Chapter 3, the CIMSS study assessed psychological distress using the GHQ-12, while the SOS studies used the GHQ-28. In order to pool the two datasets, it was necessary to harmonise the GHQ-12 with GHQ-28 measure.

Harmonisation of PROMs using common items has been tried previously in data harmonisation studies. For example, the CLESA project (Minicuci et al., 2003; Pluijm et al., 2005) discussed in Chapter 2, harmonised ADL measurement scales from six countries by summing scores from the four ADL items that were common across the studies. The harmonised four-item measure showed good reliability across countries (Pluijm et al., 2005). The disadvantage of using this approach is that important items from the different scales may be lost hence there is need to establish the psychometric properties of the common items before using the harmonised measure.

Study 3b sought to investigate whether the six selected items that are common across the GHQ-28 and GHQ-12 can be used as a harmonised measure of psychological distress in stroke survivors. In Study 3b, the psychometric properties of these six items were investigated. In this present Chapter, the methods and results from Study 3b are reported. The aims of Study 3b are stated in section 6.2; the methods that were used are described in section 6.3, and the results in section 6.4. The Chapter ends with a discussion of findings and conclusions.

6.2 Aims and objectives of Study 3b

The analyses conducted in this chapter investigated:

(1) -the dimensionality

(2) -the reliability

of the selected six common items in the GHQ-12 and GHQ-28.

6.3 Method

6.3.1 Data

Baseline data from SOS1 (n=448), SOS2 (n=585), and CIMSS (n=312) was used. The details of these studies have already been provided in Chapters 1 and 3.

6.3.2 Harmonisation of GHQ-28 and GHQ-12 Measures

Details of the GHQ-28, GHQ-30, and GHQ-12 have already been provided in Chapter 1. The items in the GHQ-12, GHQ-30, and GHQ-28 measures are shown in Table 6.1. The GHQ-30 has the whole of the GHQ-12 embedded in it and harmonising these two measures can be achieved by using the 12 items common in both measures. The GHQ-28 does not have the whole of GHQ-12 embedded in it, but has six items that are common across the two measures (Table 6.1). The six GHQ-12 items embedded in the GHQ-28 domains are: “Lost much sleep over worry”, “felt constantly under strain”, “felt that you are playing useful part in things”, “felt capable of making decisions”, “been able to enjoy day to day activities”, and “been thinking of yourself as a worthless person”. In this thesis these six items were used to harmonise the GHQ-12 and GHQ-28 measures by summing the scores from these items. As previously mentioned in Chapter 4, the SOS1 used the (0, 0, 1, 1) system for scoring the GHQ-28. The CIMSS study used the Likert scoring system (0, 1, 2, and 3) for scoring the GHQ-12. The analyses conducted in Study 3b rescored the GHQ-12 in CIMSS to (0, 0, 1, and 1).

Table 6.1 Comparison of GHQ-28, GHQ-30, and GHQ-12 measures

Item	GHQ-30	GHQ-28	GHQ-12
Somatic		x	
A1: Been feeling perfectly well and in good health		x	
A2: Been feeling in need of a good tonic		x	
A3:Been feeling run down out of sorts		x	
A4:Felt that you are ill		x	
A5:Been getting any pains in your head		x	
A6: Been getting a feeling of tightness or pressure in your head		x	
A7: Been having hot or cold spells		x	
Anxiety			
B1: Lost much sleep over worry	**x**	**x**	**x**
B2: Had difficulty in staying asleep once are off	x	x	
B3: Felt constantly under strain	**x**	**x**	**x**
B4: Been getting edgy and bad tempered		x	
B5:Been getting scared or panicky for no good reason	x	x	
B6: Found everything getting on top of you	x	x	
B7: Been feeling nervous and strung up all the time	x	x	
Social dysfunction			
C1: Been managing to keep yourself busy and occupied	x	x	
C2:Been taking longer the things you do		x	
C3:Felt on the whole you were doing things well	x	x	
C4:Been satisfied with the way you have carried out your task	x	x	
C5:Felt you were playing a useful part in things	**x**	**x**	**x**
C6:Felt capable of making decisions	**x**	**x**	**x**
C7: Been able to enjoy your normal day to day activities	**x**	**x**	**x**
Depression			
D1: Been thinking of yourself as a worthless person	**x**	**x**	**x**
D2:Felt that life is entirely hopeless	x	x	
D3:Felt that life is not worth living	x	x	
D4: Thought of the possibility that you might make away with yourself		x	
D5:Found at times you could not do anything because your nerves were too bad	x	x	
D6:Found yourself wishing you were dead and away from it all		x	
D7:Found that the idea of taking your own life kept coming into your mind		x	
Been feeling unhappy and depressed	x		x
Felt could not overcome difficulties	x		x
Been taking things hard	x		
Been losing confidence in self	x		x
Been feeling reasonably happy	x		x
Been able to concentrate on whatever you are doing	x		x
Been getting out of the house as much as usual	x		
Been feeling hopeful about your own future	x		
Been finding life a struggle all the time	x		
Been able to face problems	x		x
Been feeling hopeful about your own future	x		
Spent much time chatting with people	x		
Been finding it easy to get on with other people	x		

6.3.3 Statistical analysis

The correlations of the GHQ six common items were investigated using Spearman's rank correlation coefficient. Internal consistency was measured using Cronbach's α, and was considered acceptable if Cronbach's α was above 0.7 (Streiner et al., 2014). Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) was used to investigate the dimensionality of the six GHQ common items. Details of EFA and CFA have already been provided in Chapters 4 and 5. The Kaiser-Meyer, Olkin (KMO) measure of sampling Adequacy (Kaiser, 1974) and Bartlett's test (Bartlett, 1954) of sphericity were used to determine the suitability of conducting factor analysis in the three datasets. A KMO value of 0.06 or more and a significant Bartlett test supported the use of factor analysis. The eigenvectors (factors) were rotated in an attempt to achieve a simple structure which may be easier to interpret. The underlying factors were assumed to be correlated hence geomin rotation which is an oblique rotation was used. The best factor solution was selected based on the examination of a scree plot and goodness-of-fit indices. The criterion for goodness-of-fit was set at: root mean square error of approximation (RMSEA) ≤ 0.08; Tucker-Lewis index, ≥ 0.95 (Tucker and Lewis, 1973); comparative fit index, ≥ 0.95) (Gelman et al., 1998) and the extent of interpretability of the emerging factors. EFA for ordinal items and CFA was conducted using Mplus version 7 (Muthén and Muthén, 2012).

6.4 Results

The completion rate on all 6 items was good, ranging from 97% - 98% across the three datasets. A total of n=1316 out of 1345(97.8%) completed all six items across the three datasets.

Assumptions testing

The Kaiser-Meyer-Olkin test produced a value of 0.77 for the SOS2 dataset, 0.76 for the SOS1 dataset and 0.83 for the CIMSS dataset, exceeding the recommended value of 0.6 (Kaiser, 1974) for factor analysis, supporting strong partial correlations between the items. The Bartlett's test of sphericity was statistically significant in all the three studies: SOS2 [Chi-square = 606.06, df =15, $p < 0.001$]; SOS 1 [Chi-square = 441.85, df =15, $p < 0.001$]; CIMSS [Chi-square = 491.11, df =15, $p < 0.001$]; suggesting that the 6 items were sufficiently correlated. The data met the assumption of factor analysis thus it was suitable to use factor analysis in Study 3b.

6.4.1 Correlations of the six common GHQ items

Table 6.2 shows the correlations of the six items across the three datasets. The item correlations were > 0.3 in all three datasets.

Table 6.2 Pairwise correlations of the six GHQ common items by study

SOS1						
Item	Lostsleep	Strain	Useful	Decision	DaytoDay	Worthless
Strain	0.64	1				
Useful	0.44	0.47	1			
Decision	0.39	0.49	0.61	1		
DaytoDay	0.44	0.47	0.73	0.50	1	
Worthless	0.57	0.51	0.51	0.32	0.38	1
CIMSS						
Item	Lostsleep	Useful	Decision	Strain	Worthless	DaytoDay
Useful	0.33	1				
Decision	0.38	0.80	1			
Strain	0.66	0.38	0.41	1		
Worthless	0.48	0.53	0.55	0.59	1	
DaytoDay	0.53	0.75	0.78	0.53	0.60	1
SOS2						
Item	Lostsleep	Strain	Useful	Decision	DaytoDay	Worthless
Strain	0.64	1				
Useful	0.42	0.44	1			
Decision	0.49	0.51	0.58	1		
DaytoDay	0.46	0.42	0.68	0.52	1	
Worthless	0.43	0.66	0.53	0.46	0.53	1

6.4.2 Exploratory factor analysis of the six common GHQ items

The results from the Exploratory Factor Analysis (EFA) conducted in the SOS1, SOS2 and CIMSS datasets are shown in Table 6.3. Across all the three datasets(SOS1, SOS2, CIMSS) the Chi-square difference test showed a significant change from a one-factor structure to a two-factor structure, but no significant change from a two-factor structure to a three-factor structure suggesting a two-factor structure for the six GHQ items. The scree plots also suggested a two factor structure in all three datasets with a pronounced ‘elbow’ at two factors apparent in all three plots (Figure 6.1).

Table 6.3 Comparison of various factor solutions: Exploratory Factor Analysis of the six GHQ items

	SOS1 **Chi-Square, DF,** ***p*** **value**	**CIMSS** **Chi-square, DF,** ***p*** **value**	**SOS2** **Chi-square, DF,** ***p*** **value**
1-factor against 2-factor	30.05, df=5, $p<0.001$	135.67, df=5, $p<0.001$	38.12, df=5, $p<0.001$
2-factor against 3-factor	4.54, df=4, $p=0.337$	4.17, df=4, $p=0.384$	2.97, df=4, $p=0.562$

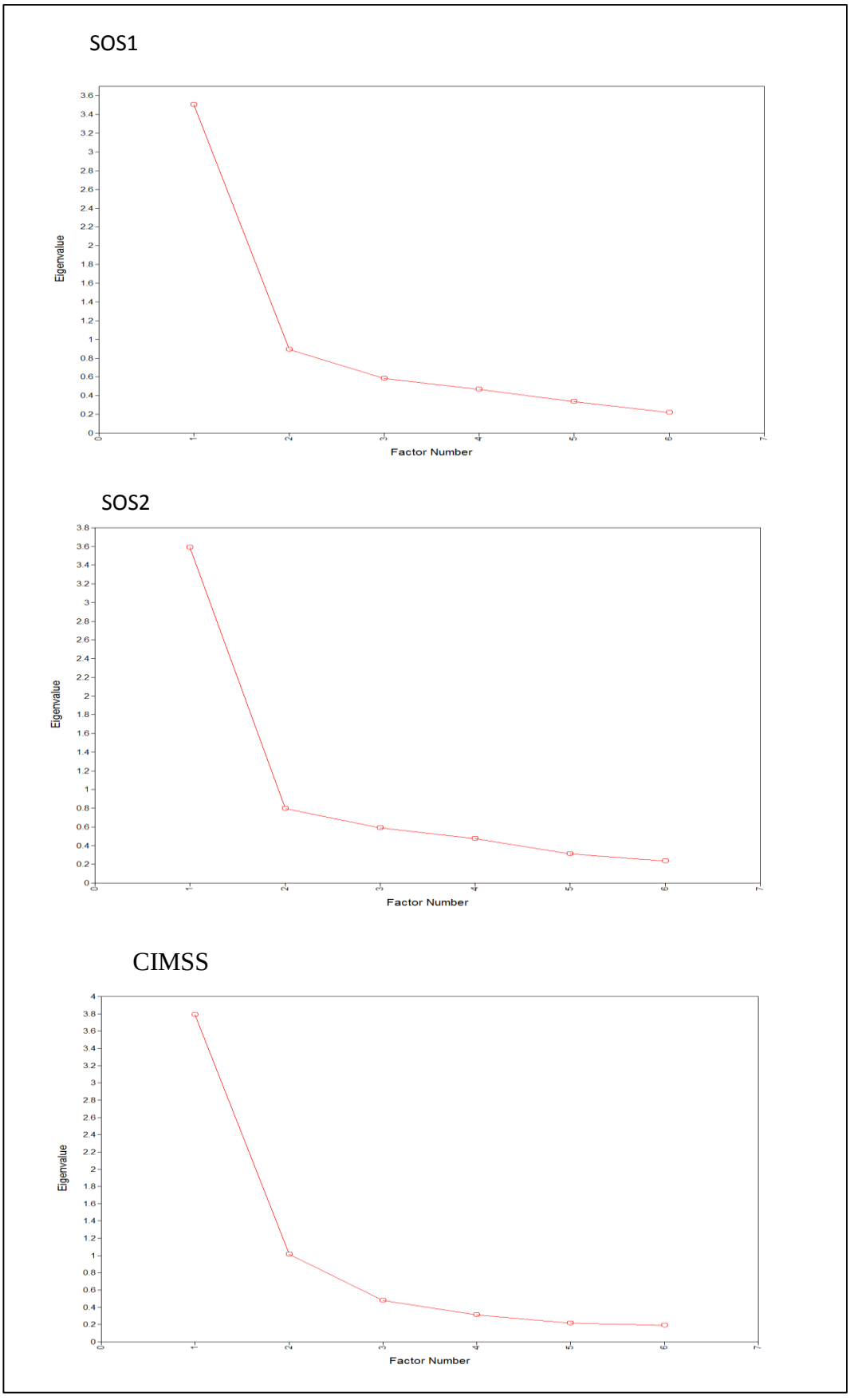

Figure 6.1 Scree plots for the six common items: SOS1, SOS2, CIMSS

The goodness-of-fit results from the EFA are shown in Table 6.4. The 2-factor structure produced acceptable goodness of fit indices, CFI >0.95 and TLI >0.95, and RMSEA < 0.05. Based on the examination of the scree plots, CFI and TLI >0.95, and interpretability of the factors, a two factor structure was preferred for the six GHQ common items.

Table 6.4 Goodness-of-fit indices: Exploratory factor analysis of the common six GHQ items

	Chi-square	RMSEA	SRMR	CFI	TLI
SOS1					
1 factor	37.70, df 9, P<0.001	0.09(0.06-0.11)	0.07	0.96	0.93
2 factor	4.263,df 4, P=0.3716	0.01(0.00-0.07)	0.03	1.00	0.99
SOS2					
1 factor	44.27, df=9, p<0.001	0.08(0.06-0.11)	0.07	0.96	0.94
2 factors	3.498,df=4,p=0.478	0.00(0.00-0.06)	0.02	1.00	1.00
CIMSS					
1 factor	169.59,df=9,P<0.001	0.24(0.21-0.28)	0.09	0.91	0.85
2 factors	3.972,df=4,p=0.409	0.00(0.00-0.09)	0.01	1.00	1.00

The Geomin rotated factor loadings of the six common GHQ items for the two factor solution are shown in Table 6.5. As expected the three items extracted from the anxiety and depression dimensions of the GHQ-28 and GHQ-12 scales loaded on to the same factor, and these were: “lost sleep”, “strain”, and “felt worthless”. This factor was named “anxiety and depression”. In SOS1 and CIMSS, the three items extracted from the social function subscales of the GHQ-12 and GHQ-28 loaded onto the same factor, and the factor was named “social dysfunction”. In the SOS2 study the item on “strain” had factor loading greater than 1 and this could be due to multi-collinearity.

Table 6.5 Geomin rotated factor loadings of the six GHQ common items by study

	SOS1		SOS2		CIMSS Project	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 2	Factor 1
Lost sleep	**0.866***	-0.012	0.363*	0.394*	0.047	**0.717***
Strain	**0.671***	0.160	**1.190***	-0.001	-0.010	**0.892***
Useful	-0.007	**0.980***	-0.006	**0.839***	**0.891***	-0.029
Decisions	0.200	**0.525***	0.175	**0.605***	**0.909***	0.004
Day-to-day	0.149	**0.671***	-0.003	**0.806***	**0.722***	0.265*
Worthless	**0.549***	0.183	0.333*	**0.492***	0.364*	**0.469***

6.4.3 Reliability

The internal consistency of the six common GHQ items measured by the Cronbach's α was: 0.81 in CIMSS; 0.71 in SOS1; and 0.73 in SOS2 suggesting acceptable reliability. The Cronbach's α for the "anxiety/depression" and "social" subscales were 0.67 and 0.77 respectively in the CIMSS, 0.61 and 0.65 in SOS1, and 0.62 and 0.64 in the SOS2. In all three datasets the reliabilities of the subscales were below the acceptable levels of greater than 0.7 recommended by Streiner (2014).

6.4.4 Confirmatory factor analysis of the six GHQ common items

Table 6.6 shows the goodness of fit results of the six common GHQ items from the confirmatory factor analysis of the two-factor structure. All the goodness of fit indices supported the two-factor structure for the six common items: CFI >0.95 and TLI>0.95 and RMSEA < 0.05. The two extracted factors showed moderate positive correlation in all three studies suggesting the existence of a higher order factor (Table 6.7). The high correlations of the two extracted factors may be suggesting the existence of a single underlying factor or unidimensionality for the six GHQ items. The two factor model could be due to the "wording effects" of the positive and negative worded items since the positive items loaded onto one factor and the negative items loaded on to the other factor.

Table 6.6 Confirmatory Factor Analysis of the six common GHQ items

	Chi-square value *P* value	RMSEA	CFI	TLI
SOS1	7.07 , p <0.01	0.00(0.00-0.05)	1	1
SOS2	20.17, p <0.01	0.05(0.02-0.08)	0.99	0.98
CIMSS	49.92, p <0.001	0.13(0.09-0.169)	0.98	0.96

Table 6.7 STDYX Standardisation factor loadings: Confirmatory Factor Analysis of the common six GHQ items

	SOS1			SOS2			CIMSS project		
	Estimate, S.E.		*P*-Value	Estimate, S.E.		*P*-Value	Estimate, S.E.		*P*-Value
Factor 1 by									
Lost sleep	**0.778**	0.057	<0.001	**0.714**	0.053	<0.001	**0.730**	0.033	<0.001
Strain	**0.801**	0.057	<0.001	**0.814**	0.046	<0.001	**0.801**	0.031	<0.001
Worthless	**0.701**	0.062	<0.001	**0.788**	0.050	<0.001	**0.806**	0.037	<0.001
Factor 2 by									
Useful	**0.898**	0.046	<0.001	**0.812**	0.038	<0.001	**0.830**	0.026	<0.001
Decision	**0.686**	0.059	<0.001	**0.736**	0.049	<0.001	**0.885**	0.036	<0.001
Day-to-day	**0.799**	0.051	<0.001	**0.787**	0.040	<0.001	**0.921**	0.021	<0.001
Factor 2 with Factor 1	**0.718**	0.057	<0.001	**0.788**	0.048	<0.001	**0.683**	0.037	<0.001

6.5 Discussion

In this chapter the dimensionality and reliability of the six common items of the GHQ-12 and GHQ-28 were investigated. The analysis was conducted in order to evaluate whether the GHQ-12 and GHQ-28 measures can be harmonised by using these six common items. Pooling GHQ scores from the SOS studies and CIMSS datasets required harmonising the GHQ-12 and GHQ-28 measures. The strength of the analyses conducted in this chapter is that the psychometric properties of the six common GHQ items were evaluated in multiple stroke cohorts assessing the reproducibility of the results across the different cohorts. Exploratory factor analysis revealed two factors that were confirmed using confirmatory factor analysis in all three studies. The positively worded items loaded on to the same factor and the negatively worded items loaded onto the other factor. This phenomenon has been termed "method or wording effects" (Hankins 2008). The two factors identified in Study 3b may be art factual grouping representing "wording effects". Wording effects of the GHQ-12 measure have been reported in other studies (Hankins, 2008; Smith, 2013; Molina, 2014). In previous studies, exploratory factor analysis of the GHQ-12 without accounting for the word effects produced two factors representing the positively and negatively worded items. Adjusting for word effects in the factor analysis of the GHQ-12 resulted in a unidimensional measure. In Study 3b, exploratory factor analysis of the six GHQ items was conducted without accounting

for word effects but confirmatory factor analysis showed that the two extracted factors were highly correlated suggesting the existence of a higher order factor or a single underlying factor for the six GHQ items thus in Chapter 7, the six common GHQ items were considered unidimensional and their summed score was used.

It is worthwhile comparing the six common GHQ items with other reduced GHQ measures proposed in literature. Smith et al. (2010) reported a GHQ-6 measure that was obtained from GHQ-12 using factor analysis. The GHQ-6 had items "Been able to face up to problems"; "Feeling reasonably happy", "Overcome difficulties"; Unhappy, depressed"; "Losing confidence". The six items that were proposed by Smith et al. (2010) are different from the GHQ-6 used in this present study to harmonise the GHQ measures. Kalliath et al. (2004) proposed a GHQ-8 item derived from GHQ-12 with the following items 4, 7,8,12, 6, 9, 10, and 11 of the GHQ-12 items. The 6 items that were used to harmonise the GHQ-12 in this present study had 4 items overlapping with the Kalliath's et al. (2004) GHQ-8 measure. The overlapping items are "Lost much sleep over worry", "felt capable of making decisions", "been able to enjoy day to day activities", "been thinking of yourself as a worthless person".

6.5.1 Limitations

Since all three stroke datasets explored in study 3b are restricted to patients with less severe strokes, generalisation to all stroke patients is not possible. More importantly this present study only assessed the reliability and dimensionality of the six common items. More research is needed to investigate the responsiveness and sensitivity of the six GHQ items. Furthermore there were other items in the GHQ-12 and GHQ-28 that had different wording but similar meaning, these could be harmonised and included together with the six common items. More work is needed to increase the number of similar items by harmonising the other similar items of the GHQ-12 and GHQ-28 measures.

6.6 Conclusion

The findings of this chapter suggested that the six common items across the GHQ-12 and GHQ-28 measures assess two subscales that were labelled "anxiety/depression" and "social function" but these two factors were highly correlated suggesting the existence of a higher order factor. The common six GHQ items showed good reliability but the subscales had only moderate reliability.

The next Chapter describes the fourth strand of research that was conducted in Study 4a of this thesis to illustrate the benefits of harmonising data and using the data for comparative research.

Chapter 7

7 PATTERNS OF EARLY DISABILITY AFTER STROKE: A MULTI-GROUP LATENT CLASS ANALYSIS

Study 4a: Patterns of early disability after stroke: Application of multi-group latent class analysis

The following publication has arisen from the preliminary work from this chapter:

> Munyombwe T., Hill. K.M., Knapp. P., West.R.M. (2014). Mixture modelling analysis of one-month disability after stroke: stroke outcomes study (SOS1). *Quality of Life Research*, 23(8), pp. 2267-2275.

As the first author Theresa Munyombwe carried out all the statistical analyses and prepared the first draft of the manuscript. The other authors provided feedback on the statistical analyses and proof read drafts of the manuscript.

7.1 Introduction

Chapter 6 reported the research that was conducted in Study 3b to investigate the psychometric properties of the six items that are common in the GHQ-28 and GHQ-12 to determine whether these items could be used to harmonise the two measures. The six GHQ items were found to have sound psychometric properties in terms of construct validity and reliability in all three datasets (SOS1, SOS2, and CIMSS) thus a harmonised psychological distress variable was derived from these six items and used in the fourth strand of research in this thesis. The fourth strand of research was to demonstrate the benefits of multi-group analysis of the harmonised datasets in Study 4a using a latent class analysis framework. Study 4a compared patterns of disability in two stroke cohorts and the factors associated with these patterns. The analyses conducted in Study 4a illustrated the benefits of harmonising datasets for comparative purposes. Preliminary results of the analyses reported in this chapter were published in

2014 (Munyombwe et al., 2014). The preliminary analyses, investigated initial disability patterns after stroke using the SOS1 dataset (n-448). The work in this chapter covers similar ground but extends the analysis by comparing the latent disability patterns and factors associated with disability across two stroke cohorts. A better understanding of disability patterns in stroke survivors and the factors associated with them is important in creating person-centred approaches to health management and outcome optimisation of stroke patients (Mayo et al., 2015). To the best of my knowledge, there are few studies that have used person- centred approaches to classify disability patterns in stroke survivors. In Study 4a, person-centred approaches provided a framework for classifying stroke patients using multiple disability measures. Findings from study 4a will add to existing literature by using an advanced statistical technique (Multi-Group Latent Class Analysis) to identify patterns of disability in stroke survivors using multiple disability measures and also comparing the latent classes across different stroke cohorts.

In this Chapter, the analyses that was conducted in Study 4a is reported. The Chapter begins by stating the aims and objectives of Study 4a. The structure of this chapter follows the order of analysis as follows: section 7.3 provides a description of the methods that were used in this Chapter, section 7.4 the results of the analyses, the chapter ends with a discussion of findings from this chapter in section 7.5.

7.2 Study aims

The primary aim of the analyses conducted in Study 4a was to investigate if the disability latent class structure within each cohort (SOS1 and CIMSS) was consistent across both datasets. Baseline factors that influence these disability patterns were also investigated.

Research questions

- Are the underlying latent disability structures consistent across the SOS1 and CIMSS datasets?
- Are the factors that influence class membership similar across the SOS1 and CIMSS datasets?

It was hypothesised that the disability patterns and the factors associated with these patterns were similar across the two datasets. Note that it was important to first

establish the number of latent classes within each dataset before multi-group analyses. Here the number was strongly guided by the Bayesian Information Criterion (BIC), LMR *p*- value (Lo et al., 2001), and confirmed with the clinical interpretability of the classes. See later in this chapter for justification of the choice of information criterion.

7.3 Methods

7.3.1 Data Sources

The analyses conducted in Study 4a used harmonised baseline (within a month after stroke) data from SOS1 (n=448) and CIMSS (n=312) studies. A detailed description of the characteristics of these studies has already been provided in Chapter 3 of this thesis. The initial analysis was conducted in separate datasets and this was followed by multi-group analysis using the combined data from the SOS1 and CIMMS studies.

7.3.2 Measures

The disability patterns investigated in this chapter were based on: physical, social, and psychological function post-stroke. Multiple disability domains were considered because stroke is a heterogeneous condition; different people have different forms of disability hence multiple measures are needed to capture a broad range of disabilities that affect stroke survivors (Kelly-Hayes et al., 1998). Disability patterns were investigated using summed scores of BI, NEADL, GHQ-28 subscales in SOS1, and BI, NEADL, GHQ-12 subscales in CIMSS. As previously mentioned in Chapter 6, the SOS1 used the (0, 0, 1, 1) system for scoring the GHQ-28, while the CIMSS study used the Likert scoring system (0, 1, 2, 3) for scoring the GHQ-12. The separate study analyses conducted in this chapter was conducted using the scoring systems in the original studies, but the GHQ-12 items in CIMSS were rescored to (0, 0, 1, and 1) for the multi-group analyses.

The multi-group analysis of the SOS1 and CIMSS was conducted using the harmonised GHQ (defined as the sum of the six common item scores), BI, and NEADL subscales. Details of the BI, NEADL, GHQ-28, and GHQ-12 measures have already been provided in Chapter 1. The psychometric properties of the harmonised six item GHQ were investigated in Chapter 6 of this thesis and the measure showed good reliability and content validity. The summed scores of the PROMs were used to determine the disability patterns and these summed scores were treated as continuous

variables. A classification of patients was based on the thresholds of the measures shown in Table 7.1 and also on the mean scores in the samples. Goldberg et al. (1998) recommended that classification of patients based on the GHQ-28 can also be guided by the mean GHQ-28 scores of the samples.

Table 7.1 Cross walking between SOS1 and SOS2 studies, Physical, Social and Psychological function measures

Measure	Commonly used cut-off points
GHQ-28 total	-Score of 4 or more points out of 28, in a (0, 0, 1, 1) scoring suggests psychological distress. (Sterling, 2011)
GHQ-28 subscales	-No established thresholds were found for subscales
GHQ-12	-Score of 3 or more in a (0,0,1,1) scoring suggests psychological distress
BI	-20-point version, ≥19/20 (independence) Kwakkel et al. (2011) -Severe 0-9 -Moderate 10-15 -Mild 15-19 -Independent >19
NEADL	-Threshold of 18 or more has been used to determine (Yohannes et al., 1998)
NEADL subscales	-No thresholds were found for both total and subscales

7.3.3 Sample size

Latent class analysis was conducted with a sample size of n=448 for SOS1 and n=312 for the CIMSS datasets. In biological studies, sample sizes of 500 -1000 have been suggested as ideal for conducting mixture modelling (Muñoz and Acuña, 1999). Finch and Bronk (2011) also suggested a sample size of n=500 for latent class analysis. In Study 4a, both the SOS1 and CIMSS datasets had sample sizes of less than n=500 but the multi-group latent class analysis of the combined SOS1 and CIMSS datasets had a total sample size of n=760. A simulation study by Wurpts and Geiser (2014) found that having more high quality indicators and a covariate that is strongly correlated with class membership compensated for having small sample sizes. In Wurpts and Geiser (2014) study, models with 4 or 5 indicators had convergence problems and regression coefficients for the multinomial logistic regression for

identifying predictors of class membership were biased for small sample sizes. In this present study, following recommendations by Wurpts and Geiser (2014), more than 5 class indicators were used to determine the best latent class structure, and a combined sample of n=760 was considered adequate for the 5 covariates that were used to predict class membership.

7.3.4 Descriptive analysis

The correlations of the summed scores of the physical, social and psychological domains were determined using Spearman's correlation coefficients since the data were considered to be ordinal.

7.3.5 Statistical modelling

Latent variable modelling was used in Study 4a to analyse baseline PROMs data. Two steps were used to accomplish the purpose of Study 4a. In the first step, the best measurement model for data was determined in each dataset separately. In step 2, the datasets from the SOS1 and CIMSS studies were analysed simultaneously using methods for analysing multiple groups. Details of these two steps are described in the next sections of this Chapter.

7.3.6 Selection of a measurement model for the SOS1 and CIMSS datasets

The initial latent variable analysis was to determine whether the underlying latent variable(s) measured by the physical function, social and psychological indicators in the SOS1 and CIMSS datasets was categorical or dimensional. The term "latent" refers to an unobserved variable. The measurement model that best fits the datasets between a latent class model (LCA) that assumes a categorical underlying latent variable and a factor analytic (FA) model that assumes a continuous underlying latent variable was determined in each study separately. Both the categorical and continuous latent variable models explain the co-variances between observed variables (Lubke and Neale, 2006). It was important to first determine the correct measurement model for the data because conducting latent class analysis may result in an over extraction of classes if the sample is homogeneous and covariance's of observed variables are due to underlying continuous factors (Lubke and Neale, 2006). Similarly, conducting an exploratory factor analysis might result in over extraction of factors if the sample is heterogeneous. Details of factor analytic models have already been provided in Chapters 2 of this thesis and in this Chapter details of mixture modelling will be provided.

7.3.6.1 Mixture modelling

Mixture modelling is a statistical technique that can be used to analyse data with unobserved heterogeneity (Muthén and Asparouhov, 2002). The advantage of using mixture modelling is that it is a person-centred approach that may yield clinically interpretable classes that can be used for targeted treatment of patients. Person-centred approaches focus on relations among individuals aiming to put individuals into groups of individuals who are similar to each other , and different from those in other groups (Lubke and Muthén, 2005). The utility of using patient-centred approaches in classifying individuals has been demonstrated in psychology (Ploubidis et al., 2007; Croudace et al., 2003) and in stroke (West et al., 2010). Patient-centred approaches have also been helpful in Randomised Control Trials (RCTS) in identifying subgroups of people for whom treatments were effective.

The finite mixture models include: Latent class analysis (LCA: (Lazarsfeld et al., 1968)) and Latent Profile Analysis (LPA:(Bartholomew et al., 2011; Pastor et al., 2007)). The class indicators are categorical for LCA and continuous for LPA (Vermunt and Magidson, 2004), The classic LCA and LPA models have a single categorical latent variable (Lubke and Muthén, 2005). Unlike traditional cluster analysis which classifies individuals according to arbitrary distances, Latent class analysis or Latent profile analysis classifies individuals based on posterior probabilities estimated from a statistical model, thus accounts for uncertainties of class membership. A diagrammatic presentation of the general latent class model is shown in Figure 7.1.

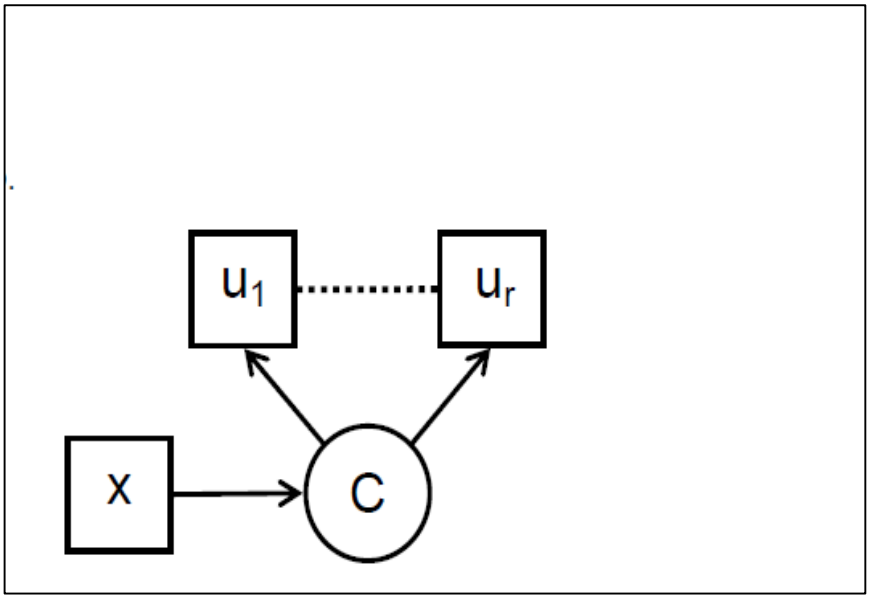

Figure 7.1 Diagrammatic representation of a latent class model with covariates

Boxes U_1 to U_r represent the observed items or indicators. These observed items or indicators can be categorical, continuous, count, censored, or nominal. The circle, "C" represents the underlying latent categorical latent variable with "K" classes. The arrows pointing to the boxes indicate that the indicators are measuring the latent variable. The box with "X" in the middle represents the covariates. In this present study the indicators were summed totals of the outcome measures, and covariates that were measured at baseline were included into the model to determine the factors that influence class membership. Details of the covariates are given in later sections of this Chapter.

The LCA model with categorical indicators/items has two types of parameters, conditional item probabilities and class probabilities. The class probabilities show the relative size of each class and the conditional item probability shows the probability of endorsing an item for an individual in a particular class. The relative class sizes indicate the prevalence of the subpopulation in the target population (Pastor et al., 2007) The parameters of the LPA models are the class sizes, means, variances and covariances.

The general structure of a finite mixture model (Vermunt and Magidson, 2004) with K classes can be expressed as follows:

$$f(\mathbf{y_i}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \qquad \text{Equation 7.1}$$

Where $\mathbf{y_i}$ is a vector of class indicator variables; K is the number of classes; π_k denote the probability of belonging to a class, $\boldsymbol{\theta}$ is the underlying latent variable, and $\boldsymbol{\Sigma_k}$ is the covariance matrix, and ($\boldsymbol{\mu_k}$) the mean vector. In this present study, the $\mathbf{y_i}$ represented the summed disability scores. Equation 7.1, states that the joint densities of $\mathbf{y_i}$ given the model parameters $\boldsymbol{\theta}$ is assumed to be a weighted mixture of class-specific densities $f_k(\mathbf{y_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$. Generally, if $\mathbf{y_i}$ variables are continuous variables the class-specific densities $f_k(\mathbf{y_i}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ are usually assumed to be multivariate Normal. The full multivariate Gaussian mixture model for continuous outcomes estimates the mean vector ($\boldsymbol{\mu_k}$) and covariance matrix ($\boldsymbol{\Sigma_k}$) for each class separately. The most complex model allows the mean vector ($\boldsymbol{\mu_k}$) and covariance matrix ($\boldsymbol{\Sigma_k}$) to vary across classes. Parsimonious models can be obtained by constraining parameters to be equal across classes. Various special cases are obtained by making restrictions on the covariance matrix ($\boldsymbol{\Sigma_k}$). The common restrictions include equal covariance of indicators across classes, diagonal covariance matrices and both equal and diagonal covariance matrices Pastor et al. (2007). Details of the various parameterisations of the covariance matrix $\boldsymbol{\Sigma_k}$ are shown in Table 7.2.

In Table 7.2 a parsimonious form of the covariance matrix is shown by model A, where variances are allowed to differ across indicators (σ_i^2) within a class, but are constrained to be equal across classes (Pastor et al., 2007). In model A the covariances are set to zero, that is the indicators are constrained to be uncorrelated both within and across classes. In the classic LCA and LPA the class indicators are constrained to be uncorrelated to each other both within classes. This is known as the conditional independence assumption. The conditional independence assumption assumes that all the co-variation between observed indicators is due to differences between classes and the observed indicators do not co vary within classes (Lubke and Muthén, 2005).

The other covariance matrix (Models B to E) shown in Table 7.2 put more constraints on the covariance matrix as shown in Table 7.2, the details of these covariance forms are provided by Pastor et al. (2007). Vermunt and Magidson (2004), provides a more detailed account of latent variable mixture models.

Table 7.2 Five different parameterisations of covariance matrix. Taken from Pastor et al. (2007)

Model	Σ_k
A	$\begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_r^2 \end{bmatrix}$
B	$\begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_r^2 \end{bmatrix}$
C	$\begin{bmatrix} \sigma_{1k}^2 & & & \\ 0 & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_{rk}^2 \end{bmatrix}$
D	$\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_{rk}^2 \end{bmatrix}$
E	$\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21k} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1k} & \sigma_{r2k} & \cdots & \sigma_{rk}^2 \end{bmatrix}$

7.3.6.2 Determining the number of latent classes

A well-known problem with mixture models is that they tend to converge on local solutions, rather than a global solution (McLachlan and Peel, 2004). To overcome problems of local solutions, several random starts values of 1000 or more can be used for model estimation. A series of models with increasingly number of latent classes are fitted and the optimal number of latent classes is determined by comparing (k-1) and k class models. Muthén (2003) recommends that when considering a plausible set of models it is wise to utilise a combination of statistical indices. Several goodness-of-fit indices are used to compare various class solutions and these include: the Akaike information criteria (AIC;(Akaike, 1998)); Bayesian Information Criteria (BIC; (Schwarz, 1978)); *p*-value from Lo-Mendell–Rubin (LMR) likelihood ratio test (Lo et al., 2001), and the bootstrap likelihood ratio tests (BLRT; (McLachlan and Peel, 2004)).

The LMR likelihood ratio test compares nested models. A significant *p*-value of the LMR likelihood test indicates a significant improvement in model fit from the (k-1) model compared to the k class model, and suggests the rejection of a (k-1) class

model in favour of the k class model. A non-significant *p*-value of the LMR likelihood ratio test suggests no significant improvement in the model, and the (k-1) class model can be kept. Similarly a significant BLRT *p*-value indicates rejection of the (k-1) class model in favour of the k class solution. For mixture models with different class solutions, the model with lowest a BIC value is considered the best fitting model. Simulation studies by Nylund et al. (2007) have shown that the BIC and BLRT perform best compared to other goodness of fit indices such as AIC. Marsh et al. (2004) recommended that the best class solution should be selected based on both statistical indices and *clinical interpretability* of the models because a model may have the best fit statistically but the emerging classes may not be *clinically meaningful*. Model fit adequacy can also be conducted using residuals or the differences between the observed and fitted values. A large number of significant residuals indicate that the model does not fit the data well.

7.3.6.3 Examining classification quality

Classification quality is also evaluated using the entropy statistic (Celeux and Soromenho, 1996) and mean posterior probabilities for each group. Higher probability values for each group indicate better classification and stronger separation. The probability π_k of each person belonging to a class is calculated using equation 7.2.

$$\pi_{k|\boldsymbol{y_i}} = \frac{\pi_k f_k(\boldsymbol{y_i}|\boldsymbol{\mu_k}, \sum \boldsymbol{k})}{\sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{y_i}|\boldsymbol{\mu_k}, \sum \boldsymbol{k})} \qquad \text{Equation 7.2}$$

Where y_i is a vector of class indicator variable, k is the number of classes, μ_k is the mean vector, and Σ_k the covariance. The entropy statistic **E** is calculated using the posterior probabilities and the formulae for **E** is shown in equation 7.3.

$$\boldsymbol{E} = 1 - \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} (-\pi_{k|\boldsymbol{y_i}} ln \pi_{k|\boldsymbol{y_i}})}{N ln K} \qquad \text{Equation 7.3}$$

Where, N is the sample size

Entropy ranges from 0 to 1 and higher values indicate better classification. Values of entropy 0.8 suggest high classification and 0.6 medium, and 0.4 low entropy (Clark and Muthén, 2009).

7.3.6.4 Factors associated with class membership

To better understand the characteristics of the latent classes, the mixture model described in equation 7.1 can be extended by including covariates to predict class membership. Once the best class solution is selected, covariates can be added to the model. The effects of covariates on predicting latent class membership can be investigated using a multinomial logistic regression model.

7.3.6.5 Application of factor analysis and mixture modelling to determine the measurement model for the data in Study 4a

Exploratory factor analysis conducted in the SOS1 dataset was based on the summed scores of BI, NEADL, and GHQ-28 subscales, while in the CIMSS dataset summed scores of BI, NEADL, and GHQ-12 subscales were used. Geomin rotation was used to extract the factors. The total scores from the different measures were considered ordinal thus a robust weighted least squares (WLSM) estimator was used for the analyses. The robust WLSMV estimator is considered to be superior to normal theory based maximum likelihood (ML) when ordinal observed variables are analysed (Li, 2014). The best factor model was selected based on the Bayesian Information criteria (BIC;(Schwarz, 1978)), scree plot, and the *clinical interpretability of the factors*. The factor models with the least BIC and with clinically interpretable factors were preferred. Factor analysis was conducted using Mplus version 7 (Muthén and Muthén, 2012).

The latent class analysis conducted in Study 4a of this thesis was exploratory with no priori hypothesis regarding the number or nature of the latent classes underlying the data. In each dataset, a series of latent class models each differing in the number of classes were fitted. In SOS1 dataset, LCA was based on the summed scores of BI, NEADL, and GHQ-28 subscales, while in the CIMSS dataset, summed scores of NEADL subscales, BI, and GHQ-12 subscales were used. Sensitivity analyses were conducted in each dataset separately to determine the effect of dropping outcome measures. In separate study analyses, LCA was first conducted with outcome measures that were collected by the studies. The second analysis used measures that were common across the studies. To ensure convergence on a global solution rather than a local solution, several random starts values of up to 500 were used for each model estimation. The latent class models were selected based on lower BIC, LMR *p*-value (Lo et al., 2001), and confirmed with the *clinical interpretability* of the classes.

The BIC was preferred instead of AIC because as mentioned earlier, simulation studies by Nylund (2007) found that the BIC performed better than the AIC. The mean class profiles of the models were examined to determine whether the best model was clinically meaningful or interpretable. The class sizes and proportions were also examined since an over-extraction of classes can result in small and non-distinct classes (Masyn, 2013). It was hypothesised that the classes would demonstrate varying levels of disability from mild to severe. The qualities of the classification were also assessed by examination of the entropy statistic and entropy values > 0.9 were indicative of good classification.

After identifying the best latent class solution in each dataset, baseline variables that were considered to be associated with class membership were included in the latent class models as covariates. These baseline covariates are shown in Table 7.3. The associations of the baseline covariates and class membership were investigated using a multinomial logistic regression and effects were reported as regression coefficients and 95% confidence intervals.

Table 7.3 Covariates used in multinomial models

Variable	Coding
Age	-
Sex	Male=1, female=2
Previous stroke	Yes=1, No=0
Living alone before stroke	Yes=1, no=0
Urinary incontinence	Yes=1, No=0

The patient characteristics across the latent classes were compared using the Pearson's Chi-square test for categorical data and Analysis of Variance (ANOVA) for continuous data. Multiple testing was corrected by using Bonferroni adjusted *p*-values. Latent class analyses were conducted in Mplus software and the details of the Mplus codes that were used are provided in Appendix F.

After identifying the best factor model and the best latent class model in each dataset, the BIC was used to decide between the two models. Lubke and Neale (2006) suggested that a comparison of the BICs from the latent class models and exploratory factor models should indicate the better model. Following these guidelines, in this present study the BIC values for the best fitting factor model and the best mixture model were compared and the model with the lowest BIC was preferred. In both the

SOS1 and CIMSS dataset, the best measurement model, based on lowest BIC were the latent class models, suggesting that the underlying latent construct measured by combining physical, social and psychological domains was better modelled by a latent categorical variable rather than a continuous factor. Thus the latent class models were used in this present study for analysing the patterns of disability using methods for multiple groups. The mixture modelling and factor analysis were conducted using Mplus version 7(Muthén and Muthén, 2012).

7.3.7 Multi-Group Latent Class Analysis of SOS1 and CIMSS datasets

After identifying the latent class model as the best measurement model for the data in both the SOS1 and CIMSS datasets, the two combined datasets were analysed using a multi-group approach, Multi-Group Latent Class Analysis (MG-LCA). A multi-group approach was selected because it can account for heterogeneity across the different stroke cohorts if it exists, and also comparisons of disability patterns across the different cohorts can be made. Simply aggregating data from different studies without assessing heterogeneity may result in biased results (Verma et al., 2009). MG-LCA also known as simultaneous latent class analysis is an extension of the single group latent class analysis for single groups. It was originally developed for the analysis of latent structures of categorical latent variables across different number of groups (Kankaraš et al., 2010). Separate study analysis to determine the appropriate number of latent classes in each study is a prerequisite of MG-LCA (McCutcheon, 2002). It is conducted to identify the number of latent classes that best fits the data in each group. MG-LCA framework is a flexible framework that can accommodate varying number of latent classes across groups, whilst still assuming measurement invariance (Kankaraš et al., 2010). A MG-LCA model with the k classes can be specified across groups even if the other group(s) have (k-1) or fewer numbers of classes. The class(s) that do not exist in the other group(s) will be empty in those particular groups. For example a five class model can be fit in a multi-group framework and a group with four classes will have no observations in the fifth class.

The main advantage of using the MG-LCA is that the framework can be used to test for homogeneity of classification patterns across groups through a series of constraints to the MG-LCA model (Steenkamp and Baumgartner, 1998). The utility of MG-LCA has been demonstrated in social sciences (Eid et al., 2003; Vandecasteele, 2010; Geiser et al., 2006) for comparative research.

Multi-Group Latent Class Models

The MG-LCA model is similar to the Multi-Group Confirmatory Factor Analysis Model (MG-CFA) model for evaluating measurement invariance which has already been described in Chapter 4 of this thesis. MG-CFA assumes that the underlying latent variable is a continuous variable, while the MG-LCA assumes a categorical underlying latent variable. A pictorial representation of the MG-LCA model is shown in Figure 7.2. The boxes with Us inside represent the observed items/class indicator variables and in this present study these were the different disability outcome measures that were used to assess disability (e.g. GHQ-28 or BI). The square with a "*g"* inside represents the known grouping variable (e.g. SOS1=1, CIMMS =2). The oval shape with "C" symbol inside represents the underlying latent categorical variable and in latent class analysis this is assumed to be categorical.

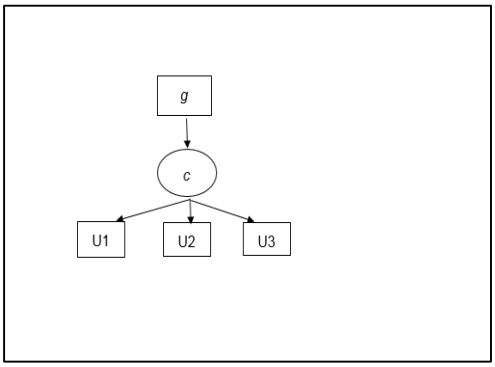

Figure 7.2 Diagrammatic presentation of a Multi-group latent variable model

Similar to the MG-CFA model specifications for testing measurement invariance described in chapter 4, there are three parameterisations that can be tested in a MG-LCA framework (Kankaras et al., 2011). The three parameterisation are: heterogeneous model, partially homogenous model, and homogenous model. The heterogeneous MG-CLA model is completely unrestricted and allows the parameter estimates to be different across groups. For continuous class indicator variables, the size of the classes, the means and variances of indicator variables in each class are allowed to vary in a heterogeneous MG-LCA model. For categorical indicator variables class-specific conditional response probabilities and class sizes are allowed to vary across groups. A heterogeneous MG-LCA is equivalent to applying a standard latent class model in each group separately (Clogg, 1985). The heterogeneous MG-LCA is comparable to configural invariance in MG-CFA. In a partially homogenous

MG-LCA model, some but not all of the model parameters are restricted to be equal across groups. The metric invariance MG-CFA model described in Chapter 4 is comparable with the partial homogeneous model in multi-group latent class analysis. The homogenous MG-LCA model fully constrains all parameter estimates to be equal across groups. It is comparable with the 'scalar invariance' model in MG-CFA that was described in Chapter 4, that constraints both factor loadings and item intercepts to be the same across groups.

To test for the assumption of equivalent underlying latent structures across groups in a MG-LCA framework, the procedure is similar to testing for measurement invariance in MG-CFA framework described in chapter 4. A series of nested, restricted models are fitted and evaluated in terms of model fit (McCutcheon, 2002). The restrictive nested models are compared using the likelihood ratio test. The difference in the likelihood ratios between two nested models represents a conditional likelihood ratio test that follows a Chi-square distribution with number of degrees of freedom equal to the difference between the degrees of freedom of the two nested models (Kankaraš et al., 2010). A non-statistical significant difference between the heterogeneous model and a partial restrictive model indicate partial invariance or that the restrictive model is no worse a fit for the data. Equivalence in the underlying structures is obtained by a non-significant Chi-square difference test between the complete homogenous and the heterogeneous model.

In some instances the MG-LCA models is not used to test for equivalence in latent structures but to identify the best model for combining data from multiple sources in comparative research. For example Vandecasteele et al. (2010) used MG-LCA to investigate country differences and the social determinants of the different poverty trajectories using data from Denmark, Spain, Germany and the United Kingdom. Vandecasteele et al. (2010) did not use the MG-LCA models to test for equivalence of latent structures across studies but to determine the best model for the pooled data analysis between the heterogeneous and partial homogenous model. The BIC and AIC goodness of fit statistics were used to identify the best model for the pooled data analysis. A lower BIC or AIC was indicative of a better fit.

The other method that can be been used for the pooled data analysis of the multiple studies in a mixture modelling framework is multilevel latent class analysis (Vermunt, 2003). Multilevel models are used to analyse data with hierarchical

structures for example patients nested in studies. Multilevel latent class analysis can also allow the parameters to vary across groups. This approach was not used in Study 4a because of the small number of studies that were being pooled. Some researchers argue that a large number of higher level units (30-50) are need for the efficient estimation of the group level variance parameters (Van der Leeden and Busing, 1994, Kreft et al., 1998), while (Busing, 1993) recommends the use of 100 or more groups. Gelman and Hill (2006) argue that the number of groups does not matter. A simulation study by Mass and Hox (2005) showed that sample sizes for higher level units as small as 10 groups produced unbiased regression coefficients but the standard errors of the level two variances were under-estimated for sample sizes of less than 100. In Study 4a, the higher level units were the two studies (SOS1, CIMSS), the number of studies was considered to be insufficient for estimating the group-level variance parameters hence the multilevel latent class analysis approach was not used.

7.3.7.1 Application of MG-LCA in this present study

The separate study analysis conducted in this thesis suggested five latent classes for SOS1 and six for the CIMSS. While the number of latent classes was not the same across the two datasets, the characterisation of the groups seemed to be similar. These findings suggested that a MG-LCA with partial measurement invariance (partially different parameters across groups) could be appropriate for the multi-group analysis of the SOS1 and CIMSS datasets. Having different class solutions for the two datasets was not problematic because as suggested by Kankaras et al. (2011) some classes can be invariant across groups and some can be group-specific. Following these guidelines from Kankaras et al., (2011), the MG-LCA analyses for the combined SOS1 and CIMSS datasets were conducted with 6 classes.

The MG-LCA of the combined SOS1 and CIMSS datasets was based on summed scores of the NEADL subscales, BI, and the harmonised GHQ measure (summed score of the six items common to GHQ-28 and GHQ-12). The psychometric properties of the harmonised GHQ measure were investigated in Chapter 6. MG-LCA models of varying constraints: heterogeneous, partial homogenous, and complete homogenous were fitted and compared. In the unconstrained or heterogeneous model the class sizes, means, and variances of baseline summed scores of the measures were free to vary in each class across studies. In the partially constrained models, the class sizes in each study were allowed to vary across studies and the means and variances of the class indicators (summed scores) were constrained to be equal across studies. The

fully constrained model was fitted by constraining the class sizes, indicator means, and variances to be equal across the SOS1 and CIMSS datasets. It was hypothesised that the latent structures across the two datasets may not be equivalent but similar thus the MG-LCA approach was not used for testing equivalence. The MG-LCA approach was used to identify the best model for the pooled data between the heterogeneous, partial homogenous and complete homogenous model. The BIC and the interpretability of the latent classes were used to select the best MG-LCA model and models with lower BIC were preferred. Mplus version 7 software (Muthén and Muthén, 2012) was used to conduct the multi- group latent class analysis. The ‘‘KNOWNCLASS’’ option in Mplus was used to allow multi-group analysis. A two-group analysis was run with a dummy for “study” LCA (SOS=1 and CIMSS=2) as the grouping (‘‘KNOWNCLASS’’) variable. The Mplus codes that were used for the MG-LCA analyses in this present study are reported in Appendix F.

7.4 Results

The number of patients in the SOS1 study was n=448 and CIMSS study: n=312. Baseline data was used in the analysis that was conducted in this study hence missing data due to attrition was not an issue. The results of the separate study analysis are presented first and these are followed by the results from MG-LCA analyses.

Participant characteristics

The baseline characteristics of patients in the SOS1 and CIMSS datasets have already been described in Chapter 3 of this thesis.

7.4.1 Descriptive analyses: SOS1 dataset

The correlations between the NEADL, GHQ-28, and BI index in the SOS 1 dataset are shown in Table 7.4. A higher score in the four NEADL subscales (mobility, kitchen, domestic, and leisure) and BI indicate a higher physical function in ADL. A higher score in the GHQ-28 subscales (Somatic, Anxiety, Social, and Depression) indicate greater psychological distress. The correlation coefficients in Table 7.4 suggested weak to moderate positive correlations between the NEADL subscales, and weak positive correlations between the NEADL subscales, and BI measure. The NEADL subscales, and BI were negatively correlated with the GHQ-28 subscales suggesting that greater independency in physical function was associated with less depressive symptoms. However these negative correlations were very small

(Table 7.4). The GHQ-28 subscales showed weak to moderate positive correlations among themselves suggesting inter-relationships among the four GHQ-28 subscales.

Table 7.4 Spearman correlation coefficients of GHQ-28 subscales, BI and NEADL subscales: SOS1 study

	Mobility	Kitchen	Domestic	Leisure	Barthel	Somatic	Anxiety	Social	Depression
Mobility	1								
Kitchen	0.55	1							
Domestic	0.55	0.50	1						
Leisure	0.56	0.33	0.31	1					
Barthel	0.17	0.12	0.02	0.16	1				
Somatic	-0.06	-0.05	-0.03	0.01	-0.10	1			
Anxiety	-0.05	-0.04	-0.02	-0.09	-0.14	0.54	1		
Social	-0.04	-0.07	-0.01	-0.08	-0.19	0.41	0.46	1	
Depression	-0.05	-0.11	-0.01	-0.08	-0.21	0.36	0.55	0.42	1

7.4.2 Mixture modelling, SOS1 dataset

The AIC, BIC , LMR *p*-value, and Entropy results of the mixture modelling that was conducted in the SOS1 study using the summed scores of the GHQ-28 subscales, BI and NEADL as class indicators are presented in Table 7.5. Several LCA models with between 1 and 8 classes were fitted. The model fit indices AIC, BIC indicated that all LCA models with larger number of classes were a significantly better fit as these had smaller goodness of fit indices (Table 7.5). Class entropy appeared reasonably good for all the models, with entropy > 0.9. The classification matrices in Table 7.6 showed high diagonal values and low off-diagonal values indicating good classification quality for all the class solutions. The LMR *p*-value was non-significant for the six class model suggesting that they was no significant improvement from a 5-class solution to a 6-class model.

Table 7.5 Model fit statistics of 2-7 class solutions for baseline severity measured by NEADL subscales, BI, and GHQ-28 dimensions: SOS1 study

Model	AIC	BIC	SSA BIC	LMR *p* value,	Entropy
LCA 2 classes	18986.85	19101.78	19012.92	0.04	0.99
LCA 3 classes	18482.56	18638.54	18517.94	0.19	0.98
LCA 4 classes	18146.67	18343.69	18191.36	0.004	0.95
LCA 5 classes	17994.72	18232.79	18048.73	0.04	0.92
LCA 6 classes	17748.54	18027.67	17811.86	0.06	0.93
LCA 7 classes	17410.95	17731.12	17483.58	0.26,	0.94
5 class with covariates	17937.23	18240.99	18006.14	0.002	0.92

AIC: Akaike Information criteria, BIC: Bayesian information criteria, LMR: Lo-Mendell-Rubin likelihood ratio test

Table 7.6 Average Latent Class Probabilities for Most likely Latent Class Membership (Row) by Latent class (Column) and class prevalence's based on estimated posterior probabilities, 2-5 class solution: SOS1 dataset

	1	2
1	0.975	0.025
2	0.002	0.998

	1	2	3
1	0.986	0.014	0.000
2	0.004	0.990	0.006
3	0.001	0.020	0.979

	1	2	3	4	5
1	0.996	0.004	0.000	0.000	0.000
2	0.002	0.956	0.028	0.000	0.014
3	0.000	0.009	0.955	0.002	0.035
4	0.000	0.001	0.009	0.979	0.011
5	0.000	0.021	0.086	0.007	0.886

Table 7.7 show the means of the class indicators for the latent class analysis that was conducted for the SOS1 dataset. The interpretability of the latent classes was conducted by examining the means of the class indicator variables from the 2-class solution up to the 6-class solution (Table 7.7). Figure 7.3 shows how class structures changed as additional classes were extracted. Based on the mean profiles shown in Table 7. 7, the two classes in the 2-class solution could be labelled as "Independent, no depressive symptoms" and "Dependent, Mild depressive symptoms. Figure 7.3 showed that from the two-class to a 3-class solution, a third class emerged which was very similar to class 2 in the two-class solution but had: poor function in ADL and

severe depressive symptoms. From the three-class solution to the four-class solution, a fourth class emerged which showed: poor function in ADL, moderate IADL, and mild depressive symptoms. From the four-class solution to the five-class solution, a fifth class emerged which showed: good function in IADL, mild depression, severe anxiety, and poor social functioning. Further extraction of classes showed no improvement from the classes that were already extracted. Furthermore the class sizes were becoming smaller when additional classes were extracted.

Based on the non-significant LMR *p*-value between a five-class model and Six-class model, and class interpretability, a five class solution was preferred for the mixture modelling of the SOS 1 in this present study. Adding covariates to the five-class solution did not change the class structures significantly (Figure 7.4). The classes showed varying combination levels of physical, social and psychological function.

Table 7.7 Prevalence's (n, %) and mean disability levels for 2-5 class solutions, SOS1 study

	2 class model		**3 class model**			**4 class model**				**5 class model**				
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**5**
% Class membership*	7.5	92.4	6.5	82.3	11.2	17.5	64.7	4.1	13.6	4.02	15.8	57.8	9.1	13.2
NEADL Mobility	6.61	15.37	5.69	15.35	15.32	7.90	16.82	4.46	16.57	4.44	7.72	16.86	16.02	16.09
NEADL Kitchen	6.95	14.74	6.55	14.71	4.40	13.50	14.82	4.61	14.69	4.52	13.43	14.83	14.35	14.87
NEADL Domestic	2.72	9.13	2.46	9.07	9.11	5.52	9.73	1.85	9.59	1.87	5.27	9.75	9.22	9.61
NEADL Leisure	7.39	12.15	7.02	12.18	11.74	8.08	12.99	6.57	12.47	6.53	7.94	13.03	11.90	12.62
BI	10.97	13.85	11.02	14.17	11.22	11.94	14.43	11.49	12.68	11.46	11.99	14.79	11.25	4.14
GHQ-28 Somatic	2.02	1.72	1.81	1.48	3.60	1.87	1.15	2.05	4.26	2.06	1.76	0.83	3.58	2.91
GHQ-28 Anxiety	1.55	1.33	1.30	0.96	4.23	1.28	0.63	1.67	4.79	1.69	1.19	0.51	4.34	4.22
GHQ-28 Social	3.45	2.61	3.27	2.33	4.87	2.74	2.11	3.41	5.01	3.45	2.62	1.87	4.27	4.95
GHQ-28 Depression	1.04	0.67	0.92	0.24	3.93	0.67	0.23	1.20	2.76	1.23	0.48	0.16	0.65	4.27

*Based on the sum of the posterior probabilities from the model

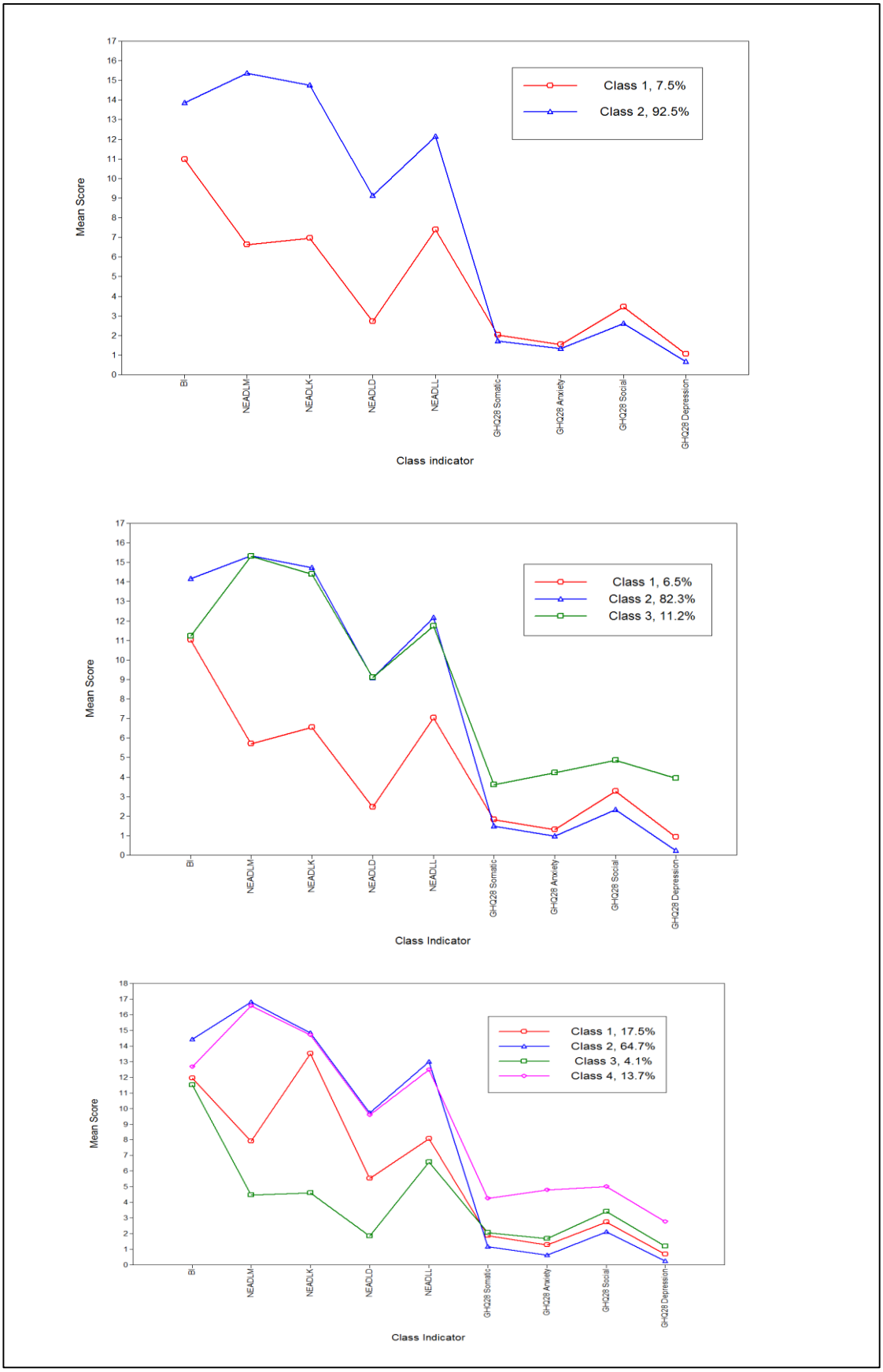

Figure 7.3. Latent class mean profiles for the 2-4 class solutions, SOS1 study

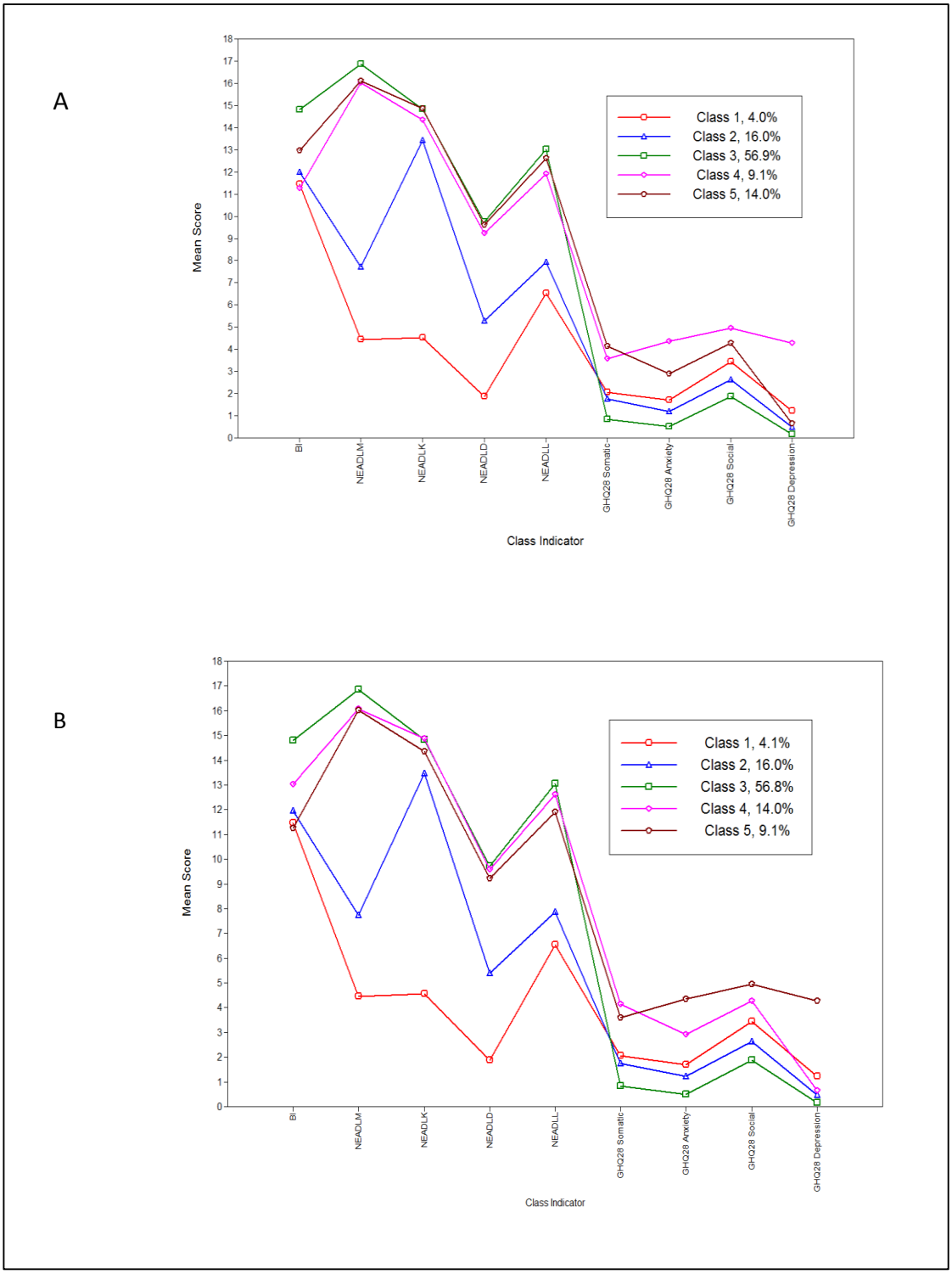

Figure 7.4 (A) Mean profiles for the five-class solution (B) Mean profiles for the five class solution with covariates, SOS1 data

7.4.2.1 Characteristics of Latent classes, SOS1 study

Class 1: There were n=18 patients in Class 1, representing about 4.1% of the patients in the SOS1 dataset. This class had the lowest mean NEADL subscales, BI scores and third highest mean GHQ-28 somatic, anxiety and social scores (Figure 7.4 B) suggesting greater dependence in performing extended activities of daily living , poor function in basic ADL, and mild depressive symptoms. Based on the mean indicator scores, class one was labelled "Poor function in ADL, Mild depressive symptoms". The results from the multinomial regression analysis shown in Table 7.8 indicated that, compared to patients in class 3 (Good physical function, no depressive symptoms), patients assigned to class 1 were more likely to be elderly, to have had a previous stroke, and less likely to have been living alone pre-stroke (Table 7.8). Descriptive statistics shown in Table 7.9 showed that, more than half of the patients in class 1 had a previous stroke, were also older (mean = 74 years) than patients in classes 3, 4, and 5 and the majority were male (61%).

Class 2: There were n=72 patients in class 2, representing 16% of the patients in the SOS1 dataset. Based on the mean scores in Figure 7.4 B, this class showed moderate function in ADL, and mild depressive symptoms, but had a very high mean score for the NEADL Kitchen subscale (Figure 7.4 B) score suggesting good physical function in kitchen tasks. Based on the mean indicator scores, class 2 was labelled "Good physical function in the kitchen". Compared to patients in class 3 (Good physical function, no depressive symptoms), patients assigned to class 2 were more likely to be elderly, and had a previous stroke (Table 7.8). Descriptive statistics shown in Table 7.9 indicated that this class had the highest mean age (78 years), and the majority of patients were female (62%).

Class 3: There were n=254 patients in class 3, representing 56.8% of the patients in the SOS1 dataset. This group was the most common; the patients had the highest mean NEADL subscale scores and BI scores, and the lowest GHQ-28 subscale scores suggesting greater independence in both basic and extended activities of daily living and no depressive symptoms (Figure 7.4 B). Based on the mean indicator scores, class 3 was labelled "Good physical function, no depressive symptoms". Compared to the patients in classes (1 and 2), patients in class 3 were younger (mean age 69 years), a smaller proportion (14%) of them had previous stroke, and a larger proportion (36.7%) of them were living alone before stroke (Table 7.9).

Class 4: There were 63 patients in class four, representing 14 % of the patients in the SOS1 dataset. This class showed greater independency in extended activities of daily living measured by NEADL subscales, moderate function in basic ADL measured by BI, and had the second highest GHQ-28 anxiety, and social dysfunction scores (Figure 7.4 B). Based on the mean indicator scores, this class was labelled "Good function in IADL, Moderate function in basic ADL, Anxious and Poor social function". The multinomial logistic regression model did not show any patient characteristics that distinguished this class from class 3 (Good physical function, no depressive symptoms) (Table 7.8). Descriptive statistics shown in Table 7.9 indicated that patients in this class were younger compared to all the classes with a mean age of 67.5 years, 45% of them had major depression measured by the Present State Examination(PSE) measure (Wing et al., 1974) an interview-based psychological assessment. The class also had similar prevalence levels of urinary incontinent (11%) with the patients in Classes (1 and 2) which had elderly patients (Table 7.9).

Class 5: There were 41 patients in class five, representing 9.1% of the patients in the SOS1 dataset. This class had high mean NEADL subscales scores similar to classes 3 and 4, but the BI scores were as low as those for class 1 ("Poor function in ADL, mild depressive symptoms" (Figure 7.4 B). Class 5 also had the highest anxiety, depression, and social dysfunction GHQ-28 scores suggesting severe mood disorder for this class. Based on the mean GHQ-28 subscale scores, Class 5 was labelled the "Poor function in basic ADL, Severe depressive symptoms". The multinomial logistic regression did not show any patient factors that distinguished class 5 from class 3 (Good physical function, no depressive symptoms). Descriptive statistics shown in Table 7.9 indicated that patients in this class were younger, similar in age to patients in class 3, but a higher proportion (24%) of them had previous stroke. This class had the largest proportion (68%) of patients with major depression as classified by the Present State Examination (Wing et al., 1974) (Table 7.9).

Generally there was good differentiation of the five classes, with wider differences revealing greater differentiation in class by the disability measure. The NEADL measure showed wider separation compared to the other measures. The BI differentiated three broad classes (poor, moderate, mild) but the separation was not as wide as that of the NEADL measure. The NEADL measure was *influenced by gender* with a large proportion of patients (62%) in the "Good physical function in the kitchen" being female, and a larger proportion of patients (61%) in class 2 being male

with poor function in the kitchen. Classes 1 and 2 distinguished between elderly male and female stroke survivors. The GHQ-28 subscales differentiated broad classes that can be labelled “no depressive symptoms”, “mild”, and “severe depressive symptoms or mood disorder”.

The emerging classes showed varying combination levels of physical, social and psychological function. The results showed that the underlying latent variable measured by a combination of these disability dimensions was not a latent continuum that can be classified from low to high because some classes, despite having high physical function in extended activities of daily living, had severe depressive symptoms. The class with severe depressive symptoms also had poor function in basic ADL measured by the BI measure.

Table 7.8 Multinomial logistic regression coefficients and p value, SOS1 study

Model covariates	**Class 1** Estimate,	SE,	*p*-value	**Class 2** Estimate,	SE,	*p*- value
Sex	0.13	0.54	0.811	0.48	0.36	0.180
Age	0.09	0.04	**0.012**	0.09	0.02	<**0.000**
Living Alone	-2.91	1.06	**0.006**	0.09	0.34	0.803
Previous stroke	2.19	0.58	<**0.000**	1.41	0.34	<**0.000**
	Class 4 Estimate,	SE,	*p*-value	**Class 5** Estimate,	SE,	*p*-value
Sex	0.13	0.41	0.742	0.02	0.37	0.960
Age	-0.01	0.02	0.552	0.003	0.02	0.860
Living Alone	0.22	0.39	0.582	0.10	0.38	0.795
Previous stroke	0.32	0.46	0.490	0.68	0.42	0.107

Reference group was class 3 (Good physical function, no depressive symptoms)

Table 7.9 Patients characteristics and one month disability levels by class. SOS 1 study

Variable	**Class 1 18(4.2%)**	**Class 2 72 (16%)**	**Class 3 (254 (56.8%)**	**Class 4 63 (14 %)**	**Class 5 (41(9.1%)**	***p*- value**
Mean Age, years	74.1(9.5)	78(9.0)	69.4(10.8)	67.5(13.6)	69.3(13.3)	<**0.001**
MMSE	23.6(3.13)	24.4(3.00)	25.9(3.05)	24.9(3.07)	24.2(3.2)	<**0.001**
Previous stroke	10(55.6)	26(35.6)	36(14.1)	12(20.0)	10(24.4)	<**0.001**
Female, n %	7(38.9)	45(61.6)	110(43.0)	27(45.0)	18(43.9)	**0.070**
Urine incontinence	2(11.1)	7(9.6%)	10(3.9)	7(11.7)	4(9.8)	**0.103**
Major depression 1 month(PSE)	7(38.9)	18(24.7)	20(7.8)	27(45.0)	28(68.3)	<**0.001**
Living Alone	1(5.6)	42(57.5)	94(36.7)	22(36.7)	16(39.0)	<**0.001**
GHQ-28 Somatic	2.61	1.74	0.83	4.14	3.59	<**0.001**
GHQ-28 Anxiety	1.69	1.21	0.50	2.92	4.35	<**0.001**
GHQ-28 Social	3.44	2.62	1.87	4.27	4.95	<**0.001**
GHQ-28 Depression	1.22	0.47	0.16	0.66	4.27	<**0.001**
NEADL Mobility	4.45	7.74	16.87	16.07	16.01	<**0.001**
NEADL Kitchen	4.55	13.46	14.83	14.87	14.35	<**0.001**
NEADL Domestic	1.87	5.39	9.72	9.58	9.22	<**0.001**
NEADL Leisure	6.53	7.87	13.06	12.60	11.90	<**0.001**
BI	11.47	11.95	14.79	13.04	11.25	<**0.001**

7.4.2.2 Summary of SOS1 mixture modelling analysis

Five classes emerged and their characteristics were as follows:

- Class 1: n=18 (4.1%): "Poor function in ADL, Mild depressive symptoms". Individuals in class one were more likely to be elderly males, who had a previous stroke, and less likely to have been living alone before stroke.
- Class 2: n=72 (16%) "Good physical function in the kitchen": Individuals in class two were more likely to be elderly females.
- Class 3: n=254(56.8%) "Good physical function, no depressive symptoms": This was the most common class. Individuals in this class were more likely to be younger stroke survivors
- Class 4: n=63 (14 %) "Good function in IADL, Moderate function in basic ADL, Anxious and Poor social dysfunction": Individuals in this class were more likely to be young and urinary incontinent.
- Class 5: n=41(9.1%) "Poor function in basic ADL, Severe depressive symptoms": Individuals in this class were young and more likely to have had previous stroke.

7.4.3 Descriptive analyses: CIMSS dataset

The correlations between the different disability dimensions for the NEADL, GHQ-12, and BI in CIMSS study are shown in Table 7.10. A higher score in the GHQ-12 subscale indicates greater psychological distress. Similar to SOS1, the correlation coefficients in Table 7.10 suggested moderate to strong positive correlations between the NEADL subscales, and weak positive correlations between the NEADL subscales and BI measure. The NEADL subscales, and BI were negatively correlated with the GHQ-12 subscale suggesting that greater independency in physical function was associated with less depressive symptoms. Similar to SOS1, physical function measured by NEADL and BI was negatively correlated with the two subscales of GHQ-12 (Anxiety and depression, social) suggesting that greater dependency in physical function was associated with poor psychological function. However these negative correlations were small (Table 7.10). The two GHQ-12 subscales were positively correlated.

Table 7.10 Spearman correlation coefficients of GHQ-12 subscales, NEADL subscales and BI: CIMSS dataset

	Anxiety/ Depression	Social	Barthel	Mobility	Kitchen	Domestic	Leisure
Anxiety/ Depression	1						
Social	0.65	1					
Barthel	-0.13	-0.13	1				
Mobility	-0.18	-0.31	0.33				
Kitchen	-0.20	-0.34	0.26	0.65	1		
Domestic	-0.23	-0.32	0.23	0.59	0.70	1	
Leisure	-0.31	-0.38	0.34	0.67	0.58	0.58	1

7.4.4 Mixture modelling, CIMSS dataset

The AIC, BIC , LMR *p*-value, and Entropy results of the mixture modelling that was conducted for the CIMSS dataset using the summed scores of the GHQ-12 depression and anxiety subscales, BI and NEADL as class indicators are presented in Table 7.11. Several LCA models with between 1 and 8 classes were fitted. The model fit indices AIC, BIC indicated that all LCA models with larger number of classes were a significantly better fit as these had smaller goodness-of-fit indices (Table 7.11). Class entropy appeared reasonably good for all the models, with entropy > 0.9. The classification matrices in Table 7.12 show high diagonal values and low off-diagonal values indicating good classification quality. The LMR *p*-value was non-significant for the 3- class model suggesting that they was no significant improvement from a 2-class model to a 3-class model.

Table 7.11 Model fit statistics of 2-8 class solutions for baseline severity measured by NEADL, BI and GHQ-12: CIMSS dataset

Model	**AIC**	**BIC**	**SSA BIC**	**LMR *p*-value,**	**Entropy**
LCA 2 classes	11372.42	11454.76	11384.98	<0.001	0.97
LCA 3 classes	11095.64	11207.93	11117.78	0.080	0.92
LCA 4 classes	10998.62	11140.86	110201.33	0.425	0.91
LCA 5 classes	10861.82	11033.99	10888.10	0.179	0.92
LCA 6 classes	10790.98	10993.10	10821.83	0.371	0.92
LCA 7 classes	10721.56	10953.62	10756.98	0.339	0.93
LCA 8 classes	10593.32	10855.33	10633.31	0.407	0.95
LCA 6 classes with covariates	10384.45	10658.78	10424.09	0.460	0.92

BIC: Bayesian information criteria, LMR: Lo-Mendell-Rubin likelihood ratio test.

Table 7.12 Average latent Class Probabilities for most likely latent class membership (Row) by Latent Class (Column) and class prevalence's based on estimated posterior probabilities, 2 to 7 class solutions: CIMSS dataset

Classification matrix	**Classification matrix**
1 2 1 0.998 0.002 2 0.010 0.990	1 2 3 1 0.943 0.010 0.048 2 0.007 0.993 0.000 3 0.029 0.007 0.964
1 2 3 4 1 0.923 0.006 0.050 0.021 2 0.013 0.987 0.000 0.000 3 0.015 0.000 0.898 0.088 4 0.009 0.007 0.023 0.961	1 2 3 4 5 1 0.936 0.045 0.019 0.000 0.000 2 0.005 0.993 0.000 0.002 0.000 3 0.001 0.000 0.951 0.004 0.044 4 0.000 0.001 0.021 0.927 0.050 5 0.005 0.003 0.023 0.014 0.955
1 2 3 4 5 6 1 0.954 0.003 0.000 0.000 0.043 0.000 2 0.000 0.934 0.010 0.028 0.000 0.028 3 0.000 0.000 0.931 0.017 0.000 0.052 4 0.000 0.022 0.021 0.843 0.000 0.115 5 0.004 0.003 0.000 0.000 0.993 0.000 6 0.005 0.010 0.011 0.018 0.003 0.952	1 2 3 4 5 6 7 1 0.957 0.000 0.000 0.040 0.002 0.000 0.000 2 0.000 0.837 0.010 0.143 0.000 0.000 0.010 3 0.000 0.010 0.931 0.059 0.000 0.000 0.000 4 0.000 0.013 0.002 0.982 0.000 0.000 0.003 5 0.027 0.000 0.007 0.032 0.933 0.001 0.000 6 0.002 0.000 0.000 0.043 0.000 0.955 0.000 7 0.000 0.015 0.002 0.151 0.000 0.000 0.831

Table 7.13 show the means of the class indicators for the latent class analysis that was conducted for the CIMSS study. The interpretability of the classes was conducted by examining the means of the indicators from the 2 class solution up to the 6 class solution (Table 7.13). Figure 7.5 shows how class structures changed as additional classes were extracted. Based on the means in Table 7.13 or Figure 7.5, the 2 classes in the 2-class solution can be labelled as “independent, mild depressive symptoms” and “dependent and depressive symptoms”. From the 2-class to a 3-class solution, a third class emerged which showed moderate physical function and depressive symptoms. The third class showed very high physical function in the NEADL kitchen subscale. From the 3- class solution to the 4-class solution, a fourth class emerged which showed poor domestic function measured by the NEADL subscale and

moderate depressive symptoms. From the 4-class to the 5-class solution, a fifth class emerged which was characterised by moderate BI and severe mood symptoms. Further extraction of classes from the 5-class solution to a 6-class solution yielded a sixth class which was similar to class 3 but had low function in kitchen activities. Classes were becoming smaller when additional classes were extracted.
Based on lower BIC and clinical interpretation of the classes, a 6-class solution was preferred. Figure 7.6 C shows the mean indicator profiles for the six classes adjusted for covariates and the prevalence and separation of the six latent classes. Adding covariates to the sixth class solution did not change the class structures significantly. The classes showed varying combination levels of physical, social and psychological function.

Table 7.13 Prevalence (n, %) and unadjusted mean disability levels by class: CIMSS dataset

	2class model		**3 class model**			**4 class model**				**5 class model**					**6 class**					
	1	**2**	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**6**
n,%Class membership*	17	83	17.5	13.4	69	10.4	11.6	13.5	64.5	9	4.7	13.7	8.7	63.7	8.3	10.7	6.45	10.4	4.75	59
NEADL mobility	3.63	14.71	6.33	3.70	16.31	8.53	2.96	7.16	16.54	4.45	5.58	13.45	13.45	16.31	4.19	8.68	13.97	6.92	2.60	15.53
NEADL Kitchen	4.18	14.46	12.89	2.73	14.64	9.69	2.17	14.36	14.78	3.32	2.08	12.96	14.03	14.64	2.85	10.55	14.79	14.50	2.02	14.82
NEADL Domestic	1.68	9.44	5.68	1.56	10.01	3.24	1.34	7.73	10.18	2.01	0.67	5.68	8.29	10.05	2.12	3.49	9.66	7.64	0.64	10.26
NEADL Leisure	4.43	13.18	7.44	4.40	14.24	7.88	3.83	8.30	14.45	4.83	3.40	7.12	10.21	14.54	4.87	8.33	11.02	7.86	3.37	14.68
BI	9.09	13.23	9.51	8.88	14.20	11.95	7.86	9.03	14.21	9.71	7.28	9.74	10.08	14.27	9.82	11.47	10.17	9.17	7.22	14.37
GHQ-12 Social	10.23	7.22	8.58	10.55	6.95	8.88	10.67	8.43	6.86	7.73	15.85	7.42	12.84	6.47	7.69	8.60	13.13	7.33	15.89	6.42
GHQ-28 Anxiety/depression	8.22	5.54	7.29	8.52	5.16	7.62	8.54	6.90	5.10	5.74	13.61	6.07	12.82	4.52	5.89	7.37	13.51	5.56	13.65	4.50

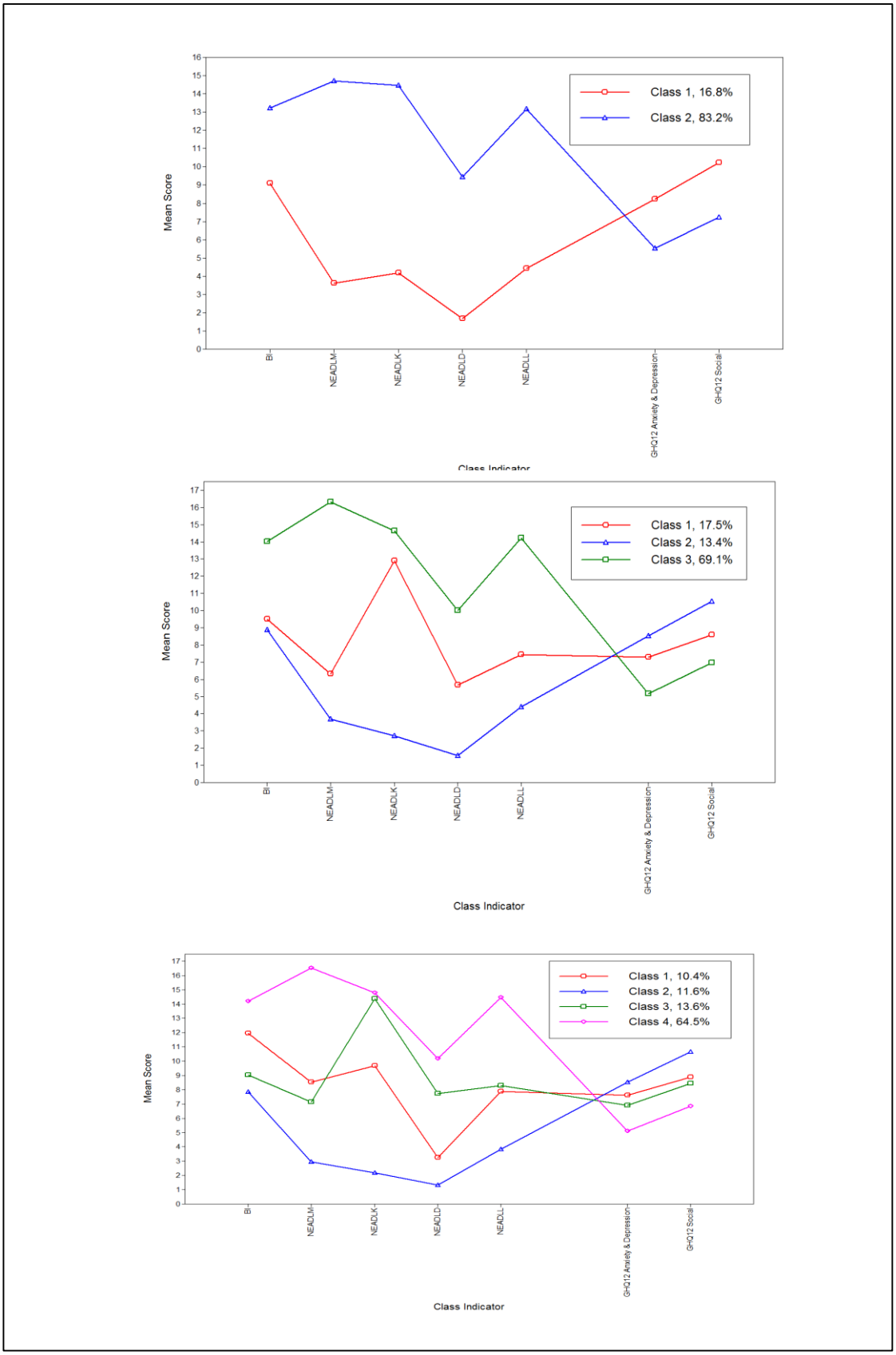

Figure 7.5 Latent class mean profiles for the 2-4 class solutions, CIMSS study

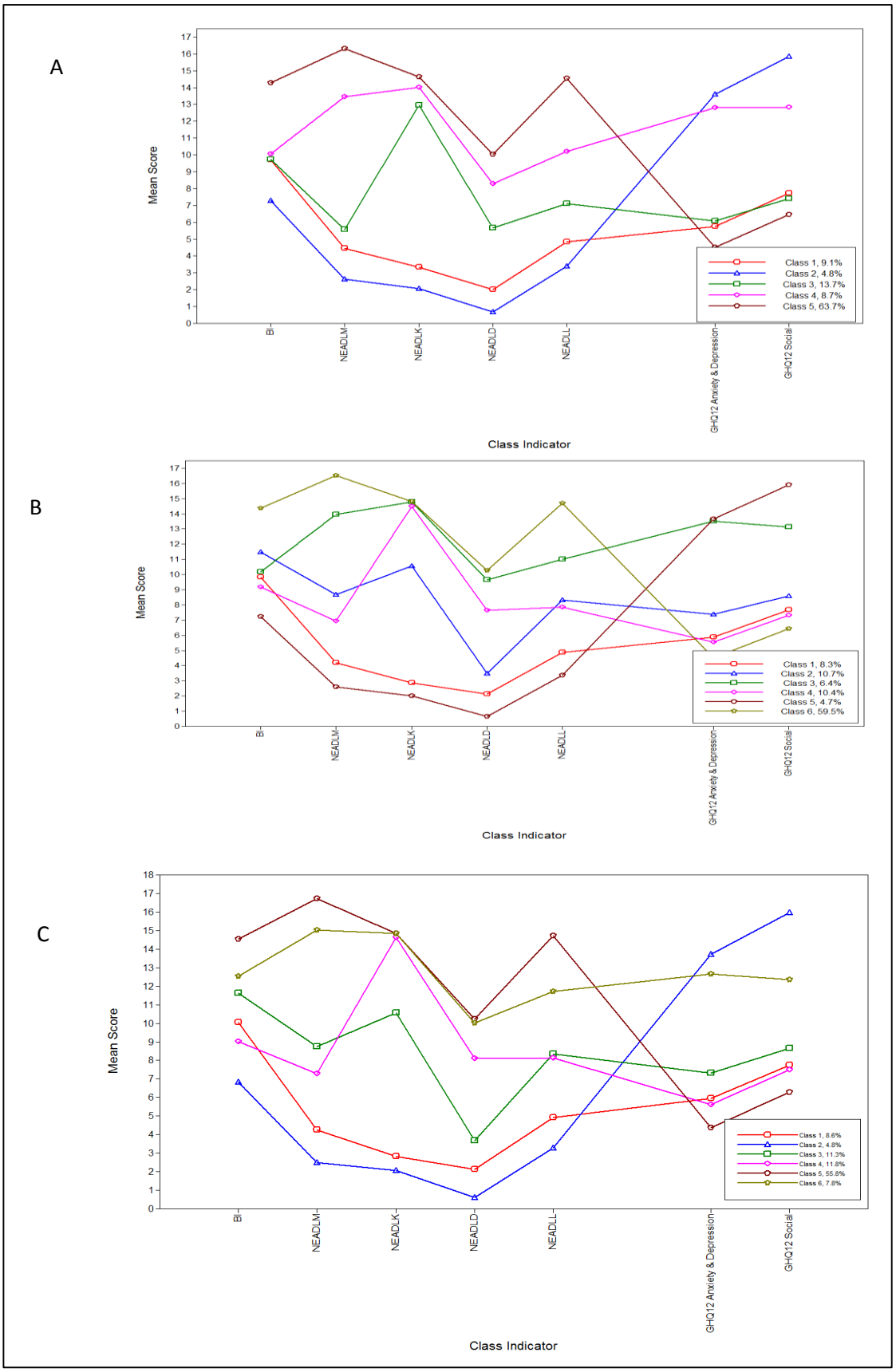

Figure 7.6 (A) Mean profiles for the five class solution (B) Mean profiles for the six class solution (C) Mean profiles for the six class solution with covariates, CIMSS dataset

7.4.4.1 Characteristics of Latent classes, CIMSS dataset

Class 1: There were n=26 patients in class 1, representing, 8.6% of the patients in the CIMSS dataset. This class had the second lowest NEADL subscale scores (Figure 7.6 C) suggesting poor function in IADL. Based on the mean indicator scores, Class 1 was labelled "Poor function in ADL, Mild depressive symptoms". The results from the multinomial logistic regression shown in Table 7.14 indicated that, compared to patients in class 5 ("Good physical function, no depressive symptoms"), patients in class 1, were more likely to have had a previous stroke (Table 7.14). Descriptive analysis shown in Table 7.16 showed that 40% of patients in this class had speech difficulties, and 32% had a previous stroke (Table 7.15). The average age for patients in this class was 71 years, similar to class 5 ("Good physical function, no depressive symptoms").

Class 2: There were n= 14 patients in class 2, representing 4.8% of the patients in the CIMSS dataset. This class had the lowest mean NEADL, BI scores, and highest GHQ-12 scores (Figure 7.6 C), suggesting greater dependence in performing both extended and basic activities of daily living and severe depressive symptoms. Class 2 was labelled "Severe dependency in ADL, Severe depressive symptoms" based on the mean indicator scores. The result from the multinomial logistic regression indicated that compared to class 5 (Good physical function, no depressive symptoms), patients in this class were less likely to be males and were more likely to have had a previous stroke (Table 7.14). Descriptive analysis in Table 7.15 showed that these patients were elderly (mean age 78) and female (80%). Forty percent of them had previous stroke, and 60% had a right weakness.

Class 3: There were n= 34 patients in this class, representing 11.3% of the patients in CIMSS dataset. Patients in this class were characterised by moderate physical function measured by NEADL and BI and moderate depressive symptoms (Figure 7.6 C). This class was labelled "Moderate physical function, Moderate depressive symptoms" based on the mean indicator scores. The results of the multinomial logistic regression shown in Table 7.14 showed that, compared to the class 5 (Good physical function, no depressive symptoms) this class was more likely to have had a previous stroke. Descriptive analyses in Table 7.15 showed that about 60% of the patients in this class had left weakness, and a 63.6% were males. Forty percent of the individuals in class 3 had speech difficulties.

Class 4: There were n = 35 patients in this class, representing 11.8% of the patients in the CIMSS dataset. Patients in this class were characterised by very high NEADL Kitchen scores suggesting very good function in kitchen tasks, moderate BI scores, and low GHQ-12 subscale (Figure 7.6 C). This class was labelled "Good physical function in kitchen tasks" based on the mean indicator scores. The results from the multinomial logistic regression analysis shown in Table 7.14 showed that, compared to class 5 ("Good physical function, no depressive symptoms"), the patients in class 4 were less likely to be males, more likely to be older , and to have had a previous stroke. Descriptive analyses shown in Table 7.15 indicated that these patients were predominantly elderly (mean age 82 years), and the majority were female (92%). About 73% of the patients in class 4 were living alone before stroke (Table 7.15).

Class 5: There were n =168 patients this class, representing 56% of the patients in the CIMSS dataset. Class 5 had the highest mean NEADL subscale scores, highest BI scores, and had the lowest GHQ-12 subscale scores suggesting greater independence in both basic and extended activities of daily living and no indication of depressive symptoms (Figure 7.6 C). This class 5 was named "Good physical function, no depressive symptoms" based on the mean indicator scores. Descriptive analyses shown in Table 7.15 showed that the patients in this class were younger compared to patients in classes 1-4 with an average age of 71 years (Table 7.15). This class was similar to class 3 ("Good physical function, no depressive symptoms") in the SOS1 study.

Class 6: There were n=23 patients in this class, representing 7.7% of the patients in the CIMSS dataset. Class 6 had the second highest mean NEADL scores and BI scores indicating good physical function, and had the second highest GHQ-28 suggesting depressive symptoms in this class (Figure 7.6 C). This class was labelled, "Poor function in basic ADL, depressive symptoms" based on the mean indicator scores. The results from the multinomial logistic regression shown in Table 7.14 showed that, compared to class 5 (Good physical function, no depressive symptoms), patients in this group were younger. Descriptive analyses shown in Table 7.15 indicated that patients in this class were younger, had the lowest mean age (61 years), and 62% were females. Fifty percent of these patients had right side stroke weakness (Table 7.15).

There was good differentiation of the 6 classes with wider differences revealing greater differentiation in class by that measure. Similar to the SOS1 study, the NEADL showed wider separation of classes (Figure 7.6 C), it differentiated very well between the low, moderate, and high function classes. The GHQ-12 measure differentiated 3 broad classes (no depressive symptoms, moderate-mild, severe depressive symptom) showing varying levels of anxiety & depression, social function. The BI measure showed 6 classes. Similar to the SOS1 study, there was a class that had severe depressive symptoms, despite having high physical function in extended activities of daily living. This class also had poor function in basic ADL measured by the BI measure.

Table 7.14 Multinomial logistic regression results, regression coefficients and *p* values: CIMSS dataset

Model covariates	**Class 1 Estimate, SE, p value**			**Class 2 Estimate, SE, *p* value**			**Class 3 Estimate, SE, *p* value**		
Sex	-0.60	0.55	0.274	-1.81	0.75	**0.016**	0.12	0.45	0.785
Age	0.01	0.02	0.756	0.03	0.06	0.591	0.05	0.03	0.073
Previous	1.76	0.58	**0.003**	2.02	0.91	**0.026**	1.19	0.60	**0.046**
stroke	-0.85	0.69	0.218	-0.76	1.06	0.474	-0.60	0.48	0.213
Living Alone	-3.23	2.01	0.109	-4.40	4.07	0.280	-6.34	2.03	**0.002**
	Class 4			**Class 6**					
Sex	-2.03	0.69	0.003	-1.11	0.58	0.055			
Age	0.08	0.03	0.003	-0.06	0.03	**0.040**			
Living Alone	1.83	0.70	0.010	1.47	0.86	0.089			
Previous	0.76	0.52	0.144	0.44	0.54	0.419			
stroke	-7.75	2.48	0.002	2.16	2.46	0.382			

Reference group was Class 5 (Good physical function, no depressive symptoms).

Table 7.15 Patient characteristics, n (%) and baseline disability levels by class, CIMSS dataset

Variable	**Class 1**	**Class 2 4.8%**	**Class 3**	**Class 4 11.2%**	**Class 5 56%**	**Class 6 7.7%**	***p*-value**
Age(years) Mean(SD)	71.8(10.05)	77.9(14.3)	76.6(11.2)	82.2(6.7)	71(12.0)	61(13.8)	<0.001
Gender Female	12(48)	12(80)	12(36.4)	34(91.9)	70(41.2)	13(61.9)	<0.001
Previous stroke	8(32)	6(40)	8(24.2)	9(24.3)	10(5.9)	4(19.0%)	<0.001
Living alone before stroke	5(20)	4(26.7)	8(24.2)	27(73.0)	63(37.1)	8(38.1)	<0.001
Weakness							
Right	8(33.3)	9(60)	4(12.1)	16(43.2)	63(37.3)	11(52.4)	<0.001
Left	13(54.2)	5(33.30	20(60.6)	17(45.9)	60(35.5)	8(38.1)	
No weakness	3(12.5)	1(6.7)	9(27.3)	4(10.8)	46(27.2)	2(9.5)	
Clinical Classification							
LACS	6(24)	5(33.3)	9(28.1)	9(25)	47(30.1)	8(40)	<0.001
PACS	11(44)	7(46.7)	6(18.8)	12(33.3)	53(34)	7(35)	
POCS	3(12)	0	8(25)	3(8.3)	25(16)	2(10)	
PACS	5(20)	3(20)	9(28.1)	12(33.3)	32(19.9)	3(15)	
Speech difficulties	10(40)	2(13.3)	14(42.4)	12(32.4)	68(40)	7(33.3)	0.397

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**	***p*-value**
Confusion at presentation	5(20)	8(57)	6(18.2)	11(29.7)	25(14.8)	2(9.5)	0.002
New urinary issues	8(32)	7(50)	5(15.2)	16(43.2)	26(15.7)	5(23.8)	<0.001
BI	10.05	6.82	11.63	9.02	14.53	12.53	<0.001
Mobility	4.24	2.49	8.73	7.29	16.72	15.04	<0.001
Kitchen	2.83	2.07	10.56	14.64	14.83	14.83	<0.001
Domestic	2.13	0.60	3.68	8.09	10.21	10.02	<0.001
Leisure	4.91	3.28	8.36	8.13	14.73	11.73	<0.001
Anxiety/depression	5.94	13.73	7.33	5.61	4.37	12.67	<0.001
Social	7.73	15.97	8.65	7.48	6.29	12.35	<0.001

7.4.4.2 Summary of CIMSS study mixture modelling

Six classes emerged and there were labelled as follows:

Class 1: n=34 (8.6%): “Poor function in ADL, Mild depressive symptoms”. Individuals in this class were more likely to have had a previous stroke and were predominantly younger.

Class 2: n=14 (4.8%) “Severe dependency in ADL, Mood disorder”: Individuals in this class were predominantly female and majority had previous stroke.

Class 3: n=34 (11.2%) “Moderate physical function, Moderate depressive symptoms”: Individuals in this class were predominantly males.

Class 4: n=35 (11.8%) “Good physical function in the kitchen”: Individuals in this class were more likely to be elderly and female.

Class 5: n=168 (56%) “Good physical function, No depressive symptoms”: this was the most common class, individuals in this class were more likely to be younger and a small proportion had previous stroke.

Class 6: n=23 (7.7%) “Poor function in basic ADL, Depressive symptoms”. These patients were predominantly younger females.

The classes showed varying levels of physical, social, and psychological function.

7.4.5 Synthesis of separate study results

The separate study analysis showed 5 latent classes for SOS1, and 6 latent classes for CIMSS. While the number of latent classes was not equal across the two datasets, there was some similarity in the characterisation of the classes even though different measures were used to identify the classes. Common latent classes were identified across the two datasets and these were labelled as:

-Good physical function, No depressive symptoms (CIMSS: 56%; SOS1: 57%)

- Poor function in basic ADL, Severe depressive symptoms (CIMSS: 8%; SOS1: 9%)

-Poor physical function, Mild depressive symptoms (SOS1: 4%, CIMSS: 9%)

-Moderate physical function, mild-moderate depressive symptoms (CIMSS: 11%, SOS1: 14%)

-A class with high physical function in the NEADL kitchen dimension emerged and the majority of these patients were women, (CIMSS: 12%; SOS1: 16%)

The factors associated with latent classes were similar across the two studies. Patients in the "Good physical function, No depressive symptoms" class were more likely to be young stroke survivors in all two studies. Patients in the "Poor physical function, Mild depressive symptoms" class were more likely to be elderly and had previous stroke. Patients in the "Good physical function, Severe depressive symptoms" class were more likely to be younger, female stroke survivors, and some had a previous stroke.

7.4.6 Multi-Group Latent Class Analysis of SOS1 and CIMSS datasets

Due to the small sample sizes in individual studies, some of the classes had very few individuals. The mixture modelling was repeated with the bigger combined SOS1 and CIMSS data (n=760). The three multi-group latent class models: heterogeneous, partial homogeneous and complete homogenous were fitted to the combined dataset. After comparing the disability latent classes across the two datasets, Multi-Group Latent Class Analysis (MG-LCA) was conducted for the SOS1 and CIMSS datasets simultaneously with six classes. In this section the results of the multi-group latent class analyses are presented.

The model fit indices for the multi-group latent class analysis of the pooled SOS1 and CIMSS datasets are shown in Table 7.16. Based on the BIC, the best multi-group model for the combined two datasets was the partially constrained six-class model which assumed measurement equivalence of means and variances across studies, but varying class prevalence. However the AIC favoured the heterogeneous model. The partial homogenous model was more parsimonious than the unconstrained heterogeneous model in which 42 additional parameters were estimated. It seemed reasonable to assume unequal class sizes across the SOS1 and the CIMSS datasets since the separate study analysis had revealed that the class prevalence were not identical across the two studies. The six-class partially homogeneous solution that was favoured by the BIC statistic was interpretable, entropy was good, 0. 95, hence the partial homogenous model was chosen.

Table 7.16 Multi-group latent class analysis based on NEADL, BI and Harmonised GHQ-6 total scores

	LogL, Scaling Factor	AIC	Parameters	BIC	SSBIC	Entropy
6 class model						
Heterogeneous	-11201.7 1.52	**22593.4**	95	23033.54	22731.9	0.92
Partial homogeneous	-11255.5 1.64	22616.9	53	**22862.49**	**22694.2**	0.95
Complete homogeneous	-11273.97 1.70	22643.9	48	22866.34	22713.9	0.95

In the partially homogeneous MG-LCA model, the SOS1 dataset contributed 59% percent of the sample, while the CIMSS dataset contributed 41% of the sample. Out of the n=312 patients in the CIMSS, 35 (12%) were in class 1, 32 (10%) in class 2, 34 (11%) in class 3, 25 (9%) in class 4, 166 (52%) in class 5, and 20 (7%) in class 6. Of the n=448 patients in SOS1 data, 12 (3%) were in class 1, 88 (20%) in class 2, 34 (8%) in class 3, 39 (9%) in class 4, 420 (56%) in class 5, and 41 (5%) in class 6. The mean indicators for the partial homogenous MG-LCA solution are shown in Table 7.17. The classes were labelled according to their mean indicator levels.

Table 7.17 Class size and class indicator means for the partially homogenous model, SOS1 and CIMSS datasets

Class	Study	Class Prevalence	Mobility	Kitchen	Domestic	Leisure	GHQ Harmonised	BI
1	CIMSS	0.12	3.7	2.4	1.4	4.8	2.4	9.0
	SOS1	0.03	3.7	2.4	1.4	4.8	2.4	9.0
2	CIMSS	0.10	16.2	14.9	5.2	12.2	1.3	15.1
	SOS1	0.20	16.2	14.9	5.2	12.2	1.3	15.1
3	CIMSS	0.11	11.4	12.0	5.7	9.8	1.7	12.8
	SOS1	0.08	11.4	12.0	5.7	9.8	1.7	12.8
4	CIMSS	0.09	6.7	14.8	6.7	7.6	1.6	10.6
	SOS1	0.09	6.7	14.8	6.7	7.6	1.6	10.6
5	CIMSS	0.52	16.7	14.9	11.4	12.9	1.4	13.8
	SOS1	0.56	16.7	14.9	11.4	12.9	1.4	13.8
6	CIMSS	0.06	7.6	8.8	3.2	7.4	2.1	11.0
	SOS1	0.05	7.6	8.8	3.2	7.4	2.1	11.0

Figure 7.7 show NEADL mobility, BI and harmonised GHQ scores across the six latent classes. NEADL mobility subscale and BI index separates four distinct groups of classes that can be labelled very poor, poor, moderate and high physical function. The harmonised GHQ-6 total score produced two groups of classes that can be labelled as "depressive symptoms" and "no depressive symptoms".

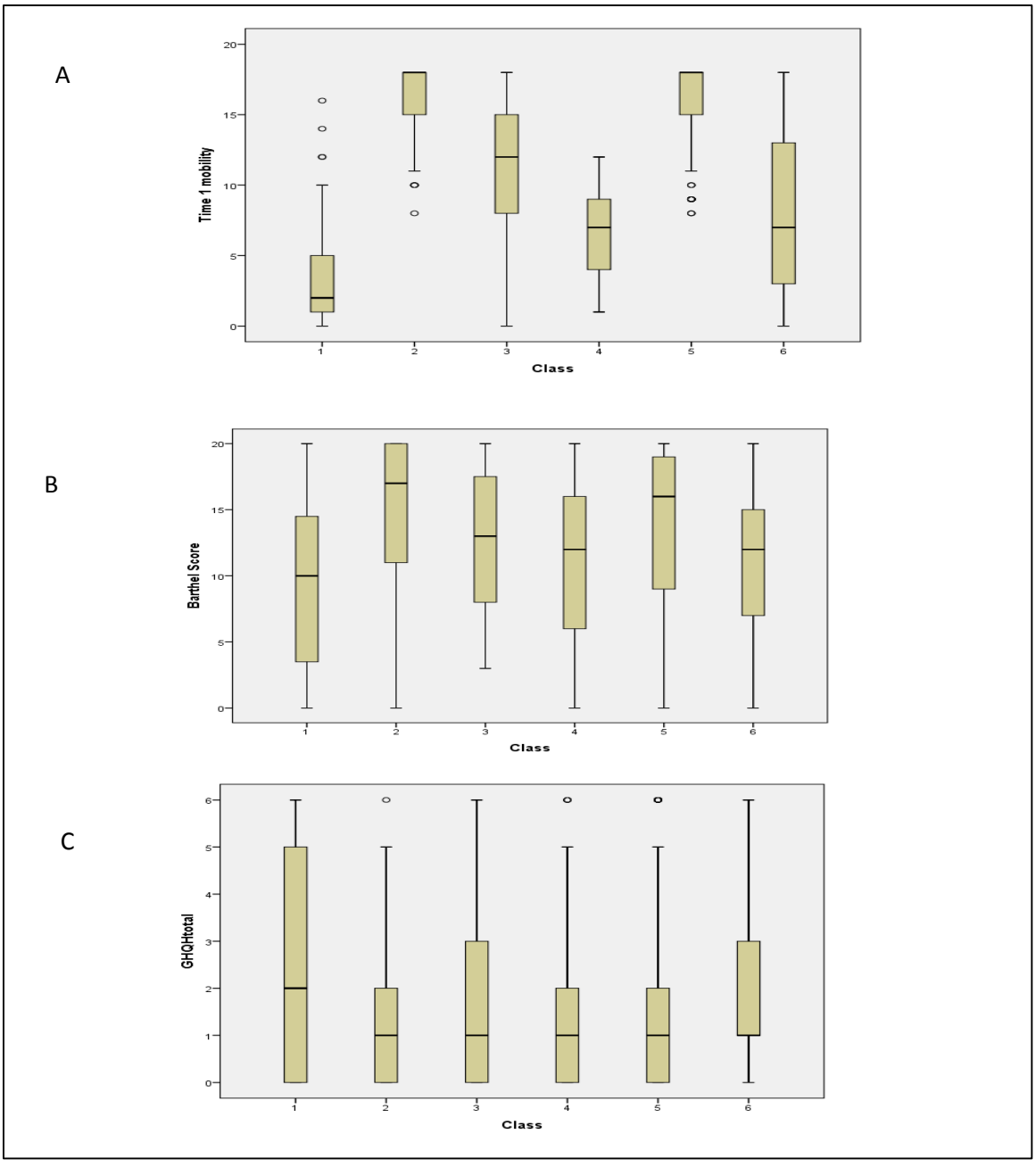

Figure 7.7 Boxplot of baseline (A) NEADL mobility, (B) Barthel Index, (C) Harmonised GHQ total, by latent class, pooled SOS1 and CIMSS datasets

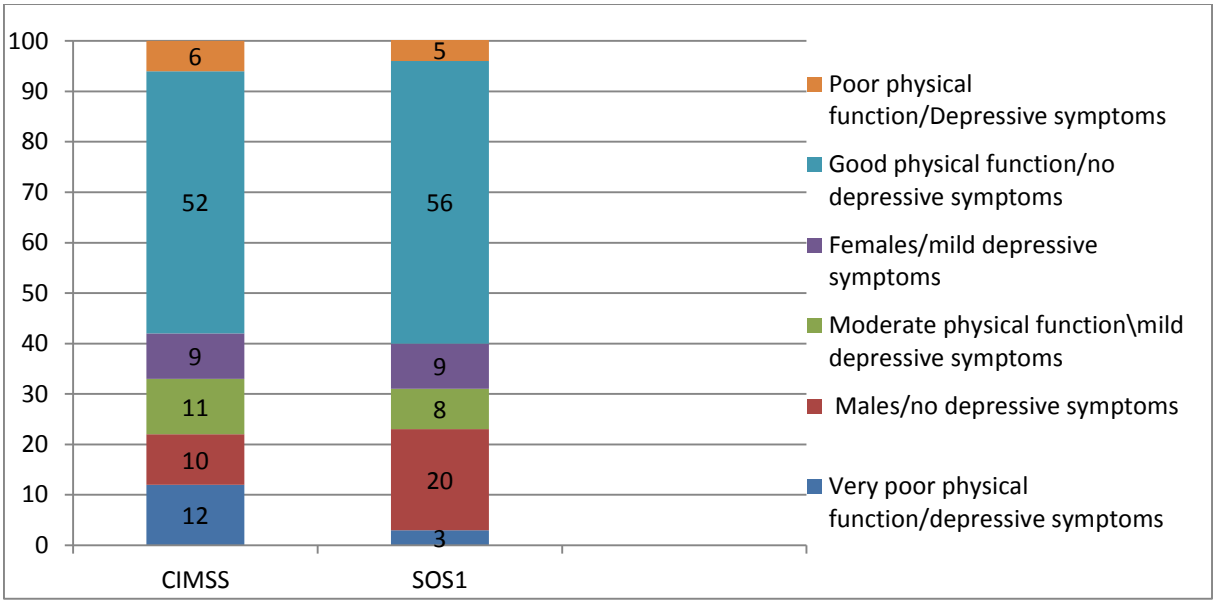

Figure 7.8 Stacked bars of class prevalence's (%) by Study

7.4.6.1 Characteristics of latent class, pooled SOS1 and CIMSS datasets

Class 1: Patients in class 1 were characterised by very poor physical function and had the highest average GHQ-6 score (Table 7.17) suggesting depressive symptoms. Based on the mean indicator scores, this class was labelled "Poor physical function, Depressive symptoms". Compared to class 5(Good physical function, no depressive symptoms) patients in this class, were more to have the following characteristics: elderly, previous stroke, urinary incontinent and were less likely to have been living alone before stroke (Table 7.18). This class had the highest proportion of patients with previous stroke (38.3%), urinary incontinent (36.2%) (Table 7.19). These patients were predominantly elderly with an average age of 75 years.

Class 2: Patients in class 2 were characterised by good physical function in both ADL and IADL, but had poor function in domestic activities (Table 7.17). The average harmonised GHQ-6 score suggested absence of depressive symptoms. Individuals in this group were predominantly males (90%). This class was labelled "Males with good physical function, no depressive symptoms". Compared to class 5 ("Good physical function, no depressive symptoms), patients in class 2 were less likely to be females and more likely to be older (Table 7.18).

Class 3: Patients in class 3 were characterised by moderate function in the NEADL mobility and NEADL kitchen subscales, and mild depressive symptoms. (Table 7.17). This class was labelled "moderate physical function, mild depressive symptoms". Compared to the

patients in class 5(Good physical function, no depressive systems) patients in class 3 were more likely to be older, less likely to be females, and more likely to have had a previous stroke (Table 7.18). The average age of individuals in this class was 73 years, 62% were males and 31% had a previous stroke (Table 7.19).

Class 4: Patients in class 4 had the second lowest average NEADL mobility scores suggesting poor mobility measured by NEADL (Table 7.17). The class also had high NEADL kitchen scores suggesting good function in kitchen tasks. This group had the highest mean age 78.5 years, and 78% of the patients in this class were females (Table 7.19). Based on the NEADL kitchen subscale average score and the demographic characteristics of individuals in this class, this class was labelled "Elderly females, mild depressive symptoms ". Compared to class 5(Good physical function, no depressive symptoms), patients in this group were more likely to be elderly females with previous stroke (Table 7.18).

Class 6: Individuals in class 6 were characterised by poor function in IADL, moderate function in basic ADL, and had the second highest average harmonised GHQH total (Table 7.17) suggesting depressive symptoms. This class was labelled "Moderate function in basic ADL, Depressive symptoms". Compared to class 5, patients in class 6 were more likely to be elderly, had a previous stroke (Table 7.18). The average age of this class was 75 years, and 29% had a previous stroke.

Table 7.18 Multinomial logistic regression results, pooled SOS1 and CIMSS datasets

	Class 1			**Class 2**			**Class 3**			**Class 4**			**Class 6**		
	RRR	***p* value**	**95% CI**	**RRR**	***p* value**	**95% CI**	**RRR**	***p* value**	**95% CI**	**RRR**	***p* value**	**95% CI**	**RRR**	***p* value**	**95% CI**
Gender Female	0.79	0.507	0.41, 1.57	0.11	**0.001**	0.06, 0.21	0.49	**0.016**	0.28, 0.88	2.01	**0.04**	1.03, 3.90	0.66	0.238	0.33, 1.31
Age	1.04	**0.012**	1.01, 1.07	1.03	**0.012**	1.01, 1.05	1.03	**0.013**	1.01, 1.06	1.07	**0.001**	1.03, 1.10	1.05	**0.002**	1.02, 1.09
Urinary Incontinence	3.74	**0.001**	1.77, 7.86	0.17	**0.017**	0.04, 0.73	0.92	0.851	0.38, 2.20	1.07	0.867	0.49, 2.28	1.09	0.861	0.42, 2.85
previous stroke	4.03	**0.001**	2.02, 8.02	1.11	0.74	0.59, 2.09	3.08	**0.001**	1.67, 5.68	3.09	**0.001**	1.63, 5.84	2.79	**0.007**	1.33, 5.88
Living alone	0.28	**0.002**	0.13, 0.62	0.35	**0.001**	0.19, 0.60	0.71	0.261	0.40, 1.28	1.15	0.66	0.62, 2.11	0.46	**0.041**	0.22, 0.97
Constant	0.01	0.001	0.001, 0.06	0.13	0.005	0.03, 0.54	0.02	0.001	0.003, 0.12	0.001	0.001	0.0001, 0.01	0.003	0.001	0.0002, 0.3

Reference group was Class 5, Good physical function, no depressive symptoms, RRR: Relative Risk Ratios

Table 7.19 Patient characteristics (n, %) by latent class: Combined SOS1 and CIMSS datasets

Baseline characteristic	**Class 1 n=47**	**Class 2 n=120**	**Class 3 n=68**	**Class 4 n=64**	**Class 5 n=420**	**Class 6 n=41**	***p value***
Age(mean, SD) years	75(11.1)	70.2(10.4)	73.0(12.8)	78.5(8.8)	69.9(12.4)	75.0(11.3)	<0.001
Gender female, n %	25(53.2%)	12(10%)	26(38.2%)	50(78.1%)	233(55.5%)	19(46.3%)	<0.001
Previous stroke	18(38.3%)	17(14.2%)	21(31.3%)	20(31.3%)	54(12.9%)	12(29.3)	<0.001
Urine	17(36.2%)	2(1.7%)	8(11.8%)	11(17.2%)	50(11.9%)	6(14.6%)	<0.001
Living alone before stroke(% yes)	10(21.3%)	21(17.5%)	24(36.4%)	39(61.9%)	183(44.3%)	13(31.7%)	<0.001

7.5 Discussion

The aim of Study 4a was to compare disability patterns across two stroke cohorts and the factors associated with these patterns, in particular the consistency of these patterns was assessed. The strength of the analyses conducted in Study 4a was that advanced statistical techniques, Latent Class Analysis and Multi-Group Latent Class Analysis were used to identify and compare patterns of disability across different stroke cohorts. Multi-Group Latent Class Analysis provided a framework for analysing combined data from multiple groups and also determining the best multi-group model (homogenous, heterogeneous or partial homogenous) for the pooled analysis of the SOS1 and CIMSS datasets. Furthermore the use of latent variable modelling approaches enabled the use of multiple disability dimensions to characterise disability patterns post-stroke and also identify factors associated with these disability patterns. The classifications of stroke patients conducted in Study 4a were based on multiple measures: physical, social, and psychological so as to provide a comprehensive classification system. This is particularly important, given the heterogeneity in disability patterns amongst the stroke patients.

The separate study analyses showed that, a five class solution was favourable for the SOS1 and a six-class solution for the CIMSS. While the number of latent classes was not the same across the two datasets, there was some similarity in the characterisation of the classes and the factors associated with class membership. Four common disability classes emerged across the two datasets: “Good physical function, No depressive symptoms”, “Poor physical function, Mild depressive symptoms”, “Poor physical function in ADL, Severe depressive symptoms”, and “Moderate physical function, Mild depressive symptoms”. In both datasets, the largest group was the “Good physical function, No depressive symptoms” class, with about 50% of the sample. Additional classes emerged as a result of the gender bias in the NEADL kitchen subscale, females showed high physical function in kitchen activities compared to males.

Based on the BIC, and the interpretability of the emerging latent classes, a partial homogeneous (equal mean indicators and variance but different class prevalence’s) six-class model was preferred for the multi-group latent class analysis of the SOS1

and CIMSS datasets. A partial homogenous model for the SOS1 and CIMSS datasets provided support for pooling these studies together and determining the disability classes prevalence's using a larger sample of n=760. The classes in the pooled data analysis were similar to those found in the separate study analysis but the classes in pooled data analysis had larger samples. Having larger classes was advantageous as this increased the statistical power for identifying the covariates that predicted class membership using the multinomial logistic regression model.

The harmonised GHQ-6 measure that was derived in Study 3b of this thesis, was less discriminating compared to the original GHQ-12 and GHQ-28 subscales. In study 4a, the total harmonised GHQ-6 score was used as the class indicator in MG-LCA instead of the subscales because the reliability analyses conducted in study 3b showed that the harmonised GHQ-6 scale had better reliability compared to the harmonised GHQ-6 subscales. Furthermore confirmatory factor analysis of the 2-factor structure showed that the factors were highly correlated suggesting the existence of a higher order or single general factor.

Results from the multinomial logistic regression in both separate study analyses and pooled data analyses showed that, age, gender female, previous stroke, and urinary incontinence were associated with class membership. The "Good physical function, No depressive symptoms" class represented young stroke survivors, while the "Poor physical function, Mild depressive symptoms" represented the elderly stroke survivors who were likely to have previous stroke and were urinary incontinent. The "Poor function in basic ADL, Severe depressive symptoms" represented younger stroke survivors who were more likely to be female with previous stroke.

The findings from the analyses conducted in study 4a were consistent with previous research. There is evidence that older age, previous stroke are associated with poor physical function (Tilling et al., 2001). The reduced physical function in older stroke survivors can be attributed to additional disabilities and comorbidities found in elderly people (Bagg et al., 2002). There is also evidence that being female is associated with depressive symptoms, and poor physical function is also positively associated with mood symptoms (Wade et al., 1987; Brown et al., 2012). A study by White et al. (2008) also concluded that feelings of dependency in basic activities of daily living contribute to low mood in younger survivors despite relatively good physical function in stroke survivors. A systematic review by Hackett et al. (2009)

also found that younger patients were more likely to experience depression post-stroke.

The disability subgroups identified in Study 4a were compared with other stroke classifications used in clinical practice or research. Some of the latent classes identified in this present study are consistent with the established AHA stroke outcome classification system (Kelly-Hayes et al., 1998) and the those reported by Sucharew et al. (2013). The American Heart Association Classification of Stroke Outcome (AHA.SOC) task force developed a global classification system that summarises the neurological impairments, disabilities, and handicaps that occur after stroke. The AHA score has five functional levels (I-V) based on basic ADL and complex levels of (IADL). Level I patients are independent in both basic ADL and IADL, Level II are patients independent in basic ADL and but partially dependent in IADL, Level III patients are partially dependent in both basic ADL and IADL. Level IV patients are partially dependent in basic ADL. Level V comprises of stroke patients who are completely dependent in both basic ADL and IADL. The profile patterns identified in this study were similar to the five classifications of the AHA classification. However in this present study identified a sixth group representing the mood disorder class.

A study by Sucharew et al. (2013) using a sample of n=2112 strokes identified six discrete profile patterns based on the 15 dichotomised NIHSS items. The median rNIHSS total scores decreased from the most severe Profile A to the mild profile F showing reduced levels of cognition. The classes were labelled: severe (profiles A and B), moderate to mild (profiles C, D, and E), and mild (profile F). Consistent with Sucharew et al. (2013), in this present study, six latent classes were also identified across the two stroke cohorts, but the characterisation was different because the classifications were based on different measurement scales. In their validation sample, Sucharew et al. (2013) found that the largest class was the mild class (Profile F) with a prevalence of 56%. This prevalence is similar to the prevalence of the mild class that was labelled “Good physical function, no depressive class” in this present study.

7.5.1 Limitations

There are limitations of the analyses that were conducted in Study 4a that warrant discussion. The datasets that were used in Study 4a excluded people with severe strokes and it could be possible that a subgroup of “very severe strokes” could have been missed in this analyses. Furthermore the samples were predominantly of

white ethnicity and both the SOS1 and CIMSS studies recruited patients from the Leeds stroke register hence findings from Study 4a may not be generalisable to the stroke population.

Another limitation was that the analysis that was conducted in Study 4a was based on baseline scores only. More research based on longitudinal data is needed to model the longitudinal developmental trajectories of patients in different latent classes for a better understanding of their recovery over time.

7.5.2 Methodological considerations

A methodological limitation of Study 4a was that the mixture modelling conducted in separate study analyses was based on the parsimonious model assuming conditional independence; other covariance matrices were not explored. These other covariance structures might yield different class structures. However, although the application of latent variable mixture modelling conducted in Study 4a was based on parsimonious models, it still identified distinct subgroups of disability patterns that represented the data reasonably well and the latent classes were clinically meaningful.

Multiple testing was corrected by using Bonferroni adjusted *p*-values, the Bonferroni correction is too conservative for large number of comparisons. However, all *p*-values for comparing patient characteristics across the latent classes were significant under the most conservative correction and any less conservative criterion would not change the interpretation of the pattern of results.

The mixture modelling conducted in Study 4a was based on cross sectional data, it was desirable to explore the utility of using latent growth mixture models using the combined datasets to identify latent classes based on the repeated measures. The use of latent growth mixture models was not possible given the differences in assessment times in the combined SOS1 and CIMSS datasets. It would be interesting in future to use mixture growth models to investigate how the latent classes found in Study 4a evolve overtime and the impact of socio-economic factors on class membership.

7.5.3 Clinical Implications

The results from this analyses presents a comprehensive classifications of disability after stroke based on multiple dimensions (physical, social and psychological function) and have the potential to impact on the care of stroke patients. For stroke survivors to receive the best care, a comprehensive stroke outcome classification system is needed to direct appropriate therapeutic interventions. The

disability subgroups identified in Study 4a and their characteristics might be clinically useful in identifying high risk stroke patients. More research is needed to determine the projected outcomes of patients in the disability subgroups identified in Study 4a so as to guide the healthcare management of these patients.

7.6 Conclusion

In conclusion, this chapter has demonstrated the utility of using MG-LCA in analysing data from multiple sources simultaneously to identify patients at high risk after stroke. Harmonisation and pooling of existing datasets was beneficial as the use of these datasets provided the opportunity to determine if latent classes and factors that determine the latent classes were consistent across different stroke cohorts. Mixture modelling provided an alternative approach for classifying stroke patients using multiple measures based on physical, social and psychological function. The pooled data analyses also provided larger samples for a better classification of disability patterns in stroke patients.

Chapter Key message

Having access to multiple stroke datasets was beneficial for comparing disability patterns across different stroke cohorts. Similar disability subgroups emerged across the SOS1 and CIMSS datasets and the factors that influence class membership were similar across the two datasets. MG-LCA seemed to be a promising approach for analysing data from multiple groups.

The next Chapter describes the fourth strand of research that was conducted in Study 4b of this thesis to illustrate the benefits of harmonising data for conducting an integrative data analysis.

Chapter 8

8 PREDICTORS OF ANXIETY AFTER STROKE: INTEGRATIVE DATA ANALYSIS OF SOS1 AND SOS2 DATASETS

Study 4b: Predictors of anxiety outcomes after stroke: Integrative data analysis of Stroke Outcomes Study 1 and Stroke Outcomes Study 2

8.1 Introduction

In Chapter 7, the benefits of analysing multiple datasets using multi-group latent class analysis for comparing the disability patterns across different stroke cohorts were demonstrated. Similar latent disability subgroups emerged across the SOS1 and CIMSS studies, and the factors associated with class membership were also similar. The fourth strand of the research that was conducted in in this thesis was Study 4b, which was an illustrative integrative data analysis of the harmonised SOS1 and SOS2 datasets with a total sample size of, n=1033. As highlighted in Chapter 1, the majority of stroke rehabilitation studies are limited by small sample sizes (Counsell and Dennis, 2001) due to difficulties in recruiting patients and attrition. Stroke is a heterogeneous condition and the effects of stroke rehabilitation outcomes are small (Counsell et al., 2001) thus large samples are needed to increase the power of statistical tests, precision of estimates and generalisability of research findings. The issues of small samples might be minimised by combining individual patient data from existing datasets (Thompson, 2009; Allen et al., 2013). These benefits have encouraged researchers elsewhere to explore the possibility of pooling existing datasets to address issues inherent in the smaller numbers of participants in each of the constituent studies (Thompson, 2009; Fortier et al., 2010); and of course it is this that has made meta-analysis a popular technique for enhancing the statistical power and analytical value of separate, smaller studies.

In Study 4b of this thesis, the benefits of pooling and analysing harmonised individual patient data from the SOS1 and SOS2 studies were demonstrated. The strength of the analyses conducted in Study 4b was to add to the existing literature on

factors associated with post-stroke anxiety by using large samples from the harmonised SOS datasets (n= 1033) and more advanced statistical technique, multi-level modelling. The effect of physical function, social function, stroke severity, somatic symptoms, and patient demographic characteristics on post-stroke anxiety was investigated in Study 4b. The majority of stroke disability outcomes studies focus on motor recovery, physical impairment and functioning. There is a paucity of studies on post-stroke depression and anxiety. Despite anxiety being frequent in stroke survivors (20% within 1 month after stroke and 24% at six months, (Campbell Burton et al., 2013)), it has been insufficiently investigated (Ferro et al., 2009). Anxiety symptoms have been shown to negatively affect long term outcomes and quality of life after stroke (Mierlo et al., 2014). As highlighted in Chapter 1, there is conflicting evidence on the effect of gender female, previous stroke, physical function in ADL on post-stroke anxiety (Menlove et al., 2015) thus more research on the effects of these factors in large and representative stoke studies is needed. It is important to investigate the factors associated with post-stroke anxiety, so as to target services to high risk stroke patients. The findings from Study 4b of this thesis might help clinicians to identify patients at high risk and improve the management of these patients.

In this present Chapter the analyses that was conducted in Study 4b is reported. The reporting format used in this Chapter followed the STROBE (Strengthening the reporting of observational studies in Epidemiology) statement (Von Elm et al., 2007). After the introduction and aims, the methods that were used in Study 4b are described in section 8.3, the results in section 8.4 and the chapter ends with the discussion of findings and conclusions.

8.2 Aims

The aim of Study 4b was to demonstrate the benefits of pooling harmonised longitudinal datasets to determine factors associated with post-stroke anxiety.

8.3 Methods

8.3.1 Data sources

The harmonised SOS1 (n=448) and SOS2 (n=585) datasets were used in Study 4b. A description of these studies and their characteristics has already been provided in Chapters 1 and 3 of this thesis. As highlighted earlier in Chapter 3, the SOS studies

were both longitudinal studies that recruited patients from the Leeds stroke register and were designed to investigate depressive symptoms after stroke and both studies recruited patients from the Leeds stroke database. Both studies collected data by face to face interviews. The sample size of the combined harmonised SOS dataset was n=1033.

8.3.2 Measures

The investigation conducted in Study 4b used measures of cognitive function measured by MMSE, psychological distress measured by GHQ-28, and functional independence by the Barthel index. The details of these measures have already been described in Chapters 1 and 3 of this thesis. As highlighted earlier in Chapter 3, these measures were administered at baseline (within 4 weeks after stroke), 1 and 2 years in SOS1, and at 3 weeks, 9, 26, 52 weeks in SOS 2. When conducting integrative data analysis, it is important to establish measurement invariance of the outcome measures across the studies so as to make valid score comparisons of outcome measures across studies (Hussong et al., 2009). In Study 2 reported in Chapter 4, measurement invariance was established for the GHQ-28 measure, but not for the MMSE, and BI measures because item data for these measurement scales were not available. Other researchers have argued that it will be excessive to assess invariance of all measures across groups and measurement invariance can be conducted for key concepts of interest (Flora et al., 2008). In Study 4b, the key measure of interest was the GHQ-28 and measurement invariance properties of this measure were established.

8.3.3 Statistical analyses

8.3.3.1 Descriptive analyses

The similarity of the categorical data across the two studies such as gender was assessed through cross tabulation and the Pearson's Chi-square test. A t-test was used to compare quantitative variables such as age. A p value of ≤ 0.05 was considered statistically significant. This was a descriptive analysis and of exploratory nature rather than to test a hypothesis. Consequently there was no adjustment for multiple testing. Individual and mean profiles of anxiety scores over time were produced to determine whether the relationship was non-linear.

8.3.3.2 Statistical modelling

A multilevel modelling approach was considered appropriate for Study 4b because the data from the two SOS studies were longitudinal data, and required non-standard statistical analyses. The repeated assessments that were made on the same

individual were not independent, hence violating the independent assumption of standard statistical models. Ignoring the dependence in the data may result in under-estimating standard errors of estimates thus inflating the type 1 errors. Type one error is rejecting the null hypothesis when it's true. The multilevel modelling framework accounts for the lack of independence in the data thus corrects the standard errors.

Another benefit for using multilevel models in Study 4b was that the two SOS studies followed patients at different time intervals, the multilevel modelling framework can be used even if studies have different patient follow-up intervals, unlike methods such as repeated measures ANOVA that excludes patients with incomplete data. The utility of using the multilevel modelling approach in stroke rehabilitation research has been demonstrated in several studies (Toschke et al., 2010; Tilling et al., 2001; Pan et al., 2008; Kwok et al., 2006). The next section provides a description of multilevel models.

8.3.3.3 Multilevel models

A multilevel model is an extension of ordinary regression model that has already been described in Chapter 5. In a multilevel modelling framework, data structures are hierarchical (Hox, 1995). For example in longitudinal studies where patients are assessed three times, the repeated measurements within a person are viewed as hierarchical as shown in Figure 8.1

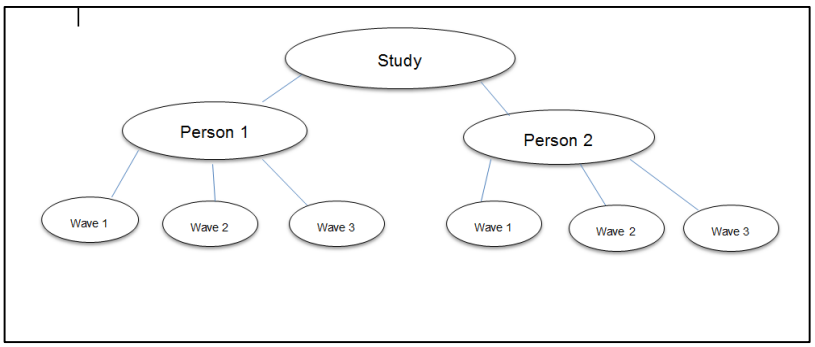

Figure 8.1 Hierarchical data structure for repeated measurements

The lowest level of the hierarchy is referred to as level 1. For example in Figure 8.1, the repeated measurements at the different waves are the level 1 units and subjects or persons are the level 2 units. In Study 4b, an integrative data analysis of the harmonised pooled SOS1 and SOS2 datasets was conducted using a multilevel Poisson model. The Poisson distribution is used to model count data and in Study 4b,

the frequency of anxiety symptoms were considered to be count data. Hox et al. (2010) p.8 provided the mathematical presentation of the multilevel Poisson model as shown in the equations below:

$$y_{ij}|\lambda_{ij} = \text{Possoin}(m_{ij}\ \lambda_{ij}) \qquad \text{Equation 8.1}$$

Where y_{ij} is the count for person i in group j and λ_{ij} is the event rate (lambda) and the model can be expanded by including a varying exposure ratem_{ij} .

The standard link function for the Poisson model is the logarithm shown in Equation 8.2

$$n_{ij} = \log(\lambda_{ij}) \qquad \text{Equation 8.2}$$

The link function links the expected value of the outcome variable "y" to the predictors.

The level 1 and level 2 models are derived using Equation 8.3 are shown below:

$$\eta_{ij} = \boldsymbol{\beta}_{0j} + \beta_{ij}\mathbf{X}_{ij} \qquad \text{Equation 8.3}$$

Where β_{0j} is the random intercept, β_{1j} the random slope and $\mathbf{X}_{ij}$ are predictors in the fixed part of the model and

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\mathbf{Z}_j + u_{0j} \qquad \text{Equation 8.4}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\mathbf{Z}_j + u_{1j} \qquad \text{Equation 8.5}$$

Where u_{0j} and u_{1j} are individual level residuals, γ_{00} is the average initial count, γ_{10} is the average rate of change and $\mathbf{Z}_j$ are predictors in the random part of the model.

Substituting β_{0j} and β_{1j}into Equation 8.3 yields Equation 8.6 shown below:

$$\eta_{ij} = \gamma_{00} + \gamma_{10}\mathbf{X}_{ij} + \gamma_{01}\boldsymbol{Z}_j + \gamma_{11}\boldsymbol{X_{ij}Z_j} + u_{0j} + u_1\boldsymbol{X_{ij}} \qquad \text{Equation 8.6}$$

The Poisson model assumes that the variance is equal to the mean. Over dispersion occurs when the variance exceeds the mean. When the variance is less than the mean this shows under dispersion and it suggests a misspecification of the model (Hox et al., 2010).

8.3.4 Application of Multilevel Poisson model in Study 4b

The statistical modelling conducted in Study 4b was exploratory data analyses to investigate the factors associated with post-stroke anxiety with no specific hypothesis to test. The analyses were conducted in two steps. In step 1, the relationships of

covariates and outcomes were investigated in separate datasets so as to determine whether these relationships were similar across studies. In step 2, the analyses were repeated in the pooled dataset using integrative data analysis. In both the study-specific and Integrative Data Analysis (IDA) of the combined SOS studies, multilevel models were used to investigate the factors associated with anxiety symptoms accounting for the hierarchical nature of the data (repeated measurements nested within individuals).

The primary outcome in Study 4b was anxiety symptoms measured by the GHQ-28 anxiety subscale. The outcome was treated as a count indicating the number of anxiety symptoms an individual reported. Due to the skewed distribution of the outcome and the many zeros, a multilevel Poisson model was preferred for the data. The Poisson model accounts for the excess zeros in the data and the skewed distribution of the outcome variable. The negative Binomial model can be used if the data violates the poison assumption (mean= variance) or if data is over dispersed. In Study 4b, the results from the multilevel negative binomial were similar to the multilevel Poisson model hence in this Chapter the results of the Poisson model are reported.

The predictors that were used in the models were the variables that were identified as relevant in the Data Schema developed in Chapter 3. These covariates have been shown to be associated with stroke disability outcomes. The predictors are shown in Table 8.1. As was highlighted in Chapter 3, pooling the two SOS datasets resulted in a loss of important clinical variables from the SOS2 study that have been shown to be associated with patient outcomes post-stroke such as: stroke side, marital status, hemianopia, aphasia, and antidepressants. The within individual factors that were explored in Study 4b were the PROMs scores that were recorded at the different time points. Assessment time was included as a dummy coded variable in the models. The individual and mean anxiety profiles suggested a non-linear relationship between time after stroke and anxiety scores hence a quadratic term for time (time squared) was included in the models.

Table 8.1 Naming and coding of variables

Predictors	Coding
Gender	Male=1, female=2
Previous stroke	Yes=1/no=0
Living alone before stroke	Yes=1/no=0
Age in years	-
Previous stroke disability	Yes=1/no=0
Urinary continence	Yes=1/no=0
Pre stroke disability	-
BI, GHQ-28 anxiety, GHQ-28 depression, GHQ-28 social dysfunction	-
Assessment time(months)	

8.3.4.1 Integrative Data Analysis: SOS1 and SOS2 datasets

As highlighted in Chapter 7, the main issue for pooled data analysis in epidemiologic studies is whether differences in the populations and methods used in original studies influence the results obtained from the pooled data analysis (Friedenreich, 1993). Systematic differences between cohorts may confound inferences drawn from the analysis of combined data (Verma et al., 2009) hence the heterogeneity between studies needs to be addressed to permit valid inferences to be drawn from the analysis of pooled data. In Study 4b, the regression coefficients from the separate analysis of the SOS1 and SOS2 datasets showed that the relationships between covariates and outcomes were similar across the two cohorts. The comparisons of the SOS1 and SOS2 studies conducted in Study 3 reported in Chapter 3 showed that these two studies were comparable in terms of their aims, study designs, sampling and data collection methods. The two studies were both observational longitudinal studies that recruited patients from the Leeds stroke database, sampling methods were highly similar and had similar inclusion and exclusion criteria. In both studies data was collected by interviews and the response rates were very high, greater than 90%. The baseline assessments were conducted within four weeks after stroke in both studies. The descriptive analysis conducted in Chapter 3, showed that the patient characteristics were also comparable in terms of: age, cognitive function measured by the MMSE, proportion of patients with previous stroke, proportion of females. There were differences in levels of disability, patients from the SOS2 dataset had a higher

average BI compared to the SOS1 suggesting that the SOS2 patients were fitter compared to SOS1.

The pooled data analysis conducted in Study 4b, raised methodological challenges of dealing with different follow-up times in the two cohorts, and the differences in the times the studies were conducted. As highlighted before, the SOS2 followed-up patients at 9 weeks, 12 weeks, 26 weeks, and 52 weeks, while the SOS 1 followed patients at 1 and 2 years. Furthermore the SOS1 was conducted between 1995-1999, while the SOS2 was conducted between 2002-2005. Failing to account for systematic differences across the two studies during pooled data analysis might confound inferences drawn from the analysis of the combined dataset. Time was harmonised by converting time intervals to months in the SOS2 so as to match with the SOS1 time units. Harmonising time in the combined dataset provided more assessment intervals for the pooled data analysis.

The pooled data analysis conducted in Study 4b followed Curran et al. (2009)'s Integrative Data Analysis (IDA) approach. The IDA approach fits statistical models directly to the pooled individual person dataset. Curran et al. (2009) and Thompson (2009) argued that if the individual data from the multiple studies were available, there are many advantages of fitting models directly to the original (raw) data rather than synthesising summary statistics. More complex analyses that might not be possible within the original studies can be conducted in the aggregated datasets and adjustment for the same confounders across studies can be made (Thompson et al., 2009). In Study 4b, the IDA approach was preferred because the individual person data from the participating studies were available, and combining the individual person longitudinal data, provided more time points than in any one study alone. IDA has been successfully used in psychological research of developmental trajectories. For example a study by Hussong et al. (2008) used IDA to examine the unique predictability of trajectories of child internalising symptoms from parental alcoholism.

In Study 4b, between studies heterogeneity was accounted for by modelling the effects of study membership directly into the model, unlike in the random effects IDA where between studies heterogeneity is modelled as a random effect. Random effects IDA models were not used in Study 4b to account for between study heterogeneity because there were only two studies that were pooled and these were considered to be insufficient for estimating the study-level variance parameter. In the fixed IDA models

that were fitted in Study 4b, cross-study effects were controlled by including a dummy covariate for the variable study (Study =1 for SOS1 and 2 for SOS2) in the model to represent the study effect. The main effects of the dummy variable compared the average outcome for SOS1 and SOS2. The main advantage of modelling study membership directly into the model is that one can also estimate interactions between covariates and study membership (Hussong et al., 2013). The interaction terms allows for testing of any differential impact of covariates on outcomes across different studies. For example a significant interaction between gender and study indicates that the relationship between gender and the outcome vary by study. For parsimony, non-significant interactions were not included in the model.

The adjusted effects of the predictors were determined using the incidence rate ratios and confidence intervals produced from the Poisson models. The results from the Integrative Data Analysis of the pooled SOS1 and SOS2 datasets were compared with results from the traditional aggregated meta-analysis of summaries from the separate study analyses.

8.3.5 Model Estimation

The statistical analyses in Study 4b was conducted using STATA version 13 (StataCorp, 2013). The meqrpoisson and menbreg procedures in STATA were used to fit the multilevel Poisson models and negative binomial models.

8.3.6 Sample size

The statistical modelling conducted in Study 4b investigated the association of patient characteristics, baseline stroke severity, clinical and socio-economic factors, physical function, social function, and depression on post-stroke anxiety. There were approximately more than 20 variables that were investigated including interactions of interest. A total sample size of n=1033 was used with repeated assessments made on individuals ranging from 2 to 5 per individual. The sample size used in Study 4b, was considered large enough to ensure reliable prediction following recommendations by Harrell et al. (1996). Harrell et al. (1996) recommended 10-20 observations per parameter estimated for prognostic models. Following this rule of thumb, a sample size of n=1033 would be able to estimate 50–100 parameters.

8.3.7 Model diagnostics

The Poisson model assumes the mean to be equal to the variance. The magnitude of the mean and variances was examined to determine whether the data were over

dispersed. The magnitudes of the mean and variances showed mild over dispersion hence the results from the Poisson models were compared with those from the negative binomial models to determine if the Poisson results were biased by over dispersion.

8.3.8 Missing data and drop out

Longitudinal studies suffer from participant drop out due to attrition. Attrition reduces the sample size causing loss of statistical power and also affects the generalisability of study findings if patients who drop out have different characteristics from those who remain in the study. If attrition is systematically related to outcomes of interest or to variables correlated with outcomes of interest then the estimates of the relationships may be biased (Banks et al., 2011). In Study 4b, factors associated with attrition in each dataset (SOS1, SOS2) were investigated. The focus was on attrition of living respondents and not attrition through deaths. Information on the possible determinants of attrition is important for a proper interpretation of findings from longitudinal data analysis (Twisk and de Vente, 2002). A multivariable logistic regression model was used to investigate the effects of age, gender, baseline stroke severity, living alone before stroke, presence of disability before stroke and baseline GHQ-28 score on the probability of dropping out of the study one year after stroke.

8.4 Results

At one year, of the n=585 participants in SOS2, 33 died, 67 were withdrawn and 485 completed the study. Sixty one of the n=448 subjects in SOS1 could not be interviewed at 12 months because, 2 moved away, 25 died, 8 were too ill, 26 refused (dropout), and at 2 years the sample size was 364, 45 died, 5 too ill, 32 refused, 2 emigrated.

8.4.1 Descriptive Analyses

8.4.1.1 Patient's characteristics

The patient characteristics of patients in the SOS1 and SOS2 datasets are shown in Table 8.2. The combined dataset of the two SOS datasets had n=1033 patients with an average age of 70 years (SD=11.8) and 460(44.5%) were females. In both SOS studies, individuals were admitted to stroke units within a month after stroke. The two SOS datasets were comparable in terms of age, proportion of females, patients with previous stroke, and baseline mean GHQ-28 scores (Table 8.2). Patients in both

samples were predominantly white but the SOS2 had a higher proportion of whites (P<0.001). The SOS2 study also had a higher proportion of urinary incontinent patients, and a higher average cognitive (measured by MMSE), and functional independence (measured by BI) (p<0.001). These descriptive statistics suggested that patients in the SOS2 study were fitter compared to those in the SOS1. The combined dataset had a wider age range (18-97) compared to the single datasets (SOS1:18-94; SOS2:22-97).

Table 8.2 Baseline patient characteristics by study

	SOS2 n=585	**SOS1 n=448**	***p*-value**	**Overall n=1033**
Gender female	253(43.2)	207(46.2)	0.343	460(44.5)
Mean Age (SD)	70.34(11.9)	70.75(11.6)	0.582	70.52(11.8)
Age range	18-94	22-97		18-97
Ethnicity white	577(98.6%)	426(95.1%)	<0.001	1003(97.1)
MMSE	26.6(2.8)	25.3(3.1)	<0.001	26.05(3.04)
Pre-stroke BI	19.5(1.5)	19.3(1.3)	0.098	19.4(1.4)
Previous stroke	129(22.0)	94(21.0)	0.627	223(21.6)
Urine	107(18.3)	30(6.7)	<0.001	137(13.3)
Living	193(32.9)	175(39.1)	0.05	368(35.6)
BI	15.5(5.2)	13.6(5.3)	<0.001	14.7(5.4)
GHQ-28	6.45(5.72)	7.70(6.22)	0.104	6.80(6.01)
Died at 12 months	33(5.6)	25(5.6)	0.189	58(5.65)

Mean (SD) are reported for continuous variables and, n and % for categorical data, urine: urinary incontinence, Living: Living alone before stroke

8.4.1.2 GHQ-28 anxiety profiles, SOS1 and SOS2

The changes in the average GHQ anxiety scores over time are shown in Table 8.3. In both the SOS1 and SOS2, anxiety symptoms improved over time (Table 8.3, Figure 8.3). Figure 8.2 shows a sample of individual anxiety score profiles. Some individual anxiety profiles showed a non-linear relationship between the anxiety scores and time post-stroke (Figure 8.2).

Table 8.3 Mean (SD) GHQ-28, anxiety scores by time and study

Occasion	Mean (SD)
SOS2	
3 weeks	1.5(1.48)
9 weeks	1.3(1.30)
13 weeks	1.0(1.02)
26 weeks	1.0(1.00)
52 weeks	0.88(0.88)
SOS1	
1 month	1.4(1.90)
1 year	1.3(1.88)
2 years	1.3(1.88)

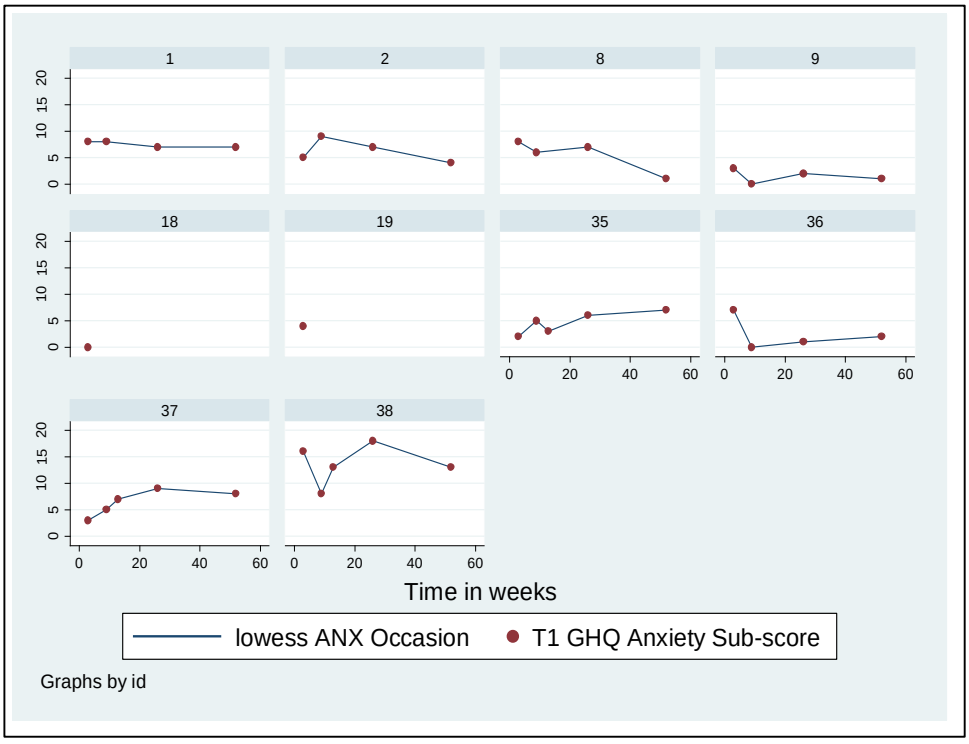

Figure 8.2 Individual Anxiety profiles measured by GHQ-28 anxiety subscale, SOS2 study

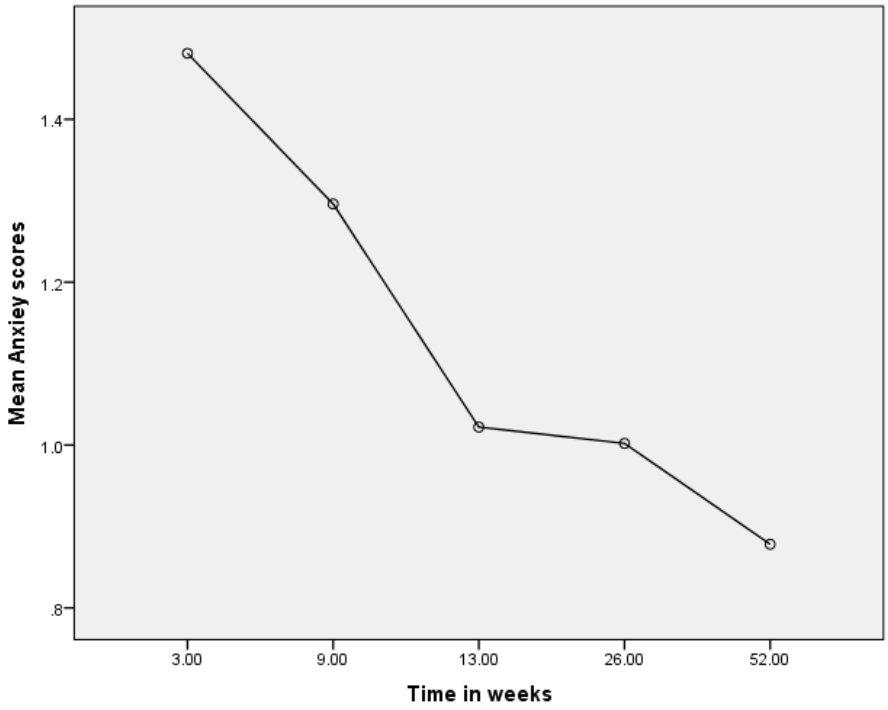

Figure 8.3 SOS2: Unadjusted mean anxiety scores by time in weeks, SOS2 study

8.4.2 GHQ-28 anxiety subscale floor effects in SOS1 and SOS datasets

In order to determine whether the GHQ-28 anxiety subscale has ceiling effects, the proportion of patients with anxiety score of zero were determined at each assessment time interval. The results are shown in Table 8.4. Histograms of baseline GHQ-28 anxiety scores in both SOS1 and SOS2 datasets were also produced and these are shown in Figure 8.4. Figure 8.4 shows that the distribution of baseline GHQ-28 anxiety scores in both studies was right skewed. The baseline mean GHQ-28 anxiety scores was 1.35 in SOS1 with variance 3.70 and mean was 1.48, variance=3.60 in SOS2. The variances were slightly higher than the mean, suggesting mild over dispersion. As highlighted earlier, the Poisson model assumes that the mean is equal to the variance.

In the SOS1, about 51.3% of the patients had GHQ-28 anxiety scores of zero at baseline and the proportion was 53.3% at 2 years. In the SOS2 about 45% of the patients had GHQ-28 scores of zero at baseline and the proportion was 65.5% at 1 year. The proportion of patients with zero scores increased overtime because of patients' recovery overtime.

Table 8.4 Proportion of patients with GHQ-28 anxiety scores=0 (floor effects by time and SOS2 study

Occasion	Proportion of patients with (GHQ-28 Anxiety =0)
3 weeks (n=585)	263(45%)
9 weeks (n=542)	273(50.4%)
13weeks(n=500)	297(59.4%)
26weeks(n=497)	302(60.8%)
52weeks(n=481)	315(65.5%)
SOS1	
1 month,448	230(51.3%)
1 year n =387	199(51.4%)
2years n=364	194(53.3%)

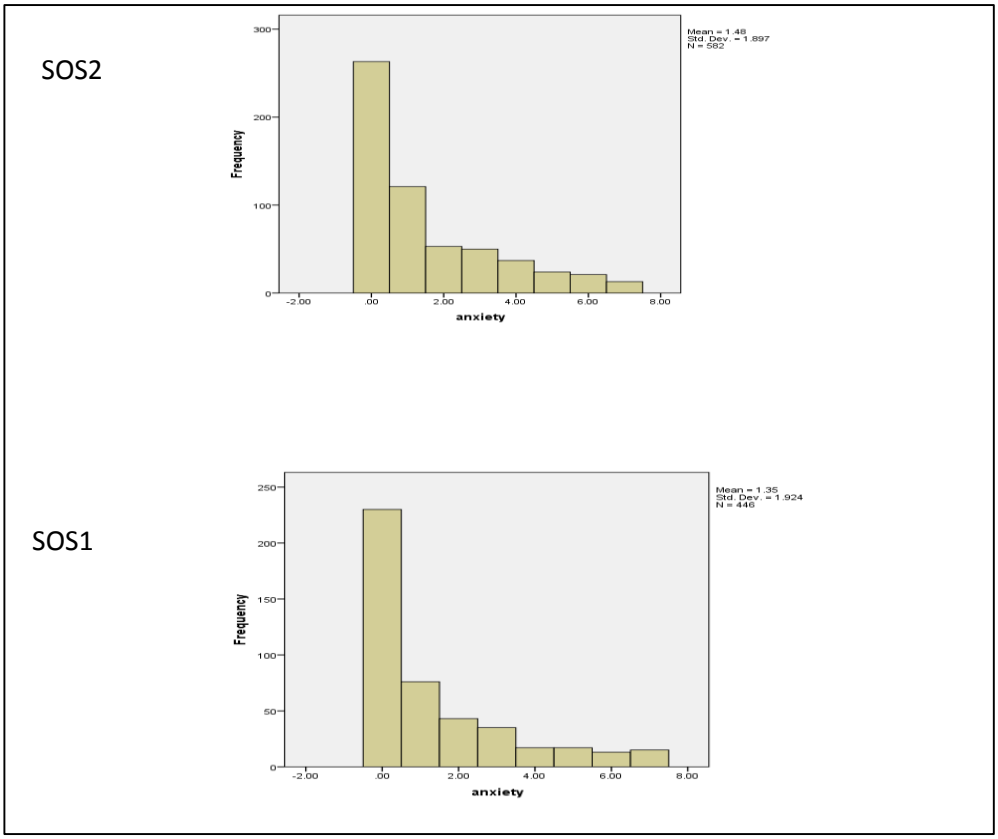

Figure 8.4 Distribution of baseline GHQ-28 Anxiety scores in SOS2 and SOS1 datasets

8.4.3 Factors associated with anxiety symptoms after stroke

The results of the statistical models that were fitted to investigate the factors associated with anxiety symptoms in SOS1, SOS2, and the combined SOS datasets are shown in Table 8.5. When data were analysed using a Negative Binomial model, the results were similar to the Poisson model hence in this Chapter the results of the Poisson model are presented. In both the study specific analyses and pooled data analysis there was no evidence of an association of time (in months) and anxiety symptoms (Table 8.5). Increased age, and increased physical function measured by BI were associated with decreased anxiety symptoms in both the study specific analyses and the IDA of the pooled datasets. Increased depressive symptoms, increased social dysfunction, increased somatic symptoms were associated with increased anxiety symptoms.

The SOS1 and SOS2 datasets produced inconsistent results on the association of gender female and anxiety symptoms. Evidence of an association of anxiety symptoms

with gender female was found in the SOS2 dataset and not in SOS1, but the direction of the effect was the same (positive) in both datasets. The integrative data analyses of the combined datasets suggested a positive association of gender female with anxiety symptoms. The standard error of the regression coefficient for gender was approximately 0.10 in both SOS1 and SOS2, and 0.07 in the integrative data analysis, showing a reduction in error of 0.03, thus the 95% confidence interval for the gender effect was more precise in the pooled data analysis.

The results of the integrative data analysis were compared with those from the aggregated meta-analysis which gives weights to the different studies and were similar (Appendix H).

Table 8.5 Predictors of anxiety symptoms post stroke by study and integrative data analysis of SOS1, SOS2 datasets

Variable	**SOS1 n=448**	**SOS2 N=585**	**Integrative data analysis SOS1 and SOS2 n=1033**
Fixed effects	**IRR(95% CI)**	**IRR(95% CI)**	**IRR(95% CI)**
constant	1.33(0.24,7.42)	0.61(0.13,2.89)	0.84(0.26,2.68)
Slope(months after stroke) Quadratic(months)	1.00(0.98,1.02) 1.00(0.99,1.00)	0.98(0.94,1.02) 1.00(0.99,1.00)	0.99(0.97,1.00) 1.00(0.99, 1.00)
Gender female	1.13(0.95,1.33)	1.32(1.13,1.55)***	1.23(1.10,1.38)***
Age(years)	0.99(0.98,0.99)**	0.98(0.98,0.99)**	0.99(0.98,0.99)***
Previous stroke	0.89(0.72,1.09)	1.05(0.87,1.27)	0.97(0.84,1.11)
Urine	0.91(0.65,1.28)	1.01(0.82,1.24)	0.98(0.83,1.17)
MMSE	0.98(0.95,1.01)	0.99(0.96,1.02)	0.98(0.96,1.01)
Pre-BI	1.00(0.94,1.07)	1.03(0.97,1.09)	1.02(0.98, 1.07)
Social	1.13(1.09,1.17)***	1.15(1.12,1.18***)	1.15(1.13,1.18)***
Depression	1.21(1.16,1.26)***	1.13(1.09,1.16)***	1.15(1.13,1.18)***
Somatic	1.21(1.17,1.26)***	1.18(1.15, 1.22)***	1.19(1.17,1.22)***
BI	0.98(0.97,1.00)**	0.98(0.97,0.99)*	0.98(0.97,0.99)**
SOS1	-	-	1.06(0.93,1.20)
Variance components			
Var(cons)	0.35(0.25, 0.48)	0.49(0.38, 0.62)**	0.44(0.36, 0.53)
Log L	-1582.64	-3035.32	-4645.52

** $p<0.01$, *** $p<0.001$, IRR: Incidence Rate Ratio

8.4.4 Missing data

Factors associated with attrition were investigated in each dataset separately. The focus was on attrition of living respondents and not attrition through deaths. Table 8.6 shows the descriptive statistics for patients who dropped out of the study at 12 months and those that didn't. In the SOS2 dataset, the mean age suggested that the patients who dropped out were older compared to those who remained in the study. In SOS1 the average age of patients who dropped out was lower than that of those who remained suggesting that patients who dropped out were younger compared to those who remained. In SOS1, patients who dropped out also had a lower baseline cognitive function compared to those who remained.

Table 8.6 Characteristics of patients who completed the study and those who did not: SOS1 and SOS2 datasets

	Completed 12 months	**Withdrawn/refused**	**Died**	***p*- value**
SOS2				
Age	69.46(12.03)	74.5(10.3)	74.9(10.6)	<0.001
Pre BI	19.5(1.5)	19.5(1.1)	19.13(1.8)	0.354
MMSE	26.8(2.9)	26.03(2.64)	25.1(3.1)	<0.001
Gender female	201(41.4)	39(58.2)	13 (39.4)	0.03
SOS1				
Age	70.5(11.4)	68.8(14.7)	77.3(7.4)	0.01
Pre BI	19.4(1.3)	19.4(1.2)	18.9(1.7)	0.164
MMSE	25.5(3.1)	23.9(3.2)	24(3.6)	0.002
Gender female	173(44.7)	21(58.3)	13(52.0)	0.244

The results of the logistic regression to investigate factors associated with attrition are shown in Table 8.7. The separate study analyses showed that in the SOS1, individuals with poor baseline physical function measured by the BI index were more likely to drop out at 1 year, while in SOS2 the older people were more likely to drop out of the study.

Table 8.7 Regression results from drop-out models (Odds ratios and 95% confidence intervals) by study

	Model for drop out (excluding deaths), SOS1	Model for drop out (excluding deaths), SOS2
Age	0.97(0.95,1.00)	1.03(1.01, 1.06)*
Gender female	1.50(0.72, 3.14)	1.63(0.93, 2.83)
Baseline GHQ-28	0.99(0.93,1.05)	1.01(0.98, 1.05)
Baseline BI	0.93(0.86,0.99)*	0.99(0.96, 1.03)
MMSE	0.89(0.79-1.00)	0.95(0.86, 1.05)
Previous stroke	1.56(0.71, 3.44)	0.92(0.47, 1.78)

*P<0.05

8.5 Discussion

In Study 4b the benefits of pooling two longitudinal studies were illustrated in an integrative data analysis of the SOS1 and SOS2 datasets to investigate factors associated with post-stroke anxiety. The strength of the analysis conducted in Study4b was that a large sample (n=1033), and an advanced statistical model Multilevel Poisson models were used to investigate the factors associated with post-stroke anxiety. The benefits of conducting an integrative data analysis of the pooled harmonised SOS datasets in Study 4b were: increased sample size, increased precision of some regression estimates, and increased representativeness of the sample. The combined dataset comprised repeated observations on n=1033 individuals assessed on between 3 and 5 occasions. In addition to increased sample size, the combined dataset had a wider range of age: ranging from 18 to 97 years. The availability of two datasets offered the advantages of investigating factors associated with anxiety symptoms using a larger dataset.

The analyses that were conducted in separate datasets showed that, generally the factors associated with anxiety symptoms were similar across the two SOS datasets, except for gender female. The SOS1 and SOS2 datasets yielded conflicting evidence of the association of female gender and anxiety symptoms. The SOS2 suggested a significant positive association of gender female and post-stroke anxiety and this association was not found in SOS1. The integrative analysis of the combined SOS datasets suggested a significant association of gender female and anxiety symptoms, demonstrating the benefits of pooled data analysis to increase sample size. In both the separate studies and integrative data analysis, increased social dysfunction, increased

depressive symptoms, increased somatic symptoms were associated with increased anxiety symptoms. Increased age and increased physical function was associated with decreased anxiety.

Consistent with findings from Study 4b, Ayerbe et al. (2013) reported a significant association of gender female and anxiety at 1 month, 1 year and 3 years in stroke survivors, increased disability, increased depression , decreased ADL with anxiety symptoms post-stroke. In this present study there was no evidence for an association of time since stroke, living alone before stroke or urinary incontinence, previous stroke with anxiety symptoms. Contrary to the findings in this present study, Lincoln et al. (2013); (Merriman et al., 2007) using sample sizes of n=220, and n=102 respectively reported an association of time since stroke and post stroke anxiety. The discrepancies with this present study could be because of differences in follow-up intervals.

Consistent with the findings in this present study, Broomfield et al. (2015) and Leppävuori et al. (2003) suggested a significant association of younger age and anxiety. Contrary to findings from this present study, Merriman, (2007) reported an association of previous stroke with anxiety symptoms. In this present study there were few people with previous stroke hence this association could have been missed due to lack of statistical power.

The missing data analysis in separate studies suggested that age, and initial physical function in ADL were associated with patients dropping out of the studies at one year. Individuals with increased initial physical function were less likely to drop out at one year after stroke. Older age was also associated with increased probability of dropping out at 1 year. These results suggest that complete case analysis assuming missing completely at random may be biased as this assumption may not be true.

8.5.1 Strengths

The analysis conducted in Study 4b had several strengths. Firstly the integrative data analysis was conducted using a larger sample (n=1033) compared to the majority of previous stroke longitudinal studies. Secondly advanced statistical modelling technique, multilevel Poisson model was used to determine the factors associated with post-stroke anxiety. The Poisson model is a unique approach for modelling skewed count data in stroke disability research. Most studies found in literature model the count data as normally distributed and this could lead to biased regression estimates

because count data is not normally distributed and cannot be negative. The multilevel modelling framework allowed me to include participants with incomplete data, thus increasing the sample size and representativeness of the sample.

8.5.2 Limitations

There are limitations to the analyses that were conducted in Study 4b that warrants discussion. While much larger than the individual studies, the number of variables in the combined dataset was smaller than the original datasets. There was a trade-off between increasing the sample size and losing variables that are known to influence patient disability outcomes. The pooled data analysis lost many important variables such as marital status, smoking, occupation, stroke type, stroke side, hemianopia, aphasia and treatments that were collected by the SOS2 study and were missing in the SOS1 study. The covariates that were dropped from the pooled data analysis were variables that could not be harmonised. It was not possible to investigate the influence of treatments on anxiety because the data was not available in both the SOS1 and SOS2 datasets. Collecting different demographic and socio economic factors makes data harmonisation of existing studies difficult or even impossible. Standardising the information collected by stroke rehabilitation studies may facilitate data sharing in stroke outcomes research. Attempts to develop a minimum dataset for stroke have been made Teale et al. (2011).

The missing data analyses conducted in Study 4b showed that missing-ness was not at random in both the SOS1 and SOS2 datasets and this might have biased the results of Study 4b. Selection bias due to attrition in longitudinal studies may affect the validity of the study if those who drop out differ in characteristics from remaining patients (Gill et al 2012).

8.5.3 Implications to stroke researchers and clinicians

Pooling data from multiple studies was beneficial but there was a trade-off between increasing sample size and losing data from important covariates and outcome measures hence there is need for stroke rehabilitation researchers to standardise data that are collected to facilitate successful sharing of existing datasets or comparisons of findings across studies. A systematic review by Menlove et al. (2015) of predictors of anxiety in stroke survivors also expressed the same views that stroke research methodologies should be standardised. Menlove et al. (2015) recommended that researchers should use standard agreed measures, cut-offs, time points of assessments, and methods of data analyses to enable better comparability

between studies. However standardising research methods, measures and times of follow up may not be practical hence the need for accurate harmonisation algorithms.

Clinicians should pay special attention to female patients, patients with previous stroke, patients with poor physical function, younger stroke survivors as they are at a significantly higher risk of post-stroke anxiety. Study 4b provided evidence of a positive association between depression and anxiety hence patients with depression need to be screened for anxiety symptoms too.

8.6 Conclusion

In conclusion, the benefits of pooling data from multiple longitudinal studies for increased statistical power, precision, generalisability have been demonstrated in Study 4b. A large sample of n=1033 stroke patients was obtained from the combined SOS datasets. The analyses conducted in Study 4b suggested a significant association of anxiety symptoms with, previous-stroke, physical function, depression, and social function. Earlier identification of patients at risk of post-stroke anxiety symptoms could lead to better treatment of these patients. Despite increasing the sample size, pooling data from different sources resulted in losing covariates that are known to be associated with stroke outcomes. To facilitate retrospective harmonisation of future stroke studies, this present study recommend that a Data Schema or a minimum dataset should be developed in stroke research and stroke research studies need to collect the minimum data that constitutes the Data Schema.

Table 8.8 Advantage and disadvantages of pooling individual person data, findings from this study

Advantages	Disadvantages
-Larger sample size hence more power and precise estimates for effect sizes -More time points hence wider coverage of recovery periods(3 to 5 occasions, within a month to 2 years) -More representative sample than in any single study -Increased Precision -More generalisability -Increased age range (18-97 years)	-Important covariates that were not collected by both studies and could not be harmonised were excluded from the analysis

Chapter 9

9 DISCUSSION

Longitudinal studies of patient-reported health outcomes for complex conditions such as stroke are often limited in size due to the difficulties of patient recruitment and retention. For stroke research, in particular, high mortality can reduce the number of patients that can be followed-up in the long term. Efforts to develop stroke prognostic models or identify care processes associated with good patient outcomes have tended to involve small longitudinal studies due to factors such as poor recruitment rates or high attrition rates. As a result, studies may lack statistical power for subgroup analyses and may not be generalisable.

Due to the heterogeneity of stroke and its many confounding factors, stroke disability outcomes studies require large samples for complex case-mix adjustment (Flick, 1999), for example using propensity score matching or statistical modelling. The problems of small sample sizes achieved by previous studies can be minimised by harmonising and pooling secondary data from similar studies to provide high quality large datasets that can be used to develop stroke prognostic models or understand factors associated with disability outcomes after stroke. Pooling results from various studies has made meta-analysis popular techniques for enhancing the statistical power and analytical value of separate, smaller studies.

However, while pooling secondary data appears an attractive solution to the small samples achieved in most stroke studies, careful attention has to be paid to differences in sampling, study designs and the measurement instruments used (Hofer and Piccinin, 2009). Combining data from multiple sources may introduce complexities both in harmonising cross-study measurements and in drawing appropriate inferences in hypothesis testing (Hussong et al., 2013).

This thesis set out to evaluate the feasibility of harmonising and pooling four different stroke datasets in order to generate large(r), high quality databases that could be analysed to inform a better understanding of stroke outcomes (in particular, the complex interplay of patient characteristics, stroke clinical factors, stroke severity,

socio-economic factors, treatments and patient disability outcomes). Challenges associated with pooling existing individual stroke patient datasets were identified and an attempt was made to address some of them using novel statistical methods. The study builds upon an extensive literature review of statistical methods that have been developed and used to harmonise data from multiple sources.

This present study makes a contribution to knowledge firstly by identifying the barriers that may hinder data sharing in stroke rehabilitation research such as measurement comparability; and secondly by investigating how some of these barriers can be overcome using the application of novel statistical methods. Recommendations from this thesis may be useful for future data harmonisation studies. Furthermore the majority of the harmonisation work undertaken in this thesis involved the study of psychometric properties of PROMs that are commonly used in stroke rehabilitation research hence the findings also contribute to the psychometric literature of these measurement scales.

In order to address the aims of this present study, a series of four research strands were conducted. These research strands were presented in the preceding chapters:

- Chapter 3 reports the first strand of research conducted in Study 1 which evaluated the feasibility of pooling the four datasets, using the DataSHaPER approach to compare the similarities and differences between the studies and also evaluated the potential to share variables across datasets.
- Chapter 4 reports the second strand of research conducted in Study 2 which was an application of multi-group confirmatory factor analysis to establish measurement invariance of the GHQ-28 questionnaire.
- Chapter 5 reports the third strand of research conducted in Study 3a which explored the utility of using regression-based models, and Item Response Theory models to harmonise the FAI and NEADL measures.
- Chapter 6 reports the third strand of research conducted in Study 3b which investigated the psychometric properties of the six common items in the GHQ-12 and GHQ-28 in order to harmonise these two measures.
- Chapter 7 reports the fourth strand of research conducted in Study 4a which was an illustrative pooled data analysis using a multi-group latent class analysis of

the SOS1 and CIMSS datasets to compare the disability patterns across the two stroke cohorts.

- Chapter 8 reports the fourth strand of research which was conducted in Study 4b to illustrate the benefits of conducting integrative data analysis using the harmonised pooled SOS datasets to investigate the factors associated with post-stroke anxiety.

The first section (9.1) of this Chapter will present a summary of findings from the different research strands; this is followed by a discussion of the findings in sections 9.2, 9.3, 9.4, and 9.6. The strengths and limitations of this present study are presented in section 9.7. This Chapter concludes with the implications of the findings from this thesis, recommendations for future research, and overall conclusions.

9.1 Summary of key findings from this thesis

Retrospective harmonisation and pooling some of the datasets used in this research was feasible, but there was a trade-off between increasing sample sizes and losing important variables. The challenges encountered in pooling the four datasets that were used in this thesis include: heterogeneity in measurement scales, heterogeneity in patient follow-up intervals, use of different scoring systems for the same measurement scales, use of different response options for similar questions, missing item data in some studies, and heterogeneity in variables that were collected by the different studies.

The heterogeneity in measurement scales used to assess similar disability outcomes raised methodological issues of harmonising the FAI and NEADL, and GHQ-12 and GHQ-28 measures. The effectiveness of using both regression-based methods and Item Response Theory (IRT) methods for harmonising these measures were explored. The results from both approaches showed that the mapping functions predicted the group means very well but individual patient level predictions were poor. Pooling the GHQ-28 scores from the SOS1 and SOS2 datasets raised methodological issues of establishing measurement of the GHQ-28 measure across the two datasets. The utility of using Multi-Group Confirmatory Factor (MG-CFA) analysis to assess measurement invariance (a prerequisite of pooling PROMS) of the GHQ-28 measure in two stroke cohorts was demonstrated providing support for pooling the GHQ-28 scores across the two SOS datasets.

Heterogeneity in follow-up intervals was a barrier to pooled data analysis of the two SOS datasets, and this was overcome by using multilevel models that can accommodate datasets with different follow-up intervals. Harmonising and pooling datasets was beneficial, larger datasets (SOS studies together, SOS1 and CIMSS together) were created that provided more representative samples and increased statistical power for subgroup analyses. Having access to multiple datasets also provided the opportunity to assess the reproducibility of results across studies. However, there is need to standardise the data collected by stroke rehabilitation studies in order to facilitate data sharing among stroke researchers.

9.2 Discussion of challenges in harmonisation and synthesising existing stroke datasets

Data harmonisation in this present study was defined as making data from different sources compatible and comparable so that the data can be combined and used in research. In Study 1, reported in Chapter 3, the DataSHaPER approach was used to compare the data collected by the SOS1, SOS2, CIMSS and Leeds SSNAP. The use of the DataSHaPER approach for retrospective harmonisation of secondary stroke datasets was beneficial; it provided a systematic approach to compare the similarities and differences across the four studies and also evaluate the potential to share data across the four datasets. The three circumstances found in multi-study analysis described by (Allen et al., 2013) were also encountered in this present study and these were: “ideal circumstances”, “less than ideal circumstances” and “circumstances that need statistical intervention”. Ideal circumstances is where common data were collected by studies and in Study 1 demographic variables such as age and gender were common across datasets and were pooled without harmonisation. Less than ideal circumstances were encountered where similar questions were asked by the different studies but studies used different response options (e.g. occupation, residential status before stroke, and ethnicity) and these variables needed recoding for harmonisation. The comparison of the four datasets that were used in this thesis showed that the datasets had similar demographic characteristics data but there was heterogeneity in the socio-economic, clinical variables, and disability data.

The main barrier to data harmonisation identified in this present study was the heterogeneity in measurement scales that were used to assess disability after stroke. Stroke is a heterogeneous condition and multiple measurement scales are used by

researchers without consensus on which is best (Duncan et al., 2000). This heterogeneity in measurement scales used in stroke studies has also been identified as the main problem for Cochrane and other reviewers (Agosti, 2008). Until there is consensus on the best measurement scales to use, retrospective harmonisation of existing stroke studies will remain a challenge. In this thesis, due to the heterogeneity in measurement scales, clinical characteristics and socio-economic data, pooling all four studies was impossible as it would have resulted in a significant loss of important covariates. The Leeds SSNAP collects data that is very different from the three research studies (SOS1, SOS2, and CIMSS) and had no disability outcomes follow-up data hence the audit data was excluded from the harmonisation work that was conducted in this thesis.

Socio-economic variables such as marital-status, education level, house ownership, and smoking were missing in the SOS1, CIMSS, and Leeds SSNAP datasets. These variables created the problem of missing data, in the harmonised datasets, and were excluded in Studies 4a and 4b of this thesis. Excluding these covariates was a draw back as these were important variables that are known to be associated with disability outcomes after stroke. For example there is evidence that smoking is associated with anxiety (Ayerbe, 2014), but in this present study the integrative analysis that was conducted in Study 4b reported in Chapter 8 excluded smoking because this data was missing in SOS1.

Effective data sharing among stroke rehabilitation researchers may be facilitated by standardising demographic, clinical and measurement scales across stroke outcomes studies. The issues of heterogeneity in measurement scales identified in this thesis may be addressed by having agreed standardised sets of core outcomes. The development of a Data Schema or a minimum dataset for stroke outcomes studies may facilitate data sharing among stroke rehabilitation researchers. However standardising data collection tools is desirable but difficult to implement. Until there is a consensus on what measures to use in stroke rehabilitation research, measurement comparability will continue to be a problem in sharing existing stroke data. In Study 1 of this thesis, a Data Schema was developed retrospectively, this Data Schema can be an initial step in the development of DataSHaPERs for future stroke collaborative studies that aim to share or exchange data. The need for a DataSHaPER for stroke has been highlighted by Fortier et al. (2011).

9.3 Was pooled data analysis beneficial?

Despite losing important covariates the analyses conducted using harmonised datasets were beneficial. The benefits of harmonising comparable datasets were demonstrated in the fourth strand of this research work which was reported in Chapters 7 and 8. Pooling the two longitudinal SOS studies was beneficial, a large dataset of n=1033 patients was obtained from the two SOS longitudinal studies, and this dataset was used in Study 4a reported in chapter 8, to investigate the factors associated with post-stroke anxiety symptoms. Longitudinal studies are important in understanding the intra-individual changes in HRQOL domains after stroke and the factors associated with these domains. The pooled SOS dataset provided a large sample with increased statistical power for subgroup analyses, and increased precision of estimates. The integrative data analysis of the two SOS studies raised statistical challenges of having different patient follow-up intervals in the two datasets but this challenge was overcome by the using multilevel modelling approaches.

The SOS1 and SOS2 had produced conflicting evidence on the association of female gender and post-stroke anxiety symptoms, with the SOS2 suggesting that females were more likely to be associated with anxiety symptoms compared to males and this association was not found in SOS1. The integrative analysis of the combined SOS1 and SOS2 datasets suggested an association of female gender and post-stroke anxiety symptoms and these findings were consistent with other previous studies that were discussed in Chapter 8. The results from Study 4a also found an association of younger age, with anxiety post-stroke. Increased depressive symptoms, social dysfunction and increased somatic symptoms were associated with increased anxiety symptoms. Increased physical function was associated with reduced anxiety symptoms. These findings were consistent with previous studies. Details of these previous studies were given in Chapter 8. Findings from the analysis conducted in Study 4b suggested the importance of developing interventions that target women, and younger stroke survivors as these were more likely to have anxiety post-stroke.

The use of Multi-Group Latent Class Analysis (MG-LCA) in Study 4a reported in Chapter 7 of this thesis was beneficial as it provided a framework for comparing disability patterns across different stroke cohorts and the factors associated with these patterns. MG-LCA provided an excellent framework for determining a suitable multi-group model for analysing independent data from different sources to compare latent

disability classes across different stroke cohorts. The MG-LCA analysis conducted in Study 4a identified six disability classes and the details of these classes were reported in Chapter 7. The preliminary findings of the research conducted in Study 4a were published in (Munyombwe et al., 2014). The findings from the MG-LCA suggested that gender female and younger age were more likely to have depressive symptoms. A latent class with severe depressive symptoms emerged in both SOS1 and SOS2. Patients in this class were more likely to be younger females. The patient characteristics identified in Study 4a could be used for creating integrated and person centred approaches to care management and outcome optimisation for the stroke survivors.

A difficult was encountered as some of the measures that were used in Study 4a were gender biased. For example the NEADL measure was *influenced by gender* with a large proportion of female patients in the "Good physical function in the kitchen class".

9.4 Discussion of statistical methods used to harmonise patient reported outcome measures

Combining Patient Reported Outcome Measures (PROMs) from the three research datasets (SOS1, SOS2, and CIMSS) posed statistical challenges of measurement comparability where studies used different PROMs to assess the similar constructs. It is a common approach to harmonise PROMs by mapping when the required outcome measure was not collected by the other study. For example in economic evaluation studies, researchers map PROMs to the EQ-5D measure. The third strand of research reported in Chapter 5 investigated the utility of using regression-based methods, item response theory models, and use of common items for harmonising patient reported outcome measures. In Study 3a, the utility of using two widely used methods of harmonising PROM: regression based-methods and IRT methods for harmonising the FAI and NEADL measures was investigated. Mapping algorithms and conversion tables for harmonising the FAI and NEADL measures were developed. The similarities between the FAI and NEADL items provided face validity for linking the two measures. The findings of Study 3a suggested that both regression-based and IRT mapping of the FAI and NEADL measures were effective in predicting the overall group means and not patient level predictions. The predicted

group means were very similar to the observed group means for both the regression based and IRT based mappings. These findings supported the use of mapping algorithms for predicting group averages and not patient level scores. The findings from Study 3a were consistent with the recommendations from economic evaluation studies where regression-based mapping is used for predicting group means, for group based comparisons but not for predicting individual patient scores. The variation of the predicted scores was less than the observed scores for both the regression based and IRT methods. This is because single regression imputations under-estimates the variation of the predicted observations hence multiple imputations are required.

The large individual person prediction errors from the mapping functions developed using regression based methods and the reduced variation in predicted scores could be due to a statistical phenomenon called "regression to the mean" (Fayers and Hay, 2014). Fayers and Hays (2014) explained that at individual level, "regression to the mean" will unfairly award patients with lower observed scores higher predicted scores closer to the mean and individuals with higher scores will be awarded lower predicted scores closer to the mean. Therefore when mapping using regression-based methods such as OLS estimators, lower scores or higher scores may become unfairly biased towards the mean thus producing poor Individual level predictions.

Other sources of poor individual person predictions could be poor conceptual overlap between measures or using a mapping function in a sample that is different from the sample that was used to develop it. Longworth and Rowen, (2011) recommended that to reduce prediction errors in mapping analyses, the mapping algorithm needs to be applied to a similar population in which the algorithm was developed and validated. If the target population is not similar to the source population, then the algorithm may not produce reliable or accurate predictions. In Study 3a, the estimation and validation sample was the same (SOS1 dataset) because there was no external dataset that collected both measures. The mapping algorithms were evaluated using wave 2 data of the SOS1 data hence there was no problem of differences between the estimation and validation samples. However, the mapping algorithm was evaluated using one year data when the model was developed using baseline (within 4 week after stroke). The long period between the baseline and follow up time points could have introduced errors in the validations but the descriptive analysis reported in Chapter 5 showed that the characteristics of the patients that

remained in the SOS1 study at wave 2(1 year post stroke) were similar to the baseline sample. There is need for more external validation of the mapping algorithms developed in this thesis.

In Study 3a reported in Chapter 5 of this thesis, IRT was also used to harmonise the FAI and NEADL measures. Similar to regression-based mapping it also produced conversion tables that were good for predicting group level statistics and not individual level predictions, despite using a common person approach or single study design (patients answering both measures) that is considered to be the best for IRT linking (Dorans, 2007). There are several reasons why IRT methods produce poor individual level predictions and these include: poor conceptual overlap between the measures, use of incorrect IRT model for calibration, sample size, number of items used for calibration. Successful IRT linking requires the measures to measure similar constructs and there should be good conceptual overlap between the measures. In Study 3a, the factor analysis of the combined NEADL and FAI items showed that the two scales measure similar constructs but the NEADL captured a wider spectrum of extended activities of daily living compared to the FAI measure and this might have been another source of poor individual level predictions for the conversion tables developed using IRT linking. There are also various IRT models that can be used to calibrate PROMS and using an incorrect IRT model can lead to inaccurate item parameters that will in turn produce inaccurate conversion tables. Furthermore, Fitzpatrick and Yen (2001) recommended sample sizes of 500 or 1000 for more precise IRT linking, in Study 3a, a sample size of 448 was used, which might not have been sufficient for IRT linking. In Study 3a, only the 2 parameter IRT model was explored, more research is needed to evaluate the utility of using other advanced multi-dimensional IRT models.

9.5 Discussion of harmonisation the GHQ-12 and GHQ-28 measures

In this thesis, the GHQ-12 and GHQ-28 measures were harmonised by using the six common items from both measures. Psychological distress scores in the harmonised SOS1 and CIMSS datasets were computed by summing the six items common in the GHQ-12 and GHQ-28. The psychometric properties of these six common items were first investigated in Study 3b reported in Chapter 6. The findings from Study 3b showed that the six common items from the GHQ-12 and GHQ-28 had good construct validity but more research is needed to externally validate the choice of

these variables. Items that were not common across the GHQ-12 and GHQ-28 were excluded, but this might have affected the content validity of the measures hence the six items common to the GHQ-12 and GHQ-28 require further investigation of their targeting properties and responsiveness. There are other items across the GHQ-12 and GHQ-28 that have similar meaning but different wording. More research is needed to harmonise these items so that they can be included in the harmonised GHQ measures.

9.6 Discussion of statistical methods for measurement invariance

Measurement invariance is a pre-requisite for pooling PROMs data from multiple studies. Combining studies raised statistical issues of establishing measurement invariance of GHQ-28, NEADL and BI across studies. In Study 2 reported in Chapter 4 of this thesis, Multi-group Confirmatory Factor Analysis (MG-CFA) was used to test measurement invariance of the GHQ-28. The findings from study 2 were published in Munyombwe et al. (2015). In this present study, establishing measurement invariance of other measures that were common across studies such as BI, and FAI was not possible as there was no item data in some studies but total scores only. Using MG-CFA in Study 2 was beneficial, it provided an elegant framework for testing various measurement invariance tests such as configural, metric and scalar invariance) via a single procedure. Study 2 established configural, metric, and scalar invariance for the GHQ-28 with respect to SOS studies, providing support for the integrative analysis of the GHQ-28 scores in Study 4b of this thesis which was reported in Chapter 8. The measurement invariance analysis of the GHQ-28 questionnaire conducted in Study 2 makes a novel contribution to the literature on the psychometric properties of a scale (the GHQ-28 measurement scale) that is widely used in stroke rehabilitation research.

9.7 Strengths and Limitations

The strength of this present study was the pooling of harmonised datasets from different stroke cohorts to create large samples which were more representative of the stroke population (wider age range), and also had more patient follow-up intervals. Due to the heterogeneity in stroke population, large samples are needed in stroke rehabilitation research to increase the representativeness of the samples, and also the precision and statistical power for subgroup analyses. The larger datasets that were created in this thesis were used to investigate the factors associated with post stroke

disability post-stroke. The strength was that advanced statistical methods (Multi-level Poisson models and Multi-group latent class analysis) were used in some research strands conducted in this thesis. The Multi-level Poisson model was useful for analysing skewed count data from the GHQ-28 anxiety subscale. Multi-group latent class analysis provided an elegant framework for analysing data from multiple sources and also comparing disability latent classes across different cohorts. A more comprehensive classification of disability patterns in stroke survivors was produced using larger samples, and multiple measures of psychological distress and physical functioning. There are a number of limitations that warrant discussion and these are discussed in the next section.

9.7.1 Literature review

The literature review presented in Chapter 2 was not meant to be comprehensive as the aim was to give an overview of data harmonisation studies, data harmonisation approaches and statistical methods that were commonly used to harmonise PROMs data. The literature review provided an overview of the most commonly used methods of data harmonisation and some of these methods were explored in this thesis.

9.7.2 Datasets

The datasets that were used in this thesis were not randomly selected and were all from Yorkshire hence these datasets may not be representative of the stroke population and may not be generalisable. However the four datasets provided the opportunity to evaluate the feasibility of harmonising such studies and to identify barriers that might prevent data harmonisation and pooling (both here and elsewhere). The four studies presented similarities (for example, all four were based in Yorkshire, UK) and differences that offered an ideal opportunity to examine a wide range of challenges and to explore various statistical methods with which to address these.

The analysis conducted in this thesis was based on research data that excluded patients with severe stroke thus the findings of this present study are therefore restricted to patients with mild or moderate stroke.

9.7.3 Pooled data analysis

The main challenges in pooling the datasets that were used in this thesis were the use of different PROMs and the heterogeneity in variables that were collected by the different studies. Despite attempting to harmonise similar measurement scales and variables it was not possible to combine datasets without a significant loss of

information. Pooling datasets retrospectively resulted in the loss of important variables that are known to influence patient outcomes after stroke. However, despite a reduction in covariates that could be used for the pooled data analysis, it was still beneficial as the pooled datasets provided sufficient statistical power for subgroup analyses and it was also possible to compare disability latent classes across different stroke cohorts using harmonised datasets.

9.7.4 Mapping PROMS

Validation of mapping/crosswalk conversion tables developed in this present study was conducted using internal data; there was no independent dataset to externally validate the models. A cross-validation would be ideal once a suitable external dataset becomes available.

9.8 Implications of the study

The current study has shown that mapping PROMs using regression-based models and IRT method could be useful for predicting group averages but not for predicting individual patient scores thus the mapping algorithms developed in this thesis could also be useful in data harmonisation studies for predicting group averages but not for making precise estimates of individual scores. The challenges identified in this thesis suggested that effective data sharing among stroke rehabilitation researchers would be facilitated by standardising demographic, stroke clinical and measurement scales across stroke outcomes studies. Big data sharing in stroke research could be achieved by developing a Data Schema or minimum dataset for stroke outcomes studies. However standardising data collection tools is desirable but difficult to implement. Until there is a consensus on what measures to use in stroke rehabilitation research measurement comparability will continue to be a problem in sharing existing stroke data.

9.9 Recommendations for future research

In common with much research, the work conducted in this thesis provides a framework and a foundation for further research. The potential of this present study could not be fully exploited due to limitations in the available datasets described in earlier chapters, and the lack of resources and time limitations of a doctoral study.

Several longitudinal stroke datasets exist but only a few studies were used in this study. With appropriate funding, the approach used in this thesis could be extended to involve other stroke rehabilitation research collaborators and develop a DataSHaPER and Data Schema for stroke rehabilitation research to facilitate data sharing and facilitate studies that would have benefit for patient care. Similar initiatives have been demonstrated elsewhere and examples of these studies were reported in Chapter 2 of this thesis.

Harmonising SSNAP registers needs further exploration. The SSNAP register has the potential to be used in stroke outcomes research to investigate the care processes associated with good patient outcomes, and raise the standard of care for stroke patients. Linking SSNAP registers with PROMs data could provide the opportunity to investigate associations of patient care processes during hospitalisation and patient outcomes. The SSNAP data that was obtained for this thesis had no patient outcomes, and due to time limitations, the present study had to continue with incomplete SSNAP data, which had no follow-up disability data. Unfortunately the audit data was eventually excluded from the harmonisation process conducted in this present study. Further research using pooled data from multiple registers is needed to demonstrate associations between processes of care and patient outcomes, and such research would be valuable to improve care for patients. As highlighted in Chapter 1, efforts to demonstrate associations between stroke care processes and patient outcomes in observational studies has proven to be difficult partly due to small sub-group analysis and inadequate case-mix adjustment. Harmonising and pooling data from SSNAP registers would provide large samples and care processes data that can be used to undertake robust studies of stroke care with increased statistical power and adequate case-mix adjustment.

Analysis of pooled SOS longitudinal datasets

In Study 4b, the pooled data analyses of the two SOS datasets investigated factors associated with post-stroke anxiety. The harmonised SOS dataset (n=1033) developed in this thesis included information on other disability dimensions such as physical function, social function and somatic symptoms. Further analysis of these various disability domains using advanced statistical methods such as structural equation modelling techniques would provide a better understanding of the inter relationships between them.

Harmonising PROMs

The harmonisation of the GHQ-12 and GHQ-28 measures conducted in Study 3b could be extended by using other IRT designs such as non-equivalent designs using the six common items as "anchor items". The anchor test design uses a subset of common test items known as anchors to link the measures and IRT method. Details of the anchor tests designs were described in Chapter 2 of this thesis.

In Study 3a, reported in Chapter 5, two methods of data harmonisation (regression-based and IRT based models) were explored. The harmonisation using regression based methods that was conducted in this present study can be extended by using multiple imputations. Multiple imputations could not be explored in this present study due to time limitations. Harmonisation using IRT methods focused on developing conversion tables between measures, more research on IRT harmonisation can be expanded by exploring harmonisation by putting the data on the same metric using suitable IRT models and then using the IRT scores in further analysis rather than developing cross walks or conversion tables. Furthermore, the IRT harmonisation conducted in Study 4a was based on unidimensional IRT models. More research based on advanced multi-dimensional IRT models could be explored. Another area for future research is harmonising more than two measures. In this present study the focus was on harmonising two measures using a common person design. More research could focus on harmonising more than two measures.

9.10 Planned publications

In addition to the two publications that have arisen from this thesis, *(Munyombwe et al., 2015; Munyombwe et al., 2014)* the following planned manuscripts will be written and submitted to appropriate journals.

- Mapping the NEADL and FAI measures of activities of daily living: Application of regression-based and IRT methods. This manuscript will be written using the findings from Study 3a reported in chapter 5.
- Harmonisation of secondary stroke outcomes datasets: Challenges and barriers that may prevent data sharing in stroke rehabilitation research and recommendations for future data harmonisation studies. This manuscript will be written using the findings from study 1 reported in chapter 3.

- Factors associated with anxiety post-stroke: Integrative data analysis of the harmonised SOS1 and SOS2 datasets. This manuscript will be written using the findings from study 4b reported in chapter 8.

9.11 Conclusion

In conclusion, this thesis has shown that pooling stroke outcomes datasets from multiple studies may offer substantial opportunities for studying patient disability outcomes after stroke. Large datasets were produced and were used to address important research questions with increased statistical power for subgroup analysis and a more representative sample. Having multiple datasets also enabled me to check the reproducibility of results across studies. There was a trade-off between increased sample size and the loss of important variables (that were missing in one or other studies). The main barrier to harmonising datasets encountered in this thesis was measurement comparability. Future work is needed to develop a Stroke DataSHaPER with a minimum dataset or Data Schema to facilitate big data sharing among stroke rehabilitation researchers. However, until there is consensus on the outcome measures that should be used in stroke rehabilitation studies and the core key variables needed in stroke outcome studies, data sharing will be difficult or impossible.

What this study Adds

- Harmonisation and pooling of similar secondary stroke datasets seems to be a promising way to deal with the small samples of many stroke rehabilitation studies.
- The main barrier to harmonising stroke outcomes studies is the heterogeneity in measurement scales used to assess disability after stroke.
- The use of the SSNAP registry data in stroke outcomes research was limited by the absence/limited availability of follow-up PROMs data.
- Multi-group confirmatory factor analysis is useful for establishing measurement invariance of PROMs. The measurement invariance analysis of GHQ-28 measure conducted in this present study contributed greater understanding about the psychometric properties of this measure that is commonly used in stroke outcomes research.
- The utility of using Multi-Group Latent Class Analysis for pooling data from multiple studies to produce comprehensive classifications for stroke was demonstrated. The characteristics of patients with a high risk of having depressive symptoms were identified.
- Mapping PROMs using regression-based methods and IRT methods was effective for predicting group averages but not predicting individual scores.

Appendix A MEDLINE SEARCH STRATEGY: LITERATURE REVIEW ON STATISTICAL METHODS FOR MAPPING OR LINKING PROMS

Searches for 1996 to May Week 3 2015, English language articles only

---

1 crosswalk.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (111)

2 co calibration.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (14)

3 mapping algorithms.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (92)

4 1 or 2 or 3 (215)

5 linking quality of life measures.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (1)

6 linking.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (62186)

7 quality of life measures.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (1842)

8 6 and 7 (8)

9 PROMs.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (223)

10 6 and 9 (7)

11 4 or 5 or 8 or 10 (229)

12 Remove 12 duplicate articles from 11 (217)

13 Remove 155 non relevant articles (62).

Appendix B MEDLINE SEARCH STRATEGY: LITERATURE REVIEW ON EXAMPLES OF DATA HARMONISATION STUDIES

Searches for 1996 to May Week 3 2015, English language articles only

---

1 data harmonisation.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (5)

2 individual patient data meta analysis.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (256)

3 integrative data analysis.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (38)

4 collaborative analysis.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (113)

5 3 or 4 (151)

6 1 or 5 (156)

7 mega analysis.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (40)

8 6 or 7 (196)

9 data harmonization.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier] (45)

10 8 or 9 (241)

11 Remove 28 duplicates from 10 (213)

12. Remove 47 non relevant articles (166)

Appendix C MPLUS SYNTAX FOR MEASUREMENT INVARIANCE MODELS FITTED IN CHAPTER 4

The following Mplus code was used to establish configural, metric and scalar invariance of the GHQ-28 measure in chapter 4.

Model 1: Configural model GHQ-28

```
DATA:
    FILE is M:\StrokeC\MIGHQ.dat ;
VARIABLE:
        Names are  id_no study sex age ghq_a1 ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
        ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3 ghq_c4
        ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6   ghq_d7;
MISSING are all (-9999) ;
GROUPING IS study  (1=SOS2 2=SOS1);
USEVARIABLES ARE ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7  ghq_b1 ghq_b2 ghq_b3
        ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5
        ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6  ghq_d7 study ;
CATEGORICAL are ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7 ghq_b1 ghq_b2 ghq_b3
        ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5 ghq_c6
        ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6  ghq_d7;
ANALYSIS: ESTIMATOR IS WLSMV; PARAMETERIZATION=THETA;
SAVEDATA: DIFFTEST=Configural.dat;
    !!! configural model for sos1 reference group
MODEL:
    ! factor loadings all estimated
     somatic by ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7*;
    [somatic@0]; somatic@1;
    ! factor mean=0 and variances=1 for identification
    anxiety by ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7* ;
    [anxiety@0]; anxiety@1;
    social by ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5 ghq_c6 ghq_c7* ;
    [depression@0]; depression@1;
    depression by ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6 ghq_d7*;
    [Social@0]; Social@1;
    ! item thresholds all free
    [ghq_a2$1 ghq_a3$1 ghq_a4$1 ghq_a5$1 ghq_a6$1 ghq_a7$1*]
    [ghq_b1$1 ghq_b2$1 ghq_b3$1 ghq_b4$1 ghq_b5$1 ghq_b6$1 ghq_b7$1*]
    [ghq_c1$1 ghq_c2$1 ghq_c3$1 ghq_c4$1 ghq_c5$1 ghq_c6$1 ghq_c7$1*]
    [ghq_d1$1 ghq_d2$1 ghq_d3$1 ghq_d4$1 ghq_d5$1 ghq_d6$1 ghq_d7$1*];
    ! Item residuals variances all fixed to 1
     ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1
```

```
 ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1  ghq_b7@1  ghq_c1@1 ghq_c2@1 ghq_c3@1 ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1 ghq_d1@1 ghq_d2@1 ghq_d3@1 ghq_d4@1 ghq_d5@1 ghq_d6@1 ghq_d7@1;
!!! configural model for sos 2 alternative group
Model SOS2:
 ! factor loadings all estimated
somatic by ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7*;
  [somatic@0]; somatic@1;! factor mean=0 and variances=1 for identification
  anxiety by ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7* ;
  [anxiety@0]; anxiety@1;
  social by ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5 ghq_c6 ghq_c7* ;
  [depression@0]; depression@1;
  depression by ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6 ghq_d7*  ;
  [Social@0]; Social@1;
  ! item thresholds all free
[ghq_a2$1 ghq_a3$1 ghq_a4$1 ghq_a5$1 ghq_a6$1 ghq_a7$1*]
[ghq_b1$1 ghq_b2$1 ghq_b3$1 ghq_b4$1 ghq_b5$1 ghq_b6$1 ghq_b7$1*]
[ghq_c1$1 ghq_c2$1 ghq_c3$1 ghq_c4$1 ghq_c5$1 ghq_c6$1 ghq_c7$1*]
[ghq_d1$1 ghq_d2$1 ghq_d3$1 ghq_d4$1 ghq_d5$1 ghq_d6$1 ghq_d7$1*];
! Item residuals variances all fixed to 1
ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1
ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1 ghq_b7@1
ghq_c1@1 ghq_c2@1 ghq_c3@1 ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1
ghq_d1@1 ghq_d2@1 ghq_d3@1 ghq_d4@1 ghq_d5@1 ghq_d6@1 ghq_d7@1;
OUTPUT: STANDARDIZED MODINDICES;
```

Model 2: metric invariance GHQ-28

```
DATA:
    FILE is M:\StrokeC\MIGHQ.dat ;
VARIABLE:
        NAMES are id_no study sex age ghq_a1 ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
            ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3
             ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6
            ghq_d7;
MISSING are all (-9999) ;
GROUPING IS study  (1=SOS2 2=SOS1);
USEVARIABLES ARE ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
             ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3
             ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6
             ghq_d7 study ;
CATEGORICAL are ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7 ghq_b1 ghq_b2 ghq_b3 ghq_b4
```

```
        ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1
        ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6 ghq_d7;
ANALYSIS: ESTIMATOR IS WLSMV; PARAMETERIZATION=THETA;
DIFFTEST=Configural.dat; ! compare against configural
SAVEDATA: DIFFTEST=MetricA.dat;! save metric info
    !!! metric model for SOS1 reference group
MODEL:
    ! factor loadings all estimated
    somatic by ghq_a2*(L1)
     ghq_a3* (L2)
     ghq_a4* (L3)
     ghq_a5* (L4)
     ghq_a6* (L5)
     ghq_a7* (L6);
    [somatic@0]; somatic@1;! factor mean=0 and variances=1 for identification
    anxiety by ghq_b1*(L7)
     ghq_b2* (L8)
     ghq_b3* (L9)
     ghq_b4* (L10)
     ghq_b5* (L11)
     ghq_b6* (L12)
     ghq_b7* (L13) ;
    [anxiety@0]; anxiety@1;
    social by ghq_c1*(L14)
     ghq_c2* (L15)
     ghq_c3* (L16)
     ghq_c4* (L17)
     ghq_c5* (L18)
     ghq_c6* (L19)
     ghq_c7* (L20);
    [depression@0]; depression@1;
    depression by ghq_d1*(L21)
     ghq_d2* (L22)
     ghq_d3* (L23)
     ghq_d4* (L24)
     ghq_d5* (L25)
     ghq_d6* (L26)
     ghq_d7*(L27);
    [Social@0]; Social@1;
    ! item thresholds all free
    [ghq_a2$1 ghq_a3$1 ghq_a4$1 ghq_a5$1 ghq_a6$1 ghq_a7$1*]
```

[ghq_b1$1 ghq_b2$1 ghq_b3$1 ghq_b4$1 ghq_b5$1 ghq_b6$1 ghq_b7$1*]

[ghq_c1$1 ghq_c2$1 ghq_c3$1 ghq_c4$1 ghq_c5$1 ghq_c6$1 ghq_c7$1*]

[ghq_d1$1 ghq_d2$1 ghq_d3$1 ghq_d4$1 ghq_d5$1 ghq_d6$1 ghq_d7$1*];

! Item residuals variances all fixed to 1

ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1

ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1 ghq_b7@1 ghq_c1@1 ghq_c2@1 ghq_c3@1ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1 ghq_d1@1 ghq_d2@1 ghq_d3@1 ghq_d4@1 ghq_d5@1 ghq_d6@1 ghq_d7@1;

!!! *metric model for SOS2 alternative group*

Model SOS2:

! *factor loadings all held now equal*

somatic by ghq_a2*(L1)

ghq_a3* (L2)

ghq_a4* (L3)

ghq_a5* (L4)

ghq_a6* (L5)

ghq_a7* (L6);

[somatic@0]; somatic@1;! *factor mean=0 and variances=1 for identification*

anxiety by ghq_b1*(L7)

ghq_b2* (L8)

ghq_b3* (L9)

ghq_b4* (L10)

ghq_b5* (L11)

ghq_b6* (L12)

ghq_b7* (L13) ;

[anxiety@0]; anxiety@1;

social by ghq_c1*(L14)

ghq_c2* (L15)

ghq_c3* (L16)

ghq_c4* (L17)

ghq_c5* (L18)

ghq_c6* (L19)

ghq_c7* (L20);

[depression@0]; depression@1;

depression by ghq_d1*(L21)

ghq_d2* (L22)

ghq_d3* (L23)

ghq_d4* (L24)

ghq_d5* (L25)

ghq_d6* (L26)

ghq_d7*(L27);

```
[Social@0]; Social@1;
! item thresholds all free
[ghq_a2$1 ghq_a3$1 ghq_a4$1 ghq_a5$1 ghq_a6$1 ghq_a7$1*]
[ghq_b1$1 ghq_b2$1 ghq_b3$1 ghq_b4$1 ghq_b5$1 ghq_b6$1 ghq_b7$1*]
[ghq_c1$1 ghq_c2$1 ghq_c3$1 ghq_c4$1 ghq_c5$1 ghq_c6$1 ghq_c7$1*]
[ghq_d1$1 ghq_d2$1 ghq_d3$1 ghq_d4$1 ghq_d5$1 ghq_d6$1 ghq_d7$1*];
! Item residuals variances all fixed to 1
ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1
ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1 ghq_b7@1 ghq_c1@1 ghq_c2@1 ghq_c3@1 ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1 ghq_d1@1 ghq_d2@1 ghq_d3@1 ghq_d4@1 ghq_d5@1 ghq_d6@1 ghq_d7@1;
OUTPUT: STANDARDIZED MODINDICES;
```

Model 3: full threshold invariance GH-Q28

```
DATA:
  FILE is M:\StrokeC\MIGHQ.dat ;
VARIABLE:
    NAMES are id_no study sex age ghq_a1 ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
    ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3 ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6 ghq_d7;
MISSING are all (-9999) ;
GROUPING IS study (1=SOS2 2=SOS1);
USEVARIABLES ARE ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
    ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3
    ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6
    ghq_d7 study ;
    Categorical are ghq_a2 ghq_a3 ghq_a4 ghq_a5 ghq_a6 ghq_a7
    ghq_b1 ghq_b2 ghq_b3 ghq_b4 ghq_b5 ghq_b6 ghq_b7 ghq_c1 ghq_c2 ghq_c3
    ghq_c4 ghq_c5 ghq_c6 ghq_c7 ghq_d1 ghq_d2 ghq_d3 ghq_d4 ghq_d5 ghq_d6
      ghq_d7;
  ANALYSIS: ESTIMATOR IS WLSMV; PARAMETERIZATION=THETA;
   DIFFTEST=MetricA.dat; ! compare against metric
  SAVEDATA: DIFFTEST=ScalarA.dat;! save metric info
  !!! full scalar model for SOS1 study
MODEL:
  ! factor loadings all estimated
  somatic by ghq_a2*(L1)
   ghq_a3* (L2)
   ghq_a4* (L3)
   ghq_a5* (L4)
   ghq_a6* (L5)
   ghq_a7* (L6);
```

[somatic@0]; somatic@1;! *factor mean=0 and variances=1 for identification*

anxiety by ghq_b1*(L7)

ghq_b2* (L8)

ghq_b3* (L9)

ghq_b4* (L10)

ghq_b5* (L11)

ghq_b6* (L12)

ghq_b7* (L13) ;

[anxiety@0]; anxiety@1;

social by ghq_c1*(L14)

ghq_c2* (L15)

ghq_c3* (L16)

ghq_c4* (L17)

ghq_c5* (L18)

ghq_c6* (L19)

ghq_c7* (L20);

[depression@0]; depression@1;

depression by ghq_d1*(L21)

ghq_d2* (L22)

ghq_d3* (L23)

ghq_d4* (L24)

ghq_d5* (L25)

ghq_d6* (L26)

ghq_d7*(L27);

[Social@0]; Social@1;

! *item thresholds all free*

[ghq_a2$1 ghq_a3$1 ghq_a4$1 ghq_a5$1 ghq_a6$1 ghq_a7$1*]

[ghq_b1$1 ghq_b2$1 ghq_b3$1 ghq_b4$1 ghq_b5$1 ghq_b6$1 ghq_b7$1*]

[ghq_c1$1 ghq_c2$1 ghq_c3$1 ghq_c4$1 ghq_c5$1 ghq_c6$1 ghq_c7$1*]

[ghq_d1$1 ghq_d2$1 ghq_d3$1 ghq_d4$1 ghq_d5$1 ghq_d6$1 ghq_d7$1*];

! Item residuals variances all fixed to 1

ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1

ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1 ghq_b7@1 ghq_c1@1 ghq_c2@1 ghq_c3@1 ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1 ghq_d1@1 ghq_d2@1 ghq_d3@1 ghq_d4@1 ghq_d5@1 ghq_d6@1 ghq_d7@1;

!!! *Full scalar model for SOS2 study*

Model SOS2:

! *factor loadings all still held equal*

somatic by ghq_a2*(L1)

ghq_a3* (L2)

ghq_a4* (L3)

```
 ghq_a5* (L4)
 ghq_a6* (L5)
 ghq_a7* (L6);
[somatic@0]; somatic@1;! factor mean=0 and variances=1 for identification
anxiety by ghq_b1*(L7)
 ghq_b2* (L8)
 ghq_b3* (L9)
 ghq_b4* (L10)
 ghq_b5* (L11)
 ghq_b6* (L12)
 ghq_b7* (L13) ;
[anxiety@0]; anxiety@1;
social by ghq_c1*(L14)
 ghq_c2* (L15)
 ghq_c3* (L16)
 ghq_c4* (L17)
 ghq_c5* (L18)
 ghq_c6* (L19)
 ghq_c7* (L20);
[depression@0]; depression@1;
depression by ghq_d1*(L21)
 ghq_d2* (L22)
 ghq_d3* (L23)
 ghq_d4* (L24)
 ghq_d5* (L25)
 ghq_d6* (L26)
 ghq_d7*(L27);
[Social@0]; Social@1;
! item thresholds now held equal if left off
! Item residuals variances all fixed to 1
ghq_a2@1 ghq_a3@1 ghq_a4@1 ghq_a5@1 ghq_a6@1 ghq_a7@1
ghq_b1@1 ghq_b2@1 ghq_b3@1 ghq_b4@1 ghq_b5@1 ghq_b6@1 ghq_b7@1
ghq_c1@1 ghq_c2@1 ghq_c3@1    ghq_c4@1 ghq_c5@1 ghq_c6@1 ghq_c7@1
ghq_d1@1 ghq_d2@1 ghq_d3@1   ghq_d4@1 ghq_d5@1 ghq_d6@1  ghq_d7@1;
```

Appendix D ADDITIONAL REGRESSION BASED MAPPING RESULTS THAT WERE PRODUCED IN CHAPTER 5 FOR MODELS WITH ITEMS AS PREDICTORS

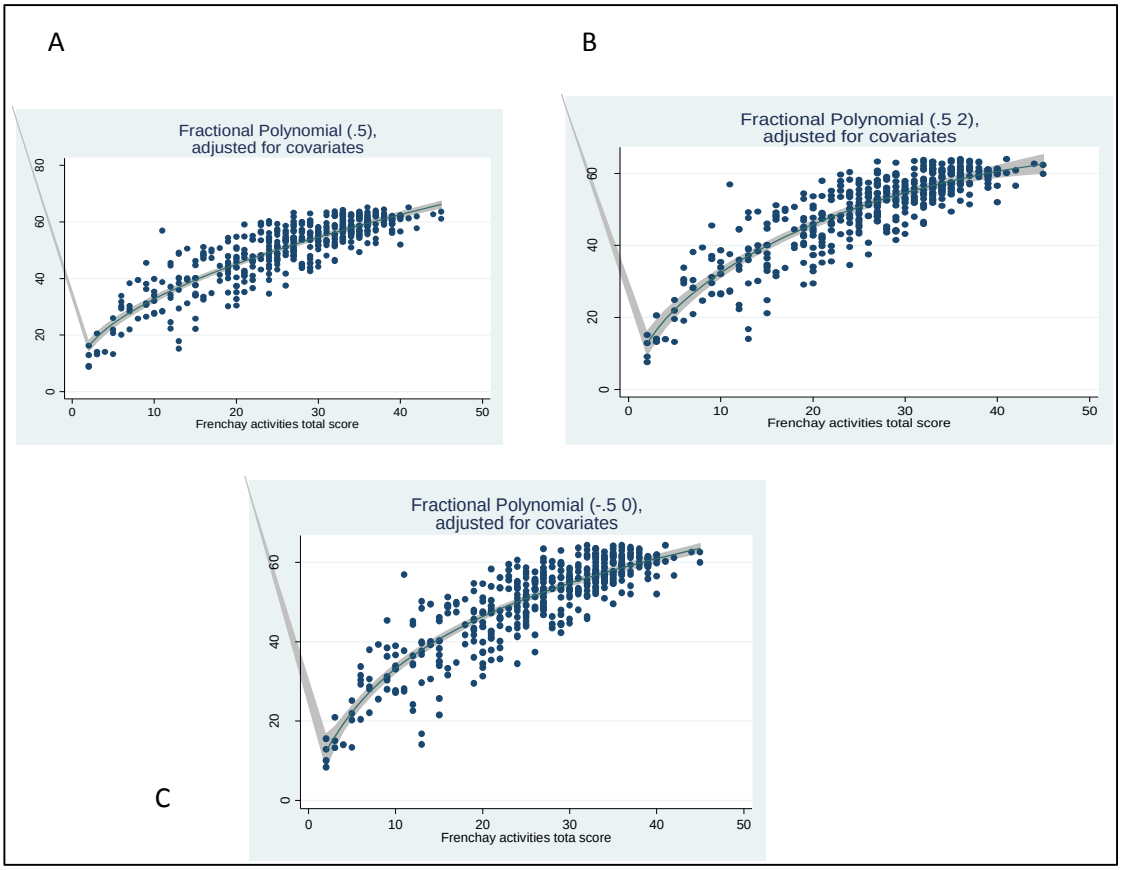

Figure 1 Fractional polynomial models, (A) OLS regression model, (B) Quantile regression, (C) Robust regression

Table 1: Fractional polynomials for predicting NEADL from FAI total score

	OLS **Powers** **FAI (0.5)** **Age(1)** **Gender female(1)**	**Q reg** **Powers** **FAI (0.5 ,2)** **Age(1)** **Gender female(1)**	**R reg** **Powers** **FAI:(-0.5, 0)** **Age (3)** **Gender female(1)**
FAI1 FAI2 + age + Female gender Constant	29.82(28.22, 31.42)*** - -0.06(-0.11,-0.02)** -1.94(-2.98,-0.91)*** 51.22(50.5,51.93) Adjusted *R* = 0.77	36.32(31.43, 41.21)*** -0.55(-0.91, -0.18)** -0.07(-0.12, -0.01)* -0.83(-2.04, 0.38) 51.69(50.76,52.62) Pseudo *R* = 0.51	17.44(9.31, 25.57)*** 26.27(22.42, 30.11)*** -0.01(-0.01, -0.003)*** -1.09(-2.08, -0.12)* 51.86(51.19, 52.54)

OLS regression model, SOS1 study

Table 2 Regression coefficients and 95% confidence interval results for mapping FAI items on to NEADL total using OLS

NEADL	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fren1_1_1	.6980854	1.072927	0.65	0.516	-1.411195	2.807366
fren1_1_2	2.385165	1.010674	2.36	0.019	.3982696	4.372061
fren1_1_3	1.920328	.8235226	2.33	0.020	.3013545	3.539301
fren1_2_1	9.650732	1.666237	5.79	0.000	6.375056	12.92641
fren1_2_2	6.236006	1.359717	4.59	0.000	3.562922	8.909091
fren1_2_3	7.612333	1.064235	7.15	0.000	5.520139	9.704526
fren1_3_1	3.78633	1.458861	2.60	0.010	.9183376	6.654322
fren1_3_2	3.213136	1.600781	2.01	0.045	.0661408	6.360131
fren1_3_3	6.223819	.7720596	8.06	0.000	4.706018	7.741621
fren1_4_1	.5723161	1.415938	0.40	0.686	-2.211294	3.355926
fren1_4_2	.3500228	1.288199	0.27	0.786	-2.182463	2.882509
fren1_4_3	-.2345575	.8991689	-0.26	0.794	-2.002245	1.53313
fren1_5_1	1.502437	1.220384	1.23	0.219	-.8967312	3.901605
fren1_5_2	2.576277	.9585749	2.69	0.007	.6918028	4.460751
fren1_5_3	1.241115	.8047569	1.54	0.124	-.3409668	2.823196
fren1_6_1	-.0117513	3.133703	-0.00	0.997	-6.172337	6.148834
fren1_6_2	2.324164	1.173116	1.98	0.048	.01792	4.630407
fren1_6_3	3.474897	.7653334	4.54	0.000	1.970319	4.979476
fren1_7_1	.9764202	1.148317	0.85	0.396	-1.281071	3.233912
fren1_7_2	2.523146	.9384858	2.69	0.007	.6781649	4.368127
fren1_7_3	4.326997	.7608453	5.69	0.000	2.831242	5.822752
fren1_8_1	2.216324	1.316567	1.68	0.093	-.3719318	4.80458
fren1_8_2	3.434432	1.324943	2.59	0.010	.8297102	6.039153
fren1_8_3	5.990918	.82435	7.27	0.000	4.370319	7.611518
fren1_9_1	1.012782	1.489874	0.68	0.497	-1.91618	3.941743
fren1_9_2	-1.626458	1.4795	-1.10	0.272	-4.535025	1.282109
fren1_9_3	1.840873	.5852809	3.15	0.002	.6902622	2.991484
fren1_10_1	4.162806	1.328116	3.13	0.002	1.551846	6.773765
fren1_10_2	4.780014	1.318398	3.63	0.000	2.18816	7.371868
fren1_10_3	5.796472	.7660954	7.57	0.000	4.290395	7.302548
fren1_11_1	.2983551	.7010506	0.43	0.671	-1.079849	1.676559
fren1_11_2	2.303626	.7101298	3.24	0.001	.9075727	3.699679
fren1_11_3	.5750168	.7695274	0.75	0.455	-.9378065	2.08784
fren1_12_1	2.643918	.7743923	3.41	0.001	1.121531	4.166306
fren1_12_2	3.004073	1.022598	2.94	0.003	.9937345	5.014411
fren1_12_3	2.895458	.7159329	4.04	0.000	1.487996	4.302919
fren1_13_1	.5130116	.8169124	0.63	0.530	-1.092966	2.11899
fren1_13_2	1.049118	.9913353	1.06	0.291	-.8997607	2.997996

fren1_13_3	1.962241	.8154523	2.41	0.017	.3591331	3.565349
fren1_14_1	3.223064	.8891235	3.62	0.000	1.475125	4.971003
fren1_14_2	2.361878	.8038129	2.94	0.003	.7816519	3.942103
fren1_14_3	.7445432	.5979174	1.25	0.214	-.43091	1.919996
fren1_15_1	1.264329	3.124522	0.40	0.686	-4.878209	7.406866
fren1_15_2	-.1689269	3.03374	-0.06	0.956	-6.132994	5.79514
fren1_15_3	.8760482	.8101069	1.08	0.280	-.7165509	2.468647
age	-.0781131	.0257575	-3.03	0.003	-.1287501	-.027476
sex_2	-1.285193	.6604015	-1.95	0.052	-2.583485	.0130983
cons	22.80402	2.257527	10.10	0.000	18.36592	27.24212

Adjusted R –squared 0.81, Root MSE 5.03

Model diagnostics, OLS

The OLS makes the assumption of normal errors. This was tested using the normal probability plots. Figure 2A shows the normal probability plot for the OLS model with outcome NEADL score and predictors FAI scores, age and gender. The plot is fairly straight suggesting normal residuals.

A

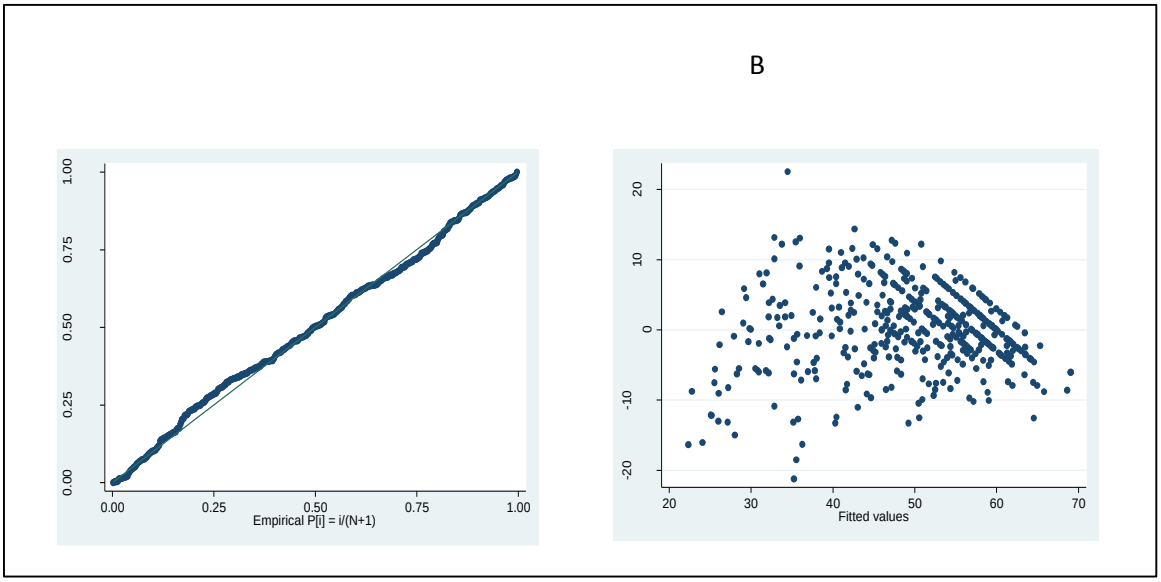

Figure 2 (A) Normal probability plot of residuals for the FAI onto NEADL mapping function, OLS, (B) Scatter plot of Residual versus fitted values

The homoscedasticity assumption was tested using a plot of residuals against fitted values for the OLS model with outcome NEADL score and predictors FAI total score, age, and gender. Figure 2B indicated non constant variance as the plot displayed a funnel shape. The error variance decreased as the fitted values increased. Homoscedasticity is indicated by a pattern less plot.

The Cook–Weisberg test was also used to assess the heteroscedacity assumption. An in significant result indicates homoscedasticity. The *p* value from the Cook–Weisberg test for the OLS model was statistically significant (Chi.sq =78.97, p <0.001) indicating heteroscedasticity

Figure 3 shows the added value plots that were plotted in STATA software to investigate the presence of outlying values. The added value for the plot of NEADL and FAI

in Figure 3 showed the presence of one outlier. Examination of the residuals showed one value with a negative residual (-3.62) and the other with a positive residual (+3.83) and the magnitude of both were > 3.5. Figure 4 also showed two points with high leverage. Sensitivity analysis by removing the outliers and re running the models did not result in significant changes in RMSE and MAE hence these cases were not removed from the data.

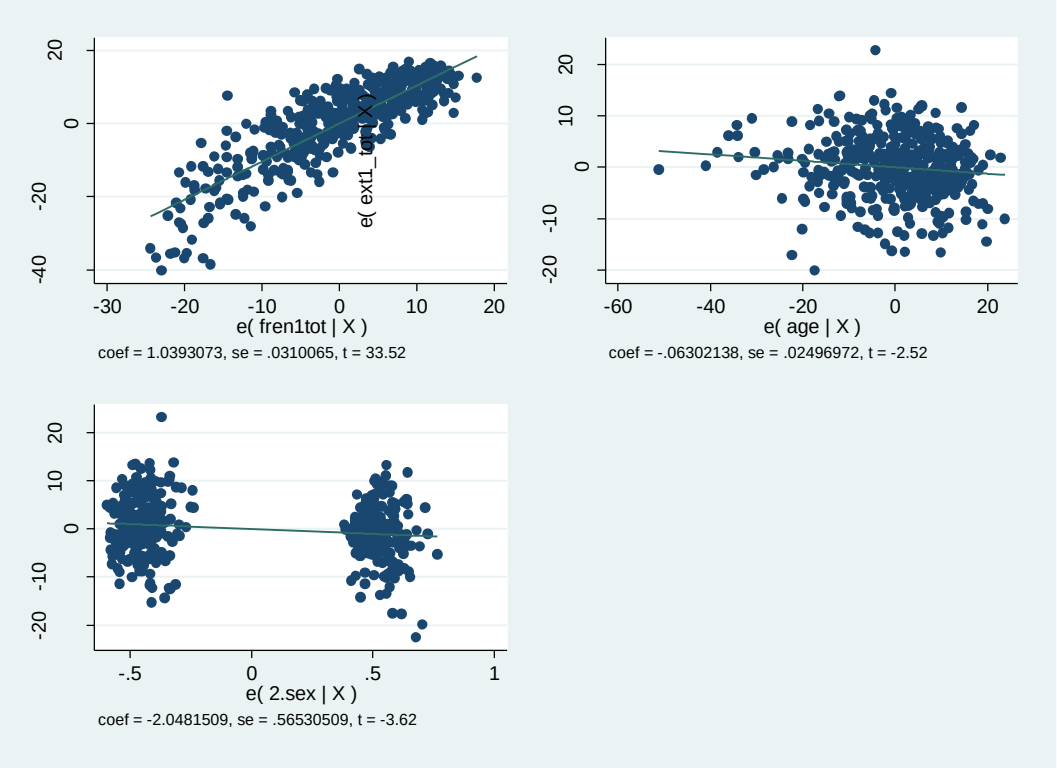

Figure 3 Added value plots, OLS mapping FAI onto NEADL questionnaire

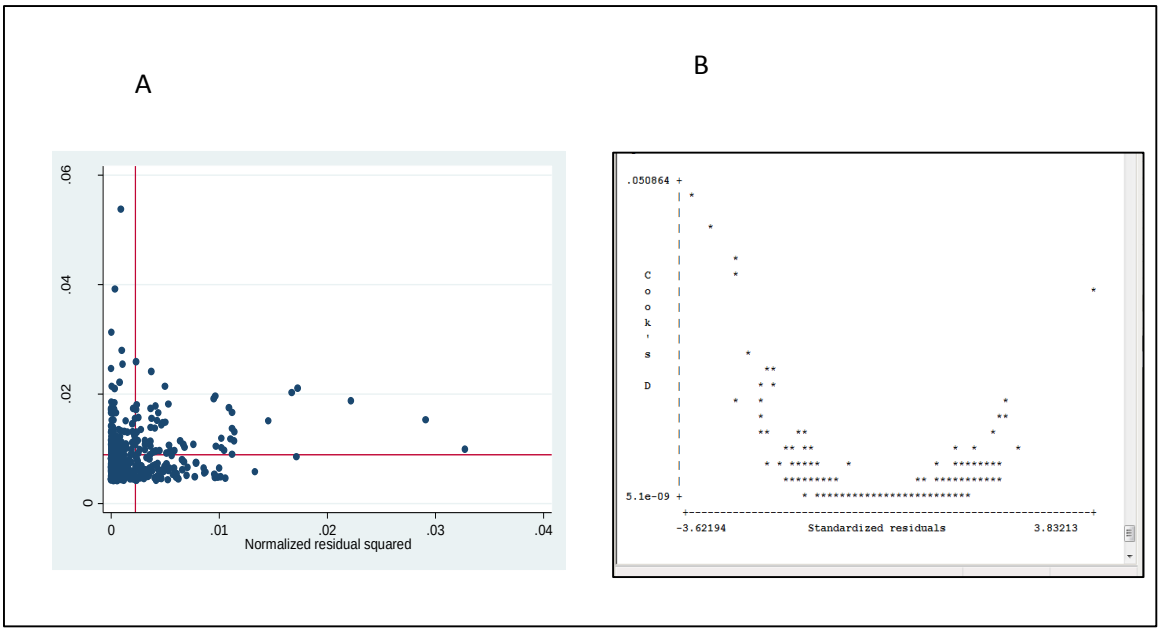

Figure 4 (A) Leverage versus squared residual plot, (B) Cooks distance plot

Cross Validation Mapping FAI-NEADL

Table 5: Five-fold validation: Mapping FAI onto NEADL

	RMSE	**MAE**
OLS FAI+FAI*FAI+age+sex	5.417541 5.696974 5.244362 5.79868 5.67789	3.787091 4.265321 4.307251 4.19425 4.540912
OLS items+age+sex	5.430917 4.959266 5.435479 5.401888 5.75866	3.35416 4.018354 4.200638 3.862025 4.467684

Table 6: Five –fold validation: Mapping NEADL onto FAI

	RMSE	**MAE**
OLS NEADL+NEADL*NEADL+age+sex	4.295964 4.849384 5.24053 4.261147 4.987681	3.898355 3.464806 3.590416 3.796449 3.936295
OLS NEADL items+age+sex	5.994952 4.346245 5.086259 4.708966 5.059642	3.553408 3.838467 4.521853 3.709249 4.975404

Table 7 Variance-Covariance matrix of OLS regression coefficients for mapping FAI onto NEADL questionnaire

e(V)	fren1tot	FAIsq	age	2.sex	_cons
fren1tot	.01587024				
FAIsq	-.00032311	6.951e-06			
age	.00008249	1.920e-06	.000552		
2.sex	.00272981	-.00009392	-.00154797	.28392669	
_cons	-.17304074	.002998	-.04198508	-.02055952	5.2640704

Table 8 Variance–Covariance matrix of OLS regression coefficients for mapping NEADL onto FAI questionnaire

e(V)	ext1_tot	neadlsq	age	2.sex	_cons
ext1_tot	.01227503				
neadlsq	-.0001423	1.707e-06			
age	-.00001313	1.497e-06	.00040667		
2.sex	.00301975	-.00003598	-.00089048	.20350877	
_cons	-.24099606	.00255328	-.03154878	-.0877158	7.6610332

Appendix E **MEASUREMENT MODEL SELECTION CHAPTER 7**

Measurement model for the SOS1 study

Table 9 Exploratory Factor analysis and latent class analysis results for baseline severity measured by NEADL subscales, BI and GHQ-28 subscales, SOS1 dataset

Model	**AIC**	**BIC**	**SSA BIC**	**LMR *p* value**	**Entropy**
Factor analysis					
FA 1f	19137.66	19248.49	19162.81	-	-
FA, 2f	18686.67	18830.35	18719.23	-	-
FA, 3f	18667.71	18840.12	18706.82	-	-
FA, 4f	18664.41	18861.44	18709.10		
Latent class analysis					
LCA 2 classes	18986.85	19101.78	19012.92	0.04	0.99
LCA 3 classes	18482.56	18638.54	18517.94	0.19	0.98
LCA 4 classes	18146.67	18343.69	18191.36	0.004	0.95
LCA 5 classes	17994.72	18232.79	18048.73	0.04	0.92
LCA 6 classes	17748.54	18027.67	17811.86	0.06	0.93
LCA 7 classes	17410.95	17731.12	17483.58	0.26	0.94
5 class with covariates	17937.23	18240.99	18006.14	0.002	0.92

AIC:Akaike Information Criteria, BIC:Bayesian information criteria, SSA:Sample Size Adjusted, LMR:Lo-Mendell-Rubin likelihood ratio test, LCA: Latent class analysis, FA: Factor analysis

Measurement model for CIMSS study

Table 10 Exploratory Factor analysis, latent class analysis results for baseline severity measured by NEADL subscales, BI and GHQ-12 subscales, CIMSS dataset.

Model	AIC	BIC	SSA BIC	LMR *p* value,	Entropy
Factor analysis					
FA 1f	11425.24	11503.84	11437.24	-	-
FA, 2f	11224.25	11325.31	11239.68	-	-
FA, 3f	11205.33	11325.11	11223.62	-	-
LCA					
LCA 2 classes	11372.42	11454.76	11384.98	<0.001	0.97
LCA 3 classes	11095.64	11207.93	11117.78	0.08	0.92
LCA 4 classes	10998.62	11140.86	110201.33	0.425	0.91
LCA 5 classes	10861.82	11033.99	10888.10	0.179	0.92
LCA 6 classes	10790.98	10993.10	10821.83	0.371	0.92
LCA 7 classes	10721.56	10953.62	10756.98	0.339	0.93
LCA, 8 classes	10593.32	10855.33	10633.31	0.407	0.95
LCA 6 classes with covariates	10384.45	10658.78	10424.09	0.460	0.92

AIC: Akaike Information Criteria, BIC: Bayesian information criteria, LMR: Lo-Mendell-Rubin likelihood ratio test, FA: Factor Analysis, LCA:Latent Class Analysis

Appendix F MPLUS CODES FOR MIXTURE MODELLING AND MULTI-GROUP CONFIRMATORY FACTOR MODELS FITTED IN CHAPTER 7

Mixture modelling

Random starts

Step 1

-Run a 2 class model, do not request tech11 and tech 14

Mplus default starts =20 5 (First number is the initial iterations; second number is number of final iterations).

The best log likelihood for k-1 and k classes real data was: -9465.424 253358

-Increase number of random starts: starts=100 20, best log likelihood was:

-9465.42915107 54

-Increase number of random starts 200 40.

The best log likelihood was -9465.42

-Making a further increase of random starts to: starts to 500 25 replicated the best log-likelihood value of -9465 that was found in the 20 5, 100 20,200 40. It was replicated 154 times.

Step 2:

-Run a 3 class solution

-Rerun it again with OPTseed from the 3 class solution in the previous model above and request tech 11

Step 3

-Rerun the 3 class solution with the same OPTseed and request tech 14

To avoid warnings when you request tech14, use lrtstarts=0 0 500 25.

Mplus syntax for Multi-group latent class analysis

Model 1: Unconstrained (heterogeneous model): The item means vary across studies, item variances vary across studies, allowed differences in class probabilities across groups.

```
DATA:
      FILE is M:\StrokeC\soslatentclass.dat ;
VARIABLE:
 NAMES ARE id_no study withdraw sex age Prevstroke living_Alone urineincon mmse
                pre_bart T1_BART T5_BART T1GHQTOT T5GHQTOT somatic1 anxiety1 social1
               depression1 T1GHQ28;
 MISSING are all (-9999) ;
 IDVARIABLE = id_no;
 USEVARIABLES ARE T1_BART somatic1 anxiety1 social1 depression1;
 CLASSES =cg (2)  c(5) ;
 KNOWNCLASS = cg(study = 1 study = 2);
 ANALYSIS: type = mixture;
  LRTSTARTS= 0 0 500 100;
  ALGORITHM=Integration;
  MODEL:
        %overall%
        c on cg;! allowing class probabilities for c to vary by study
   MODEL cg:
         %cg#1%
         T1_BART somatic1 anxiety1 social1 depression1;! allowing variances to vary across studies
          %cg#2%
         T1_BART somatic1 anxiety1 social1 depression1;
OUTPUT:TECH1 TECH8;
```

Model 2: Partial homogenous model: Model allowing differences in item means across groups, fixing class probabilities and item variances across groups and classes

```
DATA:
    FILE is M:\StrokeC\soslatentclass.dat ;
 VARIABLE:
 NAMES are  id_no study withdraw sex age Prevstroke living_Alone urineincon mmse
        pre_bart T1_BART T5_BART T1GHQTOT T5GHQTOT somatic1 anxiety1 social1
        depression1 T1GHQ28;
MISSING are all (-9999);
IDVARIABLE = id_no;
USEVARIABLES ARE  T1_BART somatic1 anxiety1 social1 depression1;
 CLASSES =cg (2)  c(5) ;
  KNOWNCLASS = cg(study = 1 study = 2);
   ANALYSIS: type = mixture;
   LRTSTARTS= 0 0 500 25;
   ALGORITHM=Integration;
   MODEL:
        %overall%
        c on cg;! allowing class probabilities for c to vary by  study
        somatic1 anxiety1 social1 depression1;
   MODEL:
   MODEL C:
         %c#1%
        [ T1_BART somatic1 anxiety1 social1 depression1];!fixing means
        %c#2%
       [ T1_BART somatic1 anxiety1 social1 depression1];
        %c#3%
       [ T1_BART somatic1 anxiety1 social1 depression1];
        %c#4%
       [ T1_BART somatic1 anxiety1 social1 depression1];
       %c#5%
        [ T1_BART somatic1 anxiety1 social1 depression1];
  OUTPUT:TECH1 TECH8;
  SAVEDATA:
  FILE is MGLCApartialmodel7.dat;
  SAVE = CPROBABILITIES;
   FORMAT IS FREE;
```

Model 3: Complete homogenous model

```
 DATA:
    File is M:\StrokeC\soslatentclass.dat ;
 VARIABLE:
  NAMES are id_no study withdraw sex age Prevstroke living_Alone urineincon mmse
              pre_bart T1_BART T5_BART T1GHQTOT T5GHQTOT somatic1 anxiety1 social1
              depression1 T1GHQ28;
 MISSING are all (-9999) ;
 IDVARIABLE = id_no;
USEVARIABLES ARE T1_BART somatic1 anxiety1 social1 depression1;
CLASSES =cg (2)  c(5) ;
KNOWNCLASS = cg(study = 1 study = 2);
ANALYSIS: type = mixture;
LRTSTARTS= 0 0 500 25;
ALGORITHM=Integration;
 MODEL:
 MODEL C:
       %c#1%
       [T1_BART  somatic1 anxiety1 social1 depression1];! fixing means
       %c#2%
       [T1_BART somatic1 anxiety1 social1 depression1];
       %c#3%
       [T1_BART somatic1 anxiety1 social1 depression1];
       %c#4%
       [T1_BART somatic1 anxiety1 social1 depression1];
       %c#5%
       [T1_BART somatic1 anxiety1 social1 depression1];
```

Appendix G R CODE FOR IRT MODELS FITTED IN CHAPTER 5

R code for fitting the grm models

```
library(foreign)
my.data=read.dta("FAINEADL.dta")
attach(my.data)
names(my.data)
library(ltm)
my.datamobility3=data.frame(ext1,ext2,ext3,ext4,ext5,ext6,ext12,ext14,ext121
,fren6,fren7,fren8,fren10,fren13,fren15)
fit11<-grm(my.datamobility3)
fit11
my.household=data.frame(fren1,fren2,fren3,fren4,fren5,ext7,ext8,ext9,ext10
,ext11,ext13,ext15)
fit12<-grm(my.household)
fit12
```

Appendix H **MULTI-LEVEL MODEL RESULTS CHAPTER 8:STUDY SPECIFIC, IDA AND TRADITIONAL AGGREGATED META ANALYSIS**

Variable	**SOS1 n=448**	**SOS2 N=585**	**Integrative data analysis pf SOS1 and SOS2 n=1033**	**Aggregate data Meta analysis, SOS1, SOS2 n=1033**
Fixed effects				
constant	1.36(-0.81, 3.54)	-0.49(-2.05,1.06)	-0.17(-1.33,0.99)	
Slope(months after stroke) Quadratic(months)	-0.01(-0.03,0.1) 0.001(-0.0003,0.001)	-0.02(-0.06, 0.02) 0.0002(-0.003, 0.004)	-0.01(-0.03,0.001) 0.0006(-0.0001, 0.001)	-0.01(-0.03,0.01) 0.001(0.000, 0.002)**
Gender female	0.13(-0.08, 0.35)	0.28(0.12, 0.44)***	0.21(0.09, 0.32)***	0.22(0.098,0.36)***
Age(years)	-0.01(-0.02, -0.01)**	-0.01(-0.02, -0.004)**	-0.01(-0.01, -0.005)***	-0.01(-0.01,-0.006)***
Previous stroke	-0.07(-0.33, 0.19)	0.05(-0.14, 0.24)	-0.03(-0.17, 0.11)	0.01(-0.14, 0.16)
Urine	0.14(-0.29, 0.58)	0.01(-0.19, 0.21)	-0.02(-0.19, 0.15)	0.03(-0.15,0.21)
MMSE	-0.03(-0.07,0.003)	-0.01(-0.04,0.02)	-0.01(-0.03, 0.01)	-0.02(-0.04,0.005)
Pre-BI	-0.01(-0.10, 0.07)	0.03(-0.03,0.09)	0.02(-0.02,0.06)	0.02(-0.03,0.07)
Social	0.11(0.08, 0.14)***	0.14(0.12,0.17)***	0.14(0.12,0.16)	0.13(0.11,0.15)***
Depression	0.16(0.11,0.20)***	0.12(0.09, 0.15)***	0.14(0.12,0.17)***	0.14(0.10, 0.17)***
BI	-0.02(-0.04, -0.01)**	-0.02(-0.03,-0.001)*	-0.01(-0.02,-0.004)**	-0.02(-0.03, -0.01)***
Somatic	0.12(0.08, 0.16)***	0.17(0.14,0.19)***	0.17(0.15,0.20)***	0.15(0.10, 0.19)***
Study	-	-	0.05(-0.07,0.18)	-
Variance components				
Var(cons)	0.72(0.54, 0.97)**	0.48(0.38, 0.62)**	0.44(0.36, 0.53)	
Log L	-1582.64	-3035.32	-4645.52	

References

Aben, I. et al. 2002. Validity of the Beck Depression Inventory, Hospital Anxiety and Depression Scale, SCL-90, and Hamilton Depression Rating Scale as screening instruments for depression in stroke patients. *Psychosomatics.* **43**(5), pp.386-393.

Adamson, J. et al. 2004. Is stroke the most common cause of disability? *Journal of Stroke and Cerebrovascular Diseases.* **13**(4), pp.171-177.

Ahn, D.-H. et al. 2015. The effect of post-stroke depression on rehabilitation outcome and the impact of caregiver type as a factor of post-stroke depression. *Annals of rehabilitation medicine.* **39**(1), pp.74-80.

Akaike, H. 1998. *Information theory and an extension of the maximum likelihood principle.* New York: Springer.

Alava, M.H. et al. 2013. The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. *Rheumatology.* **52**(5), pp.944-950.

Allen, J. et al. 2013. Integrating and extending cohort studies: lessons from the eXtending Treatments, Education and Networks in Depression (xTEND) study. *BMC medical research methodology.* **13**(1), p122.

Appelros, P. 2007. Characteristics of the Frenchay Activities Index one year after a stroke: a population-based study. *Disability and rehabilitation.* **29**(10), pp.785-790.

Ashford, S. et al. 2015. Systematic review of patient-reported outcome measures for functional performance in the lower limb. *Journal of rehabilitation medicine.* **47**(1), pp.9-17.

Askew, R.L. et al. 2013. Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. *Quality of Life Research.* **22**(10), pp.2769-2776.

Ayerbe, L. et al. 2013. Natural history, predictors and outcomes of depression after stroke: systematic review and meta-analysis. *The British Journal of Psychiatry.* **202**(1), pp.14-21.

Bagg, S. et al. 2002. Effect of age on functional outcomes after stroke rehabilitation. *Stroke.* **33**(1), pp.179-185.

Baguley, T. 2009. Standardized or simple effect size: what should be reported? *British Journal of Psychology.* **100**(3), pp.603-617.

Baker, F.B. 1992. Equating tests under the graded response model. *Applied Psychological Measurement.* **16**(1), pp.87-96.

Banks, J. et al. 2011. Attrition and health in ageing studies: evidence from ELSA and HRS. *Longitudinal and life course studies.* **2**(2), pp.101-126.

Barnes, M. and Good, D. 2013. Outcome measures in stroke rehabilitation. *Neurological Rehabilitation: Handbook of Clinical Neurology.* **110**(3), p105.

Bartholomew, D. et al. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach (Wiley Series in Probability and Statistics).* New Jersey: Wiley Hoboken.

Bartlett, M.S. 1954. A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological).* **16**(1), pp.296-298.

Bartoli, F. et al. 2013. Depression after stroke and risk of mortality: a systematic review and meta-analysis. *Stroke research and treatment.* **2013**, p1.

Bath, P.A. et al. 2010. The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. *Ageing and Society.* **30**(08), pp.1419-1437.

Bauer, D.J. and Hussong, A.M. 2009. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological methods.* **14**(2), p101.

Beauducel, A. and Herzberg, P.Y. 2006. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling.* **13**(2), pp.186-203.

Beavers, A.S. et al. 2013. Practical considerations for using exploratory factor analysis in educational research. *Practical assessment, research & evaluation.* **18**(6), pp.1-13.

Beck, A.T. et al. 1996. *Manual for the beck depression inventory-II.* San Antonio, TX: Psychological Corporation.

Bentler, P.M. 1980. Multivariate analysis with latent variables: Causal modeling. *Annual review of psychology.* **31**(1), pp.419-456.

Bhogal, S.K. et al. 2004. Lesion location and poststroke depression systematic review of the methodological limitations in the literature. *Stroke.* **35**(3), pp.794-802.

Bollen, K.A. 1989. A new incremental fit index for general structural equation models. *Sociological Methods & Research.* **17**(3), pp.303-316.

Bollen, K.A. 1998. Structural equation models. In: P Armitage, T.C. ed. *Encyclopedia of biostatistics.* Sussex,UK: Wiley pp.4363–72.

Borugian, M.J. et al. 2010. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *Canadian Medical Association Journal.* **182**(11), pp.1197-1201.

Bravata, D.M. et al. 2010. Processes of care associated with acute stroke outcomes. *Archives of internal medicine.* **170**(9), pp.804-810.

Brazier, J. et al. 2012. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). *Health Technol Assess.* **16**(32), pp.1-114.

Brazier, J.E. et al. 2010. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European journal of health economics.* **11**(2), pp.215-225.

Broomfield, N. et al. 2015. Poststroke anxiety is prevalent at the population level, especially among socially deprived and younger age community stroke survivors. *International Journal of Stroke.* **10**(6), pp.897-902.

Brown, C. et al. 2012. Post-stroke depression and functional independence: a conundrum. *Acta Neurologica Scandinavica.* **126**(1), pp.45-51.

Browne, M.W. et al. 1993. Alternative ways of assessing model fit. In: Bollen , K.A. and Long, J.S. eds. *Testing structural equation models.* Newbury Park, CA: Sage, pp.136-162.

Bryant, F.B. and Yarnold, P.R. 1995. Principal-components analysis and exploratory and confirmatory factor analysis. In: Grimm, L.G. and Yarnold, P.R. eds. *Reading and understanding multivariate analysis.* Washington, DC: American Psychological Association, pp.99-136.

Buber, I. and Engelhardt, H. 2011. The association between age and depressive symptoms among older men and women in Europe. Findings from SHARE. *Comparative Population Studies.* **36**(1), pp.77-102.

Bun Cheung, Y. 2002. A confirmatory factor analysis of the 12-item General Health Questionnaire among older people. *International Journal of Geriatric Psychiatry*. **17**(8), pp.739-744.

Burns, R.A. et al. 2011. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. *Journal of clinical epidemiology.* **64**(7), pp.787-793.

Busing, F.M. 1993. *Distribution characteristics of variance estimates in two-level models: a Monte Carlo study*. Unpublished.

Byers, K.L. 2004. *Testing the accuracy of linking healthcare data across the continuum of care*. thesis, University of Florida.

Cai, L. et al. 2011. *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* Chicago, IL: Scientific Software International.

Cameron, A.C. and Trivedi, P.K. 2005. *Microeconometrics: methods and applications.* Cambridge: Cambridge university press.

Campbell Burton, C. et al. 2013. Frequency of anxiety after stroke: a systematic review and meta-analysis of observational studies. *International Journal of Stroke.* **8**(7), pp.545-559.

Candelise, L. et al. 2007. Stroke-unit care for acute stroke patients: an observational follow-up study. *The Lancet.* **369**(9558), pp.299-305.

Carson, A.J. et al. 2000. Depression after stroke and lesion location: a systematic review. *The Lancet.* **356**(9224), pp.122-126.

Cattell, R.B. 1966. The scree test for the number of factors. *Multivariate behavioral research.* **1**(2), pp.245-276.

Celeux, G. and Soromenho, G. 1996. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification.* **13**(2), pp.195-212.

Chai, T. and Draxler, R. 2014. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions.* **7**(1), pp.1525-1534.

Chatterjee, S. and Hadi, A.S. 1986. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science.* **1**(3), pp.379-393.

Chatzidionysiou, K. et al. 2011. Highest clinical effectiveness of rituximab in autoantibody-positive patients with rheumatoid arthritis and in those for whom

no more than one previous TNF antagonist has failed: pooled data from 10 European registries. *Annals of the rheumatic diseases.* **70**(9), pp.1575-1580.

Chen, F.F. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling.* **14**(3), pp.464-504.

Chen, F.F. et al. 2005. Teacher's corner: Testing measurement invariance of second-order factor models. *Structural equation modeling.* **12**(3), pp.471-492.

Chen, G. et al. 2014. From KIDSCREEN-10 to CHU9D: creating a unique mapping algorithm for application in economic evaluation. *Health Qual Life Out.* **12**(1), p134.

Chen, W.-H. et al. 2009. Linking pain items from two studies onto a common scale using item response theory. *Journal of pain and symptom management.* **38**(4), pp.615-628.

Cheung, G.W. and Rensvold, R.B. 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling.* **9**(2), pp.233-255.

Chuang, L.-H. and Whitehead, S.J. 2011. Mapping for economic evaluation. *British medical bulletin.* **101**(1), pp.1-15.

Clark, S.L. and Muthén, B. 2009. Relating latent class analysis results to variables not included in the analysis. [Online]. [Accessed 10 January 2016]. Available from: www.statmodel.com/download/relatinglca.pdf.

Cleveland, W.S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician.* **35**(1), p54.

Clogg, C.C. 1985. Simultaneous latent structure analysis in several groups. In: Tuma, N.B. ed. *Sociological methodology 1985.* San Franscisco: Jossey-Bass, pp.81-110.

Cohen, A.S. and Kim, S.-H. 1998. An investigation of linking methods under the graded response model. *Applied Psychological Measurement.* **22**(2), pp.116-130.

Collaboration, A.T.C. 2007. Prognosis of HIV-1-infected patients up to 5 years after initiation of HAART: collaborative analysis of prospective studies. *AIDS* **21**(9), p1185.

Collen, F. et al. 1991. The Rivermead mobility index: a further development of the Rivermead motor assessment. *International disability studies.* **13**(2), pp.50-54.

Coloma, P.M. et al. 2011. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiology and drug safety.* **20**(1), pp.1-11.

Cook, R.D. 1977. Detection of influential observation in linear regression. *Technometrics.* **19**(1), pp.15-18.

Cook, R.D. and Weisberg, S. 1982. *Residuals and influence in regression.* New York: Chapman and Hall.

Counsell, C. and Dennis, M. 2001. Systematic review of prognostic models in patients with acute stroke. *Cerebrovascular diseases.* **12**(3), pp.159-170.

Croudace, T.J. et al. 2003. Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology.* **157**(9), pp.834-842.

Curran, P.J. and Hussong, A.M. 2009. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods.* **14**(2), pp.81-100.

Curran, P.J. et al. 2008. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology.* **44**(2), p365.

Daniels, B. 2012. *CROSSFOLD: Stata module to perform k-fold cross-validation.* Boston College Department of Economics: Statistical Software Components S457426.

Davenport, R.J. et al. 1996. Effect of correcting outcome data for case mix: an example from stroke medicine. *Bmj.* **312**(7045), pp.1503-1505.

de Koning, I. et al. 1998. Value of screening instruments in the diagnosis of post-stroke dementia. *Pathophysiology of Haemostasis and Thrombosis.* **28**(3-4), pp.158-166.

Dempster, C. et al. 1998. The collaboration of carers during psychological therapy. *Mental health Nursing-London-Community Psychiatric Nurses Association.* **18**(3), pp.24-27.

Dick, J. et al. 1984. Mini-mental state examination in neurological patients. *Journal of Neurology, Neurosurgery & Psychiatry.* **47**(5), pp.496-499.

Dismuke, C. and Lindrooth, R. 2006. Ordinary least squares. In: Chumney, E.C.G. and Simpson , K.N. eds. *Methods and Designs for Outcomes Research.* Bethesda: ASHP, pp.93-104.

Doiron, D. et al. 2013. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg Themes Epidemiol.* **10**(1), p12.

Donnan, G.A. et al. 2008. Stroke. *The Lancet.* **371**(9624), pp.1612-1623.

Dorans, N.J. 2004. Equating, concordance, and expectation. *Applied Psychological Measurement.* **28**(4), pp.227-246.

Dorans, N.J. 2007. Linking scores from multiple health outcome instruments. *Quality of Life Research.* **16**(1), pp.85-94.

Dorans, N.J. and Holland, P.W. 2000. Population invariance and the equatability of tests: Basic theory and the linear case. *ETS Research Report Series.* **2000**(2), pp.i-35.

Duncan, P.W. et al. 2000. Outcome measures in acute stroke trials a systematic review and some recommendations to improve practice. *Stroke.* **31**(6), pp.1429-1438.

Dunning, K. 2011. National Institutes of Health Stroke Scale. *In Encyclopedia of Clinical Neuropsychology* New York: Springer, pp. 1714-1715.

Edelen, M.O. et al. 2014. Correspondence Between the RAND-Negative Impact of Asthma on Quality of Life Item Bank and the Marks Asthma Quality of Life Questionnaire. *Clinical Therapeutics.* **36**(5), pp.680-688.

Efron, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association.* **78**(382), pp.316-331.

Eid, M. et al. 2003. Comparing Typological Structures Across Cultures By Multigroup Latent Class Analysis A Primer. *Journal of Cross-Cultural Psychology.* **34**(2), pp.195-210.

Fairhurst, C. et al. 2014. Factor analysis of treatment outcomes from a UK specialist addiction service: Relationship between the Leeds Dependence Questionnaire, Social Satisfaction Questionnaire and 10-item Clinical Outcomes in Routine Evaluation. *Drug and alcohol review.* **33**(6), pp.643-650.

Fayers, P.M. and Hays, R.D. 2014. Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value in Health.* **17**(2), pp.261-265.

Ferro, J.M. et al. 2009. Poststroke emotional and behavior impairment: a narrative review. *Cerebrovascular Diseases.* **27**(Suppl. 1), pp.197-203.

Finch, W.H. and Bronk, K.C. 2011. Conducting Confirmatory Latent Class Analysis Using M plus. *Structural Equation Modeling.* **18**(1), pp.132-151.

Fitzpatrick, A.R. and Yen, W.M. 2001. The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education.* **14**(1), pp.31-57.

Flick, C.L. 1999. 4. Stroke outcome and psychosocial consequences. *Archives of physical medicine and rehabilitation.* **80**(5), pp.S21-S26.

Flora, D.B. et al. 2008. Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Structural Equation Modeling.* **15**(4), pp.676-704.

Folstein, M.F. et al. 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research.* **12**(3), pp.189-198.

Fong, T.G. et al. 2009. Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination. *Alzheimer's & Dementia.* **5**(6), pp.492-497.

Fortier, I. et al. 2010. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *International journal of epidemiology.* **39**(5), pp.1383-1393.

Fortier, I. et al. 2011. Invited commentary: consolidating data harmonization—how to obtain quality and applicability? *American journal of epidemiology.* **174**(3), pp.261-264.

Fox, J. 2015. *Applied regression analysis and generalized linear models.* 3 ed. Thousand Oaks CA: Sage Publications.

Friedenreich, C.M. 1993. Methods for pooled analyses of epidemiologic studies. *Epidemiology.* **4**(4), pp.295-302.

Fryback, D.G. et al. 1997. Predicting Quality of Well-being Scores from the SF-36 Results from the Beaver Dam Health Outcomes Study. *Medical Decision Making.* **17**(1), pp.1-9.

Gao, F. et al. 2004. Does the 12-item General Health Questionnaire contain multiple factors and do we need them? *Health and Quality of Life Outcomes.* **2**(1), p1.

Geiser, C. et al. 2006. Separating" rotators" from" nonrotators" in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research.* **41**(3), pp.261-293.

Gelman, A. and Hill, J. 2006. *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gelman, A. et al. 1998. Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association.* **93**(443), pp.846-857.

Geyh, S. et al. 2004. ICF Core Sets for stroke. *Journal of Rehabilitation Medicine.* **36**(0), pp.135-141.

Ghatnekar, O. et al. 2013. Mapping health outcome measures from a stroke registry to EQ-5D weights. *Health Qual Life Outcomes.* **11**(1), p34.

Gibbons, P. et al. 2004. Assessment of the factor structure reliability of the 28 item version of the general Health Questionnaire (GHQ-28) in El Salvador. *International Journal of Clinical and Health Psychology.* **4**(2), pp.389-398.

Gladman, J. et al. 1993. Use of the extended ADL scale with stroke patients. *Age and ageing.* **22**(6), pp.419-424.

Goldberg, D. et al. 1998. Why GHQ threshold varies from one place to another. *Psychological medicine.* **28**(04), pp.915-921.

Goldberg, D.P. and Hillier, V.F. 1979. A scaled version of the General Health Questionnaire. *Psychological medicine.* **9**(01), pp.139-145.

Gorsuch, R. 1983. *Factor analysis, 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Graetz, B. 1991. Multidimensional properties of the general health questionnaire. *Social psychiatry and psychiatric epidemiology.* **26**(3), pp.132-138.

Graham, J.W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology.* **60**(1), pp.549-576.

Gray, A.M. et al. 2006. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making.* **26**(1), pp.18-29.

Gregorich, S.E. 2006. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical care.* **44**(11 Suppl 3), pS78.

Griffith, L. et al. 2013. *Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis.* Rockville, MD: Agency for Healthcare Research and Quality.

Griffith, L.E. et al. 2015. Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *Journal of clinical epidemiology.* **68**(2), pp.154-162.

Grootendorst, P. et al. 2007. A model to estimate health utilities index mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the knee. *The Journal of rheumatology.* **34**(3), pp.534-542.

Group, T.E. 1990. EuroQol-a new facility for the measurement of health-related quality of life. *Health policy.* **16**(3), pp.199-208.

Guadagnoli, E. and Velicer, W.F. 1988. Relation to sample size to the stability of component patterns. *Psychological bulletin.* **103**(2), p265.

Hackett, M.L. et al. 2009. Interventions for preventing depression after stroke. *Stroke.* **40**(7), pp.e485-e486.

Hackett, M.L. et al. 2005. Frequency of depression after stroke. A systematic review of observational studies. *Stroke.* **36**, pp.1330-1340.

Hadidi, N. et al. 2009. Poststroke depression and functional outcome: a critical review of literature. *Heart & Lung: The Journal of Acute and Critical Care.* **38**(2), pp.151-162.

Hair, J.F. 2010. *Multivariate data analysis (7th ed).* New York: Pearson College Division.

Hall, D. et al. 2012. Sharing heterogeneous data: the national database for autism research. *Neuroinformatics.* **10**(4), pp.331-339.

Hamilton, M. 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry.* **23**(1), p56.

Hankins, M. 2008. The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clinical Practice and Epidemiology in Mental Health.* **4**(1), p1.

Harrell, F.E. et al. 1996. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine.* **15**(4), pp.361-387.

Hatano, S. 1976. Experience from a multicentre stroke register: a preliminary report. *Bulletin of the World Health Organization.* **54**(5), p541.

Hawkins, D.M. 1980. *Identification of outliers.* London: Chapman and Hall.

Hawthorne, G. et al. 2008. Deriving utility scores from the SF-36 health instrument using Rasch analysis. *Quality of Life Research.* **17**(9), pp.1183-1193.

Hedges, L.V. and Vevea, J.L. 1998. Fixed-and random-effects models in meta-analysis. *Psychological methods.* **3**(4), p486.

Helvik, A.-S. et al. 2011. A psychometric evaluation of the Hospital Anxiety and Depression Scale for the medically hospitalized elderly. *Nordic journal of psychiatry.* **65**(5), pp.338-344.

Hill, K.M. et al. 2009. The Stroke Outcomes Study 2 (SOS2): a prospective, analytic cohort study of depressive symptoms after stroke. *BMC cardiovascular disorders.* **9**(1), p22.

Hofer, S.M. and Piccinin, A.M. 2009. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological methods.* **14**(2), p150.

Hofer, S.M. and Piccinin, A.M. 2010. Toward an integrative science of life-span development and aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences.* **65B**(3), pp.269-278.

Holbrook, M. and Skilbeck, C. 1983. An activities index for use with stroke patients. *Age and ageing.* **12**(2), pp.166-170.

Holland, P.W. et al. 2006. Equating test scores. *Handbook of statistics.* **26**, pp.169-203.

Holzner, B. et al. 2006. Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. *European Journal of Cancer.* **42**(18), pp.3169-3177.

Horn, J.L. et al. 1983. When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist.* **1**(4), pp.179-188.

Horwood, L.J. et al. 2012. Cannabis and depression: an integrative data analysis of four Australasian cohorts. *Drug and alcohol dependence.* **126**(3), pp.369-378.

House, A. et al. 2001. Mortality at 12 and 24 months after stroke may be associated with depressive symptoms at 1 month. *Stroke.* **32**(3), pp.696-701.

Hox, J.J. 1995. *Applied multilevel analysis.* Amsterdam: TT-publikaties

Hox, J.J. et al. 2010. *Multilevel analysis: Techniques and applications.* 2nd ed. New York: Routledge.

Hsueh, I.-P. et al. 2004. Rasch analysis of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke.* **35**(3), pp.721-726.

Hussong, A.M. et al. 2008. Disaggregating the distal, proximal, and time-varying effects of parent alcoholism on children's internalizing symptoms. *Journal of abnormal child psychology.* **36**(3), pp.335-346.

Hussong, A.M. et al. 2013. Integrative data analysis in clinical psychology research. *Annual review of clinical psychology.* **9**, p61.

Indredavik, B. et al. 1999. Treatment in a Combined Acute and Rehabilitation Stroke Unit Which Aspects Are Most Important? *Stroke.* **30**(5), pp.917-923.

Jolliffe, D. et al. 2001. Censored least absolute deviations estimator: CLAD. *Stata Technical Bulletin.* **10**(58), pp.13-16.

Kaiser, H.F. 1960. The application of electronic computers to factor analysis. *Educational and psychological measurement.* **20**(1), pp.141-151.

Kaiser, H.F. 1974. An index of factorial simplicity. *Psychometrika.* **39**(1), pp.31-36.

Kalliath, T.J. et al. 2004. A confirmatory factor analysis of the General Health Questionnaire-12. *Stress and Health.* **20**(1), pp.11-20.

Kankaraš, M. et al. 2010. Testing for measurement invariance with latent class analysis. In: Davidov, E., et al. eds. *Cross-cultural analysis: Methods and applications.* New York: Routledge, pp.359-384.

Kearns, B. et al. 2012. A review of the use of statistical regression models to inform cost effectiveness analyses within the NICE technology appraisals programme. *Report by the NICE Decision Support Unit.*

Kelly-Hayes, P.M. et al. 1998. The American heart association stroke outcome classification. *Stroke.* **29**(6), pp.1274-1280.

Kern, M.L. et al. 2014. Integrating prospective longitudinal data: Modeling personality and health in the Terman Life Cycle and Hawaii Longitudinal Studies. *Developmental psychology.* **50**(5), p1390.

Kersten, P. et al. 2010. The Subjective Index for Physical and Social Outcome (SIPSO) in Stroke: investigation of its subscale structure. *BMC neurology.* **10**(1), p1.

Kihç, C. et al. 1997. General Health Questionnaire (GHQ12 & GHQ28): psychometric properties and factor structure of the scales in a Turkish primary care sample. *Social Psychiatry and Psychiatric Epidemiology.* **32**(6), pp.327-331.

Kim, J.-O. and Mueller, C.W. 1978. *Introduction to factor analysis: What it is and how to do it.* Newbury Park: Sage Publications

Kim, S.-H. and Cohen, A.S. 1998. Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement.* **22**(4), pp.345-355.

King-Kallimanis, B.L. et al. 2012. Assessing measurement invariance of a health-related quality-of-life questionnaire in radiotherapy patients. *Quality of Life Research.* **21**(10), pp.1745-1753.

Kjær, J. and Ledergerber, B. 2004. Short communication HIV cohort collaborations: proposal for harmonization of data exchange. *Antiviral therapy.* **9**(4), pp.631-633.

Knoppers, B. et al. 2008. The Public Population Project in Genomics (P3G): a proof of concept. *Eur J Hum Genet.* **16**(6), pp.664-665.

Koenker, R. and Bassett Jr, G. 1982. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society.* **50**(1), pp.43-61.

Kolen, M.J. and Brennan, R.L. 2004. *Test equating, scaling, and linking.* New York: Springer.

Kolen, M.J. and Brennan, R.L. 2013. *Test equating: Methods and practices.* New York: Springer Science & Business Media.

Kollen, B. et al. 2006. Functional recovery after stroke: a review of current developments in stroke rehabilitation research. *Reviews on recent clinical trials.* **1**(1), pp.75-80.

Kreft, I.G. et al. 1998. *Introducing multilevel modeling.* London: Sage.

Kutlubaev, M.A. and Hackett, M.L. 2014. Part II: predictors of depression after stroke and impact of depression on stroke outcome: an updated systematic review of observational studies. *International Journal of Stroke.* **9**(8), pp.1026-1036.

Kwok, T. et al. 2006. Quality of life of stroke survivors: a 1-year follow-up study. *Archives of physical medicine and rehabilitation.* **87**(9), pp.1177-1182.

Langhorne, P. and Dennis, M.S. 2004. Stroke units: the next 10 years. *The Lancet.* **363**(9412), pp.834-835.

Langhorne, P. and Duncan, P. 2001. Does the organization of postacute stroke care really matter? *Stroke.* **32**(1), pp.268-274.

Lazarsfeld, P.F. et al. 1968. *Latent structure analysis.* Boston, Mass: Houghton Mifflin

Leppävuori, A. et al. 2003. Generalized anxiety disorders three to four months after ischemic stroke. *Cerebrovascular diseases.* **16**(3), pp.257-264.

Leung, S.F. and Yu, S. 1996. On the choice between sample selection and two-part models. *Journal of econometrics.* **72**(1), pp.197-229.

Li, C.-H. 2014. *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables*. thesis, Michigan State University

Lilliefors, H.W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association.* **62**(318), pp.399-402.

Lincoln, N. et al. 2013. Anxiety and depression after stroke: a 5 year follow-up. *Disability and rehabilitation.* **35**(2), pp.140-145.

Lincoln, N.B. and Gladman, J.R. 1992. The extended activities of daily living scale: a further validation. *Disability and rehabilitation.* **14**(1), pp.41-43.

Little, R.J. and Rubin, D.B. 2014. *Statistical analysis with missing data.* New York: John Wiley & Sons.

Lo, Y. et al. 2001. Testing the number of components in a normal mixture. *Biometrika.* **88**(3), pp.767-778.

Longworth, L. and Rowen, D. 2011. *NICE DSU technical support document 10: the use of mapping methods to estimate health state utility values.* [Online]. University of Sheffield, UK: Decision Support Unit, ScHARR,. [Accessed 24 August 2014]. Available from: http://www.nicedsu.org.uk/TSD%2010%20mapping%20FINAL_

Lord, F.M. 1980. *Applications of item response theory to practical testing problems.* New York: Routledge.

Lord, F.M. 1982. The standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics.* **7**(3), pp.165-174.

Lu, W.-S. et al. 2012. Smallest real difference of 2 instrumental activities of daily living measures in patients with chronic stroke. *Archives of physical medicine and rehabilitation.* **93**(6), pp.1097-1100.

Lubke, G. and Neale, M.C. 2006. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research.* **41**(4), pp.499-532.

Lubke, G.H. and Muthén, B. 2005. Investigating population heterogeneity with factor mixture models. *Psychological methods.* **10**(1), p21.

MacCallum, R.C. et al. 1999. Sample size in factor analysis. *Psychological methods.* **4**(1), p84.

MacHale, S.M. et al. 1998. Depression and its relation to lesion location after stroke. *Journal of Neurology, Neurosurgery & Psychiatry.* **64**(3), pp.371-374.

Mahoney, F.I. 1965. Functional evaluation: the Barthel index. *Maryland state medical journal.* **14**, pp.61-65.

Mair, P. et al. 2009. Extended Rasch Modeling: The R Package eRm. *PDF-Dateianhang zum Programmpaket eRm.* [Online]. [Accessed 21 February 2015]. Available from: Retrieved from http://cran.r-project.org/web/packages/eRm/vignettes/eRm.pdf

Marsh, H.W. et al. 2004. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural equation modeling.* **11**(3), pp.320-341.

Masters, G.N. 1982. A Rasch model for partial credit scoring. *Psychometrika.* **47**(2), pp.149-174.

Masyn, K.E. 2013. Latent class analysis and finite mixture modeling. *The Oxford handbook of quantitative methods in psychology.* **2**, pp.551-611.

Mayo, N.E. et al. 2015. Getting on with the rest of your life following stroke: A randomized trial of a complex intervention aimed at enhancing life participation post stroke. *Clinical rehabilitation.* **29**(12), pp.1198–1211.

Mayo, N.E. et al. 2013. Modeling health-related quality of life in people recovering from stroke. *Quality of Life Research.* **24**(1), pp.41-53.

McCutcheon, A.L. 2002. Basic concepts and procedures in single-and multiple-group latent class analysis. In: Hagenaars, J.A. and McCutcheon, A.L. eds. *Applied latent class analysis.* Cambridge: Cambridge University Press pp.56-88.

McHorney, C.A. and Cohen, A.S. 2000. Equating health status measures with item response theory: illustrations with functional status items. *Medical care.* **38**(9), pp.II-43.

McLachlan, G. and Peel, D. 2004. *Finite mixture models.* New York: John Wiley & Sons.

McNaughton, H. et al. 2003. Relationship between process and outcome in stroke care. *Stroke.* **34**(3), pp.713-717.

Mellenbergh, G.J. 1989. Item bias and item response theory. *International journal of educational research.* **13**(2), pp.127-143.

Menlove, L. et al. 2015. Predictors of Anxiety after Stroke: A Systematic Review of Observational Studies. *Journal of Stroke and Cerebrovascular Diseases.* **24**(6), pp.1107-1117.

Meredith, W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika.* **58**(4), pp.525-543.

Merriman, C. et al. 2007. Psychological correlates of PTSD symptoms following stroke. *Psychology, Health and Medicine.* **12**(5), pp.592-602.

Meyer, B.C. et al. 2002. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials prospective reliability and validity. *Stroke.* **33**(5), pp.1261-1266.

Mierlo, M.L. et al. 2014. The influence of psychological factors on Health-Related Quality of Life after stroke: a systematic review. *International Journal of Stroke.* **9**(3), pp.341-348.

Millis, S.R. et al. 2007. Measurement properties of the National Institutes of Health Stroke Scale for people with right-and left-hemisphere lesions: further analysis of the clomethiazole for acute stroke study–ischemic (class-I) trial. *Archives of physical medicine and rehabilitation.* **88**(3), pp.302-308.

Millsap, R.E. 2010. Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives.* **4**(1), pp.5-9.

Minicuci, N. et al. 2003. Cross-national determinants of quality of life from six longitudinal studies on aging: the CLESA project. *Aging clinical and experimental research.* **15**(3), pp.187-202.

Mohammed, M.A. et al. 2005. Comparing processes of stroke care in high-and low-mortality hospitals in the West Midlands, UK. *International Journal for Quality in Health Care.* **17**(1), pp.31-36.

Mokken, R.J. 1971. *A theory and procedure of scale analysis: With applications in political research.* Netherlands: Walter de Gruyter.

Molina, J.G. et al. 2014. Wording effects and the factor structure of the 12-item General Health Questionnaire (GHQ-12). *Psychological assessment.* **26**(3), p1031.

Morse, A. and General, A. 2010. *Progress in improving stroke care*. London: TSO.

Muñoz, M.A. and Acuña, J.D. 1999. Sample size requirements of a mixture analysis method with applications in systematic biology. *Journal of theoretical biology.* **196**(2), pp.263-265.

Munyombwe, T. et al. 2014. Mixture modelling analysis of one-month disability after stroke: stroke outcomes study (SOS1). *Quality of Life Research.* **23**(8), pp.2267-2275.

Munyombwe, T. et al. 2015. Testing measurement invariance of the GHQ-28 in stroke patients. *Quality of Life Research.* **24**(8), pp.1823-1827.

Muthén, B. 2003. Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran (2003). *Psychological Methods.* **8**(3), pp. 369-377.

Muthén, B. 2004. Latent variable analysis. In: Kaplan, D. ed. *The Sage handbook of quantitative methodology for the social sciences.* Newbury Park, CA: Sage pp.345-68.

Muthén, B. and Asparouhov, T. 2002. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes.* **4**(5), pp.1-22.

Muthén, B. and Muthén, L. 2012. Mplus Version 7: User's guide. *Los Angeles, CA: Muthén & Muthén.*

Nair, R.d. et al. 2011. Rasch analysis of the Nottingham extended activities of daily living scale. *Journal of rehabilitation medicine.* **43**(10), pp.944-950.

Newsom, J. 2015. *Invariance Tests in Multigroup SEM.* [Online]. [Accessed November 2015]. Available from: http://www.upa.pdx.edu/IOA/newsom/semclass/ho_multigroup.pdf

Nicholl, C. et al. 2002. The reliability and validity of the Nottingham Extended Activities of Daily Living Scale in patients with multiple sclerosis. *Multiple sclerosis.* **8**(5), pp.372-376.

Nouri, F. and Lincoln, N. 1987. An extended activities of daily living scale for stroke patients. *Clinical rehabilitation.* **1**(4), pp.301-305.

Nylund, K.L. et al. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling.* **14**(4), pp.535-569.

Organization, W.H. 2001. *International classification of functioning, disability and health: ICF*. World Health Organization.

Organization, W.H. 2007. *International Classification of Functioning, Disability, and Health: Children & Youth Version: ICF-CY*. World Health Organization.

Orlando, M. et al. 2000. Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment.* **12**(3), p354.

Osborne, J.W. and Costello, A.B. 2009. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review.* **12**(2), pp.131-146.

Pan, J.H. et al. 2008. Longitudinal analysis of quality of life for stroke survivors using latent curve models. *Stroke.* **39**(10), pp.2795-2802.

Parmigiani, G. et al. 2003. Cross-calibration of stroke disability measures: Bayesian analysis of longitudinal ordinal categorical data using negative dependence. *Journal of the American Statistical Association.* **98**(462), pp.273-281.

Pastor, D.A. et al. 2007. A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology.* **32**(1), pp.8-47.

Petrou, S. et al. 2015. Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement. *Health and quality of life outcomes.* **13**(1), p1.

Pett, M.A. et al. 2003. *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* California: Sage.

Ploubidis, G.B. et al. 2007. Improvements in social functioning reported by a birth cohort in mid-adult life: A person-centred analysis of GHQ-28 social dysfunction items using latent class analysis. *Personality and Individual Differences.* **42**(2), pp.305-316.

Pluijm, S. et al. 2005. A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. *Journal of clinical epidemiology.* **58**(10), pp.1015-1023.

Pohjasvaara, T. et al. 2001. Depression is an independent predictor of poor long-term functional outcome post-stroke. *European Journal of Neurology.* **8**(4), pp.315-319.

Prady, S.L. et al. 2013. The psychometric properties of the subscales of the GHQ-28 in a multi-ethnic maternal sample: results from the Born in Bradford cohort. *BMC psychiatry.* **13**(1), p55.

Proskorovsky, I. et al. 2014. Mapping EORTC QLQ-C30 and QLQ-MY20 to EQ-5D in patients with multiple myeloma. *Health Qual Life Outcomes.* **12**(1), p35.

Rasch, G. 1993. *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press.

Raudenbush, S.W. and Bryk, A.S. 2002. *Hierarchical linear models: Applications and data analysis methods.* Chicago: Sage.

Resche-Rigon, M. et al. 2013. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in medicine.* **32**(28), pp.4890-4905.

Riley, R.D. et al. 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ.* **340**(1), pc221.

Ripke, S. et al. 2013. A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular psychiatry.* **18**(4), pp.497-511.

Rizopoulos, D. 2006. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software.* **17**(5), pp.1-25.

Robinson, R.G. et al. 1975. Effect of experimental cerebral infarction in rat brain on catecholamine and behaviour. *Nature.* **255**, pp.332-334.

Rowen, D. et al. 2009. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship. *Health Qual Life Outcomes.* **7**(1), p27.

Royston, P. 2009. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal.* **9**(3), p466.

Royston, P. and Sauerbrei, W. 2004. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in medicine.* **23**(16), pp.2509-2525.

Rudd, A.G. et al. 1998. The national sentinel audit for stroke: a tool for raising standards of care. *Journal of the Royal College of Physicians of London.* **33**(5), pp.460-464.

Ryan, J. and Brockmann, F. 2009. *A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory.* Arizona State University.

Ryan, T.A. 1974. *Normal probability plots and tests for normality.* [Online]. State College, PA: Pennsylvania State University. [Accessed 14 May 2015]. Available from: https://www.minitab.com/uploadedFiles/Content/News/Published_Articles/normal_probability_plots.pdf

Salter, K. et al. 2007. The assessment of poststroke depression. *Topics in stroke rehabilitation.* **14**(3), pp.1-24.

Samejima, F. 1997. Graded response model. In: van der Linden, W.J. and Hambleton, R.K. eds. *Handbook of modern item response theory* New York: Springer, pp.85-100.

Santoso, S. 2014. SPSS 22 from essential to expert skills. *Jakarta: PT Elex Media Komputindo.*

Sarker, S.-J. et al. 2012. Comparison of 2 Extended Activities of Daily Living Scales With the Barthel Index and Predictors of Their Outcomes Cohort Study Within the South London Stroke Register (SLSR). *Stroke.* **43**(5), pp.1362-1369.

Schalet, B.D. et al. 2014. Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of anxiety disorders.* **28**(1), pp.88-96.

Schepers, V. et al. 2006. Responsiveness of functional health status measures frequently used in stroke research. *Disability and rehabilitation.* **28**(17), pp.1035-1040.

Schreuders, T.A. et al. 2003. Measurement error in grip and pinch force measurements in patients with hand injuries. *Physical therapy.* **83**(9), pp.806-815.

Schuling, J. et al. 1993. The Frenchay Activities Index. Assessment of functional status in stroke patients. *Stroke.* **24**(8), pp.1173-1177.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics.* **6**(2), pp.461-464.

Siddique, J. et al. 2015. Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Statistics in medicine.* **34**(26), pp.3399-3414.

Smith, A.B. et al. 2010. Research A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ)-12. *Health Qual Life Out.* **8**(1), p45.

Smith, B.H. et al. 2006. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Medical Genetics.* **7**(1), p1.

StataCorp, L. 2013. *Stata 13*. StataCorp LP., College Station, Texas, United States. http://www. stata. com.

Steenkamp, J.-B.E. and Baumgartner, H. 1998. Assessing measurement invariance in cross-national consumer research. *Journal of consumer research.* **25**(1), pp.78-107.

Sterling, M. 2011. General Health Questionnaire–28 (GHQ-28). *Journal of physiotherapy.* **57**(4), p259.

Stiles, P.G. and McGarrahan, J.F. 1998. The Geriatric Depression Scale: A comprehensive review. *Journal of Clinical Geropsychology.* (4), pp.89-110.

Stocking, M.L. and Lord, F.M. 1983. Developing a common metric in item response theory. *Applied psychological measurement.* **7**(2), pp.201-210.

Streiner, D.L. et al. 2014. *Health measurement scales: a practical guide to their development and use.* USA: Oxford university press.

Sucharew, H. et al. 2013. Profiles of the National Institutes of Health Stroke Scale items as a predictor of patient outcome. *Stroke.* **44**(8), pp.2182-2187.

Suhr, D.D. 2006. *Exploratory or confirmatory factor analysis?* Cary: SAS Institute

Sullivan, P.W. and Ghushchyan, V. 2006. Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. *Medical Decision Making.* **26**(4), pp.401-409.

Susanti, Y. and Pratiwi, H. 2014. M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION. *International Journal of Pure and Applied Mathematics.* **91**(3), pp.349-360.

Sutton, A.J. et al. 2003. Methods for meta-analysis in medical research. *Statistics in Medicine.* **22**(19), pp.3111-3114.

Sveen, U. et al. 2004. Well-being and instrumental activities of daily living after stroke. *Clinical rehabilitation.* **18**(3), pp.267-274.

Swenson, J.R. and Clinch, J.J. 2000. Assessment of quality of life in patients with cardiac disease: the role of psychosomatic medicine. *Journal of psychosomatic research.* **48**(4), pp.405-415.

Tabachnick, B. and Fidell, L. 2007. Multivariate analysis of variance and covariance. *Using multivariate statistics.* **3**, pp.402-407.

Teale, E.A. 2011. *Development of a minimum stroke dataset for electronic collection in routine stroke care*. thesis, University of Leeds.

Teale, E.A. et al. 2012. A systematic review of case-mix adjustment models for stroke. *Clinical rehabilitation.* **26**(9), pp.771-786.

Thompson, A. 2009. Thinking big: large-scale collaborative research in observational epidemiology. *European journal of epidemiology.* **24**(12), pp.727-731.

Thompson, B. 2004. *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC, US: American Psychological Association.

Tilling, K. et al. 2001. Multilevel growth curve models with covariate effects: application to recovery after stroke. *Statistics in medicine.* **20**(5), pp.685-704.

Tobin, C. et al. 2008. Health-related quality of life of stroke survivors attending the volunteer stroke scheme. *Irish journal of medical science.* **177**(1), pp.43-47.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society.* **26**(1), pp.24-36.

Tombaugh, T.N. and McIntyre, N.J. 1992. The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society.* **40**(9), pp.922-935.

Torrance, G.W. 1986. Measurement of health state utilities for economic appraisal: a review. *Journal of health economics.* **5**(1), pp.1-30.

Toschke, A. et al. 2010. Patient-specific recovery patterns over time measured by dependence in activities of daily living after stroke and post-stroke care: The South London Stroke Register (SLSR). *European Journal of Neurology*. **17**(2), pp.219-225.

Townsend, N. et al. 2012. *Coronary heart disease statistics. A compendium of health statistics.* London, UK: British Hearth Foundation.

Trialists'Collaboration, S.U. 1997. Collaborative systematic review of the randomised trials of organised inpatient (stroke unit) care after stroke. *Bmj*. **314**(7088), pp.1151-1159.

Trialists'Collaboration, S.U. 2001. Organised inpatient (stroke unit) care for stroke. *Cochrane database of systematic reviews.* [Online]. **3**. [Accessed 10 March 2014]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000197.pub2/epdf

Tucker, L.R. and Lewis, C. 1973. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika.* **38**(1), pp.1-10.

Tuokko, H. and Woodward, T.S. 1996. Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology.* **18**(4), pp.479-616.

Twisk, J. and de Vente, W. 2002. Attrition in longitudinal studies: how to deal with missing data. *Journal of clinical epidemiology.* **55**(4), pp.329-337.

Van Buuren, S. et al. 2005. Improving comparability of existing data by response conversion. *Journal of Official Statistics.* **21**(1), pp.53-72.

Van der Leeden, R. and Busing, F.M. 1994. *First iteration versus final IGLS/RIGLS estimators in two-level models: A Monte Carlo study with ML3.* Department of Psychology, University of Leiden.

Van Der Linden, W.J. and Hambleton, R.K. 1997. *Item response theory: Brief history, common models, and extensions.* New York: Springer Science & Business Media.

Van der Putten, J. et al. 1999. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. *Journal of Neurology, Neurosurgery & Psychiatry*. **66**(4), pp.480-484.

Vandecasteele, L. 2010. Poverty trajectories after risky life course events in different European welfare regimes. *European Societies.* **12**(2), pp.257-278.

Vandenberg, R.J. and Lance, C.E. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods.* **3**(1), pp.4-70.

Veerbeek, J.M. et al. 2011. Early Prediction of Outcome of Activities of Daily Living After Stroke A Systematic Review. *Stroke.* **42**(5), pp.1482-1488.

Veerbeek, J.M. et al. 2014. What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. *PLoS One.* **9**(2), pe87987.

Velozo, C.A. et al. 2007. Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. *Journal of rehabilitation research and development.* **44**(3), p467.

Verma, V. et al. 2009. *On pooling of data and measures.* Università di Siena, Dipartimento di metodi quantitativi.

Vermunt, J.K. 2003. Multilevel latent class models. *Sociological methodology.* **33**(1), pp.213-239.

Vermunt, J.K. and Magidson, J. 2004. Latent class analysis. *Encyclopedia of social sciences research methods.* London: Sage, pp.549-553.

Von Elm, E. et al. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Preventive medicine.* **45**(4), pp.247-251.

Wade, D.T. et al. 1987. Depressed mood after stroke. A community study of its frequency. *The British Journal of Psychiatry.* **151**(2), pp.200-205.

Wailoo, A. et al. 2014. Modelling the relationship between the WOMAC osteoarthritis index and EQ-5D. *Health and quality of life outcomes.* **12**(1), pp.1-6.

Wang, J. and Wang, X. 2012. *Structural equation modeling: Applications using Mplus.* Chichester: John Wiley & Sons.

Ware Jr, J.E. and Sherbourne, C.D. 1992. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical care.* **30**(6), pp.473-483.

Wei, N. et al. 2015. Post-stroke depression and lesion location: a systematic review. *Journal of neurology.* **262**(1), pp.81-90.

Werneke, U. et al. 2000. The stability of the factor structure of the General Health Questionnaire. *Psychological medicine.* **30**(04), pp.823-829.

West, R. et al. 2010. Psychological disorders after stroke are an important influence on functional outcomes a prospective cohort study. *Stroke.* **41**(8), pp.1723-1727.

White, I.R. et al. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine.* **30**(4), pp.377-399.

White, J.H. et al. 2008. Exploring poststroke mood changes in community-dwelling stroke survivors: a qualitative study. *Archives of physical medicine and rehabilitation.* **89**(9), pp.1701-1707.

Williams, B. et al. 2012. Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine.* **8**(3), p1.

Willmott, C.J. and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research.* **30**(1), p79.

Wing, J. et al. 1974. The measurement and classification of psychiatric symptoms: an instruction manual for the Present State Examination and CATEGO programme. *WHO, Geneva.*

Wright, B.D. and Bell, S.R. 1984. Item banks: What, why, how. *Journal of Educational Measurement.* **21**(4), pp.331-345.

Wurpts, I.C. and Geiser, C. 2014. Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in psychology.* **5**, p920.

Yesavage, J.A. et al. 1983. Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research.* **17**(1), pp.37-49.

Yohai, V.J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics.* **15**(2), pp.642-656.

Yohannes, A.M. et al. 1998. A comparison of the Barthel index and Nottingham extended activities of daily living scale in the assessment of disability in chronic airflow limitation in old age. *Age and ageing.* **27**(3), pp.369-374.

Yong, A.G. and Pearce, S. 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology.* **9**(2), pp.79-94.

Yoshida, K. et al. 2013. Use of data from multiple registries in studying biologic discontinuation: challenges and opportunities. *Clin Exp Rheumatol.* **31**(4 Suppl 78), pp.S28-32.

Young, J. and Forster, A. 2007. Review of stroke rehabilitation. *BMJ.* **334**(7584), pp.86-90.

Žikić, T.R. et al. 2014. The effect of post stroke depression on functional outcome and quality of life. *Acta Clin Croat.* **53**(3), pp.294-301.