Making what counts be counted: evaluating the use of preference-based outcome measures in Parkinson’s disease

Yiqiao Xin

MSc Health Technology Assessment and Management, BSc Pharmacy

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Health Economics and Health Technology Assessment (HEHTA)

Institute of Health and Wellbeing

College of Medical, Veterinary and Life Science

University of Glasgow

July 2018

Abstract

Parkinson's is a common neurodegenerative disorder that can have a significant impact on an individual's health, quality of life (QoL), and wellbeing, causing substantial economic burden on patients, their caregivers, the health service, and broader social and community services. Whilst Parkinson's wide range of QoL and financial impacts have been well documented relatively little research has explored to what extent such impacts have been appropriately incorporated into economic evaluations.

Economic evaluation is used by the National Institute for Health and Care Excellence (NICE) in the UK to guide health-care resource allocation in the NHS. It uses preference-based outcome measures to measure and value the health outcomes of different interventions. These health utilities are combined with durations to estimate quality-adjusted life-years. The important role of the preference-based outcomes requires them to be accurately capturing the benefit of interventions, otherwise the estimation of cost-effectiveness of interventions will be not be reflecting true preference/choice between interventions. This may lead to mistakes in funding decisions and insufficient allocation of resources.

Despite the importance of accurately capturing the benefit of interventions, the existing generic preference-based measures (e.g., the EQ-5D measure as recommended by NICE) are sometimes criticised for their 'health-related' nature as insufficient to capture all the QoL aspects that are affected by the disease or the intervention. This raises a question of "is the generic 'prescribed' measure appropriate for all disease areas and all interventions?" For diseases that have broad impact on people's health and wellbeing such as Parkinson's, a broadly scoped preference-based wellbeing instrument which could measure the impact of intervention beyond health may potentially fill the gap of the limited scope of the 'health-related' preference-based measures (if any). Meanwhile, there are concerns relating to their relevance and sensitivity to specific health aspects and their validity in general to be used in the healthcare context. Given this, the aim of this thesis is to examine the performance of the existing preference-based outcomes in people with Parkinson's, and evaluate the potential of using a generic preference-based capability-wellbeing measure, ICECAP-O, in this population.

This thesis conducted a systematic review of the existing preference-based measures to assess their construct validity and responsiveness in people with Parkinson's. Two empirical studies explored these properties of the ICECAP-O measure in people with Parkinson's. Construct validity and responsiveness are the two key psychometric properties relevant to preference-based measures for their use in economic evaluations. Data for both empirical analyses were obtained from the PD MED large-scale randomised controlled trial.

This thesis has identified evidence of limited responsiveness of the existing preference-based measures in people with Parkinson's and suggested that the current commonly used preference-based health-related QoL measures may underestimate the value placed on the mental and social wellbeing aspects that Parkinson's populations are affected by. This limited ability of the utility values to differentiate health states may have an impact on resource allocation decisions. Especially as this relates to the cost-effectiveness of interventions that have the capacity to influence the mental and social wellbeing aspects of people's lives. This highlights the need for consideration of a broadly scoped measure such as the ICECAP-O to incorporate such aspects in economic evaluations of diseases such as Parkinson's. This thesis established the construct validity and responsiveness of the ICECAP-O instrument and demonstrated that there are valued capability wellbeing attributes in Parkinson's beyond those quality of life attributes reflected by the EQ-5D instrument. It contributes to understanding the use of broadly scoped outcome measures for economic evaluations in Parkinson's by showing that the ICECAP-O capability wellbeing instrument was able to provide a preference-based assessment of these under-represented aspects in the Parkinson's population, without compromising its sensitivity to the clinical and specific physical QoL dimensions in this patient group. While further exploration of the role of ICECAP-O in economic evaluation and decision making through the work of assessing 'sufficient capability' is required, this thesis establishes initial foundations for the use of the ICECAP-O as a preference-based instrument to measure the impact of interventions in Parkinson's populations.

Table of Contents

List of Tables

List of Figures

Acknowledgement

This Ph.D. journey has been a rewarding and enriching adventure in my life, full of ups and downs, pride and tears. With the utmost sincerity, I wish to thank the many people who have seen me through this journey.

First and foremost, I would like to express my earnest gratitude to my supervisors Professor Emma McIntosh and Dr. Jim Lewsey for their continued and timely support, insightful advice, immense knowledge, patience and motivation throughout the process of my Ph.D. pursuit. As both my line manager for projects and PhD supervisor, besides her invaluable academic guidance, Emma is always being so understanding, listening to my needs, looking after me, and ‘protecting’ my time for PhD as much as she can. In particular, I am extremely grateful for all the time she has taken (in between her millions of projects) to improve my academic writing and correct my English, and also, after all that extra tedious work, she would tell me “Your English writing is good. I’m not worrying about it!” to protect my confidence. I cannot imagine a more patient, understanding and kind supervisor like her.

I am extremely grateful for Jim’s statistical advice that I benefited immeasurably from our discussions. Jim is like the ‘the needle that pacifies the oceans’ (‘Ding hai shen zhen’, which may sound weird, but it is a huge compliment) in Chinese idioms from the ‘Monkey King’ legend as he is always so calm, wise and incisive when I am panicking about statistics. Also, Jim’s encouragement and the positive feedback at the writing up stage of this thesis are greatly appreciated.

I would like to thank Professor Olivia Wu for opening HEHTA’s door to me five years ago when I was still a master student in Canada knowing almost nothing looking for initial opportunities of research in HTA. I also would like to thank Olivia, Professor Elizabeth Fenwick and Professor Andrew Briggs for providing me the opportunity to pursue my PhD in such a world leading, dynamic and friendly research group. An essential contribution was given by my group HEHTA, that funded this PhD.

I addition, I want to emphasize my gratitude to Emma and Olivia for their emotional support and allowing me the time to write up the thesis at HEHTA. They have been generous and inspirational mentors for me. I wouldn’t forget about

those long meetings or long chats with the burst of all the emotions at the toughest moments of my PhD process when Emma or Olivia was ‘rescuing’ my falling hope and help me re-establish my confidence.

I would also like to acknowledge the PD MED trial team, Professor Carl Clarke at the University of Birmingham, Professor Richard Gray, Professor Alastair Gray and Professor Crispin Jenkinson at the University of Oxford for providing me their expert opinions on matters regarding Parkinson’s, statistics, and outcome measurement, as well as their encouragement. Special thanks to Professor Joanna Coast at University of Bristol for her invaluable advice and feedback on the analysis of ICECAP in this thesis. Also special thanks to Dr. Philip Kinghorn at University of Birmingham for his feedback and support on this work during the Health Economics Study Group conference. I also want to thank Ms. Smitaa Patel and Dr. Caroline Rick at the University of Birmingham Clinical Trials Unit (BCTU) for provision of the PD MED data for this analysis.

HEHTA is a wonderful place to work in with the most friendly, supportive and international environment provided by my lovely and fun academic colleagues and most amazing admin team here and at the University. Thank you so much for all the get-togethers where we share our excitement, confusions, anxiety, stress and happiness at work and life. We shake the blues off, and head forward together.

Another special thanks must go to my wonderful Chinese PhD friends for their ‘family like’ support, which I treat as a remedy for my homesickness. I spent the most relaxing time with you guys and there are too many unforgettable moments through every week’s badminton and dinner, trips, and all the fun in the past four years.

I am indebted to my parents for their selfless and unconditional love. Whenever I meet obstacles, I know you are always behind me and you are the strongest back of me. Also, to my two grandmas, I wish you both know you have my deepest love.

Finally, I would like to thank Ewan, as a fellow health economist, for proof reading/peer reviewing my thesis, with the risk of ‘challenging’ me as my partner. More importantly, thank you for your love, caring, patience with my swinging mood close to the finish of my thesis, and for all the inspiring academic chats. You are the best person I’ve ever known. Thank you truly for everything.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signed:

Printed name: Yiqiao Xin

Publications, Working Papers and Presentations

The following publications, working papers and presentations were developed as part of this thesis:

Publications:

Xin Y & McIntosh E. (2017) Assessment of the construct validity and responsiveness of preference-based quality of life measures in people with Parkinson's: a systematic review. Quality of Life Research, 26(1): 1-23. (doi:10.1007/s11136-016-1428-x)

Working papers (and being drafted for submission):

Xin Y, Lewsey J, Gray R, Clarke C, Coast J, Rick C and McIntosh E. Testing the responsiveness of ICECAP-O in People with Parkinson's and a comparison with EQ-5D-3L and a Parkinson's specific quality of life measure PDQ-39. Health Economists' Study Group, Aberdeen, June 2017.

Xin Y, Lewsey J, Gray R, Clarke C, Coast J, Rick C and McIntosh E. Testing the construct validity of the ICECAP-O instrument in Parkinson's and exploring its relationship with the EQ-5D-3L and the Parkinson's specific quality of life questionnaire the PDQ-39. Health Economists' Study Group, Gran Canaria, June 2016.

Conference presentations:

[Published abstract] Xin Y, Lewsey J, Gray R, Clarke C, Coast J, Rick C and McIntosh E (2017). Too broad to be sensitive? exploring the responsiveness of the ICECAP-O capability wellbeing measure compared to the EQ-5D-3L to the change of clinical and quality of life aspects in People with Parkinson's? International Society Pharmacoeconomics and Outcome Research, Glasgow, November 2017. Value in Health, 20(9): A763.(doi: http://dx.doi.org/10.1016/j.jval.2017.08.2165)

[Published abstract] Xin Y, Lewsey J, Gray R, Clarke C.E., McIntosh E (2016). Broadening the evaluative scope of quality of life in Parkinson's: Testing the construct validity of the ICECAP-O instrument. International congress of Parkinson's disease and movement disorders. Berlin, Germany. June 2016. Mov Disord. 2016; 31 (suppl 2).

Xin Y & McIntosh E. What matters to people with Parkinson's? A systematic review of preference-based measures used in Parkinson's. Health Economists' Study Group, Glasgow, June 2014.

Abbreviations

ADL	Activities of daily living
AQoL	Assessment of Quality of Life
ASCOT	Adult Social Care Outcome Toolkit
AUD	Australian dollars
CBA	cost-benefit analysis
CEA	cost-effectiveness analysis
CS-PBM	condition specific preference-based measure
CUA	cost-utility analysis
DALY	disability-adjusted life year
DBS	deep brain stimulation
DCE	discrete choice experiments
DDI	Disability and Distress Index
EQ-5D	EuroQol 5 Dimensions
EQ-5D-3L	EuroQol 5 Dimensions – 3 Levels
EQ-5D-5L	EuroQol 5 Dimensions – 5 Levels
ES	effect size
SRM	standardised response mean
HrQoL	health-related quality of life
HTA	health technology assessment
HUI	Health Utilities Index
H&Y	Hoehn and Yahr scale
ICECAP-A	Investigating Choice Experiments for the Preferences of Adult
ICECAP-O	Investigating Choice Experiments for the Preferences of Older people
ICER	incremental cost-effectiveness ratio
MID	minimally important difference
MCID	minimally clinically important difference
MAR	missing at random
MCAR	missing completely at random
MCDA	multi-criteria decisions analysis
MI	multiple imputation
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
MNAR	missing not at random
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PbQoL	preference-based quality of life
PD	Parkinson's disease
PDQ-39	Parkinson's Disease Questionnaire – 39 items
PDQ-39-SI	Parkinson's Disease Questionnaire – 39 items Summary Index
PDQL	Parkinson's Disease Quality of Life questionnaire
PDQUALIF	Parkinson's Disease QUAlity of LIFe scale
PIMS	Parkinson's IMpact Scale
PwP	people with Parkinson's
QALY	quality-adjusted life-year
QoL	quality of life
RCT	randomised controlled trial

SF-6D	Short Form – 6 Dimensions
SD	standard deviation
SG	standard gamble
SRM	standardised response mean
TTO	time trade-off
UK	United Kingdom
UPDRS	Unified Parkinson’s Disease Rating Scale
USD	US dollars
VAS	visual analogue scale
WHO	World Health Organization
WTP	willingness-to-pay

Chapter 1 Context and rationale

1.1 Introduction

"Not everything that can be counted counts, and not everything that counts can be counted." (1)

William Bruce Cameron (not Albert Einstein) (2), sociologist, 1963

Parkinson's Disease (or Parkinson's [1]) is the second most common neurodegenerative disorder after Alzheimer's disease (3). It has a wide range of motor and non-motor symptoms which can have significant impact on patients' health, quality of life (QoL), and wellbeing (4-7) . Due to the life-changing symptoms, unclear mechanisms and the chronic progressive nature of the disease, management of Parkinson's is not merely complicated and difficult, but also costly. The cost of illness escalates as Parkinson's progresses, placing an increasing economic burden on the healthcare system, society and patients themselves (8-10).

With limited health care budget, in the UK, the health technology assessment (HTA) agencies, National Institute for Health and Care Excellence (NICE) and the Scottish Medicine Consortium (SMC) use economic evaluation methods to make recommendations to the National Health System (NHS) in England and Scotland respectively regarding resource allocation across and within budgets and judge whether an intervention is value for money (11, 12). In this process, to be able to compare the 'value' of interventions for priority setting purpose, a generic health-related quality of life (HrQoL) outcome is recommended by NICE/SMC for measuring the benefit of interventions across and within disease areas.

NICE defines HrQoL as 'a combination of a person's physical, mental and social wellbeing; not merely the absence of disease' (13). Nevertheless, many HrQoL measures, including the one that recommended by NICE (i.e. EQ-5D (14)) (15),

[1] Parkinson's UK recommends researchers to refer PD by 'Parkinson's' only and therefore 'Parkinson's' as a term to refer to PD is used throughout in this thesis.

actually measure 'self-perceived health status' and not broader 'wellbeing' (16). Wellbeing relates to the 'presence of positive emotions and moods (e.g., contentment, happiness), the absence of negative emotions (e.g., depression, anxiety), satisfaction with life, fulfilment and positive functioning' [2] (17).

Because of the narrowly implemented scope of 'HrQoL' in practice, a growing number of concerns are raised from patients, researchers and clinicians regarding whether such a generic health-related measure is capable of reflecting the value of improvements in health and wellbeing by various interventions in economic evaluation (18-23). If important domains are missing from the health-related generic measures the benefit of the intervention cannot be property measured. This would lead to cost-effectiveness estimates which may be inaccurate and consequently to errors in funding decisions. The resulting inefficient allocation of resources negatively impacts patient's health and wellbeing and society overall. Parkinson's is one particular area of concern (18).

This chapter will firstly introduce Parkinson's, its symptoms and their mechanisms, and the impact on QoL and overall wellbeing. Treatment options will be described. The economic burden caused by the disease to patients, their families and the NHS system will be reviewed. This provides a background to the next section; general principles of priority setting, market failure, and the role of economic evaluation in decision making. Whilst cost is tangible to measure, many challenges and issues are raised in outcome measurement for economic evaluation in the Parkinson's population, which provides the rationale of this thesis. The chapter concludes with the aims, research questions and structure of this thesis.

1.2 Parkinson's disease

1.2.1 Prevalence and mechanism

"With Parkinson's, it's like you're in the middle of the street and you're stuck there in cement shoes and you know a bus is coming at you, but you don't know when. You think you can hear it rumbling, but you have a lot of time to think. And so you

[2] These terms will be discussed in-depth in Chapter 2 (Section 2.3 for 'QoL' and 2.6.4 for 'wellbeing').

just don't live that moment of the bus hitting you until it happens. There's all kinds of room in that space." (24)

Michael J. Fox, Canadian-American actor

"Parkinson's is my toughest fight… it doesn't hurt. It's hard to explain." (24)

Muhammad Ali, former boxing champion

"I have a form of Parkinson's disease, which I don't like. My legs don't move when my brain tells them to. It's very frustrating." (25)

George H. W. Bush, 41th Present of the United States (1989-1993), 2012

Parkinson's is a progressive neurodegenerative condition resulting from the death of dopamine-containing cells of the substantia nigra in the brain. The worldwide prevalence of Parkinson's increases with age; a recent meta-analysis showed that 41 per 100,000 population aged 40-50 years have Parkinson's, and this number increases to 1,903 in the age group >80 years (26). Parkinson's was firstly identified as a condition by an English doctor, James Parkinson, in a monograph entitled 'An Essay on the Shaking Palsy' published in 1817 (27). It described the characteristics of Parkinson's by detailing the six patients with '*involuntary tremulous motion with lessened muscular power, in parts not in action even when supported, with a propensity to bend the trunk forward and to pass from a walking to a running pace*' (James Parkinson 1817) (p223)(27). Dopamine transmits signals between areas in the brain, and lack of dopamine causes the signal in some areas of the brain to not be transmitted properly. The most commonly affected part of the brain is the section responsible for controlling body balance and muscle movement, and thus the motor aspects of Parkinson's, such as akinesia, bradykinesia, tremor, rigidity, and postural imbalance, are the defining characteristics of the disease (28). These motor symptoms along with the non-motor symptoms are introduced in detail in Section 1.2.2.

Apart from motor symptoms, patients with Parkinson's suffer from a wide range of non-motor symptoms, such as depression, sleep problems, and bladder and bowel problems, which significantly affect their QoL. The neuropathological basis of non-motor Parkinson's is less clear than the motor symptoms. A few studies suggest that the non-motor symptoms are associated with dysfunction in non-dopaminergic processes, which may originate from the degeneration of the dopaminergic process which gradually damages other brain sections, such as the lower brain stem that affects autonomic functions and sleep (7, 29). Jellinger KA, a well-known neuropathologist, stated in a paper published in Movement Disorder in 2012 that “Parkinson’s disease….is no longer considered a complex motor disorder characterised by extrapyramidal symptoms [3] , but a progressive multisystem or-more correctly-multiorgan disease with variegated neurological and nonmotor deficiencies.” (31) This signals that the traditionally developed interventions targeting at dopaminergic process may not be effective in controlling the non-motor symptoms and these symptoms require greater attention from clinicians and researchers.

1.2.2 Symptoms

1.2.2.1 Motor symptoms

Parkinson’s is characterized by varied motor and non-motor features. Motor symptoms are the symptoms that are related to movement, the core features of which are tremor, bradykinesia and rigidity. A resting tremor is most often recognized by patients and caregivers (32-34), which occurs in approximately two thirds of PD patients (6). It can be present in the hand (pill-rolling tremor), lower limbs, toes, and jaws. The tremor can be exacerbated in stressful situations or when the patients were asked to perform a mental task (6). Compared to tremor, the other two motor features, rigidity and Bradykinesia, are considered to be more disabling (35, 36). Bradykinesia is presented as slowness of movement in the speed, gait and amplitude of a repetitive action involving voluntary movements (37, 38). Patients with bradykinesia may also demonstrate shuffling when walking, dragging one or both feet when walking, or freezing as muscle reactions may slow to the point that the muscles become immobile. Patients may also present hastening of their gait, which is described as “their walking speed increases with small, rapid steps in an effort to ‘catch up’ with their displaced center of gravity” (6, 35, 38-

[3] Symptoms that are related to biological neural network that is part of the motor system causing involuntary actions (30).

40). The third main motor feature is rigidity, which is presented as increased muscle tone or amplified resistance to a passive range of motion (6). Rigidity and bradykinesia may affect other part of the body, among which the most noticeable ones are the facial movement, which can display a 'masked' expression, as well as speaking, as patients lose their ability to speak clearly (41).

1.2.2.2 Non-motor symptoms

Besides motor symptoms, Parkinson's is associated with a broad spectrum of non-motor symptoms, which also have a substantial influence on patients' QoL (42-44). Sometimes these non-motor symptoms are more bothersome than motor symptoms, as quoted in a paper said by a patient: "I have Parkinson's. I would like you to address the following symptoms that bother me the most: sleep, pain and then my movement disorder." (7)

Non-motor symptoms can be classified into two groups based on the manifestation: physical non-motor symptoms (in contrast to the motor symptoms which affect patients' movement), and neuropsychiatric symptoms. The former includes swallow and saliva control, speech and communication issues, bladder and bowel problems, disturbances of sleep-wake cycle regulation, fatigue, dizziness, muscle cramps and dystonia, low blood pressure, sexual dysfunction etc, and the latter includes disorders of mood / apathy, depression, cognitive dysfunction, hallucination, etc (45, 46). In the Non-Motor Symptoms Scale (NMSS) developed by the Movement Disorder Society in 2006, nine domains of non-motor symptoms were identified from patients, clinicians and experts, and they are: cardiovascular, sleep/fatigue, mood/apathy, perceptual problems; attention/memory, gastrointestinal, urinary, sexual function and miscellaneous (47).

While the physical non-motor symptoms are tangible, mixed views exist for the aetiology of neuropsychiatric symptoms in Parkinson's (48-50). Whilst depression may occur reactively as a consequence of the deteriorating physical symptoms, there is consensus that these neuropsychiatric symptoms are also directly linked with the neurobiology of the illness (50-52). Pathophysiology studies suggested that the link is complex, which probably involves dopaminergic dysfunction, change of cerebrospinal fluid levels of neurotransmitter metabolites, and noradrenergic structure change (50-52). These findings provide evidence that the neuropsychiatric symptoms in people with Parkinson’s are caused by not merely the psychosocial stress of the physical disabling impact of this chronic disease, but also the intrinsic link to the dysfunction of the brain caused by the change of the neurobiological environment. Consequently, this suggests that interventions that aim to manage the physical aspects can only partly relieve the neuropsychiatric symptoms and measures need to be taken focusing on these invisible non-motor symptoms. Furthermore, owing to the connection between ‘mind’ and ‘body’ (53, 54), studies have shown that patients with major depression left untreated had faster progression of Parkinson’s: earlier initiation of dopaminergic therapy (55), greater cognitive decline (56), greater deterioration in activities of daily living (ADL) and motor complications (55, 56), and increased mortality (57). Treating non-motor symptoms is therefore of great importance to improve patients’ QoL and, thereby, measures that are capable of sufficiently reflecting the benefit of treatment are required.

Among all the motor and non-motor symptoms, depression is the most frequently identified symptom that leads to decrease of patients’ QoL, as reported in a study which summarized sixteen studies that assessing factors influencing HrQoL in people with Parkinson’s (5). A meta-analysis pooled the prevalence of depressive disorders in Parkinson’s from 36 studies and found that 17% of patients experienced major depressive disturbance, 22% had minor depression, 13% had dysthymia, and 35% of the patients presented clinically significant depressive symptoms (58). Apart from depression, the other factors that found to be significantly associated with patient’s QoL, are disease severity, anxiety, mood disorders, postural instability, insomnia, apathy, psychosis and cognitive impairment (5, 50, 59).

1.2.3 QoL and broader wellbeing in Parkinson’s

Parkinson’s is not immediately life-threatening, but, it is life-disabling. The distinguishing feature of Parkinson’s from other common chronic conditions is its wide range of symptoms, with each potentially affecting patients’ health, QoL and their overall wellbeing. These impacts are not simply tremor, involuntary movement, speech and language problems and depression, but also relate to patients’ self-perception, family relationships and social functioning. Parkinson’s had the largest impact on QoL as shown with the lowest scores on EQ-5D-3L and 15D (a PbQoL measure developed in Finland (60)) among 29 conditions, including heart failure, stroke, cancer, diabetes (4).

Individuals with Parkinson's may lose trust in their body as it can be unpredictable and uncontrollable (61, 62) and lose hope in their future with the worsening symptoms over time (63). It also can affect family relationships; the mutuality, the positive quality of the relationship as perceived by the caregiver, was shown to markedly decline in the advanced stages of the disease (64, 65). Many patients and their spouse caregivers experience stigma, feelings of shame or embarrassment about their conditions (66, 67) and fear about their future due to the increasing physical, emotional and financial burdens (10) coping with the disease (63, 68). In terms of social functioning, social isolation and degradation of social interactions were also found to be a common problem in people with Parkinson's (69), which was found to be caused by speech (70, 71) and functional communication impairment (72), progressive physical disability, mood disturbances, shrinking of social activities and secluding oneself (73). Parkinson's is a chronic progressive disease without a cure and as such the battle against Parkinson's is a long tough journey, whereby interventions to improve a patients' attitude towards the disease, their life, their general wellbeing, and even their carer's quality of life should be considered with as great importance as direct symptom-relieving interventions.

1.2.4 Management of Parkinson's

Without a cure, the goal of interventions in Parkinson's is to control the various negative impacts of its symptoms on QoL. However, the wide spectrum of its symptoms necessitates complexity of treatment. A broad range of interventions have been developed addressing different areas of symptoms, e.g. deep brain stimulation (DBS) surgery, dopamine agonists, levodopa and MAO-B targeting at motor symptoms (74, 75), anti-depressants for depression problems, modafinil for daytime sleepiness, physiotherapy to prevent falls, language and speech therapist to help patients who are experiencing problems with communications, swallowing or saliva (76). NICE recently updated its guidance on the management of Parkinson's (July 2017), where fifteen ways of managing symptoms of Parkinson's are outlined, with each component focusing on a specific area of symptoms (76). Based on the type of intervention and the target type of symptoms, these are categorised as pharmacological management of motor symptoms, pharmacological management of non-motor symptoms, pharmacological neuroprotective therapy, non-pharmacological management of motor and non-motor symptoms, DBS and levodopa-carbidopa intestinal gel, and palliative care (76).

1.2.4.1 Management of motor symptoms

In the NICE guidance, Levodopa remains the preferred first line medicine for people with troublesome motor symptoms (76). Although many motor symptoms can be initially controlled by dopaminergic drugs, over time levodopa induced side effects will start to develop, adding more complexity to the management of the disease. The traditionally established side effects include motor fluctuations, dyskinesia (involuntary movement), hallucinations and delusions (77). Motor fluctuations oscillate between “off” times, a state of decreased mobility due to losing response to medications, and “on” times, periods when the medication is working and symptoms are well controlled (78). These complications are observed in 50% of patients after five years of treatment and in 80% of patients after ten years (79).

Besides the traditional known side-effects, there has been a rising interest in the dopaminergic medication-related impulse control disorders in recent years (80) and it has been newly added to NICE’s updated guideline in 2017 as a recognised adverse effect of dopaminergic therapy (76). A large cross-sectional and case-control multicentre study in 2010 showed that impulse control disorders were observed in 13.6% of its participants (problem and pathological gambling 5.0%, compulsive sexual behaviour 3.5%, compulsive buying in 5.7%, and binge eating disorder 4.3%) (81). Treated Parkinson’s patients were found to be 25 times more likely to have pathological gambling than general hospital controls (82) and this difference was not observed in untreated patients (83). In addition, the effect of medications can often ’wear off‘ as disease progresses, which means the effect of a given dose is not maintained as long as it is supposed to (84). Increasing daily dose would also deteriorate the medication related side effects, leading to additional negative impacts on patients’ life. Due to side effects and ‘wearing off’ of dopaminergic drugs, other drugs are recommended as adjuvant treatment for motor symptoms. This includes dopamine agonists, monoamine oxidase B inhibitors, or catechol-O-methy transferase inhibitors.

In people with advanced Parkinson's, when symptoms are not controlled with best medical therapy, DBS (a type of surgical implant in brain) is recommended (76) to control the motor symptoms. DBS involves implanting a stimulation device in the patient's brain which delivers high frequency electrical stimulation to the targeted area in the brain (85). This stimulation changes some of the electrical signals in the brain that are disrupted due to the lack of neurotransmitter, dopamine. Studies have shown that DBS led to significant reduction of dyskinesias and dopaminergic medication, improvement of all cardinal motor symptoms with sustained long-term benefits, and significant improvement of QoL when compared with best medical treatment (74, 86, 87). On the other hand, some patients had operation related or stimulation induced side effects including intracerebral haemorrhage (85), visual phosphenes, nausea, dyskinesia, and dystonia (86, 88-90). As DBS targets the same mechanism as levodopa, its effectiveness is only clear in improving the dopaminergic induced motor symptoms while the impact on the non-motor symptoms remains to be elucidated (91).

1.2.4.2 Management of non-motor symptoms

Managing non-motor symptoms sometimes may be more complicated than the motor symptoms. This is not only due to the unclear pathology, the wide spectrum involved with multi organs and multi body systems, but also due to underreporting caused by the lack of awareness of the link with Parkinson's (7). Since Parkinson's has historically been recognised as a primary movement disorder, non-motor symptoms are frequently overlooked by clinicians and undertreated (7, 92). In 2010, an international survey showed that up to 62% of patients with Parkinson's do not declare symptoms such as apathy, sexual difficulty, bowel incontinence or sleep disorder, either due to embarrassment or lack of awareness of the link of these symptoms to their Parkinson's, leaving their non-motor symptoms untreated (93). As a result, NICE recent updated guideline (2017) emphasised the importance of management of non-motor symptoms (76). An increased number of studies in recent years regarding the significant impact of non-motor symptoms on patients' QoL has also been observed from the literature (7, 29, 42-44, 46, 94, 95).

To manage non-motor symptoms, a variety of pharmacological and non-pharmacological interventions are recommended depending on the specific non-motor symptom. For example, modafinil is recommended to treat excessive daytime sleepiness, levodopa or oral dopamine agonists to treat nocturnal akinesia, quetiapine to treat hallucinations and delusions in patients who have no cognitive impairment, offering a cholinesterase inhibitor for people with mild or moderate Parkinson’s disease dementia, glycopyrronium bromide to manage drooling of saliva, and others. (76).

1.2.4.3 Specialist care for motor and non-motor symptoms

Interventions involving specialist care are also recommended to control motor as well as non-motor symptoms. These include: Parkinson’s nurse specialist interventions to enable a clinical monitoring, medicine adjustment and a continuing point of contact for support; physiotherapy and physical activity intervention for patients experiencing balance or motor function problems; considering occupational therapy for those having difficulties with ADL; and speech and language therapy for those experiencing problems with communication, swallowing or saliva.

1.2.5 The economics of Parkinson’s

Parkinson’s is a major cause of morbidity and has a substantial economic impact on patients, their caregivers, the health service, and broader social and community services. The burden of illness associated with Parkinson’s comes from two main categories of costs: (a) direct costs, where payment are directly related to the treatment of disease itself, and (b) indirect costs, for which resources are lost due to the decrease in productivity that patients experience performing their everyday life task as their disease progresses and the personal cost to patients and their carers (& for employing paid carers) due to the advancing disability (96). A recent systematic review summarized the cost of illness studies in Parkinson’s and the annual cost identified from the included studies varied by countries, methods and publication year, with the majority between £15,000 and £25,000 per year in the US and Europe (8).

Direct costs cover expenditures of the NHS, social care services and private expenditure that is directly related to the treatment of the disease (96, 97). A recent UK study (2016) (9) found that annual direct medical costs to the NHS, including both primary care and secondary care, were estimated to be £2,388 per patient, which is similar to a previous survey reporting a £2,277 direct cost to the NHS (10). The cost varies by disease severity; a previous study reported that the patients at lower severity cost £2,971 per patient per year and those at higher severity cost £18,358 per patient per year. The cost also varies by countries. An Australian study (2017) found that the mean annual cost per person to the health care system was $32,556 AUD (equivalent to £18,777) and a Chinese study estimated the direct health care related cost to be $2,503 USD (equivalent to £1,870) in China. In terms of the social care, the cost was estimated to be £2,097 in the UK on average, with a marked increasing trend with age (10). Out-of-pocket private expenses towards travel and equipment for health care was reported to be £2,229 per patient, and additional living costs such as alterations in accommodation was reported to be £3,622 per household that had a person with Parkinson’s (9).

Compared to the direct health and social care cost, the cost of work ability, productivity loss and informal care due to disability in people with Parkinson’s is enormous, regardless of what approach chosen, and which country the survey is conducted. A recently completed study commissioned by Parkinson’s UK showed that the overall annual cost added up to £20,123 per person (Figure 1-1) (9), approximately half (£10,731) of which was arising from income loss and informal care. These contain direct salary lost from work days lost (£1,981), employment earnings forgone due to early retirement or unemployment due to Parkinson’s (£6,013), unpaid caring (earning loss) (£1,235), and state pension and benefit (£1,502). Another earlier study estimated the indirect cost was over £27,000 per patient per year when the care given by a family member was replaced with a professional carer (10). Similarly, the burden to society estimated in the Australian study also exceeded the health care cost, which amounted to $45,000 (£25,954) per annum per person, including formal care with nurse and personal care assistance, and informal care and meals on wheels.

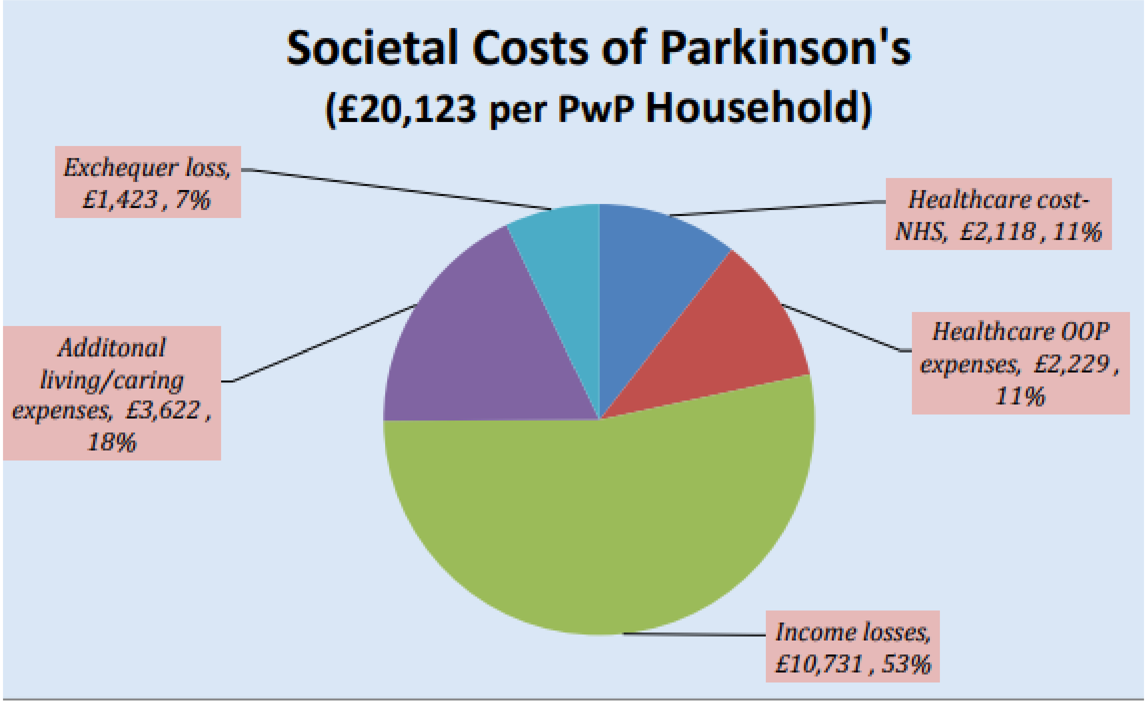


Figure 1-1: Cost of Parkinson's.

Source: Gumber A et al. Economic, social and financial cost of Parkinson's on individuals, carers and their families in the UK. Final report. Available from Sheffield Hallam University Research Archive (SHURA) 2016 (9). Re-use permission is not required according to the copyright and re-use permission information provided on SHURA website (Appendix G)

The massive indirect cost arises from the great disability caused by the symptoms of the disease and their impact on QoL. Impairment in motor functions can lead to falls and injuries, not only adding to treatment cost, but also the disability to perform work and social activities. Non-motor symptoms such as depression, urinary incontinence and cognitive decline restrict patients' independence, limit their ability to achieve their work and life roles, and lead to reliance on their carers, contributing to the psychological and economic burden of their carers and the society. As disease progresses, patients require increasing support emotionally, physically, and socially which all lead to the increasing burden on their carers (94). These demands are energy consuming, and also costly.

Once diagnosed, Parkinson's becomes a lifetime illness. As disease progresses, increasing annual cost will be cumulative until the patients' end of life. The benefit and harms of interventions on patients' symptoms and QoL may consequently translate to enormous economic impact on patients, their families and carers. Interventions that can relieve carers' emotional and physical burden, improve individual's independence, working capability and self-care ability, and improve their ability to achieve their roles, may alleviate a high proportion of cost of illness and therefore may have great value in health care decision-making.

1.3 Priority setting and economic evaluations

"We never will have all we need. Expectation will always exceed capacity… This service must always be changing, growing and improving, it must always appear inadequate." (98)

Aneuryn Bevin, Minister of Health, 1948

1.3.1 Priority setting

Resources are scarce, but demands are growing. The total healthcare expenditure in the UK has increased every year, rising from £54.9 billion in 1997 to £185.0 billion in 2015, with an annual growth rate of 8.1% between 1997 and 2009 and an average of 2.0% between 2009 and 2015 (99, 100). New health interventions (drugs, surgeries, devices, diagnostic test, preventative measures, etc.) are continually emerging which put considerable strain on the limited resources. Choices between health technologies have to be made and there is a need for healthcare decision-making and priority setting.

Making choices about allocation of health care resources implies trade-offs between the resources used to implement one intervention, and other potentially completing uses of those resources (101). The trade-offs have implications for the definition of the opportunity cost – "the opportunity cost of investing in a healthcare intervention is best measured by the health benefits (life years gained, quality-adjusted life-years (QALYs) gained) that could have been achieved had the money been spent on the next best alternative intervention or healthcare programme." (Palmer 1999) (102) The concept of opportunity cost leads to a question of how health care resources can be distributed in the most efficient, and therefore health maximising, way.

In a market system health care resources are distributed based on the choices and prevailing budget constraints of individual consumers and producers. If this is a perfectly competitive market[4] these choices would jointly lead to the most efficient, i.e. welfare maximizing, distribution of resources in health care. The market is efficient as defined by the Italian economist, Vilfredo Pareto, that an allocation of resources is efficient if it is impossible to change that allocation to make one person better off without making someone else worse off (104).

Two other types of efficiency are important when considering the distribution of resources in health care; technical/productive efficiency and allocative efficiency (105, 106). Technical / productive efficiency refers to the maximum output of production that can be generated by a given input of resources, which is close to the meaning of 'efficiency' in common English usage. Technical efficiency addresses the question of 'how to do it'. In contrast, allocative efficiency is about 'whether to do it' or 'should something else to be done instead' (107). It is achieved when it is not possible to increase the overall benefits produced by the health system by reallocating resources between interventions (107). Allocative and technical efficiency are achieved when the economy is producing exactly the quantity and type of health care that society wants and it is producing that health care for the lowest possible cost (108).

1.3.2 Market failure

In theory, a perfectly competitive market will automatically produce an equilibrium price and quantity. Whereas in reality, there are situations where a market fails. Market failure is a situation where there is an 'inefficient' allocation of resources; that is, when there exists another conceivable outcome where at least one individual may be made better-off without making someone else worse-off (109). A health care system can be a market where the patients (consumers) buy the healthcare interventions and services from the healthcare provider. In practice however, this is an example of market failure due to the special features of health and health care (110). These special features are: risk and uncertainty associated with contracting a disease, asymmetrical information between the

[4] A perfectly competitive market is when without government interference, the 'invisible hand' of the market would allocate resources optimally leading to economic efficiency (Debreu 1955) (103).

healthcare providers and the consumers, supplier-induced demand and externalities (110-112).

There is a great amount of uncertainty associated with the incidence of disease as well as progression or recovery from disease, and these uncertainties vary across individuals (110). In addition, patients may have more information about their risks and treatment than the health care providers whereas the health care providers may know more about the patient's health conditions and available interventions than the patients. This imbalance in the level of information, or 'asymmetrical information' may result in supplier-induced demand, in simpler words, over-treatment, that doctors may use a higher level of an intervention than would have been the case if the patients have had the necessary information (111). In a free insurance market, the considerable uncertainty and variability, as well as asymmetrical information, often lead to the problem of 'moral hazard' and 'adverse selection', which cause significant market failure and allocative inefficiency. Moral hazard refers to that individuals covered by insurance tend to use more health care resources and they might not take necessary precautions to stay healthy as they do not bear the cost of disease (113). This leads to a higher level of service demanded than would have been the case without the insurance cover, and thus resource allocation is inefficient (111). Another problem is adverse selection in the health insurance market leading to market failure (113). Individuals with higher risk are more likely to purchase insurance than those with lower risk while high risk individuals use more health care resources than the low risk. To compensate for higher than expected costs the insurance provider might increase the premium, which further discourages the low risk individuals from purchasing the insurance while the higher risk individuals remain, leading to bigger losses and may drive the insurance company out of the market.

Externalities are spill-over effects of consumption or production, which refers to the circumstances when the actions of one individual affects (positively or negatively) the wellbeing of another person (111) (113). In health care context, a commonly cited example is intervention such as vaccination programmes which may also protect other such as the family of the person getting immunised as they can no longer carry the infectious disease and thus there exist positive externalities. Externalities are usually not considered in decision-making in a free market since the producers of interventions are not incentivised to take into account of the effect of their actions on others (do not receive or pay compensation for these effects) (111). Ignoring externalities leads to inefficiency of resource allocation overall and market failure; the vaccination program for infectious disease may be encouraged or implemented at a higher level when the spill-over effect is considered, and vice versa (the activities that have negative externalities may be less encouraged).

Uncertainty associated with disease, asymmetry of information and externalities determine the failure of health care as a market. Market failure partly explains the emergence of non-market health care systems such as the UK NHS. In non-market systems other mechanism has to be sought to achieve technical and allocative efficiency. Economic evaluation is one such mechanism.

1.3.3 Economic evaluation

"Choices must and will be made concerning their deployment, and methods such as 'what we did last time', 'gut feelings', and even 'educated guesses' are rarely better than organized considerations of the factors involved in a decision to commit resources to one use instead of another" (114)

Michael F. Drummond, Professor of Health Economics, 2005

In a publicly funded health care system with limited budget, decisions must be made regarding the choice of which health care technology to offer. The term 'health technology' covers a range of health care interventions including devices, medicines, vaccines, procedures and systems developed to solve a health problem and improve QoL (115). In a non-market system, economic evaluation is a tool to aid prioritisation of health care technologies to make efficient and equitable decisions by comparing the costs and consequences of health technologies (97).

The last two decades have seen a rising use of economic evaluations in supporting the allocation of resources in health care agencies at national and local level worldwide. In England and Wales, since 1999, the assessment of new health technologies are conducted by NICE, through providing guidance on the effectiveness and cost-effectiveness of new and existing drugs, treatments and procedures in the NHS (116). In other countries such as Australia, New Zealand, and Scotland, economic evaluation evidence is required for all new drugs to be listed on the national drug formulary for reimbursement (116). In addition, NICE clearly recognised the importance of incorporating cost-effectiveness in its 'Methods for the development of NICE public health guidance (2012)', with the statement that 'The Public Health Advisory Committee (PHAC) are required to make decisions informed by the best evidence of both effectiveness and cost effectiveness.' (117) Choices have to be made, and economic evaluation offers a transparent process by systematically identifying, measuring, valuing and comparing the costs and benefits between different health care interventions for informing such choices.

Economic evaluation compares the costs and consequences of health technologies to support decision-making (97). This definition explicitly describes its two features. First, economic evaluation concerns itself with choices; that is, it is concerned with the incremental difference between two or more alternatives, i.e. what additional health benefit can we get for what additional cost? Second, it must be involved with both the cost and outcomes, or the input and the output when carrying out the alternative interventions. It answers the question that 'are we satisfied that the healthcare resources should be spent in this way rather than in any other way?' (98). Outcomes of interventions can be direct health benefit such as reduction of the level of disability as well as indirect benefit such as productivity and income gain, reduction of carer burden, and improvement of social and emotional wellbeing. Similar to the cost categories in the cost of illness studies mentioned earlier in Section 1.2.5, costs considered in economic evaluations may include direct medical cost (cost to the NHS and social care), direct non-medical cost (e.g. family expenditure due to the disease) and indirect cost (e.g. cost of informal care) or productivity cost (e.g. early retirement). Whilst the methods for capturing the cost of health care interventions are relatively tangible, the outcomes of health interventions however are arguably less obvious and more controversial to assess (118).

1.4 Outcome measurement

Economic evaluation aims to aid priority setting, which requires methods to facilitate comparison across different disease areas. This necessitates the need for standardisation of methods to capture the benefit of varied interventions or in populations with different conditions. As mentioned above, consequences of health care interventions can be within health (direct health benefit) and beyond health (indirect benefit), which may have considerable variations depending on the symptoms and impact of specific conditions. This raises a question as what should be ‘counted’ in the standardised measurement in economic evaluations, the keyword of this thesis.

In the UK, NICE recommends the quality-adjusted life-years (QALY) framework as the standardized approach to quantify the benefit of interventions across disease areas. It combines a person’s length of life and the person’s QoL (12, 119, 120) into a single index. Measuring length of life is straightforward however the assessment of QoL is challenging. Debates have arisen from different perspectives surrounding its identification, measurement and valuation due to the complexity of the QoL concept (this will be discussed in Chapter 2, Section 2.3). Currently, NICE recommends a generic instrument, namely EQ-5D (121), as the standardised solution to measure and value the benefit of the intervention (122). Because of this, the EQ-5D measure is considered to be the cornerstone of the QALY framework. It assesses QoL with five simple questions addressing different functioning problems, i.e. mobility, self-care, usual activities, pain/discomfort, anxiety/depression, each with three (EQ-5D-3L) or five levels (EQ-5D-5L). The EQ-5D is a ‘preference-based’ QoL (PbQoL) measure, meaning that valuation of its attributes and levels (essentially a trade-off between its states) is conducted so that its valuation system incorporates people’s preferences. The notion of QALY and ‘preference’ will be described in depth in Section 2.2.2 and 2.4. EQ-5D is preferred by NICE due to its generic nature, simplicity in its descriptive system, relevance to health-care decision makers, and the availability of large-scale validation and valuation studies (123).

Using an EQ-5D based QALY allows for cross-sector comparison. This uniform approach, not surprisingly, brings sceptical voices claiming - ‘one size does not fit all' or 'resulting in tortuous attempts to compare apples and oranges' (Robinson 1993) (124). Conditions are different, and interventions are varied in their aims, and thus the impact of different health care interventions on people's life can be varied. EQ-5D focuses on health-related QoL so its adoption beyond health interventions is argued to be limited (125). It raises the question of whether such measure is sufficient to reflect the broader impact of disease beyond narrowly defined health alone. This will be further described in depth and critically discussed in Chapter 2, Section 2.4.4.2 and Section 2.5.

The quote at the beginning of this chapter (i.e. "Not everything that can be counted counts, and not everything that counts can be counted." (1) William Bruce Cameron) has implications in the health-care context where diseases and their impact on QoL can be diverse to a substantial degree. Firstly, the aspects that are measured may not be important to patients in all disease areas; secondly, the aspects that are important may not easily be captured. Consequently, questions should be raised in each specific disease area: is this 'prescribed' measure appropriate to be used in the population with a specific disease? Are there any limitations of this measure for capturing specific benefits of different interventions? In addition, what is the desired measure for the specific disease area and how does that compare to the 'prescribed' measure? If this measure does not fit, what is the consequence to the funding decisions regarding the intervention and population affected? These questions result in a myriad of different answers and have made measuring and valuing the outcomes a substantial focus of many theoretical and empirical explorations in economic evaluations.

1.5 Rationale of this thesis

1.5.1 Assessing the use of current measures

"QALYs in their current form do not capture the positives of a treatment beyond direct health benefits. To measure the true value of a new treatment NICE must demonstrate….that it recognizes a treatment which makes somebody more independent and therefore less reliant on family or a carer. This may free up that person's time to engage with an economy or in wider society and so we would expect this is factored into any appraisal of the value of a treatment. …Parkinson's UK therefore warns NICE against becoming too tied to measuring the value of a new treatment through the use of QALYs alone and urges NICE to take into account a more individual qualitative approach" (p279)(18)

Parkinson's UK, Parkinson's research and support charity in the UK, 2014

Economic evaluation within the QALY framework relies on PbQoL measures. However, concerns are raised regarding the appropriateness of using the current PbQoL measure (i.e. the EQ-5D) in the Parkinson's population. Researchers and patient groups have argued that the comprehensive impact of Parkinson's on patients and the subsequent progressive disability (as shown in Section 1.2.2 and 1.2.3) and impacts on QoL and wellbeing may not be sufficiently captured by EQ-5D. Thus these are not given enough consideration in the decision-making process. EQ-5D focuses on HrQoL, notably physical functioning (14, 126), which, it has been argued, limits its ability to capture broader aspects of QoL (127, 128). As described in 1.2.4, there are various interventions for management of the motor and non-motor symptoms of Parkinson's, and the direct and indirect benefits of these interventions on the patients QoL are different. The potentially limited ability of EQ-5D to discriminate benefits, if there are any, may disincentivise some types of interventions when the intervention is targeting QoL aspects that are under-valued in its system. This may affect the assessment of cost-effectiveness and the ultimate allocation of resources in the Parkinson's population. This highlights a need to review and critically appraise the performance of the existing PbQoL measures including EQ-5D, in the Parkinson's population.

1.5.2 Incorporating broader aspects

The comprehensiveness of symptoms and the subsequent management raises a challenge to decision-makers regarding the comparison of the benefit between these interventions. Currently, NICE states ('developing NICE guidelines: the manual – 7 incorporating economic evaluation') , although not in its methods for technology appraisal (119) that:

"for some decision problems (such as for interventions with a social care focus) the intended outcomes of interventions are broader than improvements in health status. Here broader, preference-weighted measures of outcomes, based on specific instruments, may be more appropriate….similarly, depending on the topic, and on the intended effects of the interventions and programmes, the economic analysis may also consider effects in terms of capability and wellbeing." (NICE, 2014) (11)

Typical cases of broader benefit usually exist in public health interventions, however, for a long-term progressive disabling disease such as Parkinson's, wellbeing is important for the patients suffering from it as they have to 'accept and live with it'. As the Actor Michael J Fox (diagnosed with Parkinson's in 1992) said: "live in the moment, enjoy the day make the, most of what you have" (130).

Parkinson's specific QoL measures cannot be used in economic evaluations since none of them are preference-based and priority setting requires a generic measure. However, they were developed from the specific disease perspective and may be the best surrogate to reflect what matters to this group of people, i.e. 'what counts'. It is not hard to find that in addition to the health attributes such as mobility, broader attributes are highly prevalent in Parkinson's specific QoL scales, for example: stigma, social support, cognition and communication in 39-Item Parkinson's disease questionnaire (PDQ-39) (131); social/role function, self-image and outlook in Parkinson's disease quality of life scale (PDQUALIF) (132); social functioning in Parkinson's disease Quality of Life Questionnaire (PDQL) (133); and all of the attributes in Parkinson's impact scale (PIMS) which include self-positive (self-worth, happiness, optimism), self-negative (level of stress, anxiety or depression), family relationships, community relationships, work, leisure, travel and safety, financial security and sexuality (134). Therefore, there is scope for use of a suitable preference-based measure to appropriately capture and value these broader attributes of wellbeing in Parkinson's particularly for use within economic evaluations and priority setting.

1.5.3 Capability wellbeing and the ICECAP-O instrument

One potential approach that has been heralded for enabling such overall wellbeing evaluation in health economics is Amartya Sen's 'capability approach' (135-138). Sen's capability approach advocates the evaluation of programmes focusing on capability (what a person is able to do) in addition to functioning (what a person does) (136, 139).

Among the attempts to operationalise the capability approach, the ICECAP-O (Investigating Choice Experiments for the preferences of older people CAPability) instrument (140, 141) is the currently most well-known capability measure as an outcome for use in economic evaluation as shown in a previous study (125). It was developed with a view to expand the evaluative scope of current measures for economic evaluation and measure 'capability' wellbeing in older people (140). Research using ICECAP-O has the potential to provide rich and broad information regarding older populations' wellbeing as its attributes are of direct relevance to the older population, including: attachment, security, role, enjoyment, and control (140, 141). It could potentially provide a complement to the existing measures by providing a full picture of the impact of Parkinson's and enable a broader set of outcomes to be considered in economic evaluations across health and social care areas. ICECAP-O will be further introduced regarding its development, validation and use in economic evaluations in Section 2.7.2 and 2.7.3.

ICECAP-O is recommended by NICE's economic evaluation methods for social care interventions for capturing capability when the intervention effects are beyond health (11, 142). In NICE's manual for developing NICE guidelines, it says "depending on the topic, and on the intended effects of the interventions and programmes, the economic analysis may also consider effects in terms of capability and wellbeing. For capability effects, use of the ICECAP-O instruments may be considered by NICE when developing methodology in the future." Furthermore, ICECAP-O has been found to be the most widely applied older person specific instrument in both community and residential aged care by a recent systematic review published in 2015 (20). Given its popularity and NICE's recommendation, along with the fact that Parkinson's primarily affects elderly people aged over 60, the ICECAP-O is deemed as a strong candidate for capturing the broad benefit of interventions that are beyond health in this population.

ICECAP-O has not yet been reported in the population with Parkinson's, therefore their level of capability is unknown. Also, the feasibility of using ICECAP-O in this disease area has not been previously investigated. Among the most crucial issues for broadly defined wellbeing measures is their relevance and sensitivity to specific health and non-health aspects in the health care context (143) and therefore validation in specific populations is required.

1.5.4 Construct validity and responsiveness

Construct validity and responsiveness are two psychometric properties that can be quantitatively tested with data and are highly relevant to the use of PbQoL measures in economic evaluations (118). Construct of a measure relates to the hypothesized manifestations linking the underlying factors and a person's behaviour (298). Construct validity requires a measure to be able to differentiate between states that are different in the aspects that are measured (i.e., discriminant validity), and correlated with measures that are built with similar purpose. It is an important property for a PbQoL measure to be reliable to generate a utility value which is a key parameter in decision-analytic modelling. For example, Hoehn & Yahr (H&Y) staging (i.e. a measure of motor complication of Parkinson's) has been commonly used as a criteria to define the Markov states in cohort Markov models of disease progression (144-146). This is based on the assumption that utility weights should be able to differentiate between these H&Y staging defined health states; this assumption requires construct validity. In addition, previous modelling studies have identified that utility values were the top source of uncertainty in the studies comparing DBS and medications (144, 147). For example, Eggington et al. (2014) estimated the incremental cost effectiveness ratio (ICER)[5] to be £20,678 in base case which was within the cost effectiveness threshold set by NICE, i.e. £20,000~£30,000 (144). In their univariate sensitivity analysis however, when a different study source was used for utilities, with which the utilities in each H&Y stage were very similar to each other, the ICER increased to £64,170. In contrast, when changing utility data source to another study which reported a larger difference across H&Y stages, the ICER decreased to £18,650, which fell remarkably to below the boundary.

After construct validity has been established, responsiveness requires a measure to be sensitive to important changes in the aspects that it is designed to measure

[5] ICER: estimated difference of cost divided by estimated difference of QALYs between the alternatives. Definition will be covered in Chapter 2, Section 2.2.2 Cost utility analysis.

(148). It is essential in economic evaluation alongside any longitudinal studies that use PbQoL measures as outcomes. In these studies, a utility profile which is comprised of utility values at each assessment point is mapped out over the time horizon of the study to generate QALYs. Ultimately, the QALY difference between alternatives is affected by the ability of the PbQoL measure to detect important changes that matter to the patients in QoL. The ability of each measure was found to vary significantly across instruments and populations (149, 150). Previous studies have found that the choice of PbQoL instruments matters to the estimate of the ICER (151, 152). For example, Sach et al. (2009) (152) compared the ICER of four options for the treatment of knee pain and found that EQ-5D and SF-6D would provide opposite recommendations of which option is cost-effective. The option 'diet and strengthening exercise advice' was the most cost-effective option (ICER=£10,815 per QALY gained) when EQ-5D was used whereas this option was dominated by another option 'strengthening exercise advice only' (ICER=£9,999 per QALY gained) when SF-6D was used. The responsiveness of each instrument is different to different aspects of QoL, therefore the cost utility estimate may likely be different when an alternative instrument is used.

1.6 Research questions

The aim of this thesis is to examine the performance of the existing preference-based outcome measures in people with Parkinson's, and evaluate the potential of using a generic preference-based capability-wellbeing measure, the ICECAP-O, to incorporate broader aspects affected by Parkinson's in economic evaluations. There are two overarching research questions for this thesis:

1) Are the existing PbQoL measures appropriate to be used in the Parkinson's population? In other words, do existing preference-based generic measures capture all important aspects of QoL in People with Parkinson's?

2) Is the ICECAP-O capability wellbeing measure appropriate to capture the wellbeing impact of interventions in Parkinson's, and is it sensitive in this population?

In answering these questions the use of existing PbQoL measures was critically assessed via a systematic review and the ICECAP-O measure was empirically assessed using primary data, in the Parkinson's population. The data were obtained from a large-scale long-term randomised controlled trial (RCT) in Parkinson's in the UK, namely the PD MED study. The PD MED is the first study collecting ICECAP-O data in the Parkinson's population and thus this thesis will, for the first time, provide the information regarding the capability wellbeing in this population, and how its validity compared with existing measures in this context. To my knowledge, up to the submission of this thesis (March 2018), the PD MED study is the only study that reported to have collected ICECAP data in the Parkinson's population. Given its large scale nature which has recruited 1620 patients with early Parkinson's and 500 with advanced Parkinson's, and a broad range of patient profiles representing the general Parkinson's population, it is deemed to be more practical than collecting primary survey data. To answer the two overarching research questions, this thesis is split into three main empirical works.

The first research question is addressed by a systematic review of studies which used PbQoL measures in people with Parkinson's. This was conducted to identify and determine how PbQoL measures have been used in people with Parkinson's. Construct validity and responsiveness of the identified measures were assessed with the secondary data provided in each included study. Given mapping is recommended by the NICE to generate EQ-5D-3L score when it is not directly measured, studies that mapped from non-preference based measures to EQ-5D-3L were included in the review.

The second research question is addressed through two case studies, both using data from the PD MED RCT in Parkinson's, each focusing on one of the two key psychometric properties important for PbQoL measures, construct validity, and responsiveness.

The first case study explored, cross-sectionally, the impact of Parkinson's on capability-wellbeing and assessed the construct validity of ICECAP-O in people with Parkinson's in terms of its discriminant ability between groups and convergent validity with measures with similar construct. The second case study further explored the impact of progression of Parkinson's over time on patients' capability-wellbeing and assessed, longitudinally, responsiveness of the ICECAP-O in people with Parkinson's to the change of patients' overall and various aspects of QoL, and clinical health status. To aid understanding of the results in the context of PbQoL measures and understand implications for decision-making, the psychometric properties of the existing measure recommended by NICE (122), the EQ-5D-3L, were also tested. In these measurement tests, the most widely validated Parkinson's specific QoL measure, the PDQ-39, was assumed as the 'gold standard' to measure the QoL in this population.

In each empirical chapter, specific objectives are also defined under the primary aim and addressed separately within each section.

1.7 Structure of thesis

Following this introduction to Parkinson's, priority setting, use of economic evaluation, issues in preference-based outcome measurement, rationale for this thesis, and research questions, this chapter concludes with an overview of the thesis. This overview is visualised in Figure 1-2. This figure will be shown at the beginning of each chapter to highlight how it fits within the overall thesis structure.

Chapter 2 overviews and critiques the essential theories and developments in measuring and valuing health outcomes for economic evaluations. It includes the three primary economic evaluation frameworks, the concept and measurement of health, QoL, HrQoL, utility and the QALY, a critique of the QALY method particularly in outcome measurement, followed by a critical description of alternative approaches, then finally focused on one of the proposed approaches, the capability approach, as operationalised using the ICECAP-O instrument. Drawing on this, chapter 2 provides an overview of the use of health outcomes in economic evaluation, serving as a foundation for the empirical work of this thesis.

Chapter 3 introduces the methods used for assessment of measurement properties to determine whether an instrument is appropriate to be used in populations in given health states. In particular, this chapter focuses on the definition and assessment methods of construct validity and responsiveness, the two properties that are employed to address the overarching research questions of this thesis.

Chapter 4 presents a systematic review which addresses the overarching research question 1, which is to identify PbQoL measures in people with Parkinson's and assess their appropriateness in terms of construct validity and responsiveness in this population. This chapter details the methodology of systematic review and the assessment methods for construct validity and responsiveness and provides the results of the research and assessment. The identified economic evaluations in Parkinson's were reviewed and the challenges arising in relation to valuing outcomes were summarized. This chapter demonstrated a justification for further exploration of the construct validity and responsiveness of the PbQoL measures in the Parkinson's population due to possible lack of considerations in the social and mental wellbeing attributes in the EQ-5D measure important to people with Parkinson's.

Chapter 5 provides a brief further justification based on the findings in Chapter 4 for the next two chapters of the empirical case studies. In addition, given the data for the case studies both come from the PD MED trial, this chapter provides an overview of the trial and the key outcomes data collected. Since there are a number of challenges when applying the classic psychometric testing methods to answer the research question of the case studies, this chapter concludes by discussing the challenges related to the assessment of these properties, which have important implications in the methods chosen and interpretation of results in Chapter 6 and 7.

Chapter 6 and 7 present two empirical works assessing the broadly defined measure, the ICECAP-O in terms of its construct validity and responsiveness, respectively. These two chapters directly address the second overarching research question by investigating the impact of Parkinson's on capability-wellbeing and assessment of the appropriateness of ICECAP-O. Drawing on the methods reviewed in Chapter 3, Chapter 6 details the construct validation methods and results with hypotheses tested in regards to the ability of the ICECAP-O to differentiate between groups that are expected to differ, and its correlation with other measures with similar construct. Similarly, Chapter 7 details the methods and results for the assessment of responsiveness to examine the extent to which the ICECAP-O is sensitive to the change of various health, QoL and wellbeing aspects. Chapter 7 also explores the impact of missing data handling strategies on the results of the responsiveness assessment result. To aid the interpretation of the results, the EQ-5D-3L was also tested and compared with the assessment results of ICECAP-O in each chapter. Discussions surrounding the findings and the methods used are provided at the end of each chapter.

Finally, Chapter 8 summarises the main findings and contributions of the thesis by revisiting the two overarching research questions. The challenges and recommendations arising from the practical application of the assessment methodologies in this thesis are discussed and summarized. This chapter also places the findings from this thesis within the context of wider literature in relation to incorporating broader aspects into consideration for healthcare policy making in Parkinson’s. Lastly, Chapter 8 provides the overall conclusions drawn from the research conducted and scope for future research.

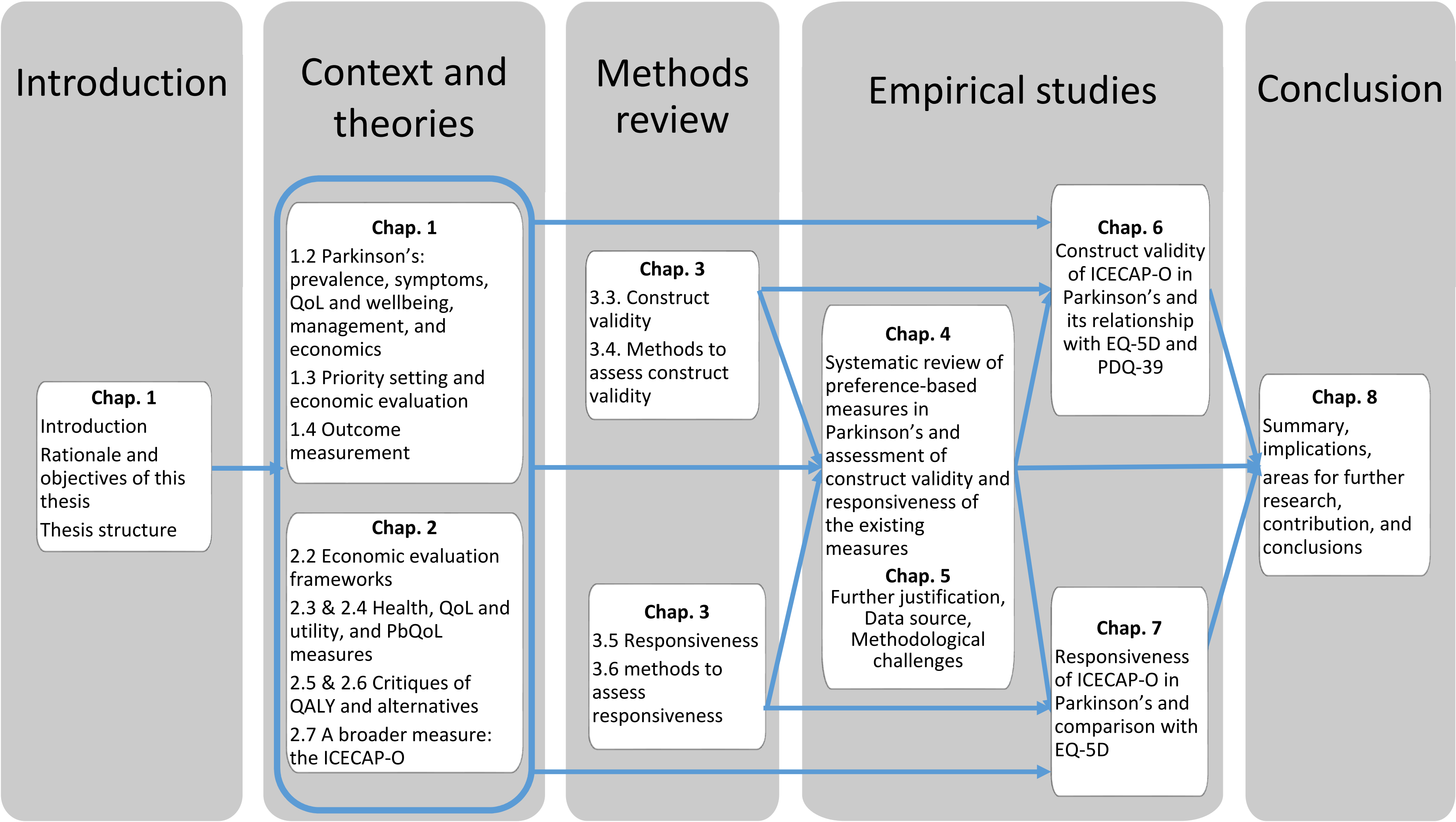
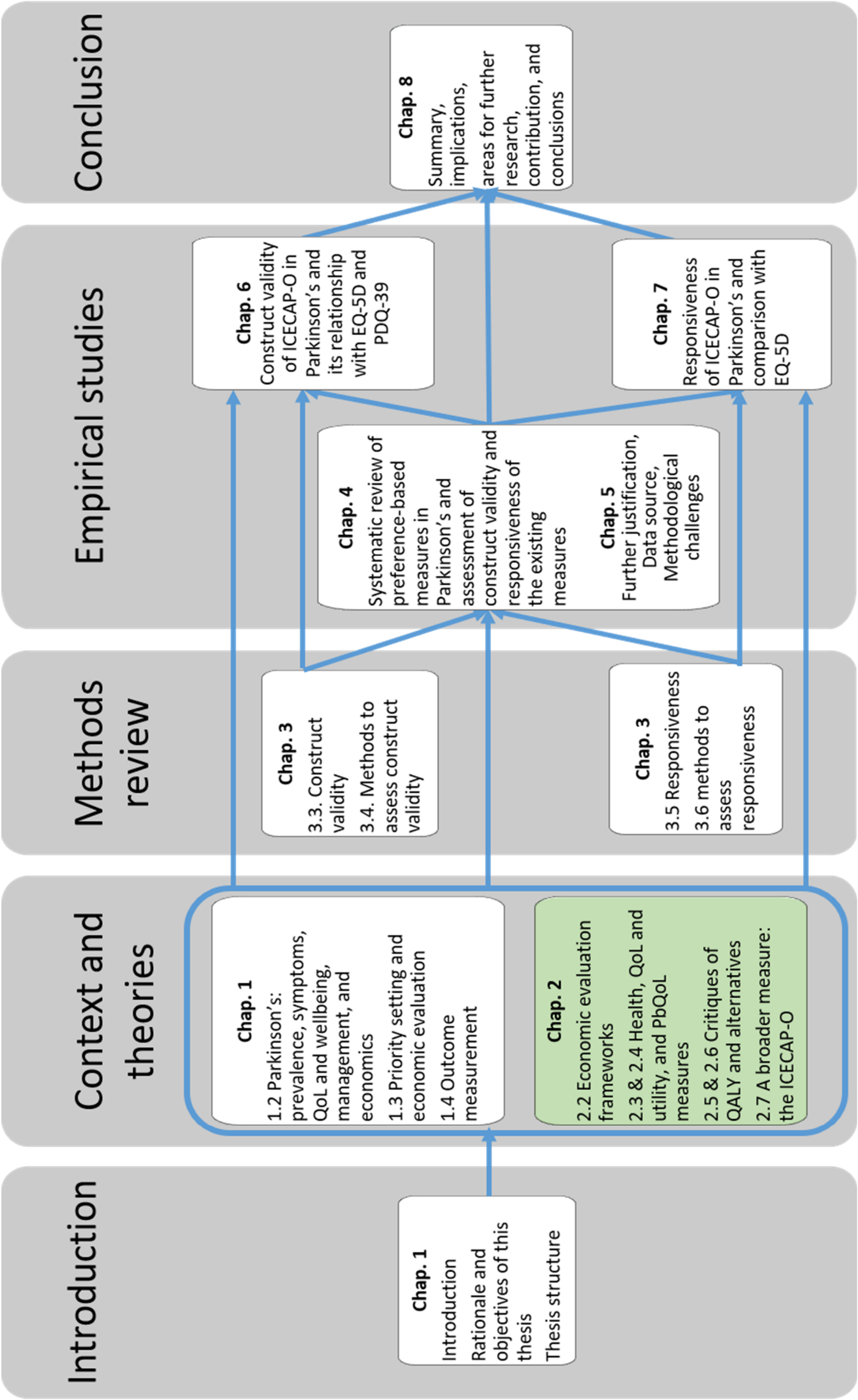

Introduction
Context and theories
Methods review
Empirical studies
Conclusion
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Chap. 8
Summary, implications,
areas for further research, contribution, and conclusions

Figure 1-2 Visualisation of the thesis structure

Chapter 2 Outcome measurement in economic evaluations



Introduction
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Context and theories
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Methods review
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Empirical studies
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Conclusion
Chap. 8
Summary, implications, areas for further research, contribution, and conclusions

2.1 Introduction

As mentioned in last chapter (Section 1.4), whilst the cost of health care interventions is relatively straightforward for measurement and valuation, challenges arise regarding how to measure and value the benefits of healthcare interventions (118).

This chapter describes and critiques the essential theories and developments in measuring and valuing health outcomes for economic evaluations. As described in Chapter 1 (Section 1.3.3), economic evaluation is the comparative assessment of the costs and consequences of alternative health care interventions (97). This chapter begins by introducing three main economic evaluation frameworks, cost-effectiveness analysis (CEA), cost-utility analysis (CUA) and cost-benefit analysis (CBA). The distinguishing feature of these analyses is the expression of outcomes and thus an in-depth description of relevant outcomes is further provided as well as their roles, and the limitations of use of each method in economic evaluations. This includes health, QoL, health utility and the QALY. The benefit of health care interventions is often classified into two key dimensions, life expectancy, and how well an individual lives, i.e. QoL. The most widely applied approach to generate health utility is through the use of PbQoL measures and thereby this chapter then introduces the preference-based measures, and one of the best-known examples among them, the EQ-5D-3L/5L instruments (153) as mentioned previously in Section 1.4. Accompanied by the growing recognition of the usefulness of the QALY in healthcare resource allocation, is the increasing debate regarding the relevance of the QALY, among which a well-known controversial issue is the recommendation of using EQ-5D as a cornerstone in the QALY framework. Therefore, this chapter discusses the issues in the PbQoL measures used in the QALY framework and reviews the alternatives to the use of current generic preference-based measures. This is followed by further examining the theoretical basis and role of wellbeing measures as alternatives to the current health-related preference-based measures, with a particular focus on the relatively new ICECAP-O instrument, a preference-based capability-wellbeing measure for older people (141).

2.2 Economic evaluation in healthcare

2.2.1 Cost effectiveness analysis

CEA compares the costs and consequences of the alternative interventions and the consequences are quantified through a single natural unit, e.g. life years gained, number of falls reduced, number of cases avoided in preventative interventions etc. CEA results are presented in terms of incremental cost per unit of health gain, such as cost per additional HIV child prevented for a HIV screening program, or cost per hospitalisation avoided for people with Parkinson's, cost per LDL cholesterol unit decreased, etc. The outcome measure is typically the primary outcome measure used in a study and this is assumed to appropriately capture the effect of interest and be specific to the condition (124). For example, a physiotherapy intervention to prevent falls in people with Parkinson's may assess cost per fall averted, and a diagnostic technology intervention to increase the sensitivity and specificity of the detection of cases may assess cost per additional case found.

However, while appropriate for the specific intervention or issue, the CEA framework is not useful for decision-making across different disease areas as the CEAs are incompatible. CEA is helpful to some extent for decision makers to rank interventions for the same condition (or symptom) using the same cost per natural unit improved. Nonetheless, it does not provide information on whether the intervention is value for money. This would require a valuation of the natural unit of the specific outcome and a further comparison to a societal consumption value of health benefit (154).

In addition, when the intervention has an impact on more than one aspect in the population, CEA is unable to include the full impacts together as only one measure of outcome at a time can be used in any given CEA analysis (118). It is argued that CEA is only considered to be appropriate when that outcome is the major objective of therapy. However, this is unlikely the case in most circumstances as any specific health outcomes (e.g. falls, hospitalisation, LDL cholesterol) are usually linked with other aspects of life, either in life expectancy, or QoL, or both (97). Even when the clinical outcome is survival / mortality, it is likely that the intervention affects QoL as well, or the reduction in mortality may be at the expense of reduction in QoL and hence the patients would be concerned with QoL. Therefore, only measuring a specific clinical outcome will probably leave some impact of the intervention overlooked in economic evaluation.

2.2.2 Cost utility analysis

CUA is the most frequently used form of economic evaluation for decisions making to aid health care resource allocation (97). It is often seen as a special case of CEA as the unit of effect is 'one year in full health'. The most widely used outcome in CUA is QALYs and the result of CUA would be incremental cost per QALY gained. As previously mentioned in Chapter 1 Section 1.4, QALY combines length of life and QoL. One QALY means a person lives for one year in perfect health (i.e. full utility). It is generated by multiplying a person's life expectancy by the health utilities in each period and summing the products from each period together (118). Health utilities represent the preferences (i.e. desirability) of individuals for a health state as valued against length of life or risk of death (155). Therefore, although much of the literature on economic evaluation does not differentiate between CEA and CUA, the outcomes used in CUA incorporate public preferences which distinguishes it from CEA and constitutes one of the necessary conditions accounting for its wide use in health care decision-making.

Health utilities are typically between 1, corresponding to optimal health, and 0 corresponding to a health state judged to be equivalent to death. Health utilities can be negative, indicating a health state worse than death. These values are often estimated through administering a standard questionnaire to get a description of an individual's health state (i.e. health profile) and then typically 'off-the-shelf' preference weights are attached to the described health state. Those preference weights for each standard questionnaire are elicited through specific valuation techniques from a sample of the general population. The standard questionnaires along with their preference weights form preference-based measures. Health utilities, preference-based measures and the valuation techniques will be defined and discussed in greater depth in Section 2.4.

The preference-based measures are intended to be general and relevant to all conditions in order to enable comparisons of cost-effectiveness across interventions and disease areas (will be covered in Section 2.4.3). The composite outcome, incremental cost per QALY gained, also called ICER (incremental cost effectiveness ratio as briefly mentioned in Chapter 1, Section 1.5.4), regardless of interventions and disease areas, permits comparison across the wide range of healthcare programmes which makes CUA of particular use to aid resource allocation by decision-making agencies (12, 119). Despite the ideal intention to be relevant to all conditions, the QALY has been criticised for being insensitive or irrelevant to some specific conditions (156, 157) and therefore the use of typical CUA methods is not without sceptical critical voices in real clinical and patient decision-making setting (158). This will be further discussed in Section 2.5.

In health care decision-making process, the ICER of a programme is compared against a threshold, which represents the maximum willingness to pay (WTP) of the health care system for an additional QALY gained. The threshold also reflects the opportunity cost in terms of the foregone benefits because other interventions cannot be provided (159). An intervention with an ICER below the threshold would be considered cost-effective in comparison to the alternative and would be likely to be funded, while one above the threshold would be considered not cost-effective and less likely to be funded. This decision rule ensures that the total QALYs generated from a given budget is maximized (118).

In the UK, the threshold value or national accepted ceiling ratio for the ICER established by NICE is considered to range between £20,000 - £30,000 per QALY gained (119). NICE Methods for Technology Appraisal (2013) states that when ICER is below £20,000/QALY, judgements about the use of a technology are based primarily on ‘the cost-effectiveness estimate and the acceptability of a technology as an effective use of NHS resources’. When the ICER is between £20,000 and £30,000 per QALY gained, the following factors are considered: the degree of uncertainty around the ICER calculation, whether the change in HrQoL has been adequately captured, the innovative nature of the technology especially when the benefits brought by the innovative nature cannot be adequately captured in the reference case QALY measure; and aspects that relate to non-health objectives of the NHS including broader benefits beyond health and costs and benefits incurred outside the NHS and personal and social services. Notably, the 2013 guide for the first time explicitly recognizes the possible inappropriateness of the HrQoL measure in some conditions and populations (15, 160). This change in NICE’s guideline reflects the development of methodology in the last ten years as well as the increasingly important methodological issues and debate in outcome measurements.

In 2015, Claxton et al. published a study which empirically estimated the NICE cost-effectiveness threshold based on routinely available data and suggested a much lower threshold practiced by NICE, which is estimated to be £12,936 per QALY surrounded by considerable uncertainty. This work has generated an uproar in the press (161, 162) as well as challenges from other researchers (e.g. Office of Health economics (OHE) (163, 164)) and NICE (165), as it will eventually mean there will be a great proportion of new interventions be rejected by NICE unless the price of new interventions are greatly reduced, as said by NICE’s chief executive Sir Andrew Dillon (165). Despite the methodological and implementation issues in debate surrounding this work, it is the first meaningful attempt at empirically estimating the threshold. If a lower threshold is adopted by NICE, it would lead to more fierce competition of interventions for NHS funding, and appropriately capturing the benefit of these interventions would become even more vital.

2.2.3 Cost benefit analysis

Cost-benefit analysis (CBA) requires programme consequences to be valued in monetary units, alongside the cost (114). The consequences not only include those financial consequences but also intangible outcomes such as survival and QoL. The use of monetary units as output measurements allows for comparisons between the return on investment in health and return from elsewhere in the economy (129). An intervention is considered worthwhile if the monetary valuation of all the benefits exceeds the costs (i.e. positive net benefit) (97).

In CEA and CUA, the results indicate the price of achieving a particular health goal (e.g. incremental cost per QALY gained) while information is not given on whether the price is worth paying to achieve such a goal. For this reason, CEA and CUA must rely on an external criterion of value to determine whether or not an intervention is cost-effective, such as the aforementioned NICE threshold £20,000 - £30,000 per QALY in the UK. CBA, on the other hand, incorporates the monetary valuation of the outcomes in the evaluation process and thus can inform us whether a goal is worth achieving given the social opportunity costs (i.e. foregone benefits) of the programme that would be displaced (97).

A number of techniques for obtaining monetary valuation of health benefits have been proposed, among which the 'stated preference' method or more specifically contingent valuation method is the most commonly used approach in applied microeconomics. The contingent valuation method asks respondents to determine how much they would be 'willing to pay' (WTP) for an intervention in a survey, through a variety of formats (e.g. closed ended, open-ended, payment scales (166)), although they are not required to pay (which contrasts the 'revealed preference' approach) (167) . WTP has the potential to allow all the benefit (not just health) of an intervention to be considered by the respondent in the preference elicitation process and thus it would be suitable for use in the scenario where wider benefits beyond health are expected from the intervention (168).

However, there are a number of concerns with the use of WTP methods. Firstly, in a publicly funded health care system, people may not have an accurate sense of the value as they do not pay for health care out of pocket (169). In a more practical way, this may lead the method being vulnerable as it is open to manipulation and hence NICE puts less weight to it when making decisions than other methods using patient self-reported preference-based outcomes (169). In addition, a strong relationship was found between income and WTP, whereby people with low income provide low valuations (170). This may lead to measurement bias and equality issues, affecting reliability of WTP method to be used in economic evaluations especially when effect of intervention is related to income (171).

2.3 Health, Qol and HrQoL

As mentioned in last section, the distinguishing feature of different economic evaluation frameworks is how the outcomes are measured (and valued). A prerequisite to critique these outcomes is the understanding of the concept that they are measuring. There are considerable confusions regarding the use of the terms health, QoL, and HrQoL and those terms can be used interchangeably in some situations but they may refer to different concepts in other situations.

2.3.1 Definition of Health

What is health? The definition of health is evolving over the years with the development of health care and wellbeing but still under debate (172). One of the key controversial issues is its scope. The 'Father of western philosophy', Aristotle (384-322 BC) discussed that extremes in the bodily condition should be avoided and maintaining a proper balance is a virtue. He considers 'endaimonia' (wellbeing) the final goal and 'final good for man' (173, 174). This represents the typical historical view of health about the human potential to be in a state of balance and the aim of developing oneself to achieve wellbeing. This view remained highly influential in western medicine and thinking of what health is over 15 centuries (173) and its impact continues to this day. From the 16th century, a 'microscopy' way of interpreting human health and disease began with the invention of microscope, marked as the milestone for the development of modern medicine. Disease is no longer to be explained by misbalance of nature but a result of the changes in physical body detected by modern technology and correspondently (173), health is defined as the absence of disease (175).

In the 20th century, the World Health Organisation (WHO) was founded after the Second World War in 1948 and the definition of health in its Constitution is perhaps the current most well-known and enduring one. It explicitly clarifies its broad scope as "a state of complete physical, mental and social wellbeing and not merely the absence of disease or infirmity." The breadth and ambition of this definition was a 'radical development in its day' (173) from the dominant definition among physicians of the 'absence of disease' over the previous four centuries (176). In 1986, WHO elaborated its definition in the 'Ottawa Charter for Health Promotion' as "a state of complete physical, mental and social wellbeing, an individual or group must be able to identify and to realise aspirations, to satisfy needs, and to change or cope with the environment." Health is regarded as a resource for everyday life, not the objective of living and health promotion goes beyond the healthy life-styles to wellbeing (177).

This broad definition of health has been subject to criticisms. It is argued that health is one of the determinants of social wellbeing but social wellbeing is not part of health (178). One of most cited argument in the area of health economics comes from the views by Evans and Wolfson (1980) (179). They argued that the WHO definition is indistinguishable from the concept of utility and that a public system ‘wishes you well, but not necessarily happy’ (180, 181). They suggested to “conceptualise health status for inclusion within the utility function in its narrow, negative, but more or less objectively measurable form” (180, 181). This narrow definition, on one hand, offers the advantage of easy measurement of health outcomes as a result of health care interventions. On the other hand, this narrow approach brings concerns over its limited ability to understand the underlying causes of disease (182, 183) and the impact of disease, the outcome measure developed under which is criticised for failing to capture the full spectrum of the impact of intervention (184).

Related to the scope debate is the argument proposed by Alex Jadad and Laura O’Grady that the ‘absolute’ or ‘complete’ health state in the WHO definition makes it impracticable for what the health care systems can achieve, unattainable for people with chronic illness and disabilities, and lowering the threshold for unnecessary intervention (185). They argued that the requirement for complete health “would leave most of us unhealthy most of the time” (186). Therefore, they proposed a new definition of health in 2009 as “the ability to adapt and self-manage” in the face of social, physical, and emotional challenges. This new concept has its advantage of emphasizing on human-beings more than their illness and its focus on their strength rather than their weakness, yet it requires substantial personal input, as not all people were believed capable of providing such input (173).

2.3.2 QoL and HrQoL

2.3.2.1 QoL

The WHO broad definition of health, in particular, the term ‘wellbeing’, means that the measurement of health and the effects of health care must include not only an indication of changes in the frequency and severity of diseases but also an estimation of wellbeing (187). This is believed to be one of the most important traced root to the development of the concept of QoL and has been very influential in the development of QoL measures (188). Following the WHO’s definition of health, the majority concepts of QoL developed in health sciences encompassed at least three dimensions, namely physical function, mental status, and the ability to interact with society and achieve roles. For example, in the dictionary of epidemiology, QoL is "the degree to which persons perceive themselves able to function physically, emotionally, mental, and socially." (Hartge 2015 p234) (189)

A discriminatory feature between the concept of QoL and health lies in the subjectivity of QoL. A person's QoL is about how health is perceived by that individual. WHO defines QoL as "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns." It is a broad ranging concept which encompasses "individual responses to physical, mental and social effects of illness on daily living which influence the extent to which personal satisfaction with life circumstances can be achieved" (Bowling 2005) (190). Therefore it can be seen that QoL is a subjective concept with a focus on people's perception and reaction to their health status (191), which means it may vary substantially between individuals (192). Due to this subjectivity, the measurement of an individual's QoL exhibits complexity and cannot be replaced by the perception of other people. A substantial body of evidence has demonstrated the significant differences in the perception between professionals on patients' health status and the patients themselves (193-195). Consequently, the importance of measuring how patients perceive their health status themselves is becoming increasingly recognized in order to reflect the actual experience of the disease and the intervention (196).

2.3.2.2 QoL vs. HrQoL

In the literature, the use of QoL may refer to varied scope and there is overlap between QoL and health (16, 188, 197). Underlying the issue of the mixed use of terms is an important implication for policy in terms of what should be counted as a benefit and what could be given less weight in the decision-making process (118). Ware argues that the definition of the QoL in health science should be aligned with the aim of the health care system: "the goal of the health care system is to maximize the health component of quality of life, namely health status. Measures of health outcomes should be defined accordingly." (198) What echoed Ware's argument is the development of the term "health-related quality of life

(HrQoL)" (Torrance 1987) as a pragmatic approach to QoL for the resource allocation purpose (Dolan & Olsen 2002). Torrance (1987) defined HrQoL as the subset of QoL, relating only to the health domain of that existence. Health is an important factor among the many that contribute to a person's QoL (199).

However, as a compound word derived from the multi-dimensional concept of health and QoL, the problem appears when defining HrQoL. Some of the definitions are more closely linked to health definition while others are closer to QoL (16). An example of the former is such as Torrance's original definition which suggests that HrQoL includes only those factors that are part of an individual's health (199). In contrast, some definitions resemble QoL: 'those aspects of self-perceived wellbeing that are related to or affected by the presence of disease or treatment' (200). It brings the issue to discriminate between what aspects of QoL are affected or not affected by health, especially when the indirect influence is considered (e.g. health affects income and education) (16, 201). Guyatt et al. argued that although clinicians focus on HrQoL, 'when a patient is ill or disabled, almost all aspects of life can become health related' (202). Perhaps an explanation to this issue would be that all the aspects of QoL could be affected by health, but except for health, they may or may not be affected by other factors as well. As such, the term 'health related' in the HrQoL concept may not be the clearest description of what this term is intended to be. In health economics literature, HrQoL sometimes refers to the utility values assigned to different health states which are used to calculate QALYs (203). The values can be elicited from a range of preference-based measures, some of which are broad while some have a focus on health, and therefore it remains unclear if all of the measures that could produce values to be combined with length of life are eligible to be called HrQoL measures.

Due to the complexity of the concept, there is considerable confusion regarding the use of the term QoL and HrQoL. It was recommended that researchers be as specific and clear as possible about the concept and operationalisation of QoL in the studies and the audience should inspect the context those terms are used (188). In summary, the essential concept of QoL is subjective, multi-dimentional and encompassing broad ranging aspects of wellbeing, whereas the HrQoL is also subjective but focuses on dimensions that are primarily determined by health and is more related to the objectives of the health care system for resource allocation purposes.

2.3.3 QoL outcome measures

Due to the subjectivity of the QoL concept, QoL outcomes are subjective measures concerning with how patients perceive themselves about their health and wellbeing. The development of QoL measures can be traced back to the earliest stage of measuring function status as an extension beyond clinical outcomes (dates back to 1937) (204). An example is a single numerical scale developed by David Karnofsky in 1948, to measure the performance status of cancer patients (205). It gave scores between 0 and 100 for a combination of three factors: the ability to carry out normal activities, the need for custodial care, and the need for medical care. In the 1970s, several highly influential publications (206) showed that the subjective indicators could be measured, “enabling examination of the ‘soft data’ for QoL” (204). The earliest instruments specifically aiming at measuring QoL appeared in medical literature are the Vitagram Index (207) and Life Units (208) in the 1970s. The first QoL measurement to become popular was Priestman and Baum’s 1976 Linear Analogue Self Assessment Scale (204, 209) whereby the subjects were asked to place a mark corresponding to their feelings, on a visual analogue scale. Since the late 1970s, the researchers began to construct QoL measures with attributes. Examples are Index of Wellbeing, Index of Psychological Affect and Index of Overall Life Satisfaction, all developed by Campbel, Converse and Rodgers (206). The measurement of QoL became officially acknowledged with the requirement for QoL data as one of the ‘key efficacy parameters’ in clinical trials for new anti-cancer agents by the FDA in the US in 1985 (204), followed by the incorporation of QoL in outcome assessment for new health technologies by the UK Department of health in 1992 (210).

Depending on the scope of the applicability of instrument, the QoL instruments can be categorised to either generic or specific. Generic measures are intended to be relevant to all conditions, many of which may be applicable for use within the general population, such as the EQ-5D-3L/5L (153) and Short Form -36 items (SF-36) (211). The specific measures are developed to measure QoL in people with a specific condition such as the Parkinson’ specific questionnaire, PDQ-39 (131) (will be introduced in Chapter 5 Section 5.3.4) and the cancer specific questionnaire EORTC QLQ-C30 (212).

The QoL measures that are not valued are not suitable for use in economic evaluation directly because they simply measure the amount of limitation a patient is experiencing compared to a perfect health; it does not contain weights corresponding to people's preference for each amount (i.e. levels) of limitation for each attribute. The weights are elicited through valuation process. If the valuation of the QoL measure is completed, the measure becomes a preference-based measure. The preference-based measures can be used in CUA. These will be discussed in depth in the next section.

2.4 Health utility and preference-based QoL measures

In the 50-year history of health economics, one of the most important innovations in economic evaluation has been the development of the QALY (213). It was initially introduced in 1968 by Herbert Klarman and colleagues in a study on chronic renal failure, where for the first time the life-year gained was calculated with the QoL adjustment in an economic evaluation (214). The Q in the QALY comes from utility values attached to the health state, which is usually measured indirectly with preference-based outcomes. This section will introduce the definition of health utility, how the utility values are elicited, the features of preference-based measures and the most widely used preference-based measure, the EQ-5D instrument.

2.4.1 Health utility

Health utility (also called health state preference values (215)) is used as a preference weight to adjust the length of life in the calculation of QALYs. In microeconomic theory, utility represents the degree of satisfaction experienced by the consumer through the consumption of a good or services. In Alfred Marshall's book 'Principle of Economics', it states that 'utility is taken to be correlative to Desire or Want.' (p78) (216) Utility cannot be directly measured; however, economists suggest it can be indirectly revealed as "the price, which a person is willing to pay for the fulfilment or satisfaction of his desire" (Marshall 2013) (216).

Adapted from the traditional economics theory, utility in health economics refers to the degree of desirability by the individuals or society for any particular set of health outcomes (e.g. for a given health state, or a profile of states through time) (97). The more desirable (i.e. more preferred) health outcomes will be attached with larger health utility values on the scale, while the less desirable health outcomes will be attached with smaller values on the scale. In simpler words, 'a health state that is more desirable is more valuable' (Weinstein 2009) (120) and vice versa. As such, health utility measures can be differentiated from the other measurement of health as it represents a valuation. Consequently, CUA allows health outcomes to be 'valued according to their desirability' (97).

As mentioned in Section 2.2.2, utility values conventionally fall between 0 and 1, where 0 indicates the valuation of death and 1 indicates the valuation of a state of perfect health (215). In some scoring system a negative utility value is also possible (e.g. EQ-5D-3L (217)), which indicates that a health state that is less desirable than death (118). Utility values are on an interval scale, on which the same change means the same irrespective of the part of the scale being considered (e.g. a change in health from 0.2 to 0.3 is equivalent to a change from 0.8 to 0.9) (218). Utility can be compared but cannot be multiplied or divided (e.g. a utility value of 0.6 does not mean the desirability for this health state is twice as much as another health state with a utility value of 0.3).

In the literature, utility is often used interchangeably with the term 'value' and 'preference'. Some consider the value is equated with preference or desirability (which is the core concept of utility as mentioned above) (120) while the others argue that there are differences between them (97). Preference is regarded as the umbrella term (97) describing trade-offs between outcomes. Whether it is 'value' or 'utility' depends on how the question is framed in the preference measuring (or elicitation) process. It has been suggested that when the question is framed under uncertainty which is usually involved with probability or risk, 'utility' is elicited, whereas 'value' is elicited when the question is framed under certainty (97). However, in practice, their meanings are usually not differentiated, for example, in the NICE glossary, utility is the 'the measure of the preference or value that an individual or society gives a particular health state'. In this thesis, 'utility' or 'utility values' refers to the index score of a preference-based measure or the result score from direct preference elicitation, which to be combined with length of life in the QALY calculation. 'Value sets' refers to the readily used preference weights attached to each state defined by the preference-based measure. 'Preference' is used in more general circumstances, to refer to ordering of people's desirability between health states, and used to describe a measure for which the health states have gone through the preference-elicitation process and have the value sets attached, i.e. preference-based measures.

2.4.2 Preference elicitation

The methods by which preferences are elicited vary, but fall into two main categories: scaling based methods such as the visual analogue scale (VAS) and choice-based methods such as the standard gamble (SG), time trade-off (TTO), the discrete choice experiment (DCE) and best worst scaling (BWS). The choice-based methods are more commonly used and preferred by health economists compared to the scaling based method since the former incorporate 'trade-off' in the valuation (219). The choice-based method is also recommended by NICE in its guide to the methods of technology appraisals (122).

2.4.2.1 The scaling based method

The scaling based method is to ask participants first to rank health outcomes from most preferred to least preferred, and then, to place the outcomes on a scale such that the distance between placements corresponds to the differences in preference (97). Scores generated from the scaling based method provide the information of the ordering of health outcomes and the relative degree of preferences between these outcomes. The utility score for a health state is a proportion of its placement as relative to where death and full health is marked on the scale. However, rating scales do not satisfy the axioms of expected utility theory[6], nor do they require 'trade-off' between length of life and QoL by the participants (221). In addition, this technique is associated with specific

[6] The expected utility theory states that decision maker chooses between risky or uncertain prospects by comparing their expected utility values (220).

measurement biases, such as the end aversion bias (222), and context bias (223, 224). The end aversion bias reflects the fact that participants are reluctant to place health states at the extreme ends of the scale (97). Context bias refers to the fact that the VAS score for a state depends on its relative place compared to the other states presented at the same time (222). A higher value will be given for a state if it is included along with many worse states, and a lower value will be given if the state is presented along with many better states. Despite models were invented to adjust the biases (225) to some degree, the scaling based method is less preferred by health economists compared to the choice-based method (97, 120) due to aforementioned intrinsic limitations in valuation. It is more often replaced by the latter or only used as a 'warming up' exercise (218).

2.4.2.2 The standard gamble

The SG method is based directly on the third axiom of expected utility theory about continuity of preferences, first presented by von Neumann and Morgenstern (1944) (226, 227) and sometimes called 'the von Neumann-Morgenstern standard gamble' (228). It asks participants to trade-off between two alternatives. In alternative one the participant stays in a chronic state (*i*) for lifetime; alternative two is a treatment with two possible outcomes attached with different probabilities: for outcome one the participant returns to perfect health and lives for an additional set number (*t*) of years (probability *p*), or for outcome two the participant dies immediately (probability 1-*p*) (97). Probability *p* is varied in the exercise until the participant is indifferent between the two alternatives, at which point the required utility for state *i* for *t* years is equal to *p*. In simpler words, utility can be understood as the probability of full health in the gamble that makes the participant indifferent between the two choices, staying in chronic state, or going for the gamble. There are two disadvantages associated with SG. It was found that SG results could be affected by risk attitude – risk-seeking respondents tend to choose gamble while risk-averse respondents tend to choose staying in chronic state (229, 230). In addition, unlike the scaling-based method, it complicates the task by incorporating trade-off and uncertainty into the process (97). This leads to an issue that participants may find the concept of probability difficult to grasp. Despite the development of visual aids, an incorrect understanding may still exist to some degree, thereby causing measurement bias.

The process of administering the exercise may also being time-consuming for participants.

2.4.2.3 The time trade-off

The TTO technique is considered easier to understand than SG. The TTO was developed specifically to be used in health care (228, 231). It asks participants to choose between two alternatives, living in full health for a given period of time (*x*) followed by death, versus, living in a worse health state (*i*) than full health for a longer period of time (*t*) followed by death. Time *x* is varied in the task until the respondent is indifferent between the two alternatives, at which point the required preference score for state *i* is given as the ratio of *x* divided by *t*. It requires participants to choose between alternatives. The less complex alternatives in TTO overcomes the difficulty of explaining probabilities to respondents in SG. A key criticism of the TTO is the bias caused by time preference, as it is argued that TTO can be contaminated by the variation of time preference of each individual, i.e. individuals have higher preference for health now over future health all else being equal (232). It was also found some respondents were unwilling to sacrifice any of their life expectancy, leading to difficulty of administering the task (233).

2.4.2.4 Discrete choice experiment and best worst scaling

Besides the VAS, TTO and SG, two commonly used alternative preference-based approaches in health economics are DCEs and ranking methods such as BWS. These alternatives can establish the degree of preference for one alternative over another directly. They do not establish the indifference point of the individual respondent in a single question as the SG and TTO do. The DCE is a survey method asking respondents to choose between two or more alternatives which vary on level for each attribute or characteristics. In a typical BWS, respondents are asked to indicate the best and worst attributes with levels for one single profile at a time. Compared to the SG and the TTO methods, the DCE and BWS (or ordinal methods in general) require less abstract reasoning and are thus less cognitively demanding. Nonetheless, the ordinal techniques have an important limitation attributed to their ordinal nature that the elicited values require rescaling to be anchored to death so that the measure can be used for QALY calculation (118).

2.4.3 PbQoL measures

Compared to the time/cost consuming direct preference measurements such as the SG or TTO for each time a CUA is conducted, the off-the-shelf preference-based instruments are most widely used in health economic evaluations to obtain utility values for the calculation of QALYs due to its simplicity. Such measures usually comprise two components: a questionnaire formed by a number of descriptive attributes and levels regarding a person's QoL status, and an algorithm to calculate the value attached to each health status described by the questionnaire (i.e. health profile). The algorithm contains weights for each attributes and levels, derived from health state valuation tasks (as introduced in Section 2.4.2) where a sample of the general public's preferences for different combination of health states are elicited.

After an instrument is developed, validation tests should be conducted to examine their measurement properties in order to determine their appropriateness to be used in future studies (118). Chapter 3 will introduce a range of important measurement properties, among which the construct validity and responsiveness will be discussed in depth and will be assessed in the case studies in Chapter 6 and 7.

The traditional PbQoL measures are generic to enable comparisons of CUA results across areas when making decisions. However, a growing body of evidence has been published expressing concerns on the degree of sensitivity of the generic PbQoL measures to some specific conditions (156, 234-236). A way to address this criticism is the research of developing condition-specific preference-based measures (CS-PBMs) in the last decade and therefore the preference-based measures now can be either generic or specific (237). The advantages and limitations of CS-PBMs will be discussed in section 2.6.2.

2.4.3.1 Examples of generic PbQoL measures

Examples of generic PbQoL measures include the SF-6D, the Health Utilities Index (HUI) and the EQ-5D-3L/5L. The SF-6D is developed based on the longer SF-36 QoL instrument by reducing it to a six-dimension classification and the preference elicitation process transforms its scores to utility values (238). The six dimensions are: physical functioning, role limitations, social functioning, pain, mental health and vitality. The UK scoring model for the SF-6D was developed using SG technique on a sample of 836 among general population. The scores are on a conventional dead-perfect health 0-1 scale, with the worst state with a score of 0.345. In the UK, it has been used as the primary health utility measure by NICE for CUA analysis of pharmacological treatments such as for Alzheimer's, low platelet count, peripheral arterial disease, and gout (239).

The HUI consists of two systems, HUI2 (240) and the newer HUI3 (241). Scoring algorithms are both based on SG measured in the general public. The two systems shared some attributes: emotion, cognition, and pain. Additionally, HUI2 contains sensation (as one dimension), mobility, and fertility. HUI3 removed 'fertility', spilt 'mobility' into 'ambulation' and 'dexterity' to increase the structural independence, and expanded the sensation into three attributes: vision, hearing, and speech. It is suggested HUI3 should be used as the primary analysis and HUI2 in a secondary role with the exception of circumstances that focus on self-care, worry/anxiety, and fertility (97).

The EQ-5D-3L is NICE's preferred instrument for cost-utility evaluations in healthcare technology assessments. It will be introduced along with its newly developed variant, EQ-5D-5L, in the next part.

2.4.4 The EQ-5D instrument

2.4.4.1 Introduction

The EQ-5D-3L is a generic preference-based health-related QoL measure that has been widely used worldwide (153, 217). It was developed by a multidisciplinary group of researchers from five western European countries, the EuroQol group, in the late 1980s. The EuroQol group selected the 'core' domains common to other generic PbQoL measures and which reflected the most important concerns of the patient based on the group's expertise and evidence from literature (14, 118, 242). It was initially comprised of six dimensions: mobility, self-care, main activity, social relationships, pain, and mood (14). The instrument was further modified to a standard five-dimensional format which has since remained unchanged (153).

The five dimensions are: mobility, self-care, daily activities, pain and discomfort, anxiety and depression.

EQ-5D has two forms: the classic EQ-5D-3L and the recently developed EQ-5D-5L. The EQ-5D-3L contains three levels for each dimension: 'no problem', 'some problems' and 'a lot of problems', which defines 243 possible health states. By March 2017, it has been translated into 172 languages (243). The country-specific value sets have been elicited in approximately 20 countries and regions using a mixture of TTO and VAS technique (118, 244). In the UK, the valuation work was undertaken by the UK Measurement and Valuation of Health (MVH) group at York who applied the TTO technique with a random sample of 2,997 members of general public selected using the national postcode address file from England, Scotland, and Wales (245). By applying the scoring values, the EQ-5D-3L health states can be converted into utility values. The UK EQ-5D-3L values ranged from - 0.59 to 1, with 0 representing death.

2.4.4.2 Limitations

As use of the EQ-5D-3L has become common, voices both criticising and endorsing its use have been heard in a growing body of literature over the last decade (234-236, 246-248). Whilst in many applications the EQ-5D-3L has been shown to be a valid and reliable measure of QoL (example such as (249), (250)), its limitations are raised to an increasing volume of literature which could be mainly summarized by two points. The first concern deals with the sensitivity of the EQ-5D-3L to small changes. It was found that in some contexts the EQ-5D-3L may lack responsiveness to small changes especially when people have milder conditions (251). Related to this is the exhibited ceiling effects (i.e. the proportion of respondents reporting the best possible health is high (typically >15%) who are therefore unable to record any improvement in health status (252)) in both general and disease-specific populations (127, 253-255), leaving less room for improvement over time in response to an intervention.

Another concern is that the scope of EQ-5D dimensions may fail to capture important aspects of QoL in certain condition areas, for example mental health (234), schizophrenia (246), cancer (247), Alzheimer's disease (236) and dementia (248). One suggestion is that the generic attributes making up these measures may not be sufficiently relevant to the specific populations (256). Longworth et al. (235) valued three condition-specific 'bolt-on' attributes as extensions to the EQ-5D related to hearing, tiredness and vision, and found that the 'bolt-on' attributes had a significant impact on the values of the health states.

To address the limitation in sensitivity, in 2011 the EuroQoL group developed the EQ-5D-5L which contains five levels for each dimension (255). The two additional levels are 'slight problems' between the existing 'no problems' and 'moderate problems', and 'severe problems' between the existing 'moderate problems' and 'extreme problems'. This version now describes a total of 3,125 distinctive health states and thus should be more sensitive than the EQ-5D-3L version to detect minor changes. While valuation work for the EQ-5D-5L is underway, an interim value set was developed from 3,691 respondents with broad-ranging level of health in six countries by mapping (cross-walk) from the EQ-5D-3L (257). This is also relevant to those wishing to achieve consistency with previous studies using the EQ-5D-3L (118). The English value sets for EQ-5D-5L have been developed in 2016 using the TTO and the DCE techniques with data provided from 996 participants (258), although NICE recently chose not to recommend this new value set owing to the concerns on consistency with the 3L version (259). Value sets for other countries are under construction at the time of writing this thesis and not available yet. Up-to-date information can be found from the Euroqol official website: www.euroqol.org. Although EQ-5D-5L has been shown in several studies with an improved sensitivity and reduced ceiling effect compared to EQ-5D-3L (260-263), the relevance of its dimensions and scope to some specific conditions still remains questionable.

2.5 Critiques of the use of the QALY in outcome measurement

2.5.1 Evaluative scope

As mentioned in the example of the EQ-5D above, QALYs have been criticized for not encapsulating all the relevant attributes of health care and being too narrowly focused on health in its narrow meaning (97). Related to this, the concerns raised regarding the methods of the QALY, for being not sensitive enough to the health change brought by an intervention. Empirical evidence regarding this ‘generic vs. specific’ debate can be traced back to the 1980s. Donaldson et al. showed that the one of the earliest generic PbQoL instrument, the Rosser disability and distress scales (264), is less responsive to the changes in elderly people’s health states compared to the specific measures in stress and life satisfaction, using data from a trial of long-term care for elderly people (184).

A review in 2014 of the use of generic and condition-specific HrQoL measures in the context of NICE decision-making found that the EQ-5D-3L’s performance was poor in hearing impairments and varied in vision according to aetiology (235). Qualitative research suggested that the EQ-5D and the SF-36 (and subsequently, SF-6D) have limitations in capturing most of the concerns for patients with mental health problems (265). An overview of reviews published in 2017 assessed the appropriateness of five commonly used PbQoL measures, including EQ-5D, SF-6D, HUI3, 15D and AQoL (19). In this overview, the performance of these measures varied across conditions. The EQ-5D was found to perform poorly in hearing impairments, multiple sclerosis, personality disorders, schizophrenia and dementia. SF-6D showed poor performance in cardiovascular, respiratory disease, and neoplasms and HUI3 for some subpopulations of neoplasms.

Furthermore, this limitation in evaluative scope may cause problems in evaluation of the public health and social care interventions where the social and medical considerations overlap, since the benefits of these interventions are often beyond health and may also fall in other sectors such as empowerment, education, and crime. Therefore, QALYs and their associated PbQoL measures like the EQ-5D or SF-6D are likely to underestimate or overlook the relative benefits of public health interventions when compared to health care interventions (171).

As a response to this criticism, a two-and-a-half-year research project called ‘extending the QALY’ led by University of Sheffield has begun in May 2017 to review the way QoL is measured across health and social care. It aims to assess if the current measures miss the important benefits of treatments beyond HrQoL, such as independence, or improved relationships with family, friends and carers (266). If the results of the review demonstrate a gap, the research team said “NICE would consider whether and how to include any new QoL measure in its work.” (266)

2.5.2 Comparison of different preference-based measures

Another issue of the QALY framework is the discrepancies in utility values when measured with different preference-based instruments in the same patients and this explains why NICE recommends specifically the use of EQ-5D as the preference-based outcome in economic evaluations. The discrepancies have been shown by a substantial body of evidence across many different conditions (127, 267-270). The discrepancies not just lie in the absolute magnitude on a scale but also the relative direction. Richardson et al. (2015) compared the utilities obtained by the EQ-5D, SF-6D, HUI 3, 15D, Quality of Wellbeing (QWB) and Assessment of Quality of Life (AQoL-8D) and it was found that the agreement between these measures was substantially varied with interclass correlation coefficient values ranging between 0.34 and 0.82 (271). For the head-to-head comparison, substantial differences between EQ-5D and SF-6D have been reported widely (118). For instance, Brazier noted that a full score in EQ-5D could have as low as 0.56 on SF-6D in the scatter plot of the pairs (127). The difference also has been shown to have an implication on the results of QALY gained and affect the cost-effectiveness results. Xie et al. found the difference between the utility values generated from the two measures is 0.14, which yielded a difference of $10,000/QALY in ICER estimation (272).

The difference of these values may come from three aspects: differences in dimensions and items, the number of levels, and valuation methods (118). These measures differ in their coverage; e.g. EQ-5D, HUI3 and 15D are mainly concerned with physical aspects, SF-6D have special wide coverage of the sensations. Difference in the number of levels also cause the incompatibility of the measures. for instance, after adding two levels, the EQ-5D-5L was found to lead to smaller incremental QALY gain compared to EQ-5D-3L from effective health technologies and therefore interventions may appear less cost-effective (273). Furthermore, it has been suggested that SG would generate higher values than TTO due to risk aversion and positive time preference, and TTO values would exceed VAS in most of the studies due to measurement bias of VAS (118).

These differences have implications in utility measurement and the result of economic evaluations, which may lead to the varied degree of sensitivity in different population and for use with evaluating different interventions. This ‘unfair’ treatment by different measures may cause concerns about their validity to measure true preferences. If the benefit of an intervention for a certain disease is relatively under-estimated this may cause it to be ‘unfairly’ assessed in NICE’s decision-making process.

2.5.3 Whose preferences matter?

Another debated issue is that of whose preference should be valued. The values from most of the preference-based instruments for estimating QALYs use samples of the general public (as recommended by NICE) rather than the specific population who are currently experiencing the health state of interest. This poses a question of whose preferences to elicit in health-state valuation (274). The general public may or may not be in the health state at the time of assessment and they have to try to imagine what the state would be if not. The arguments for the use of general public are the insurance principle (i.e., the public are payers), the social contract principle (i.e., health system benefits all members of society), and the concern about bias associated with patient valuations, practical issues with obtaining patient samples, and to ensure comparability across different studies (203). However, some suggest that the preferences should be elicited from the patients (275). It was argued that the general public does not have the same experience of the disease as patients and thus cannot reveal the true preference of the specific population being evaluated (276).

The values can vary with the source. A number of empirical studies have shown that higher values tend to be placed on disease state by the patients who are experiencing the disease than the public (275, 277). However, this was not supported a review of studies which did not find consistent difference between the values from patients and general public (278). This review suggests that patients tend to give higher values on severe health state but lower values on milder health state than general public.

There are three key factors leading to the differences: different understanding of the health state description, ‘adaptation’ to disease, and incorporation of self-interest from different perspectives (279). It was suggested that health state description might not fully capture the patients’ experience of a specific health state due to lack of scope, which may cause different understanding between the patients' experience and what is intended to value (280). The fact of placing higher values can be explained by 'adaptation effect' where the patients have adjusted themselves to their limited health state physically and psychologically. On the contrary, when healthy people are asked to imagine the hypothetical impaired health state, they tend to 'focus on the negatives' and thus lack ability 'to look at the bigger picture of life' (279). Another factor led to the difference comes from the different perspectives held by the patients and the general public. Kahnemann (281) previously described the general public as the 'seller of health' and patients as 'buyer of health' however this analogy may not be correct in the UK NHS context. General public is a 'payer' rather than 'seller' as they pay for the health care through the tax system, while patients are 'consumer' rather than 'buyer' as they obtain the benefit from the health care services and product. Payer would assign a lower value to the services and products as they do not get the benefit of them, and on the contrary, the consumer would assign a higher value as they do not need to pay.

2.6 Alternatives to health-related generic preference-based measures

Due to the limitations of generic preference-based measures which may be insensitive or irrelevant to some specific conditions or interventions, alternative methods have been proposed to 'bypass' this issue in CUA. The alternative methods can be classified to two types: a) condition-specific approaches which include mapping from a non-preference based (usually condition-specific) QoL measure to a preference-based measure, valuation of a condition-specific QoL measure, and adding 'bolt-on' items to EQ-5D; and b) incorporating broader aspects to preference-based measures.

2.6.1 Mapping from non-preference based measures

Preference-based measures are the key instruments to value the impact of the intervention in economic evaluations which enable the decision bodies such as NICE to judge whether an intervention is value for money, however, they are not always included in clinical studies. For example, studies of new interventions sometimes only include the condition-specific measures to demonstrate if the intervention is working as what it claims to do, since those studies are mainly conducted to inform licensing decisions (118). Preference-based measures are less often seen in studies conducted in the countries where CUA is not used to inform decision-making. One practical solution to make use of these studies for decision-making is to conduct mapping (also called cross-walking) to predict utilities from non-preference based measures.

Mapping is the process of development and use of an algorithm, typically a regression equation, to predict the primary outputs of generic PbQoL instrument using data on other measures or indicators of health (282). The regression equation is then the mapping algorithm which can be used to predict the PbQoL value in a dataset which contains the source measure but not the PbQoL measure. There is a growing trend of exploring and applying mapping algorithms where the utilities are not directly measured in studies. Mapping is recommended by NICE to estimate EQ-5D utility data when EQ-5D data are unavailable in the study dataset (283). Around one quarter of the QALY estimations informing recent NICE appraisals in England and Wales involved the implementation of a mapping algorithm (284). A database of mapping studies has been developed by Health Economics Research Centre (HERC) at University of Oxford (current version 5.0) which provides a readily-accessible collection of all studies mapping to EQ-5D (285).

As use of mapping algorithms becomes increasingly common, a growing number of researchers show their concerns regarding its development, reporting and application in practice. In 2015, the ‘MAPS’ (Mapping onto Preference-based measures reporting Standards) statement has been developed which is a checklist to promote transparent reporting of mapping studies. However, besides poor quality of reporting, there are many fundamental issues related to mapping that are not yet been addressed which can be summarized to three aspects: inaccuracy of utility predictions for poor health states, lack of instructions on the generalisability of the mapping algorithms from the authors, and failure to capture uncertainty around means and the variability across individuals (284, 286, 287). In addition, the mapping function relies on statistical association which is based upon the conceptual overlap between the source measure and the target measure. It was argued that mapping may not be appropriate for measures that have different construct, such as between the HrQoL measures and the wellbeing measures. An example of this is to map the Adult Social Care Outcomes Toolkit (ASCOT) / ICECAP to EQ-5D, which may not be appropriate since EQ-5D would be unable to reflect many of the outcomes captured by the wellbeing measures (169). These issues are likely to introduce important biases when using those mapped utilities in economic evaluations to compare alternatives.

2.6.2 Valuation of condition-specific QoL measures

Besides the mapping, another attempt to overcome the limited sensitivity of the generic measures in some specific populations is to construct condition-specific PbQoL measures (CS-PBM) (237). These measures can be developed from existing QoL measures in a specific area or developed de novo. Examples include the AQL-5D developed from the Asthma Quality of Life Questionnaire (AQLQ) for asthma (288) and the EORTC-8D developed from the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC-QLQC30) for cancer (212).

Although such CS-PBMs are able to achieve great precision and sufficient coverage to reflect what is suffered by the patients in the specific condition, they also face criticism. Researchers are concerned that CS-PBMs are sometimes insensitive to measuring the side-effects which have different symptoms and impact on QoL from the condition, and lack of comprehensiveness in people with comorbidities due to the narrow scope (234, 289). These may cause issues leading to bias in the values elicited. One such issue is the preference interaction whereby other important aspects of QoL that are not included in the CS-PBM may interact with the included aspects thus causing the coefficient (weights for dimension and level) to change (290). This implies that a preference-based measure should contain all of the important aspects of QoL into its descriptive system, which might be unattainable for a CS-PBM. Another issue is focusing effect whereby respondents overemphasize the dimensions included and ignore other aspects of life (291). In addition, in the same way as the other solutions that go for a ‘condition-specific’ approach, CS-PBMs share the same criticism that they would lose comparability across disease areas (289). However, some argued that comparability should be achieved by the use of a common numeraire such as money or a year of full health

(118). This means the comparability is not affected as long as the valuation was conducted using the same technique with common anchors, and from the same type of respondents (118). Despite the criticisms, the development of CS-PBM is argued to be valuable as it enriches the database of utilities measured by different approaches in a disease area where there exist limitations with current methods (289) and may provide valuable supplements to existing generic measures (291).

2.6.3 Bolt-on attributes

Besides mapping and CS-PBM, another 'specific' approach is to add condition-specific 'bolt-on' attributes to the generic measures (235). These 'bolt-on' attributes are designed to cover the dimensions missing from the generic measures such as EQ-5D-3L without compromising the comparability across disease areas. Cognition (292), sleep (22), vision (235), hearing (235) and tiredness (235) have been explored as bolt-on dimensions to the EQ-5D-3L in the literature.

Longworth et al. (2014) developed three 'bolt-on' items related to hearing, tiredness and vision to the EQ-5D-3L and valued them along with three health states (i.e. mild, moderate and severe) defined by EQ-5D-3L using the TTO method (235). They found that each of the bolt-on items had a significant impact on at least one EQ-5D-3L health state. The magnitude and direction of the impact varied according to the relative level (i.e. severity) of the bolt-on item compared to the health state to which it was added. The addition of a relatively severe 'bolt-on' tends to lead to a decrease of the health state values and addition of a relatively milder 'bolt-on' would result in an increase of the health state values.

For the comparability issues, the bolt-on approach is claimed to have a lower degree of inconsistency than the CS-PBM by retaining the EQ-5D as the core basis for measurement and by using a common valuation methodology. The research on bolt-ons is still at early stage and hence it is not yet clear what the valuation approaches should be the best, e.g. whether a full valuation of the EQ-5D plus bolt-on is required for each new bolt-on item (235), what the capacity of a valuation model is for the 'bolt-on' items if many items have to be added (169). Another issue is double counting as the bolt-on dimension may have already been captured to some extent by the existing generic dimensions of EQ-5D (169), e.g. vision can affect mobility, usual activities and self-care. A more fundamental limitation is related to the evaluative scope of the EQ-5D which cannot be addressed by the addition of one or more missing dimensions (169).

2.6.4 Wellbeing measures

The specific approaches as alternatives to the generic preference-based measures should improve the sensitivity of the measure to the specific conditions, yet they cannot avoid the criticism on comparability issues across programmes. Another option is the use of generic wellbeing measures. The wellbeing measures broaden the scope of the measurement which can capture the full impact of health from an overarching level. Although it is argued that the impact of health on wellbeing has been considered in the valuation process, evidence has shown that the respondents have limited ability to predict the impact of the health state on wellbeing in the preference elicitation process (293, 294). As a result, some argue that a more direct measurement of wellbeing is required (294).

Subjective wellbeing (SWB) has been described under three headings: hedonism (pleasure), fourishing theories (fulfilments) and life satisfaction (295). Although there are a number of wellbeing measures, only a limited amount of research has been done to explore how to use the wellbeing measures in economic evaluations. Very few of the wellbeing measures are preference-based, among which are the capability measure ICECAP (Investigating Choice Experiments Capability measure) (140) as mentioned previously in Section 1.5.3 and later in Section 2.7 and the Adult Social Care Outcomes Toolkit (ASCOT) (296).

ICECAP capability measures are based upon Sen's capability theory which focuses on what a person is able to do rather than what a person actually does (136). The ICECAP measures cover a broad range of psychological wellbeing along with enjoyment. The measures (ICECAP-O for older people, and ICECAP-A for adults) are valued using BWS (i.e. best worst scaling) method (141, 297). Section 2.7 will discuss in depth the capability approach and ICECAP. Also based on Sen's capability approach, the ASCOT is a social care-related QoL measure which aims to assess the extent to which an individual's social care needs and wants are being met. It contains eight domains: control over daily life, personal cleanliness and comfort, food and drink, personal safety, social participation and involvement, occupation, accommodation cleanliness and comfort, and dignity (296). Similar to the ICECAP, ASCOT preference score is developed using the BWS method, which can be anchored onto the QALY scale with 0 representing being dead using the TTO technique, and thus ASCOT can be used in economic evaluations (296).

The inclusion of wider aspects into economic evaluation depends on the debate about whether the NHS should be primarily concerned with promoting health, or some broader notion of wellbeing for the purpose of its resource allocation. On one hand, it has been argued in a substantial amount of literature that the QALY should not be health-only, since this will overlook the benefits of the interventions beyond health, such as freedom, strength of relationships, etc (298). On the other hand, some argue that SWB may suffer from memory bias or involving too much subjectivity, making it less useful to be used for health resource allocation (299). In addition, some evidence showed that SWB was not as responsive to the health status changes to the same extent as the generic HrQoL preference-based measures, leading to more doubts on the use of wellbeing measures for economic evaluations (299, 300).

2.7 The ICECAP-O instrument

As introduced in Section 1.5.3 and discussed in Section 2.6.4, a possible solution to the limitation in evaluative scope of the current generic PbQoL measures is the wellbeing measures that are developed based on Sen's capability approach. An attempt to measure capability in health and social care is the ICECAP instrument, which is recommended by NICE's latest guidelines on social care and public health interventions to measure broader benefit (117, 142). It expands the evaluation space to consider whether a programme enhances an individual's capability and wellbeing. The next section will introduce Sen's capability approach and one of the ICECAP measures, the ICECAP-O for older people. The application of the ICECAP-O in economic evaluations will be discussed at the end.

2.7.1 Sen's capability approach

Sen's capability approach advocates the evaluation of programmes focusing on capability (what a person is able to do) rather than functioning (what a person does) (136, 139). Sen has argued that actual achieved wellbeing can be assessed by ‘functioning’ measures, which should be differentiated from a person’s ability to achieve. Bleichrodt and Quiggin (2013) (301) explained that the capability approach distinguished between ‘means’ and ‘ends’; that is, only ends are important and means are instrumental in reaching the ends. Good health therefore has two-fold meanings: it is a means to achieve other functionings, such as working and leisure, but it is also an end itself. This constitutes a key distinguishing feature between the expression form of the capability approach (e.g. ICECAP-O questionnaire) and the QALY approach (e.g. the EQ-5D-3L/5L). The former views health as a means to achieve, for example, the ICECAP-O include dimensions like attachment (love and support from family and friends) and role (doing things that make you feel valued), while the latter focuses on health as an end goal, for example, the EQ-5D -3L/5L include attributes such as mobility and pain/discomfort.

Capability was defined by Sen as freedom of choice to achieve functionings, and can be viewed as the set of potential combinations of functionings from which an individual could choose to live (139, 302). Cookson (2005) (303) argued that the emphasis on the choice for functioning differentiates the capability approach from the conventional welfare approach as it relaxes the assumption of rational self-interest. That means, in the capability approach, an individual does not necessarily choose the option with the best value. Wellbeing can be improved when additional choices are provided even if the option with the maximum value already exists, while in the welfare approach, the utility of a set of functioning is determined by its most valued element and thus in this case utility won’t change with the additional less-valued option (135, 304). A widely quoted example is the ‘fasting-starving’ distinction (305, 306). Someone voluntarily fasting may have the same nutritional intake as someone who is starving, however, the person who is fasting has the freedom to choose to fast or eat whereas the starving person has no choice. The notion of capability considers the freedom of choice for achieving actual functioning rather than whether the functioning has been achieved.

Although Sen’s capability approach has been criticised for being highly conceptual (307), it has contributed to several theoretical and practical development in health economics (125). It influenced the development of ‘extra welfarism’ (308) with enriched evaluative space. The main features of the extra welfarism are permitting outcomes other than utility and taking into account the sources of valuation from other people other than the affected individuals (309). This is distinguished from the traditional 'welfare' which focuses on maximizing a social welfare utility function. In practice, it is argued that this complex and broad extra-welfarist approach has been narrowly implemented with utility replaced by health and function as the only outcome but losing other and broader outcomes (138, 310). The focus of CUA on health represented by the generic preference-based HrQoL measures limits the broader outcomes of the health care and social care intervention to be incorporated into economic evaluations (171).

Given that 'extra-welfarism' in practice has not incorporated broader benefit, Sen's capability approach has been promoted in the last decade as an alternative to broaden evaluative space and several instruments have been developed, including the ICECAP and ASCOT (as mentioned in last section). In addition, criticisms are raised that beside 'health', the capability, i.e. the ability of achieving functioning, is also of importance to people. This concern was reflected with the empirical evidence shown by Grewal et al. (2006) that older people in the UK appeared to be concerned about their (lack of) ability to meet particular 'functionings' (140). This has led to the theoretical and empirical development of ICECAP-O which draws directly on the capability approach (138) in contrast to the utility approach.

2.7.2 ICECAP-O

ICECAP-O considers wellbeing in a broader sense than health itself and therefore could potentially be used in economic evaluations across health and social areas in which a broader set of outcomes is considered (140, 311). It contains five capability attributes which are identified through qualitative in-depth interviews with older members of the British public (140). The attributes are:

- *Attachment* which incorporates feelings of love, friendship, affection and companionship, sources of which appear to include partners, family, friends, and pets'

- *Security* which 'incorporates ideas of feeling safe and secure, not having to worry and not feeling vulnerable'

- *Role* which 'incorporates the idea of having a purpose that is valued, either by the individual and /or others'

- *Enjoyment* which 'pulls together notions of pleasure and joy, and a sense of satisfaction, sources of which include personal and communal activities'

- Control which 'involves being independent and able to make one's own decisions'.

By asking respondents whether they 'can have…' or 'are able to..', the ICECAP-O is aligned with Sen's capability approach of focusing on the freedom of choice. The breadth of the dimensions reflects its wellbeing theme. Each attribute contains one question with four levels (no, a little, some, and a lot of capability), thereby distinguishing 1,024 possible 'capability states' (140). The value set was developed using the BWS method from the UK older adults whereby the respondents were asked to choose the best and worse scenarios from a selection of methods. From these choices values for the capability were derived. The values were anchored between 0 (no capability) and 1 (full capability) and did not make assumptions about where death fell on this scale (141).

A variety of studies have evaluated the psychometric properties of ICECAP-O in different populations. Its construct validity was tested and demonstrated among the general population in the UK (312), post-hospitalized older people in the Netherlands (313), in the general population (314) and the older post-acute patient population in Australia (315), in a falls prevention clinical setting in Canada (316), and among dementia patients at a nursing home in Germany (317). Its face validity was assessed and demonstrated in hip and knee arthroplasty patients (318). Recently, a study from the Netherlands assessed the ICECAP-O and demonstrated its test-retest reliability, construct validity and responsiveness in frail older adults (319).

2.7.3 Use of the ICECAP-O in economic evaluation

The ICECAP-O was developed to capture broader benefit for economic evaluations, however little guidance is provided on how such measure should be used to aid healthcare resource allocation decisions. A fundamental issue is how to combine

ICECAP-O with length of life (320). The reason why the preference-based measures such as the EQ-5D can be combined with length of life to calculate QALY is that the valuation of health states is against death which is given the value of 0. This allows the interpretation of QALY to be that 'zero QALY means years of death' since it is generally accepted that the absence of health is the same as the absence of life. However, the ICECAP-O is valued with zero representing 'no capability' rather than death, which means ICECAP-O adjusted length of life cannot generate the QALY with the same conventional meaning. Rather, it generates a new concept, called Years of Full Capability (YFC). Zero YFC means years of no capability rather than years of death. Some suggest that it could be assumed that those who die have no capability, or more conservatively, death is among the states that have no capability (320). With this assumption, a person who dies would have zero YFC.

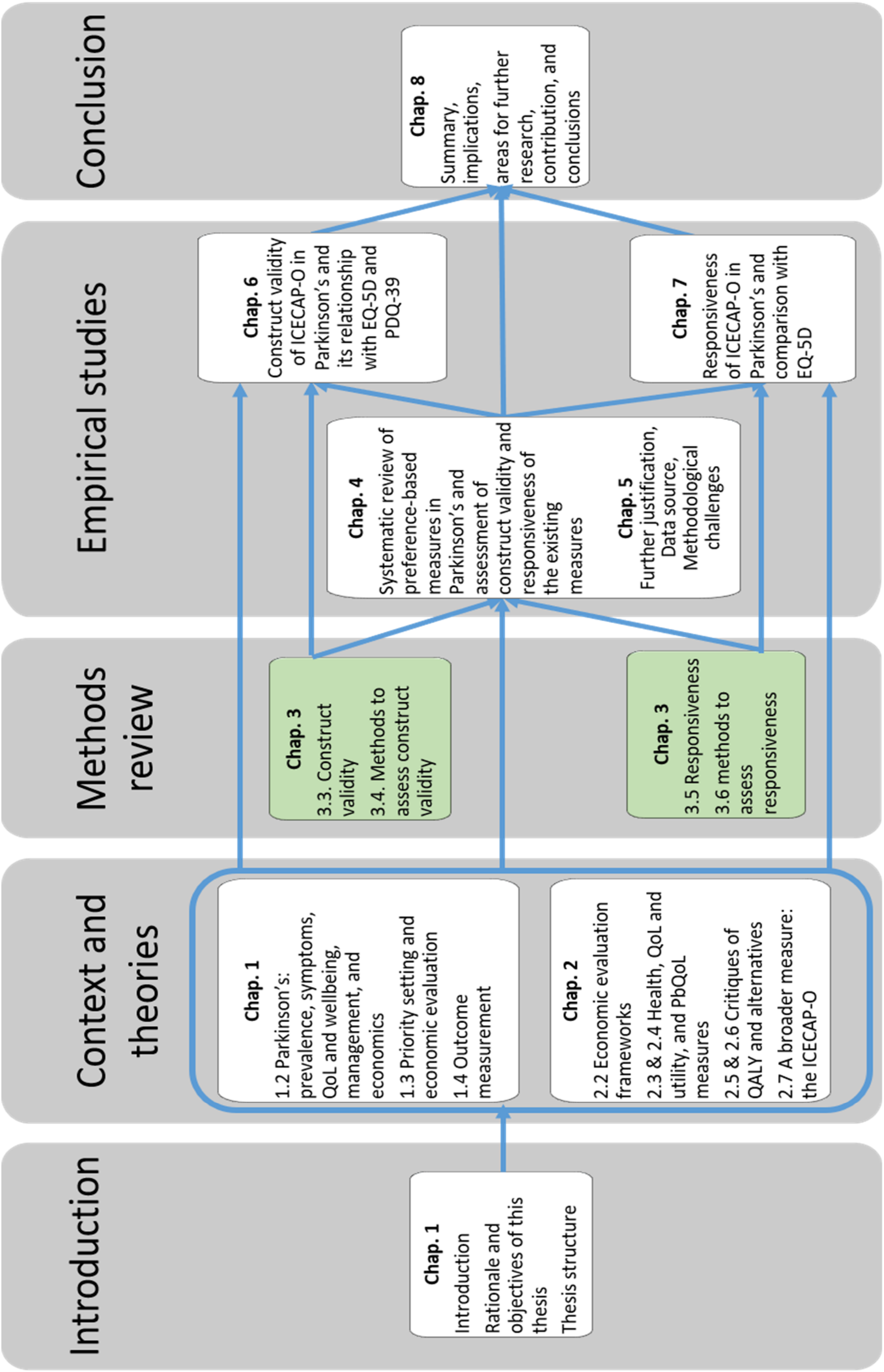
Another concept undergoing development is the Years of Sufficient Capability (YSC), whereby the length of life is adjusted by the amount of capability that deemed to be sufficient for the consideration of equity (321). Existing approaches to economic evaluation focus on maximising outcomes, irrespective of the distribution of outcomes within society. In contrast, the capability approach particularly has been concerned with equity as it focuses on what a person is able to do rather than what a person actually does (321). For example, a better-off person may not do a lot of leisure things but he/she has the ability to do it while a poor person does not have the ability to do them. Therefore, decision making using the capability approach might aim to provide a "decent minimum level of capability for as many people as possible, and thus focus on the distribution of capability not its maximisation" (Coast 2008) (125, 137). Kinghorn conducted a qualitative study applying deliberative methods to establish a sufficient level for capability and found the sufficient capability to be 33333, i.e. level three (feel capable in *many* areas) for all attributes (322). This distinguishes the concept of 'sufficient capability' from the 'full capability' (level four for all attributes) and established the basis for its use in decision-making in the contexts of public health and social care.

Once the YFC or YSC is generated, it would require a decision threshold to judge if the intervention is value for money. For QALY, the cost-effectiveness threshold in the UK set by NICE is £20,000 to 30,000 per QALY gained. This threshold reflects the amount of willingness to pay by the society for each additional QALY provided by a new intervention. Establishing societal willingness to pay for a year with sufficient capability would be more difficult due to the complex nature of the concept. This work is ongoing by Dr. Kinghorn at the University of Birmingham (320).

2.8 Chapter summary

This chapter provides an overview of outcome measurement for economic evaluations. Upon reviewing the forms of economic evaluations, concept of health, QoL and utility and existing approaches employing the generic preference-based HrQoL measures, limitations around their use are discussed and alternative methods including the capability approach are provided. The ICECAP-O instrument offers a broader evaluative space than the current HrQoL preference-based measure EQ-5D, which shows potential to be an alternative as a preference-based outcome in populations with diseases such as Parkinson's that have a broad impact on people's wellbeing or complex interventions such as those in public health and social care. The next chapter will introduce the criteria for assessing the appropriateness of outcome measures which will be applied in the empirical works of this thesis.

Chapter 3 Construct validity and responsiveness: theory and assessment methods



Introduction
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Context and theories
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Methods review
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Empirical studies
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Conclusion
Chap. 8
Summary, implications, areas for further research, contribution, and conclusions

3.1 Introduction

Chapter 2 introduced a variety of outcome measures used within the different economic evaluation frameworks, along with their merits and limitations. There are varied opinions on the subjective concept of QoL and what should be included in PbQoL measures being considered for resource allocation purposes. Consequently, the choice of the measures to be used in the intervention studies has often become a point of debate. Are the PbQoL measures suited to capture the health, QoL and wellbeing aspects that are valued by people? In other words, are the PbQoL measures appropriate to be used in the population of interest? This echoes the overarching research questions of this thesis.

Construct validity and responsiveness are important properties for preference-based measures to exhibit (118, 148). A PbQoL measure with limited construct validity or responsiveness would generate unreliable utility values for different health states, which would eventually affect QALY calculations in economic evaluations. In particular, responsiveness of a PbQoL measure means that the utility profile is able to reflect the change in health state caused by the intervention that are deemed to be important for the patients. Lack of responsiveness of the PbQoL measure may lead to a false judgement of an effective intervention being not cost-effective, whereas the truth may be that the PbQoL measure could not fully capture the intended benefit of the intervention. Consequently, the rigor of the economic evaluations will be undermined and its role in health care decision-making will be weakened. NICE clearly emphasizes the importance of assessment of these two properties when there is a doubt on the use of EQ-5D in specific populations in its ‘Guide to the methods of technology appraisal’ (119). It states that construct validity and responsiveness in a particular patient population should be investigated through a synthesis of peer-reviewed literature to support the claim that the EQ-5D may not be the most appropriate in some circumstances (323).

Cautions should be made when applying the classic psychometric testing methods to the PbQoL measures (148, 246). The purpose of a PbQoL instrument is to measure all differences or changes in health state that are important to patients and valued by public. As introduced in Chapter 2 (Section 2.4), for a PbQoL instrument, an improvement in a defined state being ‘important’ means that public would like to trade their length of life for this positive change, or accept higher risk of death for this positive change. The PbQoL measures are developed and valued incorporating people’s preference and thus assumptions are made in the testing regarding the resulting values rather than simply aggregated scores (i.e. weighting attributes equally and equal difference between levels). The interpretation of results should take account of all the required assumptions in the valuation process. Methodological considerations will be discussed in this chapter following the description of each method and will be further summarized in Chapter 5 (Section 5.4) before their applications in the case studies in Chapter 6 and 7.

This chapter will start with a brief introduction of psychometric properties in general and then focus on construct validity and responsiveness. Assessment methods for these two properties introduced in this chapter are used in Chapter 4 as criteria for critically appraising the existing preference-based measures identified through the systematic review, and also Chapter 6 and 7 for empirically testing these two properties of the ICECAP-O capability measure.

3.2 Overview: psychometric properties

Whether a QoL measure is appropriate to be used in a given context depends on its psychometric properties. Psychometric validation tests whether a QoL instrument is performing in the way expected. These methods were initially developed in the field of psychology and used in areas such as behaviour testing, personality, and beliefs, and they now extend to measures of QoL. Their importance is increasingly emphasized by health economists to evaluate PbQoL measures (148). Measurement of PbQoL can be described as “the process of linking abstract concepts to empirical indicants” (Carmines 1979) (305, 324) given the intangible, patient self-reported nature of the concept of PbQoL. As a result, testing the psychometric properties of the instruments measuring such abstract concepts are important for PbQoL measures to be trusted when being used in economic evaluations for decision-making.

Psychometric validation is the “process by which an instrument is assessed for reliability and validity through the mounting of a series of defined tests on the population group for whom the instrument is intended.” (Bowling 2014) (325). Reliability and validity are the two basic domains for assessment, while there are additional domains proposed in the literature such as practicality (118), and responsiveness (326).

3.2.1 Reliability

Bowling (2014) defined reliability as meeting two criteria: ‘reproducibility’ (the degree to which it is free from random error) and ‘internal consistency’ (the homogeneity of the instrument) (325). Mokkink et al. (2012) also included these two criteria but defined reliability more generally as “the degree to which the measurement is free from measurement error” (327). Mokkink et al. further clarified the use of the criteria of ‘internal consistency’ that it is only relevant when the measure is constructed in a ‘reflective model’, but not when the measure is constructed in a ‘formative model’.

3.2.1.1 Reflective model versus formative model

‘Reflective model’ and ‘formative model’ are specific types of measurement models, which describe the relationship between a construct and its indicators / items (328). The terminology of formative and reflective models was introduced into the health sciences in the 2000s by Fayers and Hand for the measurement of QoL (329). In a reflective model, all items are a manifestation of the same underlying construct (330) and hence they are expected to be highly correlated and homogeneous (330, 331). In contrast, a formative model applies to the construct in which the items together form a construct and thus it is not necessary for the items to be highly correlated (327). A way to differentiate between the two types of framework is to judge whether the items would change when the overall construct changes (327). Change in overall construct is expected to lead to changes in *all* items in a reflective model but not in a formative model. It was suggested that instruments measuring perceived health or HrQoL are usually in the form of a formative model and hence the test of internal consistency would not be relevant (327).

For example, to measure HrQoL, EQ-5D instrument contains mobility, pain/discomfort, anxiety/depression and so on, where these dimensions represent the multi-faceted nature of the HrQoL concept and are not interchangeable. The items are not highly correlated (e.g. a person with extreme anxiety and depression may have no physical discomfort or disability at all). This explains why the EQ-5D is constructed in a formative model rather than a reflective model and only 'reproducibility' should be assessed but the 'internal consistency' is irrelevant. This is in line with Brazier et al's (2017) definition of reliability for testing PbQoL measures, which states that "reliability is the ability of a measure to reproduce the same value on two separate administrations when there has been no change in health". This definition eliminates the 'internal consistency' and keeps 'reproducibility' only (118).

3.2.1.2 Reproducibility

Three aspects of reproducibility are usually assessed: test-retest, inter-rater, and intra-rater reproducibility. Test-retest assesses the stability of the measure over a period of time during which what is measured is not expected to change (332). It is examined by presenting the same data repeatedly within a period of time to a single rater. Good test-retest reliability is represented by the same or highly similar measurements. The key to this methodology is to ensure that there is no actual change over the period of time so that any discrepancy between the two measurements can only be attributed to error.

Both Interrater and intra-rater reproducibility (commonly called interrater and intra-rater reliability) concern with raters: interrater examines the degree to which the results obtained from two or more raters agree with each other while intra-rater assesses the agreement between the repeatedly obtained results from the same rater (333). Interrater reliability investigates the (in)consistency among individuals since human observers may have variable individual experience and thus interpret the phenomena differently. Examples include scoring injuries by different observers using MRI grading and prognostic parameters (334), counting 2-minute push-up repetitions that meet the push-up protocol by different raters (335), etc. Intra-rater reliability examines if an individual interprets and records the data the same when the exactly same data are presented. For example, the study of assessing the 2-minute push-up test examined the intra-rater reliability by making a single rater counting video-taped push-up repetitions repeatedly with a minimum 1-week apart (335). The within-individual agreement in this study was found to be not ideal, ranging from 41.8% to 84.8%.

3.2.2 Validity

Validity is the extent to which an instrument measures what it purports to measure. The word ‘valid’ is derived from the Latin word ‘validus,’ meaning strong (336). In that sense, validity requires an instrument to reflect strongly what it claims to measure. Validity has been more complexly defined as “an overall assessment of the degree to which evidence and theory support the interpretation of the scores entailed by proposed uses of the instrument” (Krabbe 2016) (p113) (337). In some literature this conventional meaning is referred to as internal validity to distinguish from external validity which refers to the generalisability of the research findings to the wider population (338). Validity testing may not seem to be complicated for the measures of observable outcomes, e.g. temperature measured by a thermometer, however, for the unobserved concept, for instance, QoL, life stress, testing validity is a prerequisite to their use. This is because the measurement of these factors is dependent upon their definitions, which may vary according to individual’s perceptions or preferences and the way the perceptions or preferences are being measured (339). This may lead to different results yielded by different instrument although they may claim to measure the same concept, raising the question of which instrument is valid.

3.2.2.1 Face validity and content validity

Validity has many different components. Face validity and content validity both assess whether the descriptive system of the measure is relevant, logical and sensible for the population. *Face validity* is more ‘superficial’ which is, according to some, e.g. Bowling (2014), based on investigators’ subjective assessments (325) and to others, e.g. Holden (2011), based on the respondent’s perspective (340). In contrast, *content validity* assessment is usually conducted by an expert panel using a more systematic approach which also assesses the comprehensiveness of the instrument in addition to the relevance of the items (325).

3.2.2.2 Criterion validity

Criterion validity assesses the correlation of a measure with another measure of the same trait under study, ideally, a 'gold standard' accepted in the field (339). Sometimes, the criterion may have drawbacks such as being expensive and invasive for a diagnostic measure, or time-consuming for a questionnaire, and as such a new measure is intended to reduce these burdens. Criterion validity has two types: concurrent validity and predictive validity; the distinguishing feature between the two is the timing of administering the criterion (339). To assess *concurrent validity*, the criterion and the new measure are given at the same time whereas for *predictive validity*, the criterion is given at a later time to examine how well the new measure can predict the criterion. Examples of application of criterion validation include medical diagnostic measures using concurrent validation to test if the new diagnostic procedure under scrutiny can provide the same diagnosis as the reference standard, and in school admission context to test if the criteria for admission can predict the performance in the school using predictive validity. However, in the field of QoL, criterion validation has limited application due to its requirement of an existing 'gold-standard' measure. The new ankle-brachial pressure index designed to detect arterial disease in the leg can be compared against the gold standard of venography (341). Similarly, students' performance in school can be measured by their scores of standardised exam with the highest score representing the best performance for each subject. In contrast, due to the different views on the definition of QoL as mentioned in Section 2.3, individuals have varied perceptions of what representing the ideal status of QoL and thus there is no standardised criteria.

3.2.2.3 Construct validity

Another important component of validity is *construct validity*. It considers whether the instrument is measuring the underlying concept it purports to measure. Construct of a measure relates to the hypothesized manifestations linking the underlying factors and a person's behaviour (339). The underlying factors are referred to as hypothetical constructs (339). In psychology, these hypothetical constructs are explanatory variables which are not directly observable. This is distinguished from other sciences where a construct is a real existence, for example, the natural sciences contain constructs such as gravity, temperature, and pressure whereas the behavioural science contain construct such as motivation, intelligence, self-esteem, etc (342). In QoL studies, construct refers to the unobservable / hypothetical factors that contribute to the concept of QoL (343).

All the components mentioned in this section can be organized to ‘three Cs’ according to Landy’s ‘trinitarian’ point of view (1986), i.e. content validity, criterion validity and construct validity (339). The three Cs are seen as three relatively independent attributes of a measure. Among the three Cs, only construct validity can be empirically and quantitatively tested for a QoL measure. This will be described in-depth in section 3.3 and 3.4.

3.2.3 Other properties

Besides the above-mentioned components in validity and reliability, a number of other issues have also been purported to be important concerns for assessing the performance of any measurement instrument (339), and are deemed important for PbQoL measures (118). Practicability considers the acceptability of the descriptive system to the respondents and the cost of administration (118). Responsiveness refers to the ability of an instrument to be responsive to change. Some argue it is a special form of validity and should be covered under the umbrella term of validity since the ability to measuring change is essentially a discriminant validity between different states of what is being measured (327). Responsiveness will be primarily focused in Section 3.5 and 3.6 in this chapter.

An issue related to responsiveness is floor and ceiling effects. Ceiling effect occurs with tests or scales “that are relatively easy, when a substantial proportions of individuals obtain either maximum or near-maximum scores” (Uttl 2005) (344). On the contrary, floor effect occurs when the test is difficult and as a result a substantial proportion of individuals produce the minimum possible score (345). The existence of ceiling or floor effect of a PbQoL measure will result in score distributions that are compressed at the upper or lower end of the scale and thus cannot reveal any differences among the individuals that scored the highest or lowest of their utility values. The EQ-5D-3L is an example which has been shown to exhibit ceiling effect (full score recording perfect health) in both general and disease-specific populations, leaving less space for improvement in response to an intervention (346, 347). In contrast, SF-6D (introduced in Section 2.4.3.1) is often found to exhibit floor effect for patient groups in severe health where a significant number of patients report the lowest level of health possible for some dimensions (348-350). Ceiling effects can be also understood as “when one can be better than can be captured by the measure’ and floor effect is ‘when one can be worse than the lowest score in the range of the measure” (Feeny 2012) (351). When a measure has ceiling or floor effects, its responsiveness to change would be threatened since there is no space for the score to move up when the baseline value is the highest on the scale or vice versa.

3.2.4 COSMIN checklist

As shown above, the literature in psychometric properties contains varied opinions regarding what criteria are important when selecting an instrument. To address this, in 2010, a group of international experts reached consensus on the criteria to evaluate the performance of health related patient reported outcomes, and developed a critical appraisal checklist in a Delphi study, named the COSMIN checklist (352). COSMIN stands for Consensus-based Standards for the selection of health Measurement Instruments, which contains standards for evaluating the methodological quality of studies related to the measurement properties of health related patient reported outcomes (326, 327). Three assessment domains are distinguished in COSMIN checklist, i.e. reliability, validity, and responsiveness. Figure 3-1 presents the components under each domain.

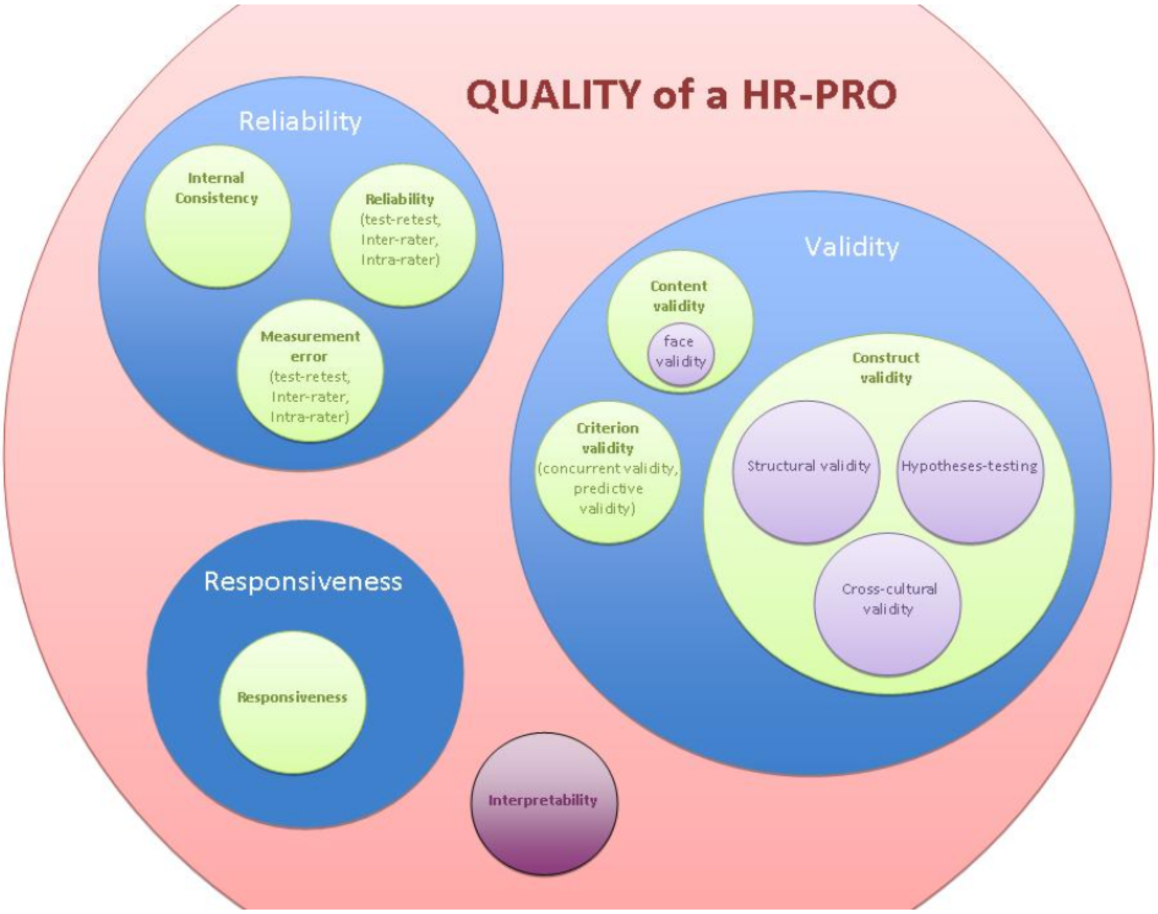


Figure 3-1: Relationship between measurement properties.

Source: Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. J Clin Epidemiol 2010;63:737-745. 2012.(352) Permission to include this figure in this thesis has been granted from the publisher, Elsevier (Appendix G).

The 'reliability' domain contains internal consistency, reliability, and measurement error. The 'validity' domain covers three measurement properties: content validity which also contains face validity, criterion validity and construct validity. The 'responsiveness' domain contains only one measurement property, which is also called responsiveness. Besides these measurement properties, interpretability is also listed as an important characteristic of a measurement instrument which considers the degree to which one can easily interpret the quantitative score by clinical or commonly understood connotations. The definition of 'construct validity' and 'responsiveness' defined by COSMIN checklist will be mentioned in Section 3.3 and 3.5. COSMIN checklist provides a detailed guidance on how these measurement properties should be evaluated in terms of study design and statistical analysis and critically discussed the associated issues. It can be used when selecting a measurement instrument and designing or reporting a study on measurement properties.

An important note from the authors is that this checklist is to evaluate the *quality of studies* on measurement properties of a HrQoL instrument, rather than determining what constitutes *good* measurement properties of *instrument (326)*. The criteria for determining the adequacy of measurement properties were discussed in the Delphi discussion however consensus was not achieved. This relates to the limitation in the interpretation of the results of case studies of this thesis and the proposition of future research in that a clear guidance is needed to determine the degree of construct validity or responsiveness is adequate. The lack of consensus among the international experts indicates that determining what degree of the properties should be judged adequate is very challenging. This will be further discussed in Chapter 8 (Section 8.6.3).

Among the many psychometric properties, content validity, construct validity and responsiveness are recommended by NICE to judge the appropriateness of the currently recommended EQ-5D for specific populations (323). Furthermore, construct validity and responsiveness can be empirically tested and are essential for economic evaluation which requires an instrument to be able to differentiate between different health states or responsive to the change of health states over time (118).The following sections, 3.3-3.6, will further discuss construct validity and responsiveness along with their testing methods.

3.3 Construct validity

As introduced in Section 3.2.2, in psychometrics, construct refers to the unobservable objects that are used to represent or explain a concept, and construct validity represents the ability of an instrument to measure the underlying concept it intends to measure (332, 353). Cronbach and Meehl (1955) stated that “construct validity is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined” (353). Consideration of construct validity is therefore necessary “whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured.” This section will introduce a brief history of the emergence of construct validity theory and the various views around the definition of construct validity.

3.3.1 History of construct validity

Construct validity is the last developed measurement test in history compared to criterion and content validity owing to the gap identified in psychometrics (339). Assessment of validity was dominated by criterion validity prior to 1950s however it cannot be used in an area without a criterion (354). Content validity has limitations as it does not provide inferences quantitatively about the validity of test scores (355). Consequently, there was a gap in methods to assess the usefulness of a scale in clinical psychology where there is no criterion but uses quantitative scores. In health, scales for physical symptoms are objective as the symptoms are mostly directly observed in contrast to the scales for psychological aspects which are subjective as the aspects are invisible, such as attitudes, feelings, depression. To fill this gap, Cronbach and Meehl (1955) introduced the concept of construct validity as “a framework of hypothesis testing based on the knowledge of the underlying construct” (p230) (339). In other words, when there is no ‘gold standard’, criterion testing is replaced by hypothesis testing about the relationship between underlying construct and the observed outcomes, which constitutes the basis of construct validation. Since then, construct validity, together with content validity and criterion validity, gradually became the three key criteria for testing of an instrument. Zumbo and Chan (2014) conducted a systematic review to identify the trend in the number of publications of validation studies since 1960s. Figure 3-2 shows that a clear increasing trend was identified in both overall number and the number in life satisfaction, wellbeing and QoL area (356).

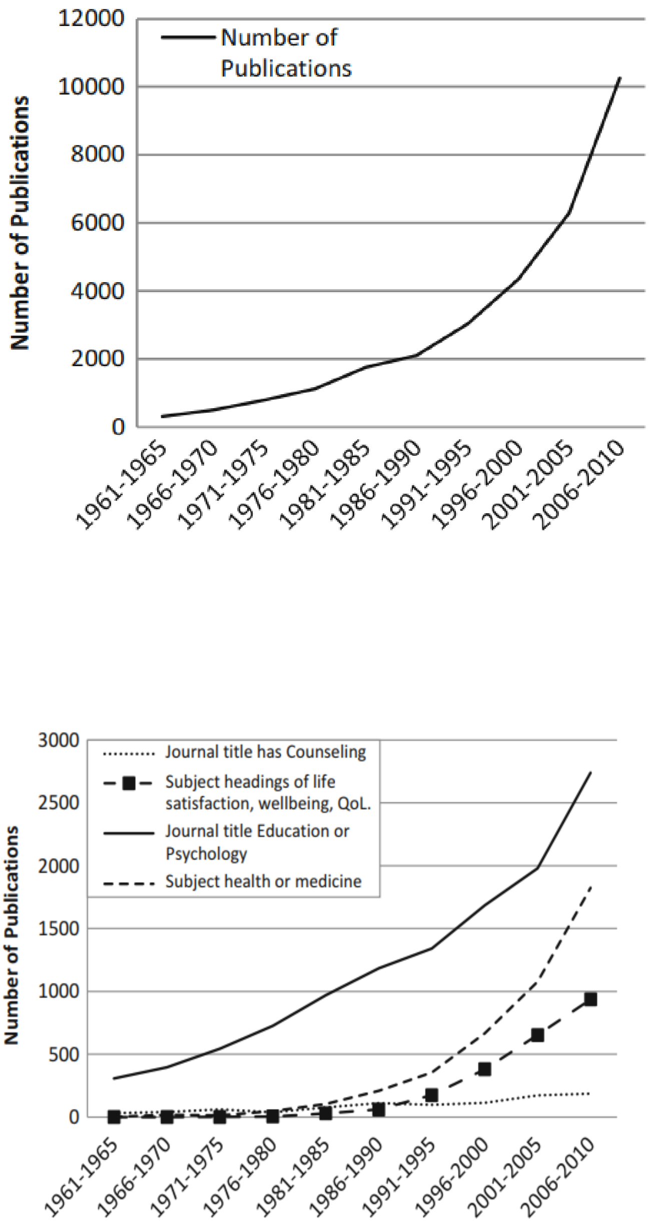


Figure 3-2: Trend line of number of publications of validation studies.

Source: Zumbo BD, Chan EKH. Setting the Stage for Validity and Validation in Social, Behavioral, and Health Sciences: Trends in Validation Practices. In: Zumbo BD, Chan EKH, editors. Validity and Validation in Social, Behavioral, and Health Sciences. 1 ed: Springer International Publishing; 2014. (356) Permission to include in this thesis has been granted from the publisher, Springer (Appendix G)

Construct validity of a measure is context-specific, i.e. a measure exhibiting construct validity in one population does not guarantee is construct validity in another population, but this was not seen until the late 1960s (339). Previously it was viewed as an intrinsic property of a scale rather than a varying property in different populations. In 1971, Cronbach (1971) shifted the focus of the interpretation of validity testing results from the measure's property to the characteristics of the people who were being assessed (357). Landy (1986) interpreted this focus change led by Cronbach as "validation process are not so much directed toward the integrity of tests as they are directed toward the inferences that can be made about the attributes of people who have produced those test scores" (p1186) (358). In other words, validation process is about the inferences, claims, or decisions that one can make based on the scores rather than whether the measure is valid itself (356). Validation process provides information about how the measure performs in the population being assessed.

3.3.2 Definition of construct validity

In the psychometrics literature, there is no consensus regarding the scope of the definition of construct validity. Some support the traditional view of treating it as a component of validity in the three C model (along with content validity and criterion validity) as originally published by Cronbach and Meehl (1955) (326, 353). Others, sometimes called revolutionary theorists (including Cronbach himself in later years (337)), developed novel views whereby construct validity is the overarching concern of validity research, encompassing all the other types of validity evidence (359).

Among the pioneers for modern views, Messick described construct validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on instrument scores." (355) Following Messick, the American Psychological Association handbook of research methods in psychology, written by Grimm and Widaman (2012), also considers construct validity as a comprehensive concept which is formed by two major axes, internal validity and external validity (360). Internal validity represents the extent to which an instrument measures the intended construct. It includes content validity, dimensionality, reliability and discrimination. External validity focuses on the relations between test scores and external criteria. It consists of criterion-related validity, convergent and discriminant validity, change validity, score interpretation and consequences (360). Grimm and Widaman's multi-faceted definition of construct validity reflects Messick's view of the inclusiveness of construct validity in that all measurement tests are based on hypothesis and construct interpretation underlies all score-based inferences (355, 361). From this sense, all the tests that claim to examine ‘validity’ are by nature testing the ‘construct’ of a measure and thus become construct validation and all the components, including content validity and criterion validity, can be seen as types of construct validity. Hypothesis testing is therefore a method for all validity testing regardless of the types.

Others maintained the traditional narrower scope of construct validity as a component of three C under validity as this framework is considered to be sufficient and clear (337) and for the convenience of understanding the different testing methods (339). With the narrow scope, construct validity is usually defined as having two components, convergent validity which examines how well it correlates with another measure of the same construct, and discriminant validity which examines whether it is possible to differentiate between groups thought to differ in the characteristics that the new instrument is supposed to measure(p185) (118, 332).

The COSMIN checklist is consistent with the three C model whereby construct validity is one of the three elements within validity. It defines construct validity as “the degree to which the scores of an HR-PRO (health related-patient reported outcome) instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the HR-PRO instrument validly measures the construct to be measured” (352). Construct validity is considered to include three aspects: structural validity, hypotheses-testing and cross-cultural validity, as shown in Figure 3-1 (327). Structural validity examines “the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured”. It is not considered to be relevant in a ‘formative model’ (as mentioned earlier in Section 3.2.1), in which the items together form the construct, such as the case of most QoL measures including ICECAP or EQ-5D-3L, in contrast to a ‘reflective model’ in which all items are a manifestation of the same underlying construct. Cross-cultural validity considers the external generalisability that “the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument”. It is only relevant to the translated or culturally adapted instrument.

Hypotheses testing is the basic spirit of construct validity which examines “the degree to which the scores of an instrument are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured”. (327) The research question for hypothesis testing asks ‘does the hypothesis of this validation study make sense in light of what the scale is designed to measure? ’ (339)

Regardless of the scope, as mentioned in the last section, validity is about the interpretation of the scores in context. From this sense, the validation process is an ongoing process in which various types of evidence are accumulated and synthesized to support the construct validity of interpretation of an instrument (356). It is further argued that any conclusion about a construct validity test is not dichotomous but is a question of degree (362). It is a question of *how* well or *poorly* the measure performs in the population rather than whether or not it performs.

3.4 Methods to assess construct validity

Assessment of construct validity is to test the hypotheses which are made in relation to the underlying construct. A key concern for Cronbach and Meehl (1955) was that theories concerning inferred constructs be tested with rigor (339). Rigor generally refers to soundness of method, design, and test construction. In this scenario, it refers to the quality of assessment methods, which contains the hypotheses one tests about a theory, the methodology for testing and the statistical methods to generate inferences (363, 364).

3.4.1 A five-step model for construct validation

Smith (2005) (365) proposed a five-step model for construct validation. This model is shown in Figure 3-3. In practice, the five-step model can be applied to any validation tests as they all require hypothesis testing (355). This model has been used in previous studies validating the ICECAP questionnaire (305).

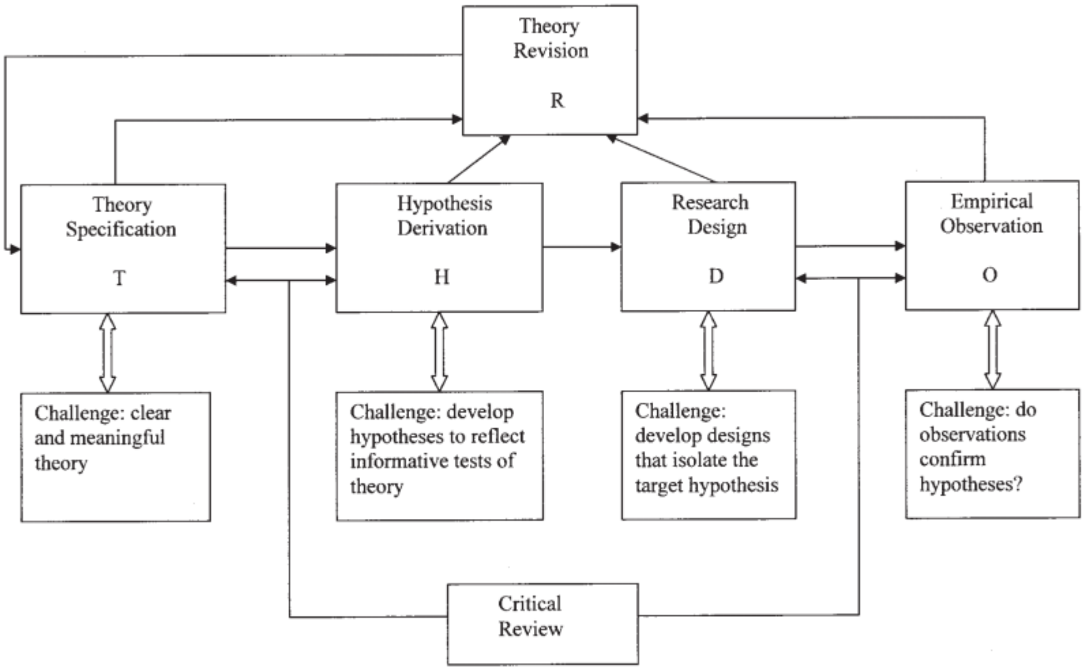

Theory Revision
R
Theory Specification
T
Hypothesis Derivation
H
Research Design
D
Empirical Observation
O
Challenge: clear and meaningful theory
Challenge: develop hypotheses to reflect informative tests of theory
Challenge: develop designs that isolate the target hypothesis
Challenge: do observations confirm hypotheses?
Critical Review

Figure 3-3: A five-step model for construct validation.

Source: Smith GT. On construct validity: issues of method and measurement. Psychol Assess. 2005;17(4):396-408.(365) Permission to include this figure in this thesis has been granted from the publisher, American Psychological Association (Appendix G).

The steps are:

(1) Theory specification (T): careful specification of the theoretical constructs in question,

(2) Hypothesis derivation (H): articulation of how the theory of the construct is translated into informative hypotheses,

(3) Research design (D): specification of appropriate research designs to test one's hypotheses,

(4) Empirical observation (O): articulation of how observations from samples pertain to one's prediction, and,

(5) Revision (R) of the theory and the constructs.

First, theoretical background for the construct of a measure to be tested should be clearly specified. This can be informed by the past empirical and theoretical work. Second, underpinned by the theoretical basis in step 1, hypotheses should be stated clearly, well justified, and be able to provide informative evidence about the test measure's ability. A good hypothesis should be able to "facilitate the ongoing process of critical evaluation that is the hallmark of science" (Weimer 1979) (366). Third, designing an appropriate study is crucial, as they should reflect what the hypothesis is intended for test. Inappropriate design will cause deviation from the hypothesis and misinterpretation of the results.

Fourth, empirical observations require the application of appropriate statistical methods to be able to make inferences about the test measure. The classic statistical method proposed in psychological testing for construct validity is the multitrait and multimethod matrix (MTMM) method firstly proposed by Campbell and Fiske (1959) (367), which is essentially correlation testing. In health economics, the primary approaches are the 'known-group' method and convergent validity test proposed by Brazier (1999) adapted from psychometric literature (148). All these methods will be described in depth in the next section. The statistical test should allow the evaluation of the degree to which empirical observations conform to hypotheses.

In the end, revision of the theory reflects that the "construct validation process is an ongoing, iterative process in which new findings and new theories clarify and alter existing theories, thus requiring new measures and new theory tests" (Weimer 1979) (366) . Assessment of an instrument's validity is gradually built up through accumulative evidence, contributing to people's understanding of the actual construct that an instrument can truly measure. The revision process allows the interpretation of the test measure in the context and thus reflects the value of the construct validation process.

3.4.2 Statistical methods for testing construct validity

Step 3 and 4 in Smith's five-step model is to design methods to empirically test the construct validity based on observed data. The classic method of testing construct validation in psychology is by examining discriminant validity and convergent validity (325, 353). *Discriminant validity* requires that a measure should not correlate with other dissimilar, unrelated measures and *convergent validity* examines the extent to which a measure correlates with another measure with the similar construct (353, 368). To examine these, the classic statistical approach is the MTMM method as mentioned above and the factor analysis method (332, 337, 360, 369, 370).

PbQoL measures are unique from non-preference based HrQoL instruments as discussed in depth in Chapter 2 (Section 2.4) since the former additionally contains a value set which is to be attached with the responses from people on the levels for each attribute in questionnaires. Owing to this, the index values of PbQoL measures are reflecting relative preferences that patients and others place on the dimensions and the items, rather than simply score aggregates from the ordinal responses of level to the dimensions. As such establishing the validity of preference-based measures was described by Williams as “chasing will o’ the wisp, and probably equally unproductive” (371). However, construct validity is all about hypothesis testing and thus construct validation should still be rigorous and meaningful providing the hypothesis is constructed in a certain way to incorporate assumptions about preferences (118).

As mentioned earlier, two approaches are commonly used to empirically test PbQoL measures: ‘known-group’ and convergent validity (148). Both approaches begin with setting up hypotheses. The COSMIN checklist states that specific hypotheses to be tested should have been formulated a priori, which ‘concerns expected mean differences between groups or expected correlations between the scores on the instrument and other variables, such as scores on other instruments, or demographic or clinical variables.’

The classic methods along with the known-group method and convergent validity approaches are introduced below.

3.4.2.1 Classic psychometric approach

The MTMM assessment is based on four sets of correlation coefficients, which aims to show that measures of the same construct should yield similar results (convergent validity) while measures of different constructs should produce different results (discriminant validity) (337, 367). The first set assesses the correlation of using the same method to measure the same construct (monotrait-monomethod) at two separate occasions, which is essentially a correlation test for test-retest reliability. The second set of coefficients assesses the correlation of using the same method to measure different traits (heterotrait-monomethod). Discriminant validity is established if this set of correlation coefficients is low. The third set examines the correlation of using different methods to measure different constructs (heterotrait-heteromethod), and the last set examines the correlation of using different methods to measure the same construct (monotrait-heteromethod). Convergent validity is established if the monotrait-heteromethod coefficient is high.

However there are several limitations of the MTMM methods as summarized by O'Leary-Kelly and Vokurka (369). It was argued that there is no precise standard for determining the degree to which the correlation criteria are met and the original recommendation of visual inspection is subjective. Furthermore, there lacks assessment for the underlying assumption that the traits are all equally influenced by the different methods used to measure them. Related to this is that there is no way to separate out the variance that is attributable to the traits, vs. the methods, vs. random error (369). Also, finding a measure with similar construct is straightforward but finding a measure with a different construct is less pragmatic since theoretically there could be unlimited options.

Besides the MTMM method, another classic method to test construct validity is factor analysis (325). Measurement of variance contains both shared and unique variance across variables. Factor analysis is concerned with the variance that each variable has in common with other variables. It is used to determine the number and nature of latent constructs within a set of observed variables and cluster highly interrelated variables into factors. Researchers can use it to group similar questions together from a lengthy questionnaire (337). In construct validation, it can be used to group the correlated attributes together from different measures to aid the understanding of the overall correlation between measures. Another related method, known as principal component analysis, extracts factors based on the total variance of the variables, aiming to find the fewest variables that explain the most variance. Principle component analysis decompose a set of data with correlated variables to independent (i.e. uncorrelated) components and thus it is sometimes used by researchers to reduce a large number of variables to a smaller, more manageable number of components (337). Therefore, factor analysis and principle component analysis are different processes to understand the structure of the construct; the former looks for the shared factors while the latter looks for the independent components to form the whole construct, and as such they can be used for determining whether two measures can be substitute or complement for each other.

3.4.2.2 The known-group method

Another method to demonstrate construct validity is the known-group method. The relatively new measure is administered to two groups that are known to or logically should be different in the feature that is expected to be captured in the construct of the instrument (372). The hypothesis is that there is difference in the scores of the instrument between the two groups (368). Therefore the known-group method is to determine the extent to which the instrument can differentiate between groups that are expected to differ in terms of the concepts of interest being measured. Good evidence of construct validity is demonstrated by a statistically significant difference of the scores of the instrument between the groups. Given that statistical significance is dependent on sample size, weak evidence of construct validity is also considered if a statistically significant difference is nearly shown.

Known-group validity has been widely tested for validating PbQoL instruments (19, 21, 373-379). Brazier et al. (2014) assessed the known-group construct validity of the EQ-5D and SF-6D in populations with mental health problems (380), Stavem et al. (2011) assessed known-group construct validity of the 15D and EQ-5D in a community sample of people with epilepsy (381), Maddigan et al. (2004) assessed known-group construct validity of the SF-12, HUI-2 and HUI-3 in type 2 diabetes (382). In particular, the known-group method has also been used in previous works of testing the construct validity of ICECAP-O and ICECAP-A (305, 383). For example, Makai et al. (2013) used it to test the ICECAP-O in a population of post-hospitalised older people in the Netherlands (313). Al-Janabi et al. (2013) used it to test the ICECAP-A and found that its responses and scores could differentiate between different health and socioeconomic groups but not across individuals with different levels of local deprivation (383).

One issue to be noted in the known-group method is the criteria for groupings. The choice of criteria can influence the results of the known-group tests; usually the higher correlation between the criteria and the test measure, the more favouring the known-group results will be to the test measure (118). When using measures that only have a weak relationship to QoL, the results of assessment should be interpreted carefully. For example, when grouping using clinical measures such as cholesterol level, in which case the known-group results may fail to show the difference in the mean score of the QoL measure between people who have a high and low cholesterol level, but this cannot conclude that the QoL measure has weak construct validity. Any negative results towards the QoL measure may likely be caused by the unsuccessful choice of criteria and hypothesis rather than weakness in validity with the measure. Similarly, any positive results could also be due to the choice of criteria and hypothesis, rather than the strength in validity with the measure.

Another issue relates to the use of a non-preference based measure as criteria to test the construct validity of a preference-based measure. As mentioned earlier, hypothesis of the known-group difference has to be made relevant to the preference values (118). Using the example above, the hypothesis is not simply 'patients with low and high cholesterol level have differed QoL', but would be 'patients would prefer having low cholesterol level than having high cholesterol level'. The 'prefer' here means that patients would like to trade some of their length of life for some decrease in their cholesterol level.

Given the above issues, care must be taken to scrutinize the criteria being used to establish known-group differences, and difference in preferences must be assumed in the hypothesis if the test measure is a preference-based measure.

3.4.2.3 Convergent validity

Convergent validation is another test of construct validity defined as the extent to which one measure correlates with another measure of the same or similar construct (353). The direction and magnitude of the correlation are important for understanding the association between the measures. Convergent validity is demonstrated if the test measure is highly correlated (correlation coefficient (r) $\geqslant 0.5$) with a measure with similar construct (324). A perfect correlation or very high correlation (r>0.8) would mean the test measure and the criteria measure are measuring the same construct.

Like the known-group method, the choice of the criteria measure is important for interpretation of the result. A very high correlation is not expected if the test measure and criteria measure are essentially different but related, such as physical health and general wellbeing. Physical health is one determinant of general wellbeing but not equal to general wellbeing. A perfect correlation means the measure of physical health and the measure of wellbeing are measuring exactly same thing, which is incorrect. Therefore, the test of convergent validity is about whether the strength of the correlation meets expectation based on the assumed overlapping concept between the test measure and the criteria measure, rather than simply expecting a high correlation.

Convergent validity is commonly assessed together with known-group validity or occasionally independently to provide evidence for construct validity of PbQoL measures (21, 246, 375-377, 379, 384). For example, Papaioannou et al. (2011) assessed convergent validity of EQ-5D and SF-6D in patients with schizophrenia. Ratcliffe et al. (2017) assessed convergent validity of EQ-5D-5L, and the preference-based dementia specific QoL measures, DEMQOL-U and DEMQOL-Proxy-U, in a post-hospitalisation population of frail older people living in residential aged care (379). Lorgelly et al. (2015) assessed the convergent validity of the cancer-specific preference-based measure EORTC-8D in cancer patients (384). It also has been used in the recent validation work of ICECAP measures. For instance, Sarabia-Cobo et al. (2017) assessed convergent validity of the Spanish version of the ICECAP-O in nursing home residents with dementia (385) and Goranitis et al. (2016) assessed convergent validity by exploring the correlation between the ICECAP-A and the EQ-5D-5L (386).

3.5 Responsiveness

Responsiveness refers to the ability of an instrument to measure change. Responsiveness is a relatively new term that has been introduced to the field of psychometric evaluation in the past 20 years (339). There is a debate on how it is related to the classic categories of psychometric properties, i.e., reliability and validity (337). Theorists often regard it as a longitudinal construct validity (387, 388) whereas there is new proposition that responsiveness should be considered as the third essential measurement property of an instrument, primarily in the area of HrQoL (389). The term responsiveness is also often used interchangeably as 'sensitivity to change' and there are mixed views of its definition. This section will introduce why responsiveness is an important property of a measure and why it is crucial for economic evaluations. This is followed by a summary of the various views towards how responsiveness should be defined. The definition of responsiveness is embedded with the concept of minimally important differences (MID) and thus MID will also be discussed.

3.5.1 Why we measure change

The ultimate goal of healthcare interventions is to induce positive change in the population's health status (339). Therefore measuring whether the status of patients – physical and mental health, QoL and wellbeing - has changed over time (either due to effect of intervention, or natural health status change), and to what extent the change has happened is of great importance in clinical practice and health research.

The measurement of change can be directed at three different goals (339, 390). The *first* goal is to measure the differences between individuals in the amount of change. This aids to differentiate between the individuals who have larger changes and those who have little change when receiving the same intervention. So the first goal is to identify individual variability in terms of the magnitude of change, which has received renewed attention in precision medicine (391). Because of the differences across individuals, an intervention that, on average, has been shown statistically significantly effective in a large group of sample may not lead to the same amount of change to every individual. The distribution of change is hoped to be deciphered by the research of precision medicine which analyses person's genes, lifestyle and environment to investigate explanations of 'what works for whom' and tailor the treatment (392).

The *second* goal is a logical follow-up objective after the first one. When the individual difference in change is identified, researchers may then be interested in identifying the factors that are associated with this change. This will help understand the reasons leading to the difference and can subsequently take measures to adjust the change stimuli (e.g. health intervention) according to population characteristics. The *third* goal of measurement of change is to infer treatment effects from group differences. This goal is mostly relevant to any intervention study with two or more groups which differs in the intervention received. For example, in a RCT, a treatment effect is determined by comparing the average change from baseline until endpoint between the different groups.

Simply put, the first goal is to identify if there are individual differences in their response to treatment, followed by a second goal to identify the factors that are contributing to the differences if the differences are identified. Then the last goal is to identify the treatment effect of intervention by comparing the change between treatment group and control group.

3.5.2 How responsiveness should be defined

PbQoL instruments are developed to measure change and to what degree the PbQoL measure is sensitive to change is what the 'responsiveness' property is about. The literature contains various definitions of responsiveness, and the differences between them are instructive, leading to a number of parameters proposed in the literature to assess responsiveness. The common basic framework to define responsiveness is 'the ability to measure change' or 'sensitivity to change' and the differences between definitions are usually surrounding the meaning of 'change'. For example, the change could be defined as 'clinically important changes' (393-395) such as in Guyatt et al's original definition that the "instrument must detect clinically important changes over time, even if those changes are small" (396). Alternatively, the change could be 'in the construct to be measured' such as in the definition by COSMIN checklist that responsiveness refers to "the ability of a health-related patient-reported outcome to detect change over time in the construct to be measured"(327). Terwee (2003) reviewed literature published between 1985 and 2002 and categorised the varying definitions for the concept of responsiveness to three groups (397). In the first group, responsiveness is defined as the ability to detect change in general while in the second group it is defined more specifically as the ability to detect clinically important change, such as the example of Guyatt's definition above. The third group expands the focus on clinical area in the second group to ‘the concept being measured’ (397).

The many definitions are primarily distinguished upon two traits: whether the change is meaningful / important or not, and whether the change is specific to clinical or health in general or in the concept being measured. The first trait differentiates between the term ‘responsiveness’ and the term ‘sensitivity to change’. Although literature sometimes uses ‘responsiveness’ and ‘sensitivity to change’ interchangeably, the meanings of the two terms differ according to Liang (2000); ‘sensitivity to change’ makes no judgement about whether the change is important or not but ‘responsiveness’ does, in some of its many definitions (394). Confusion upon the second trait can be traced back to the evolving process of the term ‘minimally clinical important difference’ (MCID) and MID (i.e. minimally important difference). MCID was introduced by Jaeschke and Guyatt in 1989 as a way to translate changes in instrument scores into clinically meaningful terms (398). The focus of MCID on clinical arena, however, limits its use in HrQoL instruments which emphasize on patients’ experience, and subsequently the ‘C’ is removed from the original MCID and MID was born (399). This will be further discussed in the next section. This broader scope of the definition of ‘change’ expands the meaning of responsiveness accordingly.

This thesis considers responsiveness as a context-specific term and adopts the definition in the third group outlined by Terwee et al.’s review (397); responsiveness refers to the ability of an instrument to detect important change over time in the construct to be measured. A prerequisite of the test of responsiveness therefore lies in the interpretation of the change that occurs in its construct; a question of what is an important change in the context. This change may, for example, allow an individual to achieve walking without assistance, or live with a more manageable level of stress, or just simply is perceived by an the individual as important. The next section will continue from last paragraph on the discussion around MID and the question as to what degree a change is considered important.

3.5.3 Minimally important differences

MID and its earlier form, MCID, were developed to aid the interpretation of a change score on an instrument. MID is defined as “the smallest difference in score in the outcome of interest that informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management” (Schünemann and Guyatt 2005) (399). This definition did not stress that the change must occur in one specific domain; rather, a general change could occur as long as it is the outcome of patients’ interest and perceived important by the patients.

In clinical trials, where QoL instruments are being used increasingly as the primary outcome measure, determining if the change is statistically significant is easy, but this does not, however, inform clinicians whether the change is meaningful to patients or not. Also, any change, no matter how small, can be statistically significant with a large enough sample size. However, statistical significance will not inform if those small changes are meaningful to patients or, from a health economist’s perspective, can lead to any difference in preferences. To fill this gap, MID (or MCID) was introduced to place the magnitude of change in a measure in a context which can be detected and is valued by a patient (398).

Criticisms have been raised regarding the research on obtaining MID for the PbQoL measures, such as the EQ-5D. Walters and Brazier established the MID for the SF-6D and EQ-5D-3L to be 0.041 and 0.074, respectively (400). They defined the MID in this context as “the smallest change in utility scores that can be regarded as important”. However, it can be argued that utility scores have their own meaning since they are ‘preference-based’ and as such they represent the trade-offs between health states and length of life. Owing to this, any difference in utility scores could be quantitatively translated to a difference in length of life. In simpler words, because of the valuation process, the utility values represent for how much length of life an individual would trade for, or how much higher/lower risk of death an individual could accept. For example, despite being small, a 0.01 absolute difference of utility value for one year means a difference of living for 3.65 days with full health. This forms the fundamental basis for the utility values to be able to combine with length of life to generate QALYs. On the contrary, MID is only relevant to the instruments with a score that does not have a meaning. The preference elicitation process (the valuation stage, as introduced in Chapter 2 Section 2.4.2) has already incorporated the concept of ‘importance’, as the larger coefficient of the health state defined by the dimension and level, means more importance, which is related to the degree of how it is preferred by the public. Therefore, any difference in utility value could be considered to be important and the remaining issue is whether the difference is statistically significant to guard against a chance finding. Besides, the PbQoL measures are developed for the purpose of economic evaluations, where what is important is not whether there is a meaningful change in the preferences but whether there is a difference in the combined cost-effectiveness, or the ICER (401) and therefore eliciting a MID for a PbQoL measure may not be relevant for economic evaluations.

3.5.4 Why responsiveness is important for PbQoL instruments

Responsiveness is an important property for PbQoL measures, given their role in economic evaluations. As covered in Section 1.4, 2.2.2, and 2.4.3, PbQoL measures are used to obtain utility values which are to be combined with length of life to calculate QALYs in economic evaluations and the magnitude of incremental cost per QALY (or, ICER) will affect the funding decisions of new interventions. Therefore, responsiveness is an essential property of an instrument for comparing the outcomes of health care interventions as well as measuring longitudinal change over time (396, 402). For example, in clinical trials where the effectiveness of an intervention is demonstrated by the condition-specific QoL measures with the assumptions that these changes are deemed important to both patients and public, if this is not appropriately reflected on the change of the PbQoL measures, the treatment effect may be underestimated or overestimated in the QALY calculation for the intervention arm. In the case of the former, i.e. the underestimated QALY, the aspects where benefit shows cannot be fully captured and valued by the PbQoL measure, while for the latter case, those aspects may have been over-emphasized by the PbQoL measure at the cost of compromising the value of other domains to overall QoL. In both cases, low responsiveness of the PbQoL measures may cause error to the ICER estimation of the alternative interventions, which may affect health care decision-making.

3.6 Methods to assess responsiveness

The definition of responsiveness suggests its assessment would require another measure to identify the happening of the important ‘change’, i.e. whether the patients have improved or worsened over time, regardless of whether the change is meaningful clinically or on the concept of interest over time (403). This is commonly known as the ‘anchor-based’ method (403). It explores the relationship between the change in scores of a measure and the same or similar concept measured by an independent anchor.

Another method that is commonly seen in literature is the ‘distribution method’ which uses statistical parameter such as effect sizes of a sample to estimate change (404). It is considered as an alternative to anchor-based methods when an appropriate anchor is not available. However, the application of ‘distribution-based’ method in the assessment of responsiveness should be treated with caution. The effect size statistics is concerned with both the size of the real change and the ability of detecting change of the test measure (327). In the absence of an external reference point to confirm the magnitude and direction of the change in the population, it is unknown whether the small effect size is the consequence of ineffective treatment / small real change, or due to the poor responsiveness of the measure (327, 404). Examples of such misuse of the distribution-based method are, unfortunately, not hard to find in previous literature (405, 406). Given the limitations, distribution based approaches are not recommended by FDA guidance to play a primary role for patient reported outcome measures (404). A solution to this limitation is that the distribution-based method could be applied in conjunction with the anchor-based method, in which the effect size statistics are calculated for each of the anchor group, rather than the whole population.

In the anchor-based approach, anchors are used as an external surrogate ‘criterion’ to identify the change. It is hypothetical because there is no real criterion in QoL area. This indicates that testing responsiveness also follows the five-step model introduced in 3.4.1 which starts from setting up hypothesis and testing the hypothesis (118). For testing a PbQoL measure, in the same way as for a hypothesis set for construct validity, the researcher must assume the change of preferences. The hypothesis could be, for example, the preferences captured by the new measure should be responsive to the change of the degree of pain, or in other words, when the patients feel less pain over time, it is expected to see a higher utility value (measured by the test measure) towards the later state with less pain.

Once an anchor is selected, either clinical or QoL-centred, the anchor is then used to assign participants into groups reflecting some degree of change according to the size and direction of the change between baseline and follow-up in the anchor measure (403). The groups could be no change, improvement, worsened, or if sample sizes allow further stratification, such as no change, small or large improvement, small or large worsened. After the grouping is completed, statistical methods are then used to determine the direction and magnitude of the change in the test measure in relation to the variance of the change for groups of patients with a confirmed experience of change.

The most commonly used statistical parameter to assess responsiveness is the ‘effect size’, where the mean change in score is divided by either the standard deviation at the baseline or the standard deviation of the change (402). The effect size statistics indicate the relative size of the ‘signal’ in comparison to the underlying ‘noise’ in the data (148). Good responsiveness of a test measure to the concept measured by an anchor is demonstrated if the change in the scores of the test measure in each group of participants is in the expected direction as indicated by the change in the scores of the anchor measure (403, 404). In other words, it is expected to see the improvement of the score on the test measure in the ‘improvement’ group defined by the anchor measure, no change of the test measure in the ‘no change’ group, and worsening of the test measure in the ‘progressed’ group.

This section will introduce how the anchors are selected and how the change groups are formed, followed by the effect size statistics and other statistical methods to aid the understanding of responsiveness.

3.6.1 Anchor selection

The anchor should be a validated measure with the same or related concept as the test measure. It is widely recommended to use multiple anchors (403, 404, 407, 408). The anchor(s) may be a clinical objective measure, or a subjective measure reported by patients (409). Condition-specific scales, being more focused and tailored towards problems of particular importance to the target patient groups, are generally more sensitive in the specific context than generic health-status measures (410) and thus more commonly chosen as anchor measures.

In general, the choice of anchor is a function of the strength of correlation between the anchor and the test instrument, and the degree to which it would increase understanding and is of interest to the researchers (403, 404). Anchors can be justified when it is shown to have a theoretical or proven association with the test measure. The association can be informed by initial assessment of the correlation of change scores of the anchor and the test measure. An acceptable correlation threshold is taken to be 0.3 (403, 404), while a lower correlation thresholds may still be acceptable in some situations (403). Cross-sectional correlations at baseline and follow-up between the measures can also be considered. Besides, anchors can also be chosen if any theoretical or methodological reasons can be provided, or when analysis using the anchor would be of interest to investigators and researchers.

Two cautions should be noted for the selection of anchor. First, before testing the correlations, the measurement properties of the anchor measure or the comparator instruments should be adequate (403). Otherwise, it is difficult to decide afterwards whether negative results are due to lack of responsiveness of the instrument under study or poor quality of the anchor. Second, where multiple anchors are selected, differences in their constructs are expected (403). The benefit of choosing multiple anchors is to enable testing the construct of the new measure from different angle, e.g. whether the new measure is responsive to pain, or whether it is responsive to change in QoL. Therefore, it is important to choose a series of anchors with different correlated constructs with the test measure.

3.6.2 Anchor group formation

Once anchors are selected, they are then used to assign participants into ‘change groups’. As mentioned earlier, the groups could be no change, improvement, worsened, or in more detailed stratification. Depending the type of anchor, four methods to form change groups are identified from literature. The first method is the global rating method where groups could be formed by directly asking patients whether or not they feel some degree of change. The second method is used where there is an intervention with proven effectiveness. Groups can be formed naturally by patients receiving intervention vs. not receiving intervention. Groups can also be formed using anchors such as clinical measures or QoL measures; the former would require clinician's opinions on the interpretation while the latter would need MID to infer how many points of change on the score would translate into a meaningful change experienced by the patient. These methods are described below.

3.6.2.1 Global rating of change scales

In a classic anchor-based approach, the change groups are formed naturally by using Global rating of change scales, or simply put, Global rating scales (GRS), as the external anchor (411). GRSs are designed to quantify the magnitude of the improvement or deterioration of a person's health status over time (411). It asks a person to rate his or her current health status compared to a previous time-point on a multi-point GRS. For example, to assess responsiveness of the Anterior Cruiciate Ligament Quality of Life Measure, Lafave et al. (412) reported that patients were asked to select one of seven categories of change on the 7-point GRS and were grouped accordingly: 7, significantly better; 6 much better; 5, somewhat better; 4, no change; 3, somewhat worse; 2, much worse; 1, significantly worse. Similarly, Greco et al. (413) created three change groups based on the 7-point GRS with slightly different description, including: the 'the improved' group consisted of individuals who rated themselves on GRS as 'much better' or 'somewhat better', 'unchanged group' consisted of 'slightly better', 'not changed' or 'slightly worse', and 'worse' group for GRS rating of 'somewhat worse' or 'much worse'. Although GRS provides a convenient path for forming the anchor groups, it is acknowledged that GRS should not be considered as gold standard as its reliability and validity is not established (327, 409). A prominent criticism is the potential for recall bias. Studies have found that people tend to link to their current status when asked to recall a prior state, leading to retrospective judgements of change vulnerable to bias (414). On the other hand, reliable and accurate information from the GRS scale places considerable cognitive demand on the patient (411).

3.6.2.2 Intervention group with proven effectiveness

The second approach is to divide the groups according to the different interventions assigned, where the effectiveness of the interventions has been proven to be distinctive (415). For example, to assess the responsiveness of the EQ-5D-3L and SF-6D to the change of inflammatory arthritis, Harrison et al. (416) grouped the patients according to their treatment arms in clinical trials based on the expectations on the effectiveness of the intervention, and the natural progression/deterioration of the control arm, e.g. the patients receiving treatments which inhibit the action of TNFα were assigned to the 'improvement group' as the treatment was expected to dramatically improve the outcomes. However, this method is subject to limitations in the 'proven effectiveness' of the intervention since response to an intervention may be varied across individuals. It also neglects the fact that responsiveness is about 'degree' rather than 'yes' or no' since 'proven effectiveness' alone would provide no information on how much change one should expect on both the health status of the patients and the test measure.

3.6.2.3 Distinct health stages defined by objective clinical measures

The third approach is to use objective clinical measures as external criteria, which sometimes are combined with the global rating scale to substantiate the patients' subjective assessment (409). For instance, due to the lack of a gold standard to assess patients with heart failure, two clinical objective assessments and one cardiologist completed GRS were used instead as external indicators of heart failure status change in Eurich et al's study (417), which evaluated the relative responsiveness of several QoL measures to the clinical change of heart failure. One clinical objective measure used in this study was cardiologist's assessment of the patients' New York Heart Association (NYHA) classification at baseline and at endpoint. This is an ordinal measure which produced easily defined groups and therefore subjects were classified to five change groups: improved/deteriorated two classes of NYHA, improved/deteriorated one classes of NYHA and no change. Another objective measure was the six-minute walk test (6 MW) which produces the travelled distance within six minutes as a continuous number. The authors classified the change of the travelled distance between baseline and endpoint to seven mutually exclusive categories, on the basis of previous research which used physician-assessed global rating of change as criterion to quantify the number of minutes between each category of change.

3.6.2.4 Minimally important difference

The last approach is to use the MID of the condition-specific anchor measure to define the groups providing the MID information is available (418). People who experienced a change equal to or greater than the MID are categorised into the change groups, i.e. improved or deteriorated group. For example, Keeley et al. (418) assessed the responsiveness of ICECAP-A and one of the anchors used was the EQ-5D-3L. By using the MID value of EQ-5D-3L from one previous research (400), 0.074, three subgroups were formed: improved/deteriorated group (patients who had improved/worsed by larger or equal to 0.074), or no change group (patients who had a change with a smaller size than 0.074). Although the meaningfulness of MID for preference-based measures is controversial as mentioned in 3.5.3, this is a useful example of grouping by MID when the anchor is a PbQoL measure.

The approach of using MID to defining groups has its own limitations in application in addition to the issue of interpretation of MID for PbQoL measures. First, it relies on the robustness of the previous study of testing the MID and the generalisability to the current study. Second, using a universal MID to generate both improved and worsened groups may not be appropriate in some clinical areas. It was found that the MID generated from the ‘somewhat better’ group and ‘somewhat worse’ group were different for SF-6D (mean difference: 0.079, p=0.02) in people with back pain and EQ-5D-3L (mean difference: 0.275, p=0.001) in people with osteoarthritis in knee (400). Therefore, care should be taken before generalizing the results of MID to another study. Lastly, it does not provide any inferences on the threshold between the ‘a little change’ and ‘a lot of change’, since MID by its definition only concerns with the difference of the measure between ‘minimal change’ and ‘no change’.

Nevertheless, using MID as criteria may be a more robust approach to assessing the responsiveness of a ‘preference-based’ measure compared to the other approaches. As mentioned earlier, Brazier and Deverill (148) are concerned that the common approach of using the non-preference-related instrument to confirm the change (such as using clinical measures), cannot reflect the changes in 'preferences', or the degree of importance of a change to patients. The MID method incorporates 'importance' to some extent into the measurement of change. To test MID, typically, studies would ask patients whether there is no change, a little change or a lot of change in their overall health. In this process, when a patient is considering whether what happened over the study period could be called a 'change', the aspects that are related to a patient's QoL are being weighted with importance to some degree in his/her mind.

3.6.3 Statistical methods for testing responsiveness

After grouping is completed, statistical methods are then conducted to determine the degree of responsiveness of the test measure to the change of the anchor. Terwee et al. reported there were as many as 31 different responsiveness statistics (397). Methods that are relevant to the testing of PbQoL measures within the anchor-based approach are described below.

3.6.3.1 Effect size statistics

Effect size statistics are recommended as the primary method for assessing responsiveness of patient-reported outcome measures (148, 397, 404, 415). They quantify the magnitude of change based on variation in the scores of the measure.

The standard effect size (ES) is also called Cohen's effect size, which was invented by Cohen in 1988 (419). It is calculated by dividing the mean change between baseline and endpoint by the standard deviation (SD) of the baseline scores (Formula 1) (415) .

$ES = Mean_{change} / SD_{Baseline}$ (Formula 1)

The standardised response mean (SRM) is a variant of ES, which was suggested by McHorney and Tarlov in 1995 (420). It is calculated by dividing the change between baseline and endpoint with the SD of this change (415).

$SRM = Mean_{change)} / SD_{Change}$ (Formula 2)

Compared to the ES, the SRM is more closely linked to the paired t-test. In a paired t-test, T is calculated by dividing the mean change by standard error. Standard error is calculated by dividing standard deviation by the root square of sample size (√n). And therefore, SRM is simply the T divided by √n (339). (Formula 3)

T= $\text{Mean}_{\text{change}}$/ $\text{SE}_{\text{Change}}$= $\text{Mean}_{\text{change}}$ / ($\text{SD}_{\text{Change}}$/√n) = $\text{Mean}_{\text{change}}$/$\text{SD}_{\text{Change}}$*√n=SRM*√n (Formular 3)

And therefore, SRM=T/√n

Both of the methods are based upon means and SDs, which implies an underlying assumption that the data distribution of the outcome measure follows a normal distribution. Many QoL scales have a non-normal distribution, in which case unfortunately little work has been carried out into how to test the responsiveness when the assumption is not met.

Cohen (419) provided a rule of thumb for the cut-off values to interpret the magnitude of the ES. A score below 0.2 represents very small ES, 0.2 to 0.5 – small, 0.5 to 0.8 – medium ES, larger than 0.8 - large ES (421). These cut-offs, however, are argued to be problematic when being applied to interpret SRM (422). The main reason is due to the different SD used for ES and SRM, SD of baseline value is used for ES but that of the change value is used for SRM and thus using Cohen’s threshold to determine the magnitude of the effect size may not be accurate. Middel et al. have shown some estimates based on Cohen’s threshold applied to SRM values being either over- or underestimation of an intervention-related effect (423). Sivan suggested to use the method proposed by Middel & Sonderen, which applies the correlation coefficient between the repeated measurements (i.e. baseline and endpoint) to Cohen’s threshold (424).

However, all the rules above are aiming for the situation where effect size is used as a measure for treatment effect, rather than a measure for testing the responsiveness to a change. Therefore focus should not be put on finding measures with the largest responsiveness statistics or determining if a measure can produce a ‘large’ effect size statistics when the aim is to test the responsiveness of a test measure to a hypothetical anchor (327). Within the anchor-based approach, the expected size of the effect size statistics is conditional on the relationship between the test measure and the anchor. As mentioned earlier, when the anchor and the test measure are designed for different purposes, expecting a large ES is inappropriate. Therefore, hypothesis should be carefully set describing what the expected ES/SRM based on the relationship between the anchor and the new measure or relative size of ES/SRM if multiple new measures are compared.

3.6.3.2 Paired t-test

The paired t-test tests the null-hypothesis that there has been no change in the mean response of the new measure. In the anchor-based approach, the paired t-test is conducted within each change groups, e.g. no change, small improved, small deteriorated, etc (415, 425). A weakness of the t-test is that it is highly dependent on the sample size included in the measure and thus its result only plays a supportive role in determining responsiveness.

3.6.3.3 Correlation method

The correlation between change scores is the preferred method of the COSMIN group for comparing changes in the test measure with changes in an anchor if the scores on the test measure and the anchor are both continuous (327). The correlation method provides a useful indication of the extent to which the change score of the anchor and test measure are associated; a stronger correlation typically means a stronger responsiveness of the test measure to the anchor (415). The correlation coefficient describes both the strength and direction of the relationship.

There are two types of correlations: Pearson product moment correlation, and Spearman rank-order correlation, the choice of which is dependent on whether there is a linear relationship between the scores of the two measures. Pearson’s correlation coefficient is used when a linear relationship is demonstrated, typically by visual inspection in a scatter plot of the two variables or a more sophisticated regression method. In a regression between the two variables (regardless which one is the dependent or independent variable), a linear relationship is shown if the regression residuals (fitted value – observed value) are normally distributed and do not show skew (426). Two approaches are proposed to judge if the residuals are normally distributed: the kernel density plot, and the

Shapiro-Wilk W test. The kernel density plot is plotted to allow a visual comparison of the distribution of the residuals against an overlaid normal distribution (427, 428). The Shapiro-Wilk W test is performed for significance testing of the assumption that the distribution is normal (427, 429). When this linear relationship assumption does not meet, Spearman's rank correlation coefficient should be used instead (430). Spearman's rank correlation is a nonparametric measure of rank correlation, which is similar to the Pearson correlation but using the rank values of the scores of the measures.

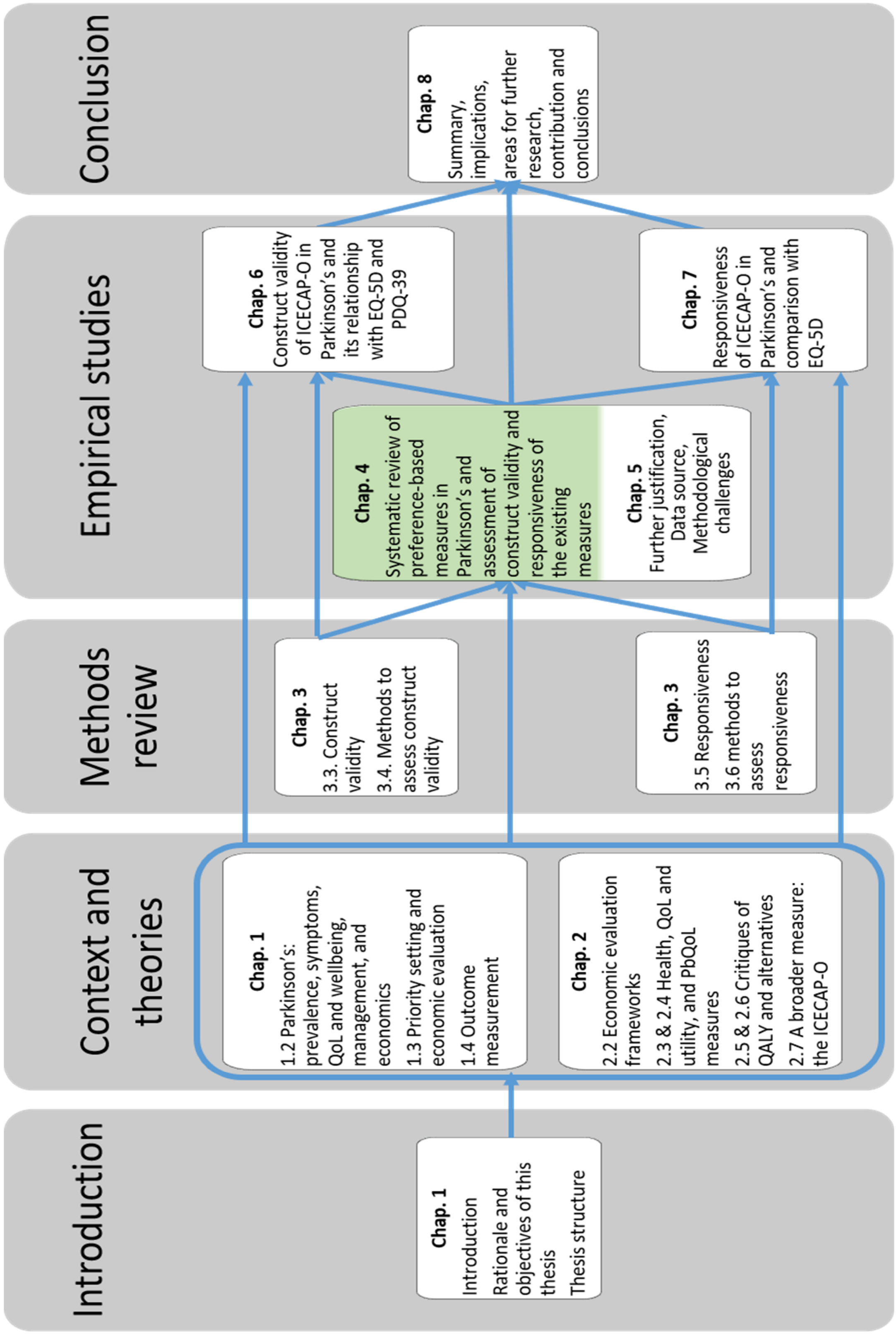
3.6.3.4 Regression methods

Regression methods can be used to explore the relationship between the change of the anchor and the change of the test instrument after adjusting for potential confounders (415). It can identify the key determinants for the change in the test measure. When multiple anchors are used as independent variables to predict the test measure, the regression result should be able to aid the interpretation of the results from other aforementioned tests, such as why the test measure is more responsive to some anchors than the others.

3.7 Chapter summary

Both construct validity and responsiveness are crucially important for any measurement. This chapter provides an overview of psychometric properties of validity, reliability and responsiveness, discussed the definition and features of construct validity and responsiveness, summarized the methods through which they can be assessed, and critically discussed the challenges when applying classic psychometric testing methods to the assessment of PbQoL measures. These assessment methods will be used in Chapter 6 and 7 for the testing of construct validity and responsiveness of ICECAP-O in people with Parkinson's. Prior to the case studies, the practical challenges and special considerations that were discussed in this chapter will be summarized in Chapter 5 Section 5.4, which provides justifications for the methods chosen and assumptions for the case studies.

Chapter 4 A systematic review of the use of preference-based measures in Parkinson's and assessment of their construct validity and responsiveness



Introduction
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Context and theories
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Methods review
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Empirical studies
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Conclusion
Chap. 8
Summary, implications, areas for further research, contribution and conclusions

4.1 Introduction

Economic evaluation using the QALY framework relies on PbQoL measures. However, as described in Chapter 2 (Section 2.5), current PbQoL measures are often criticised for being insensitive or failing to capture important aspects of QoL in specific populations. In people with Parkinson's, the patient group Parkinson's UK have expressed their concerns on the use of EQ-5D as it may not be sufficient to capture the impact on QoL from all motor and over 40 types of non-motor symptoms of Parkinson's (see quote on p19). These symptoms have a broad influence on patients' physical, emotional and social wellbeing, and there is some evidence as summarized in Chapter 2 (Section 2.4) that generic PbQoL measures like EQ-5D with its five generic health-related dimensions may have limited ability to fully measure the broad impact of diseases on QoL and wellbeing (128, 431, 432). For example, in schizophrenia, Mulhern et al. (2014) found that the responsiveness of EQ-5D and SF-6D was weak, as shown by a smaller than 0.2 SRM (below the clinically significant range) while the clinical measure of schizophrenia has large SRM (128). Jenkinson et al. (1997) compared SF-36 and EQ-5D-3L with condition specific measures in a RCT of transurethral resection of the prostate with laser vaporization prostatectomy for benign disease and found that although the condition specific measures showed statistically significant difference between the arms (which indicating the effectiveness of the intervention), the PbQoL measure, EQ-5D-3L failed to show any difference (431). The insensitivity of PbQoL measures often found in other disease areas raised a concern about whether they are sensitive enough to capture the broad impact of Parkinson's.

Imagine a situation when people with Parkinson's are unable to control their limbs due to the involuntary movement, suffer from social isolation because of stigma and face the fact that their symptoms can only slowly get worse without a cure in the future. When there are new interventions available to improve these situations, is the population willing to trade some length of life for these improvements? If so, to what degree is the population willing to trade for each of these improvements? These questions require systematic valuation of the health aspects against risk of death or length of life, which have not been conducted yet (further discussion in Chapter 8 Section 8.6.2). However, an intuitive answer to these questions is yes, and yes to all of the three aspects of improvement because the health aspects as well as wellbeing aspects are all highly prevalent in all Parkinson's QoL questionnaires as shown in Chapter 1 (Section 1.5.2) indicating their importance to patients, clinicians and researchers.

After determining the importance of those aspects, the next question is whether such important improvements in health and wellbeing are being adequately reflected in the PbQoL measures? In other words, to what degree can the existing generic PbQoL measures capture the comprehensive impact of Parkinson's on people's life? As discussed in Chapter 1, the consequence of underestimation of any important QoL aspects in the PbQoL measures is that the benefit of interventions targeting on such aspects would not be captured by the PbQoL measures and thereby the interventions may appear ineffective and have smaller QALY gains than they should have. This will result in the intervention appearing less cost-effective and impacting on funding decisions. This highlights a need to review and critically appraise the performance of the existing PbQoL measures including EQ-5D, in the Parkinson's population.

A brief scoping search of the literature identified two published reviews of QoL measures in Parkinson's which assessed the use of several PBQoL measures (433, 434), however both reviews are not specific to PBQoL measures hence insufficient to provide an overall critical assessment of the PBQoL measures. Martinez-Martin et al. (433) classified the generic and specific HrQoL scales to three groups ('recommended,' 'suggested,' or 'listed') by summarizing the existing evidence of psychometric properties from other studies. EQ-5D-3L and 15D were the only two PbQoL measures in their assessments; the former was assessed to be 'recommended' and the latter was grouped to the 'suggested' category due to lack of validation studies. This study, however, did not assess the properties using a pre-defined methodology, instead, the recommendation was established upon reviewing the reported conclusions from the existing validation studies. Another study from Dodel et al. (434) reviewed approaches to evaluate cost of illness, cost effectiveness, and discussed the utility instruments in Parkinson's. In this study, EQ-5D-3L, SF-6D, 15D, and HUI were compared upon six criteria of psychometric properties, which were adapted from two previous studies published in 2001 (435) and 2005 (436). Although the authors recommended the use of EQ-5D-3L and HUI over 15D and SF-6D along with the direct valuation method, a gap was identified in this study which necessities the assessment of psychometric testing of these measures in Parkinson's. In particular, it pointed out that the responsiveness of EQ-5D-3L required further validation and there was inadequate amount of validation evidence for HUI, SF-6D and 15D in Parkinson's (434).

As discussed in Chapter 3, psychometric properties are context-specific and psychometric testing is an iterative process, whereby evidence is gradually accumulated to lead to an increasing understanding of to what degree a measure is suitable for use in a certain population. The above reviews point towards a need to conduct a systematic review by collecting all existing evidence regarding the use of PbQoL measures in Parkinson's and critically analysing the identified evidence to assess their psychometric properties.

In addition to directly using a PbQoL instrument to measure preferences for economic evaluation, there is a growing trend of applying mapping algorithms to predict EQ-5D-3L utilities where a PbQoL measure is not used, as recommended by NICE. These mapping algorithms are generated through applying statistical methods to explore the relationships between a non-preference based measure and EQ-5D-3L using cross-sectional measurements of both. Accompanied with the increasing number of mapping studies are the growing voices to strengthen the methodological quality of these studies (see Section 2.6.1 for details). One factor affecting quality is the conceptual relationship between the measures on the two ends of the mapping algorithm. When mapping from dimensions of a non-preference based measure to EQ-5D-3L, whether or not each of the dimensions was included or to what extent the inclusion of the individual dimensions is in the mapping algorithms can affect the weight of the individual dimensions in the EQ-5D-3L. The answers to these questions are closely linked to the construct validity of a measure. Given this, there is a need to conduct an overview of the existing mapping studies to EQ-5D-3L in the Parkinson's, critique their study quality and compare their results.

This chapter will start by introducing the objectives of the systematic review and assessment methods of the PbQoL measures. The theoretical basis, definitions, assessment methods of the measurement criteria have been introduced in Chapter 3. This is followed by describing how the search was conducted, eligibility criteria and data extraction for the methodological systematic review. Following this is the result section which contains search results, assessment results, a summary of the identified economic evaluation studies and the mapping algorithms. A summary of results, discussion, and a summary of this chapter are provided at the end. Notably, given the methodologies used in this chapter and the case studies in Chapter 6 and 7 are similar, the strength and limitations are summarized altogether in Chapter 8.

4.2 Objectives

This systematic review has four objectives. They are:

1) To describe the use of PbQoL measures in studies in the Parkinson's population;

2) To critically assess the construct validity and responsiveness of the identified PbQoL measures in Parkinson's;

3) To critique the use of PbQoL measures in the included economic evaluations of interventions in Parkinson's; and

4) To summarize the mapping studies from condition specific QoL measures to EQ-5D-3L in people with Parkinson's identified in the literature search in terms of their data, methods, and the generated mapping algorithms.

The first objective is to investigate how frequently each PbQoL measure was used in the literature in the Parkinson's population, and summarize the purposes (study design, country, patient characteristics) that these measures were used for. Meanwhile, except for PbQoL measures, this study will also summarize the use of each Parkinson's specific measure used in the included studies, to facilitate the second objective below (i.e. the assessment of the generic PbQoL measures).

Through analysing the summary statistics provided in the included studies for both the PbQoL and another QoL measure, this chapter will also critically assess the construct validity and responsiveness of the identified PbQoL instruments in Parkinson's. As mentioned in Chapter 3, these two properties are essential for PbQoL measures to provide accurate utility values associated with the benefit of interventions. As mentioned previously (Section 1.5.4, 3.1), NICE in its current guideline for technology appraisal recommends investigating these two properties when determining the appropriateness of PbQoL measures (119). In particular, this chapter will assess (1) to what degree the PbQoL measures are able to differentiate between groups that are expected to differ, i.e. known-group construct validity (see 3.4.2.2 for details), (2) to what degree they are correlated with a measure with similar construct, i.e. convergent validity (see 3.4.2.3 for details) and (3) to what extent they are responsive to the 'known' changes that they are expected to detect, i.e. responsiveness (see 3.5 and 3.6 for details).

The third objective is related to the assessment of responsiveness as above but putting this assessment in a real economic evaluation context. It will contrast the results from the Parkinson's specific measures and the PbQoL measures, and discuss the implications of consistency or inconsistency between them.

The last objective is to compare the mapping algorithms. As outlined in the introduction section above, mapping is a NICE recognised avenue to generate EQ-5D-3L values. However the quality of these mapping studies varies, the original patient population where the mapping formula generated varies, and their resulting algorithms can vary accordingly, which necessities an overview of them.

4.3 Methods

4.3.1 Assessment criteria

Construct validity and responsiveness of the PbQoL measures used in the included studies were assessed. Methods for this assessment through the format of systematic review were adapted from the empirical assessment methods introduced in depth in Chapter 3 (Section 3.4 and 3.6). These methods have also been commonly used in previous similar reviews that aiming to assess the appropriateness of PbQoL measures (21, 235, 246).

For example, Longworth et al. (2014) conducted a systematic review of psychometric properties of three commonly used generic PbQoL measures, EQ-5D, SF-6D and HUI-3 in four broadly defined conditions: visual impairment, hearing impairment, cancer and skin conditions (235). They assessed the (a) known-group construct validity, i.e. the extent to which the measure can differentiate between groups defined according to severity or between people with or without the condition, (b) convergent validity, i.e. the strength of correlation, and (c) responsiveness, i.e. the extent to which the change (size and statistical significance) shown on other measures has been observed in PbQoL measures, and vice versa, i.e. the extent to which the PbQoL measure shows no change when no change was shown on other measures (although they called this as 'reliability' in the original text). Similarly, an earlier study conducted by Papaioannou et al. (2011) (246) and a more recent study conducted by Yang et al. (2015) (21) also assessed known-group construct validity, convergent validity and responsiveness of PbQoL measures through systematic reviews; the former assessed EQ-5D and SF-6D in people with schizophrenia, and the latter assessed EQ-5D, HUI-3 and SF-6D in patients with skin conditions.

4.3.1.1 Reference measures

Assessment of convergent validity and responsiveness requires at least one reference measure, or anchor measure. As introduced in Chapter 3, convergent validity was based on the expectations on the relationships between the PbQoL measure and the reference measure. When the reference measure has a very similar construct with the test measure, the relationship is expected to be highly correlated; when the reference measure is related but not with similar construct, then a high correlation is not expected. In addition, a reference measure is required in the examination of responsiveness to confirm the happening of change over time. In this context, the reference measure has to be condition specific which is assumed to be sensitive to the change in Parkinson's patients. It could be another PbQoL measure (although in the Parkinson's population, no CS-PBM is available), non-preference based QoL measure, or commonly used clinical measures in Parkinson's.

For the clinical measures, The UPDRS and H&Y are commonly used clinical measures in Parkinson's to assess disease severity. The UPDRS assesses clinical status of Parkinson's in four domains including, mood and cognition, ADL, motor symptoms severity, and complications of treatment (437). The H&Y describes progression of motor function in Parkinson's population, ranging from stage I (mildest) to stage V (most severe) (438).

4.3.1.2 Construct validity

As described in Chapter 3, construct validity represents the ability that an instrument measures the construct it is intended to measure (353, 439), and is typically assessed by the known-group method (Section 3.4.2.2) and convergent validity (Section 3.4.2.3) (235, 260, 353, 440-442).

The known-group method tests the extent to which a measure can discriminate between groups that are theoretically known to differ (353, 368). This review examined to what extent the index scores distinguished between patients with different characteristics of Parkinson's, with the premise that the mean utilities of the different patient groups were expected to differ. The characteristics that were used to define the groups were examined prior to the performance of PbQoL measures to determine the expectations on the mean difference of the PbQoL scores. Good evidence of construct validity deemed to be demonstrated by a statistically significant difference (e.g., t test) of the mean utility values between the 'known' groups that were expected to differ.

However, simply relying on statistical significance may bias the results given that sample size may have a great influence on the statistical significance; a large sample size may give statistical significance to very small effect, and a small sample size may fail to achieve statistical significance to a large effect. Therefore, when sample size is relatively small, appropriate size of difference with near significance was also considered as evidence for 'known-group' validity. In addition, as mentioned in Chapter 3 (Section 3.4.2.2), another issue is regarding the use of a non-preference based measure as reference measure to test the construct validity of a preference-based measure (118). Assumptions have to be made when assessing the PbQoL measures regarding people's preferences in the groups defined by the reference measures in that their preference for the two state has to be different, i.e. patients would trade different amount of their length of life for the two states. This assumption regarding people's preferences have to be made for the assessment of convergent validity and responsiveness as well.

Convergent validation examines the extent to which one measure correlates with another measure of the same or similar construct (see Section 3.4.2.3 for details)

(324, 353, 368, 440). If the PbQoL measure is highly correlated (correlation coefficient (r) ≥0.5) with a reference measure of similar concept then convergent validity is determined to be adequate. A moderate correlation (0.3 < r < 0.5) is expected if the PbQoL and the reference measure are convergent to some degree but not strongly. As mentioned in the 'reference measure' section, when the reference measure is not a similar concept as the test measure, a high correlation (r ≥ 0.5) is not expected as the PbQoL and the reference measure are designed to measure different concepts and this would not treat as negative evidence for the performance of PbQoL measures.

4.3.1.3 Responsiveness

Responsiveness is the ability of an instrument to accurately detect a 'known' change on its construct over a longitudinal time period (443, 444). This study examined the extent to which PbQoL measures were able to detect changes in some characteristics over time as confirmed by clinical measures or Parkinson's-specific QoL measures, i.e. reference measures. The change could be due to the health intervention or natural progression of Parkinson's.

As with the known-group method, responsiveness is determined to be adequate when the change/difference between the baseline and follow-up time point is statistically significant different or nearly statistically significant, if the happening of the change is confirmed by a reference measure of similar construct, or when the change on the reference measure is expected to associate with a change in the test measure. Similarly, when the reference measure shows no change, and no change led by other factors is expected to happen on the test measure, responsiveness of the PbQoL measure is determined adequate if no change happening on the PbQoL measure as well.

In addition, correlations between the change scores of the PbQoL instrument and the reference measures were also examined when they were reported in the study. The correlation method was another recognised method to assess responsiveness (327, 415), which was introduced in Chapter 3 Section 3.6.3.3. As with convergent validity, a moderate to high correlation coefficient was expected when the reference measure was with similar construct with the PbQoL measure (e.g. both are HrQoL measures, or both are wellbeing measures). When the PbQoL and the reference measure were with dissimilar construct, e.g. a clinical measure and a QoL measure, a small correlation was acceptable.

It is worth noting that although the assessment methods of responsiveness used in this systematic review were adapted from the methods introduced in Chapter 3 (Section 3.6), there are some differences between them. Firstly, the methods in Chapter 3 are for empirical validation studies which are applied on actual individual patient data, while the method used in this chapter using secondary summary statistics reported from the literature. Secondly, the main statistics for testing responsiveness are the effect size statistics which may not be reported in every study included in this review; it may be that only studies that are designed for the purpose of assessment of responsiveness would report these statistics. Therefore, this review did not use this as the main method for the assessment based on the secondary data.

4.3.2 Databases and search strategy

PbQoL measure is the outcome of health economic research as an interdisciplinary science that is established on both the theory of economics and health measurement. Therefore, use of multiple databases across social science and health science to search for literatures relevant to PbQoL measures would benefit maximizing the number of relevant results. Due to the differences in coverage of journals and search systems, more than one database was searched for literature in the area of biomedical science and social science.

In total, nine databases were searched to identify studies that used at least one PbQoL instrument to measure preferences in people with Parkinson’s. The databases were: biomedical databases including MEDLINE (Ovid and Pubmed)) and EMBASE (Ovid), nursing database CINAHL, behavioural and psychology database PsycINFO, Social science databases including Applied social sciences Index and Abstracts (ASSIA) and Social service abstracts (SSA) (ProQuest), AgeInfo, Database of Abstracts of Reviews of Effects (DARE), and NHS Economic Evaluation database (NHS EED). These databases are briefly introduced in the following paragraphs. In addition, the aforementioned (Section 2.6.1) database of mapping studies developed by the Health Economics Research Centre at University of Oxford

(Database version 5.0, based on search conducted in April 2016) (285) was also checked for mapping studies that were not identified in the primary search.

Pubmed and MEDLINE (Ovid) both provide access to the database MEDLINE. MEDLINE (1946 – present) provides more than 15 million articles published in more than 5600 biomedical periodicals (445). EMBASE (1947 – present) covers the same subjects as MEDLINE with an additional focus on drugs and pharmacology, medical devices, clinical medicine, and basic science relevant to clinical medicine (446). EMBASE includes all of MEDLINE’s citations plus 2,500 journals not currently indexed in MEDLINE (446). CINAHL (1937 – present) contains 5400 journals which covers health science in a broader sense including nursing science, paramedical science, education, behavioural science, and health administration (445). Search in MEDLINE, EMBASE and CINAHL was expected to identify original intervention studies or health determinant studies that used PbQoL measures.

PsycINFO (1967 – present) specializes in behavioural science and social science, produced by the American Psychological Association (447). ASSIA (1987 – present) contains records from over 500 journals in social science and health from the practical and academic perspective (448) and SSA focuses on social work, social welfare, social and health policy and community development (449). Through PsycINFO, ASSIA, and SSA, it was expected to identify the additional psychometric literature regarding the PbQoL measures and the use of PbQoL measures in social care interventions that were not covered by the above major biomedical databases. Parkinson’s mostly affects elderly people and hence Ageinfo was searched which focuses on social gerontology. In addition, two HTA focused databases produced by the NIHR center for Reviews and Dissemination (CRD) at the University of YORK, DARE (1994- March 2015) and NHS EED (1968 - March 2015), were searched to identify relevant systematic reviews and economic evaluations (450).

A search strategy was developed together with an expert information scientist from University of Glasgow library to maximize the chance of retrieving potential relevant studies. Search filters (pre-tested strategies) for economic study developed by the Scottish Intercollegiate Guidelines Network (SIGN) were reviewed and discreetly selected to aid the development of search strategies for the aim of this study (451). It was developed initially in MEDLINE (Ovid) and adapted for other databases (Appendix A). Databases were searched from inception until November 2013 and the search was updated in July 2015. The database of mapping studies was checked in March 2017.

4.3.3 Eligibility criteria and data extraction

Studies were included when meeting the following criteria:

- a PbQoL instrument was used to measure preferences in people with Parkinson's; and

- sufficient data were provided to allow the assessment of construct validity and/or responsiveness (the details are provided as follows).

Studies that were eligible for the assessment of convergent validity and responsiveness must also contain a reference measure. Besides, for the assessment of 'known-group' validity, at least two groups of patients that were differed in their characteristics had to be available, divided based on the score of the reference measure. PbQoL measure index scores had to be available for those groups. For convergent validity, correlation coefficients should be reported between the PbQoL measure and the reference measure. For responsiveness, at least two measurements or difference over a period of time (e.g., baseline and primary end point) of both PbQoL measure and the reference measure should be reported.

There was no limit on study types so both RCTs and observational studies were included. Conference abstracts were excluded as they are usually not peer-reviewed and thus difficult to judge validity of the results. All results were limited to English. In addition, mapping studies from non-preference based measures to preference-based measures in Parkinson's were also included.

Studies were excluded if the population being measured were patients without a confirmed diagnosis of Parkinson's; the utilities of patients were not measured, measured but not reported, not appropriately presented (e.g., EQ-5D index value not on a '0 (death) -1 (full health)' scale), or not adequately presented for the assessment purpose; or a full result published later covering the shorter time period result in earlier papers.

After two-step screening based on title & abstract and full-text, included studies were reviewed and study characteristics were extracted. They contained: first author and publication year, country, study type, number of participants, clinical characteristics, and length of follow-up (when applicable). For the purpose of assessing psychometric properties, study objectives, methods, the measures used, and their scores were also extracted. The characteristics of the mapping studies were also extracted to enable the critique, including: author, year, country, the QoL instruments involved (i.e., the condition specific measure to map from and the generic PbQoL measure to map to), sample size for the estimation of the algorithm and validation of the algorithm, the mapping model(s) used, measure of model performance, and the final mapping algorithm.

4.4 Results

4.4.1 Search results

A total of 2,758 records were retrieved after removing duplicates. The number of records identified from each database is presented in Appendix A. Titles and abstracts were initially screened based on eligibility criteria and 2,536 records were excluded. Full text of the remaining 222 studies was further screened from which 22 studies were included in this review for the assessment of construct validity and responsiveness, and five studies were included for the review of mapping. A flowchart of the screening process with the reasons for exclusion in the full-text screening stage is shown in Figure 4-1.

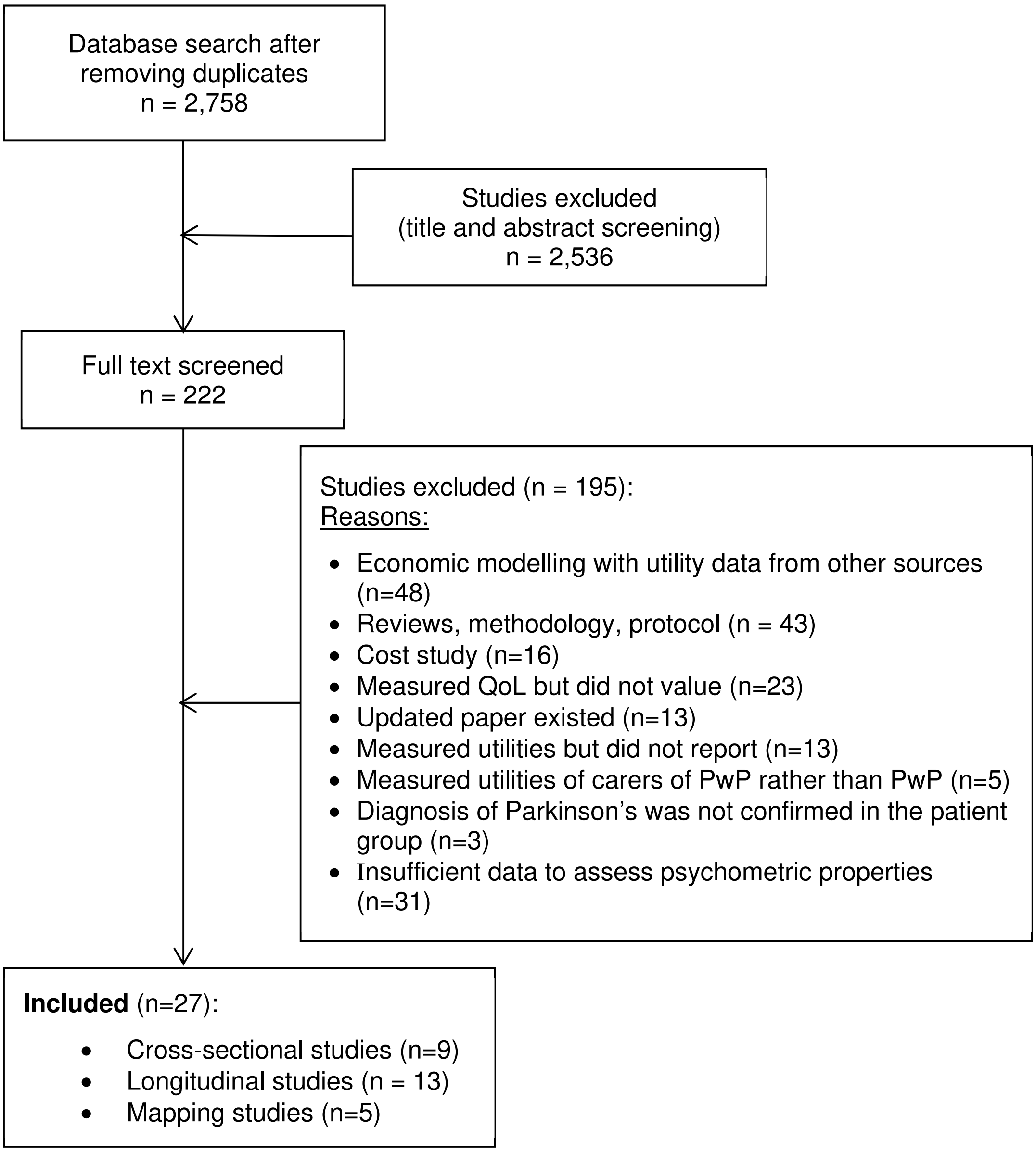

Database search after removing duplicates
n = 2,758
Studies excluded
(title and abstract screening)
n = 2,536
Full text screened
n = 222
Studies excluded (n = 195):
Reasons:
• Economic modelling with utility data from other sources (n=48)
• Reviews, methodology, protocol (n = 43)
• Cost study (n=16)
• Measured QoL but did not value (n=23)
• Updated paper existed (n=13)
• Measured utilities but did not report (n=13)
• Measured utilities of carers of PwP rather than PwP (n=5)
• Diagnosis of Parkinson's was not confirmed in the patient group (n=3)
• Insufficient data to assess psychometric properties (n=31)
Included (n=27):
• Cross-sectional studies (n=9)
• Longitudinal studies (n = 13)
• Mapping studies (n=5)

Figure 4-1: Flowchart of study screening process

Included studies were classified into three groups based on their study type for the assessment: Group A: cross-sectional studies (250, 452-459) for assessing 'known-group' and convergent validity (n = 9); Group B: longitudinal studies (251, 460-471) for assessing responsiveness (n = 13); Group C: mapping studies (472-476) (n = 5).

Among the included studies, one focused on people with early Parkinson's (465), three focused on advanced Parkinson's (466, 468, 471), and the remaining studies covered a wide range of severity levels. Among the cross-sectional studies, five explored the relationship between QoL and specific symptoms of Parkinson's, including apathy (452), depression (454, 458), life stress (454), presence of dyskinesia (250), presence of ‘wearing off’ period of drugs (250), sweating dysfunction (459). The remaining studies examined the association between QoL and more general Parkinson’s status, as measured by H&Y stages (453), MDS-UPDRS domains (455), SCOPA-AUT for automatic dysfunction (457), and the presence of Parkinson’s in general (456).

Among the longitudinal studies, there were seven RCTs (460, 462, 463, 465, 466, 468, 470), five prospective self-comparison studies (251, 461, 464, 469), and one cohort study (467). Two studies measured patients’ natural progression over a period (251, 464) and the remaining eleven studies evaluated the effect of an intervention. The interventions included: drugs (461, 465, 466, 468), provision of community-based nurse specialists (462), provision of instructions of clinical guidelines to neurologists (463), standardised pharmaceutical care (467), adherent therapy (460), deep brain stimulation surgery (471), and multidisciplinary rehabilitation (469, 470). Among the intervention studies, three studies conducted CUA (465, 466, 471) and one study conducted cost-consequence analysis (470).

EQ-5D (3L & 5L) was the most commonly used PbQoL instrument, which was reported in 19 studies (250, 251, 434, 452, 453, 457-465, 467-471). Meanwhile, HUI-3 was reported in two studies (454, 456), HUI-2 in one (458), 15D in two (453, 466), and the Disability and distress index (DDI) (often referred to as the Rosser Index) in one (458). EQ-5D, HUI-3 and HUI-2 have been introduced in 2.4.3.2 in Chapter 2. The DDI, developed by Rosser and colleagues in 1970s, is comprised of eight levels of disability (loss of function and mobility) and four levels of subjective distress, describing 29 disability/distress states (264, 477). One single index score is available for each state, which is generated through a valuation process using ranking and relative magnitude of severity exercise (478). The 15D is a less commonly used instrument developed in Finland (60). It was chosen in the Norwegian and Swedish studies due to its wider spectrum aspects of QoL, higher sensitivity with five levels on each attribute and availability of value sets in the specific country where the study was conducted (479, 480).

Among the non-preference based QoL measures identified as reference measures for the assessment of psychometric properties, the PDQ-39 was the most widely used Parkinson’s-specific QoL measure, reported in 9 studies (251, 458-460, 462, 463, 466, 470, 471), followed by the short version of the PDQ-39, the PDQ-8 in 5 studies (250, 453, 455, 464, 467), the PDQUALIF in one study (465), the PDQL (251) in one, and the generic QoL instrument, the SF-36 in one (470). The measures used in each of the included studies are presented in Table 4-1. The characteristics of all the identified QoL (including both Parkinson's specific and generic PbQoL measures) in the included studies are summarized in Table 4-2.

Table 4-1: Measures used in the included studies

Study	PbQoL instruments						Non-preference based QoL instruments					Common clinical measures	
	EQ-5D[c]	EQ-VAS	HUI-3	HUI-2	15D	DDI	PDQ-39	PDQ-8	PDQUALIF	SF-36	PDQL	UPDRS	H&Y
Studies for assessment of 'known-group' and convergent validity (n=9)													
Benito-Leon et al. 2012 (452)	✓	✓										✓	✓
Garcia-Gordillo et al. 2013 (453)	✓[b]	✓			✓			✓					✓
Jones et al. 2009 (454)			✓										
Luo et al. 2009 (250)	✓	✓						✓				✓	✓
Martinez-Martin et al. 2014 (455)	✓	✓						✓				✓[a]	✓
Pohar et al. 2009 (456)			✓										
Rodriguez-Blazquez et al. 2010 (457)	✓	✓											✓
Siderowf et al. 2002 (458)	✓			✓		✓	✓					✓	
Swinn et al. 2003 (459)	✓	✓					✓					✓	✓
Studies for assessment of responsiveness (n=13)													
Daley et al. 2014 (460)	✓						✓					✓	
Ebersbach et al. 2010 (461)	✓											✓	✓
Jarman et al. 2002 (462)	✓						✓						
Larisch et al. 2011(463)	✓	✓					✓						✓
Luo et al. 2010 (464)	✓	✓						✓					✓
Noyes et al. 2006 (465, 481)	✓	✓							✓			✓	
Nyholm et al. 2005 (466)					✓		✓					✓	✓
Reuther et al. 2007 (251)	✓	✓					✓				✓	✓	✓
Schröder et al. 2012 (467)	✓	✓						✓					✓
Stocchi et al. 2011 (468)	✓	✓										✓	✓
Trend et al. (469)	✓	✓											✓
Wade et al. 2003 (470)	✓	✓					✓			✓		✓	
Zhu et al. 2014 (471)	✓						✓					✓	

EQ-VAS EuroQol Visual Analogue Scale, *HUI-3* Health Utilities Index – Mark 3, *HUI-2* Health Utilities Index – Mark 2, *15D* 15 Dimensions, *DDI* Disability and Distress Index, *PDQ-39* Parkinson's Disease Questionnaire-39-item, *PDQ-8* Parkinson's Disease Questionnaire-8-item, *PDQUALIF* Parkinson's Disease QUAlity of LIFe scale, *SF-36* Short-Form 36-item, *PDQL* Parkinson's Disease Quality of Life questionnaire, *H&Y* Hoehn and Yahr scale, UPDRS Unified Parkinson's Disease Rating Scale

[a] Movement disorder society - UPDRS

[b] EQ-5D-5L

c refers to EQ-5D-3L if no other notation.

Table 4-2: Characteristics of the health-related QoL instruments in the included studies

Name	Generic or Parkinson 's specific	Possible score range (UK value)	Dimensions (D) / attributes
PbQoL measures			
EuroQoL EQ-5D-3L (14)	Generic	-0.594 (worst) ~ 1 (full health)	5D: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression
HUI-2 (Health Utilities Index – Mark 2) (482)	Generic	-0.03 (worst) ~ 1 (full health)	6D: sensation, mobility, emotion, cognition, self-care, and pain
HUI-3 (Health Utilities index – Mark 3) (483)	Generic	-0.36 (worst) ~ 1 (full health)	8D: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain
15D (15 Dimensions) (60)	Generic	0 (being dead) ~ 1 (full health)	15D: mobility, vision, hearing, breathing, sleeping, eating, speech, elimination (bladder and bowel function), usual activities, mental function, discomfort and symptoms, depression, distress, vitality, and sexual activity.
DDI (Disability and distress index, or Rosser Index) (477)	Generic	–1.486 (worst) ~ 1.0 (full health)	2D: disability and distress
Non-preference based QoL measures			
SF-36 (Short-Form 36-item) (211)	Generic	Physical summary: 0 (worst) ~ 400 (full health) Mental summary: 0 (worst) ~ 400 (full health)	8D: physical functioning, role physical, bodily pain, general health perceptions, vitality, role emotional, social role functioning, and mental health
PDQ-39/8 (Parkinson's Disease Questionnaire - 39/8-item) (131)	Specific	0 (best) -100 (worst)	8D: mobility, ADL, emotions, stigma, social support, cognition, communication, and bodily discomfort
PDQUALIF (Parkinson's Disease QUAlity of LIFe scale) (132)	Specific	0 (best) -100 (worst)	7D: social/ role function, self-image/ sexuality/sleep, outlook, physical function, independence, urinary function and one global health-related quality of life item
PDQL (Parkinson's Disease Quality of Life questionnaire) (133)	Specific	37 (worst) -185 (best)	4D: Parkinsonian symptoms, systemic symptoms, emotional functioning, and social functioning

4.4.2 Assessment of the construct validity and responsiveness

The assessment relied on the expectations of the relationship between the PbQoL instrument and the group defining criteria, or between the PbQoL instrument and the reference measure. Given this, in addition to describe what was reported in each of the included studies for each assessment, the results section also provides the explanations of the expectations and whether the results of PbQoL instrument met the expectations. The degree to which the result of PbQoL measure met the expectations are marked in the tables as 'assessment result'.

4.4.2.1 Known-group validity

Four studies provided sufficient evidence (i.e. reference measure available, groups that differed in characteristics are defined, PbQoL scores for each group available, see Section 4.3.3 for details) for the assessment of the known-group validity of the EQ-5D-3L (250, 452, 458, 459), two studies for the HUI-3 (454, 456), one study for the EQ-5D-5L and 15D (453), and one study for the DDI and HUI-II (458). The characteristics of these studies are shown in Table 4-3 along with the assessment results.

EQ-5D-3L index scores achieved statistically significant differences between the groups defined by the presence of apathy ('with' vs. 'without': 0.64 (0.26) vs. 0.83 (SD 0.17), p=0.001) (452), and in a case-control design comparing people with Parkinson's with sweating disturbances' and healthy controls (459). These results were expected given there were large differences between the groups shown in the reference measure. The study investigating apathy showed that there were large differences between the groups ('with' vs. 'without') in terms of the UPDRS motor score (p<0.001), disability and disease severity, as such the EQ-5D-3L was expected to be distinguishing between the groups. The other study was a case-healthy control design, which determined that there must be large difference in utilities between the groups given the large impact on QoL by the disease of Parkinson's as introduced in Chapter 1 (Section 1.2.3).

Inconsistent results were found for groups defined by the presence of dyskinesia ('with' or 'without') and the presence of 'wearing off' periods ('with' or 'without')

(250, 458). Statistically significant differences for both were shown in one study which reported a 0.28 difference of EQ-5D-3L between groups defined by dyskinesia (p=0.009), and 0.18 difference between groups defined by 'wearing off' (p<0.0001) (250). In contrast, the differences detected in the study by Siderowf et al. were not statistically significant: 0.09 (p= 0.43) for dyskinesia and 0.14 (p= 0.29) for the 'wearing off' period (458). The Siderowf study did not identify any difference in HUI-2 and DDI for the above groups either. The inconsistent result may be due to the smaller sample size in Siderowf study, but a closer investigation of the literature lowered the expectations on the strength of the relationship as well. Other literature also failed to reach consistent conclusions regarding the QoL and these two characteristics: Pechevis et al. (2005) found dyskinesia substantially affect patients' QoL measured by SF-36 and PDQL (484), while Schrage and Quinn did not find any difference between patients with / without motor fluctuations, or with/without dyskinesias, measured by PDQ-39 (79).

Moreover, Siderowf et al. (458) found a limitation in EQ-5D-3L and HUI-2's ability but not DDI to differentiate groups with mild Parkinson's defined by total UPDRS score. It showed that all of the three measures could differentiate between groups with upper (severe) and lower (mild) halves of UPDRS score (p < 0.001) and between first (mildest) and fourth (most severe) quartiles (p < 0.001); however, no difference was found in the EQ-5D-3L and HUI-2 between groups with first and second quartiles of UPDRS scores (mean difference = -0.009, p = 0.88 for EQ-5D-3L; mean difference=-0.008, p = 0.85 for HUI-2) whereas a statistically significant difference was shown in the DDI (p = 0.03). This should be considered as negative evidence for the 'known-group' validity of EQ-5D-3L and HUI-2 since the UPDRS were with high correlations with EQ-5D-3L (r=-0.61) and HUI-2 (-0.59), therefore their relationship was expected to be strong (this is reported in the Section 4.4.2.2 result of convergent validity).

In the same study, all three measures were found to be sensitive to symptoms including falling, freezing, visual hallucinations and depression with a statistically significant unadjusted mean difference between groups divided based on these symptoms (p < 0.05), although HUI-2 did not show difference between groups with and without swallowing difficulty (p = 0.20) (458).

For the HUI-3, both studies showed a statistical significant difference between the groups, with relative large magnitude of difference (454, 456). This was as expected given the known groups in these two studies were characterised by the aspects that have been known to affect QoL in a large way. The first study was a case-control study, which demonstrated a large difference between people with Parkinson's and the general population, with the HUI-3 score being 0.56 (95% CI 0.48, 0.63) and 0.87 (95% CI 0.87, 0.88) respectively (456). Besides the presence of Parkinson's, their QoL was expected to differ given there was a 20 years age difference between the groups as well as the difference in their number of medical conditions. The other study, by Jones et al. (2009), divided the Parkinson's patients by whether or not they had depression, and they reported that the HUI-3 values for those who had depression was 0.20 (95% CI 0.03, 0.37) and those who did not have depression was 0.49 (95% CI 0.39, 0.59); the difference was statistically significant after adjusting for several confounders such as age, sex, duration of Parkinson's etc (454). This study also evaluated the impact of life stress on HUI-3 utility values and identified statistically significant adjusted mean difference between not at all/not very stressful and quite a bit/extremely stressful (adjusted mean difference 0.19 ($p < 0.05$)), but no difference found between a bit stressful and quite a bit/extremely stressful groups (0.14, $p < 0.05$) (454).

One study reported EQ-5D-5L and 15D values for groups with varied severity of Parkinson's stratified with H&Y (stage 1&2 vs. stage 3&4) and found the mean values were statistically significantly different between the defined groups for both groups (453). This was also as expected given numerous evidence has shown that patients with advanced Parkinson's (usually H&Y stage equal or larger than 2.5) had substantially decreased QoL compared with the patients with early stages of Parkinson's (H&Y stage equal of less than 2).

Table 4-3: Characteristics of included studies – assessment of ‘known-group’ validity (n=7)

Study	**Year**	**Country**	**No. of participants**	**Study eligibility criteria**		**Study type**	**Group define criteria(C) and groups (G)**	**Evidence for ‘known-group’ validity: mean (standard deviation)**		**Assessment result[b]**
				Stage of Parkinson’s (Early or Advanced	**Other characteristics**			**Reference measure[a]**	**Preference-based measure**	
Benito-Leon et al. (452)	2012	Spain	557	Both	Recently diagnosed with Parkinson’s, duration <2 yrs, age ≥ 30	Cross-sectional	C: presence of apathy defined as Lille Apathy Rating Scale. G1: Noapathetic; G2: Apathetic	**UPDRS motor**; G1: 17.1 (8.5); G2: 24.8 (11.3); *p* < 0.001. **H&Y**; Higher proportion of early stages in G1; *p* < 0.001	**EQ-5D-3L**; G1: 0.83 (0.17); G2: 0.64 (0.26); *p* < 0.001. All attributes of EQ-5D-3L showed sig	✓

Garcia-Gordillo et al. (453)	2013	Spain	133	Both	Able to answer questions independently, age > 18	Cross-sectional	C: H&Y.
							G1: H&Y stages 1-2;
							G2: H&Y stages 3-4
								PDQ-8;
								G1: 18.30 (11.83);
								G2: 31.58 (19.56);
								p < 0.001
									EQ-5D-5L;
									G1: 0.70 (0.18);
									G2: 0.53 (0.28);
									p < 0.001.
									15D;
									G1: 0.81 (0.10);
									G2: 0.70 (0.17);
									p = 0.001
										✓

Jones et al. (454)	2009	Canada	259	Both	Self-reported Parkinson’s in a Canadian Community Health Survey	Cross-sectional	C: depression.
							G1/G2: without/with depression.
							C: life stress.
							G1’/G2’/G3’: not at all/ a bit/extremely stressful
								NA	**HUI-3**;
									G1: 0.49 (95% CI 0.39, 0.59);
									G2: 0.20 (95% CI 0.03, 0.37);
									p (G1 vs. G2) < 0.05.
									G1’: 0.42 (95% CI 0.29, 0.55);
									G2’: 0.38 (95% CI 0.24, 0.51);
									G3’: 0.23 (95% CI 0.10, 0.36);
									p (G1’ vs. G3’) <0.05;
									p (G2’ vs. G3’) >0.05
										o

Luo et al. (250)	2009	Singapore	135	Both	Without severe disabilities, Chinese MMSE score > 20	Cross-sectional	C: presence of dyskinesia. G1: no dyskinesia; G2: with dyskinesia. C: presence of ‘wearing off’ periods. G1’: no ‘wearing off’; G2’: with ‘wearing off’	NA	**EQ-5D-3L**; G1[C]: 0.80 (0.65, 1.0) G2[C]: 0.52 (0.52, 0.73) p (G1 vs. G2) < 0.01. G1’[C]: 0.80 (0.71, 1.0); G2’[C]: 0.62 (0.52, 0.78); p (G1’ vs. G2’) < 0.0001	✓

Study	Year	Country	No. of participants	Study eligibility criteria		Study type	Group define criteria(C) and groups (G)	Evidence for 'known-group' validity: mean (standard deviation)		Assessment result[b]
				Stage of Parkinson's (Early or Advanced	Other characteristics			Reference measure[a]	Preference-based measure	

Pohar et al. (456)	2009	Canada	261	Both	- Data from Canadian Community Health Survey	Cross-sectional, case-control	C: presence of Parkinson's.
							G1: With Parkinson's;
							G2: general population
								Age;
								G1: 68.9 (95% CI 66.6, 71.2);
								G2: 44.8 (95% CI 44.8, 44.9);
								$p < 0.05$.
								No. of medical conditions;
								G1: 3.0 (95% CI 2.5, 3.4);
								G2: 1.5 (95% CI 1.5, 1.5);
								$p < 0.05$
									HUI3;
									G1: 0.56 (95% CI 0.48, 0.63);
									G2: 0.87 (95% CI 0.87, 0.88);
									$p < 0.05$
										✓

Siderowf et al. (458)	2002	US	97	Both	Without cognitive impairment	Cross-sectional	C: total UPDRS score. G1 and G1': upper and lower halves; G2 and G2': 1st and 2nd quartiles; G3 and G3': 1st and 4th quartiles. C: depression. G4 and G4': with and without depression; and a various motor & non-motor symptoms	NA	**EQ-5D-3L**; Diff (G1vs.G1'):0.24; $p < 0.001$; Diff (G2vs.G2'):-0.009;$p = 0.88$; Diff (G3vs.G3'):0.40;$p < 0.001$; Diff (G4vs.G4'):0.26;$p < 0.001$. **DDI**; Diff (G1vs.G1'):0.09;$p = 0.007$; Diff (G2vs.G2'):0.01;$p = 0.03$; Diff (G3vs.G3'):0.17;$p = 0.02$; Diff (G4vs.G4'):0.17;$p < 0.001$. **HUI-II**; Diff (G1vs.G1'): 0.15;$p = 0.001$; Diff (G2vs.G2'):-0.008;$p = 0.85$; Diff (G3vs.G3'):0.25;$p = 0.001$; Diff (G4vs.G4'):0.17;$p < 0.001$.	o

Swinn et al. (459)	2003	UK	77	Both	Patients with sweating disturbances, without marked cognitive impairment or confusion	Cross-sectional, case-control	Case-control.
							G1: PwP with sweating disturbances;
							G2: healthy controls
								PDQ-39;
								G1: 41.7 (19.5);
								G2: NA
									EQ-5D-3L;
									G1: 0.47;
									G2: 0.85;
									$p < 0.005$
										✓

a. Reference measure could be either another PbQoL measure, Parkinson’s-specific QoL measure, or (if the former two not available) clinical measures.

b. Assessment result for discriminant validity: ‘✓’ evidence available to demonstrate that the PbQoL measure was able to show statistically significant difference between the known groups that were expected to differ as shown by the reference measure; ‘o’ some evidence available but still uncertain whether PbQoL measure can show statistically significant difference between the known groups that were expected to differ; ‘×’ – evidence showing the PbQoL measure failed to differentiate between the known groups.

c. median (inter-quantile).

Abbreviations: *MMSE* - Mini-Mental State Examination, *H&Y* Hoehn & Yahr scale, *HAD* Hospital Anxiety and Depression Scale, *SCOPA-Motor* Scales for Outcomes in Parkinson’s disease – Motor examination, *UPDRS* Unified Parkinson's Disease Rating Scale, *Diff* mean difference between groups, *sig* statistically significance, *C* criteria, *G* group, *NA* not available, PwP people with Parkinson’s.

4.4.2.2 Convergent validity

Five studies reported correlation coefficients between a PbQoL measure and a reference measure for the assessment of convergent validity (250, 453, 455, 457, 458). Among them, three studies examined the correlation between EQ-5D-3L and other measures (250, 455, 457), whereas two studies examined multiple PbQoL measures (one for EQ-5D-3L and 15D (453), another for EQ-5D-3L, DDI and HUI-II (458)) in regards to their correlation with other measures. The characteristics of these studies are shown in Table 4-4 accompanied by the evidence for assessment and the assessment result.

The EQ-5D-3L score showed strong correlation with (in the order of correlation coefficient from strongest to weakest) the PDQ-8 summary score (*r* = -0.75) (250), Movement disorder society – UPDRS (MDS-UPDRS) motor score (r = -0.72) (455), MDS-UPDRS non-motor score (r = -0.63) (455), UPDRS total score (r = -0.61) (458), and Non-Motor Symptoms Scale (NMSS) score (r=-0.57). It showed moderate to strong correlation with H&Y staging (*r* = -0.32 (250), *r* = -0.53 (455)), and moderate correlation with the Scales for Outcomes in Parkinson’s disease – Autonomic (SCOPA-AUT) (r = -0.49) (457) and UPDRS motor score (r = -0.39) (250).

The above results met expectations to some degree. EQ-5D-3L was expected to show the strongest correlation with the Parkinson’s disease QoL measure, PDQ-8. However, the correlation with the UPDRS motor score was unstable: 0.72 with MDS-UPDRS motor score, which was halved in another study with UPDRS motor score, given the similarity between the UPDRS and MDS-UPDRS scale (which was adapted from UPDRS scale).

Two studies compared multiple PbQoL measures in terms of their correlations with Parkinson’s-specific QoL measures, and the results were mixed (453, 458). Garcia-Gordillo et al. (453) found that the correlation between the 15D and the PDQ-8 summary score were stronger than that between the EQ-5D-5L and PDQ-8 summary score, with coefficients being -0.710 and -0.679, respectively. The authors explained that this could be due to the broad attributes of 15D such as leisure activities, housework, communication, worries about the future, which were likely to be substantially affected by Parkinson’s (453). As with the authors, this result was expected given both 15D and the PDQ questionnaire contains broadly scoped dimensions (Table 4-2) (60) (131). Siderowf et al. (458) compared DDI, EQ-5D-3L, and HUI-II and found that the utility score from EQ-5D-3L correlated most strongly with PDQ-39 while DDI showed the weakest correlation. Regarding the specific PDQ-39 dimensions, they found that the EQ-5D-3L correlated most strongly with the ADL attribute (*r* = -0.69) and weakly with social support (*r* = -0.27), HUI-II correlated most strongly with mobility (*r* = -0.62) and weakest with stigma (*r* = -0.12), and DDI correlated most strongly with mobility and ADL (*r* = -0.42 for both) and weakest with stigma (*r* = 0.067) (458). These results also met expectations given the three PbQoL measures all have a focus on daily functioning, rather than psychological and social wellbeing, as shown in the coverage of their dimensions summarized in Table 4-2.

Table 4-4: Characteristics of included studies – assessment of convergent validity (n=5)

Study	Publication year	Country	No. of participants	Study eligibility criteria		Study type	PbQoL measure (s)	Evidence for convergent validity		Assessment result[b]
				Stage of Parkinson's (Early or Advanced)	Other clinical characteristics			Reference measure [a]	Correlation coefficients (r)	
Garci-Gordillo et al. (453)	2013	Spain	133	Both	Be able to answer questions independently, age > 18	Cross-sectional	EQ-5D-5L, 15D	PDQ-8	15D/PDQ-8: -0.710. EQ-5D-5L/PDQ-8: -0.679	✓
Luo et al. (250)	2009	Singapore	31	Both	Well enough to complete surveys	Before and after self-comparison, 4 yrs	EQ-5D-3L	PDQ-8 SI, H&Y, UPDRS motor	EQ-5D-3L/PDQ-8: -0.75. EQ-5D-3L/H&Y: -0.32. EQ-5D-3L/UPDRS motor:-0.39	✓
Martinez-Martin et al. (455)	2014	Argentina, Cuba, Mexico, US, and Spain	435	Both	Spanish native speakers, at any age and severity of Parkinson's	Cross-sectional	EQ-5D-3L	H&Y, NMSS, MDS-UPDRS-non motor, MDS-UPDRS-motor	EQ-5D-3L/H&Y: -0.53. EQ-5D-3L/NMSS: -0.57. EQ-5D-3L/MDS-UPDRS-non motor: -0.63. EQ-5D-3L/MDS-UPDRS-motor: -0.72.	✓
Rodriguez-Blazquez et al. (457)	2010	Spain	387	Both	Age ≥ 30 at disease onset, with a main carer	Cross-sectional	EQ-5D-3L	SCOPA-AUT	EQ-5D-3L/SCOPA-AUT:-0.49	✓
Siderowf et al. (458)	2002	US	97	Both	Without cognitive impairment	Cross-sectional	EQ-5D-3L, DDI, HUI-II	PDQ-39 all sub-attributes, UPDRS	EQ-5D-3L/PDQ-39 all attributes: from -0.27 (social support) to -0.69 (ADL). EQ-5D-3L/UPDRS total: -0.61. HUI/ PDQ-39 all attributes: from -0.12 (stigma) to -0.62 (mobility). HUI/UPDRS total: -0.59. DDI/PDQ-39 all attributes: from 0.067 (stigma) to -0.42 (mobility/ADL). DDI/UPDRS total: -0.40	o

a. Reference measure could be either another PbQoL measure, Parkinson's-specific QoL measure, or (if the former two not available) clinical measures.

b. Assessment result for convergent validity: '✓' evidence available to demonstrate that PbQoL measure and the reference measure were highly related (r ≥ 0.5); 'o' the PbQoL measure and the reference measure were moderately correlated (0.3 ≤ r < 0.5); '×' the PbQoL measure and the reference measure were weakly correlated (r < 0.3).

Abbreviations: *NMSS* Non-Motor Symptoms Scale, *SCOPA-AUT* SCales for Outcomes in PArkinson's disease – AUTonomic, *ADL* Activities of Daily Living, *H&Y* Hoehn & Yahr stage, , r correlation coefficient

4.4.2.3 Responsiveness

Thirteen studies provided required information according to the eligibility criteria (Section 4.3.3) to allow an assessment of responsiveness of the PbQoL measures, including twelve studies for the EQ-5D-3L (251, 460-465, 467-471) and one study for the 15D (466). The evidence for the assessment of responsiveness is provided in Table 4-5. The expectations on the mean change and direction of PbQoL score were established based on the mean change and direction of the reference measure, as well as the relationship between the PbQoL and the reference measure.

Overall, there is some evidence supporting the responsiveness of the PbQoL measures. The one 15D study, by Nyholm et al. (466), demonstrated improved QoL in the duodenal levodopa infusion arm compared to conventional oral polypharmacy arm on both PDQ-39 and 15D (both $p < 0.01$); agreement between the Parkinson's specific QoL measure PDQ-39 and the 15D supported the responsiveness of the 15D. Among the twelve EQ-5D-3L studies, half (n=6) showed consistency between the EQ-5D-3L and the reference measures in terms of the evidence for whether there was a statistically significant change over time; the reference measures included UPDRS part II ADL (461), PDQ-39 (462, 463, 471), PDQ-8 and H&Y (464), and the Hospital Anxiety and Depression scale (469).

Concerns are raised to various degrees regarding the agreement between the EQ-5D-3L and reference measures in the remaining six studies (251, 460, 465, 467, 468, 470). Among them, four (251, 460, 467, 468) studies (Group A as below) showed a change in the reference measures but not in EQ-5D-3L whereas in the other two studies (Group B as below) (465, 470), change was not found in the reference measure but in EQ-5D-3L.

Group A: change shown in reference measures but not in EQ-5D-3L (n=4)

Among the four studies, Daley et al. (460) reported statistically significant higher QoL as shown on PDQ-39 summary score, mobility, ADL, emotional wellbeing, cognition, communication and bodily discomfort after adherence therapy as compared to routine care in a RCT, but the change in EQ-5D-3L was small and not statistically significant (mean difference 0.07, 95% CI -0.1, 0.2). Nevertheless, given the sample size of this study is small (n=76), the assessment result was determined to be 'uncertain'. Similarly, Schroder et al. (467) detected an improvement (difference = -3.3, *p* = 0.034) in PDQ-8 score in the group with standardised community pharmaceutical care for eight months and deterioration (difference = 4.4, *p* = 0.019) in the group with usual care. However, this treatment benefit was not only not replicated in EQ-5D-3L score for either group, but the direction of change was the opposite, although the change was not statistically significant (difference for intervention = 0.02, p=0.29; difference for control = -0.03, p=0.13; sample size: n=161). In both of the above cases, given the change detected in the specific QoL measures, it was expected that change was also shown in the PbQoL measure, which was not the case.

In addition to the PDQ, the inconsistency was also found when using the UPDRS clinical measure as reference measure. Stocchi et al. (468) compared adjunctive ropinirole prolonged release and immediate release in a RCT and reported an improved UPDRS total motor score (*p* = 0.022), but a non-significant improved UPDRS ADL score (*p* = 0.270) and EQ-5D-3L score (difference = 0.03, *p* = 0.165). Although the difference was not statistically significant, given the unstable correlation between the UPDRS and the PbQoL measures as identified in the result of convergent validity (see Section 4.4.2.2), the non-significant result for EQ-5D-3L score was not considered as evidence rejecting the responsiveness of EQ-5D-3L.

One study (251) reported a counterintuitive result between the clinical measures and the QoL measures. Reuther et al. (251) evaluated the change in QoL and clinical measures over one year without any study intervention (i.e. before – after comparison) in 145 patients. They found that clinical scores deteriorated (H&Y, *p* = 0.000, and UPDRS, *p* = 0.019); however the scores of PDQ-39 and PDQL improved (PDQ-39, difference = -3.8, *p* = 0.000, and PDQL, difference = 4.2, *p* = 0.030), and there was no difference in the EQ-5D-3L (difference = 0.01, *p* = 0.488). In addition, all of the PDQ-39 sub-dimensions in their study showed an improvement, including the dimensions that may have an overlapped concept with EQ-5D-3L dimensions such as mobility, ADL, emotional wellbeing, and bodily discomfort. The authors briefly explained that this could be due to the bias in repetitive measurement or other factors but did not provide any details. Although the result was inconsistent, the Parkinson's QoL measures were judged to be a more suitable measure assessing PbQoL instruments given they both measure the concept of QoL, compared to the clinical measures which do not. As such the EQ-5D-3L was expected to show a larger difference given all the dimensions of PDQ-39 showed an improvement. This expectation was not met based on the above results.

Group B: change not shown in reference measures but shown in EQ-5D-3L (n=2)

In contrast to the above results (Group A) where EQ-5D-3L was not responsive to a confirmed change, two studies showed statistically significant change over time in the EQ-5D-3L but not in the reference measures (465, 470). Noyes et al. (465) compared pramipexole and levodopa in a RCT with 301 patients over four years. Although a difference in PDQUALIF was detected it was not statistically significant, whereas EQ-5D-3L showed a difference between the arms from year 2 to 3 (difference = 0.048, p = 0.03) and year 3 to 4 (difference = 0.071, p = 0.04). Wade et al. (470) compared multidisciplinary rehabilitation program versus usual care in 94 patients, in which difference was shown between the arms in the SF-36 physical score and EQ-5D-3L score, albeit the difference was small (0.026 for EQ-5D-3L, p=0.026), while no difference found for PDQ-39 (0.5 on a 0-100 scale, p=0.687) and SF-36 mental score (0.5 on a 0-400 scale, p=0.655). Given that the reference measures (SF-36 and PDQ-39) were not consistent in term of confirming the happening of the change, and the fact that the difference shown in EQ-5D-3L was small in size, albeit statistical significant, this study was considered as 'uncertain' evidence for the assessment of responsiveness.

Table 4-5: Characteristics of included studies - assessment of responsiveness (n=13)

Study	Publication year	Country	No. of participants	Study eligibility criteria		Study type and time horizon	Intervention (I) and comparator (C) or; before (B) and after (A)	Evidence for responsiveness – *change* from baseline to primary endpoint: Mean change (standard deviation)		Assessment[c]
				Stage of Parkinson's (Early or Advanced)	Other clinical characteristics			Reference measure[a]	PbQoL measure	

Daley et al. (460)	2014	UK	76	Both	On anti-parkinsonian drug(s), no dementia	RCT, 12 wks	I: adherence therapy;
							C: routine care
								PDQ-39;
								I : -6.8 (6.4);
								C: 2.3 (7.4);
								Diff: -9.0 (95% CI -12.2, -5.8);
								$p < 0.001$
									EQ-5D-3L;
									I: 0.04 (0.3);
									C: -0.03 (0.3);
									Diff: 0.07 (95% CI -0.1, 0.2);
									$p = 0.055$
										o

Ebersbach et al. (461)	2010	Germany	61	Both	Responsive to levodopa, had not responded to or did not tolerate entacapone, age 30-80, H&Y 2-4, on stable medication for ≥ 4 wks	Before and after self-comparison, 4 wks	B and A: tolcapone targeting sleep quality	**UPDRS part II (ADL)**;
								B[e]: 15.1 (7.1);
								A[e]: 10.8 (7.0);
								$p < 0.0001$
									EQ-5D-3L;
									B[e]: 0.562 (0.234);
									A[e]: 0.678 (0.206);
									$p = 0.0001$
										✓

Jarman et al. (462)	2002	UK	1859	Both	On anti-parkinsonian drug(s)	RCT, 2 yrs	I: provision of community based nurses specialists; C: no provision. B and A: Also analysed deterioration over 2 yrs' of all participants	**PDQ-39**; B and A: all sub-attributes: $p < 0.05$; Diff[b]: 0.47 (95% CI -2.72, 3.66); $p = 0.77$	**EQ-5D-3L**; B and A: -0.10 (-0.12, -0.08); $p < 0.001$; Diff[b]:-0.02 (95% CI -0.06, 0.02); $p = 0.30$	✓

Larisch et al. (463)	2011	Germany	386	Both	Not reported	Cluster RCT, 9 mths	I: providing instructions of clinical practice guidelines to neurologists;
							C: without instructions
								PDQ-39;
								I: 1.8 (11.2);
								C: 1.1 (11.5);
								p[d]=0.7591
									EQ-5D-3L;
									I: -0.001 (0.195);
									C: 0.007 (0.209);
									p[d]=0.5148
										✓

Luo et al. (464)	2010	Singapore	31	Both	Well enough to complete surveys	Before and after self-	No intervention	**PDQ-8 SI**;
								B[e]: 17.74 (14.17);
								A[e]: 35.08 (17.43);
									EQ-5D-3L;
									B[e]: 0.76 (0.23);
									A[e]: 0.52 (0.33);
										✓

Study	Publication year	Country	No. of participants	Study eligibility criteria		Study type and time horizon	Intervention (I) and comparator (C) or; before (B) and after (A)	Evidence for responsiveness – *change* from baseline to primary endpoint: Mean change (standard deviation)		Assessment[c]
				Stage of Parkinson's (Early or Advanced	Other clinical characteristics			Reference measure[a]	PbQoL measure	
						comparison, 4 yrs		$p < 0.0001$. **H&Y**; B[e]: 2.09 (0.38); A[e]: 2.40 (0.70); $p = 0.0133$	$p = 0.0014$	
Noyes et al.(465, 481)	2006	US	301	Early	Age ≥ 30, duration with Parkinson's ≤ 7 yrs, H&Y 1-3, required dopaminergic anti-Parkinson's therapy	RCT, 4 yrs; cost-utility analysis	I: pramipexole; C: levodopa	**PDQUALIF**; Diff over 4 yrs:0.040; $P = 0.45$. Diff from yr 2 ~3: 0.015; $P = 0.36$. Diff from yr 3~4: 0.036; $P = 0.25$	**EQ-5D-3L**; Diff over 4 yrs: 0.149; p=0.11. Diff from yr 2 ~3: 0.048; $P = 0.03$. Diff from yr 3~4: 0.071 p=0.04	o

Nyholm et al.(466, 481)	2005	Sweden	24	Advanced	Experiencing motor fluctuations and dyskinesia	Crossover RCT, 2 three wks trial plus 6 mths follow up; cost-utility analysis	I: duodenal levodopa infusion (DLI) as monotherapy;
							C: conventional oral polypharmacy
								PDQ-39;
								I[e]: median 25 (range 10-42);
								C[e]: median 35 (range 16-55);
								$p < 0.01$
									15D;
									I[e]: median 0.78 (range 0.64-0.95);
									C[e]: median 0.72 (range 0.58-0.88);
									$p < 0.01$
										✓

Reuther et al. (251)	2007	Germany	145	Both	Not reported	Prospective self-comparison non-intervention, 12 mths	No intervention	**PDQ-39**;
								B[e]: 29.4 (17.5);
								A[e]: 25.6 (16.2);
								$P = 0.000$.
								PDQL;
								B[e]: 118.6 (27.5);
								A[e]: 122.8 (26.1);
								$P = 0.030$.
								H&Y;
								B[e]: 2.81 (1.16);
								A[e]: 3.13 (1.04);
								$P = 0.000$.
								UPDRS;
								B[e]= 48.1 (33.3);
									EQ-5D-3L;
									B[e]: 0.61 (0.30);
									A[e]: 0.60 (0.28);
									$P = 0.488$
										×

Study	Publication year	Country	No. of participants	Study eligibility criteria		Study type and time horizon	Intervention (I) and comparator (C) or; before (B) and after (A)	Evidence for responsiveness – *change* from baseline to primary endpoint: Mean change (standard deviation)		Assessment[c]
				Stage of Parkinson's (Early or Advanced	Other clinical characteristics			Reference measure[a]	PbQoL measure	
								A[e]= 53.1 (34.0); *P* = 0.019		
Schröder et al. (467)	2012	Germany	161	Both	On anti-parkinsonian medication(s), age > 35, sufficient physical and cognitive ability to complete questionnaires without assistance	Cohort study, 8 mths	I: standardised community pharmaceutical care; C: usual care	**PDQ-8**; I: -3.3 (95% CI -6.3, -0.3); p[f]=0.034. C: 4.4 (95% CI 0.8, 8.1); p[f]=0.019	**EQ-5D-3L**; I:0.02 (95% CI -0.02, 0.06); p[f]=0.29. C:-0.03 (95% CI -0.08,0.01); p[f]=0.13	✗
Stocchi et al. (468)	2011	Bulgaria, Canada, Czech Republic, France, Hungary, Poland, Romania, Spain, UK.	177	Advanced	Age ≥30, H&Y 2-4, not adequately controlled on L-dopa (3-12 hrs of daily awake time spent as 'off' time)	RCT, 24 wks	I: adjunctive ropinirole prolonged release; C: immediate release	**UPDRS total motor**; Diff: -2.30 (95% CI -4.27, -0.33); *P* = 0.022. **UPDRS ADL in 'off' state**; Diff: -0.77 (95% CI -2.13, 0.60); *P* = 0.270. **UPDRS ADL in 'on' state**; Diff: -0.69 (95% CI -1.51, 0.13); *P* = 0.100	**EQ-5D-3L**; Diff: 0.03 (95% CI -0.01, 0.08); *P* = 0.165	o

Trend et al. (469)	2002	UK	118	Both	Score of at least 7/10 on Hodkinson's mini-mental test, no cognitive impairment	Before and after self-comparison	B and A: intensive multidisciplinary rehabilitation	**HAD anxiety**:
								B[e]: 5.51 (3.31);
								A[e]: 5.19 (3.43);
								p value not sig.
								HAD depression:
								B[e]: 6.06 (2.88);
								A[e]: 5.57 (2.80);
								P = 0.029.
								p value of all of the other motor and non-motor scales achieved sig.
									EQ-5D-3L;
									B[e]: 0.55 (0.24);
									A[e]: 0.63 (0.22);
									P = 0.001
										✓

Study	Publication year	Country	No. of participants	Study eligibility criteria		Study type and time horizon	Intervention (I) and comparator (C) or; before (B) and after (A)	Evidence for responsiveness – *change* from baseline to primary endpoint: Mean change (standard deviation)		Assessment[c]
				Stage of Parkinson's (Early or Advanced	Other clinical characteristics			Reference measure[a]	PbQoL measure	

Wade et al. (470)	2003	UK	94	Both	Without severe cognitive losses	Crossover RCT, 24 wks; vost-consequence analysis	I: multidisciplinary rehabilitation program;
							C:usual care
								PDQ-39;
								B: 25.5 (10.7);
								A: 26.0 (12.7);
								P = 0.687.
								SF-36 physical;
								B: 29.5 (11.1);
								A: 27.28 (10.9);
								P = 0.046.
								SF-36 mental;
								B: 51.0 (8.4);
								A: 50.5 (10.3);
								P = 0.655
									EQ-5D-3L;
									B: 0.72 (0.22);
									A: 0.66 (0.21);
									P = 0.026
										o

Zhu et al. (471)	2014	HK (China)	13	Advanced	Disabling or troubling motor symptoms, dopa responsive, clear understanding risk of and realistic about surgery outcomes, age<70	Prospective before and after self-comparison, 2 yrs;
						cost utility analysis (before-after)
							B and A: deep brain stimulation surgery	**PDQ-39**;
								B[e]=39 (13);
								A[e]=27 (14);
								P = 0.019.
									EQ-5D-3L;
									B[e]=0.504(0.24);
									A[e]=0.662(0.13);
									P = 0.033.
										✓

ADL activities of daily living, *H&Y* Hoehn & Yahr scale, *HAD* Hospital Anxiety and Depression scale, *UPDRS* Unified Parkinson's Disease Rating Scale, *I* intervention group, *C* control group, *B* before, *A* after, *Diff* difference of scores between the ***changes*** of the two comparative groups over the trial period, *yrs* years, *mths* months, *hrs* hours, sig significant

[a] Reference measure could be either another PbQoL measure, Parkinson's-specific QoL measure, or (if the former two not available) clinical measures.

[b] Difference between the intervention group and the control group at endpoint (no difference was found between two groups at baseline.)

[c] Assessment result for responsiveness: '✓' evidence available to demonstrate that PbQoL measure and the reference measure were consistent; 'o' weak evidence available but uncertain; or the PbQoL measure and the reference measure were not always consistent; '×' the PbQoL measure and the reference measure were inconsistent.

[d] Hypothesis testing if the difference in change over time between the intervention and the control group equals to zero

[e] Score at either baseline or endpoint, instead of change over time

[f] Hypothesis testing if the change within group over time equals to zero

4.4.3 Performance of PbQoL measures in economic evaluations

Since the results of PbQoL measures are usually reported alone in a separate CUA study while the results of Parkinson's specific QoL measures and clinical measures are usually reported in the main clinical study report, the number of CUA studies that met our eligibility criteria (which required a reference measure together with a PbQoL measure) for the assessment purpose was limited. Three studies conducted CUAs of health interventions (465, 466, 471, 481).

The first CUA was one of the two studies in the Group B described in the last section 4.4.2.3, where there was statistically significant change over time in the EQ-5D-3L but not in the reference measures (465, 481)[7]. The CUA used the identified EQ-5D difference in a four-year economic evaluation model and determined that the probability that the intervention of pramipexole compared with levodopa was cost effective was 0.57 when the WTP threshold was USD 50,000. However, the sensitivity analysis on the QALY gained revealed great uncertainty, which is expected given no difference was identified in the Parkinson's specific QoL measure. The authors varied the QALY profiles following drop-out of participants and found that the ICER could be varied up to USD 233,025 per QALY with the probability for the intervention to be cost effective decreasing to 0.14, and down to USD 29,759 per QALY with the probability increasing to 0.88.

The other two CUAs both come from the studies that supported the responsiveness of PbQoL measures, one with 15D evaluating the effect of duodenal levodopa infusion for advanced Parkinson's in Sweden (466), the other with EQ-5D-3L evaluating the DBS for advanced Parkinson's in Hong Kong (471). Despite this, the 15D study reported that the change in 15D was among the parameters that had the greatest impact on the cost per QALY, although no detail regarding the amount of impact from varying 15D was provided. The EQ-5D-3L study did not report any sensitivity analysis so the uncertainty around the EQ-5D-3L estimate on the ICER was not determined, although the size of improvement of PDQ-39 and EQ-5D-3L

[7] The CUA was reported in a separate paper from the other outcomes.

were both relatively large (mean difference 0.203, p=0.013 for the first year; mean difference 0.158, p=0.033 for the second year) (471).

4.4.4 Mapping algorithms

Five mapping studies were identified from the literature search which used mapping from the non-preference based measure to the preference-based measures (472-476). This includes two studies mapping from the PDQ-8 to the EQ-5D-3L (473, 474) and three studies mapping from the PDQ-39 to the EQ-5D-3L (472, 475, 476). Their characteristics and the resulting mapping algorithms are presented in Table 4-6 and Table 4-7, respectively.

Table 4-6 showed that four out of the five studies estimated a prediction model using regression approaches in an original dataset and then validated their derived algorithm in one or more validation dataset(s) (473-476). They compared different regression models based on model fit informed by statistical indicators. The remaining one study used a Markov blanket-based approach, which is a method for learning multi-dimensional Bayesian network classifiers to identify the relationships within multi-dimensional classification systems (472).

Sample size varies in the original datasets for deriving the different models. The largest dataset comes from Kent and colleagues (2015) (475) which contains 9,123 pairs of observations for estimation of the algorithm and 719 pairs of observations for validation. Linear regression, beta regression, mixtures of linear and beta regressions, and multinomial logistic regression were compared based on model fit indicators including mean error, mean absolute error, and mean square error. The estimated regression model incorporated adjustment for age and sex (Table 4-7).

For the mapping results shown in Table 4-7, mobility, ADL and bodily discomfort were included in all mapping algorithms, and emotional wellbeing was included in all but one (474) algorithms. However, four of the five mapping algorithms did not include half of the PDQ dimensions which are related to mental health and overall wellbeing aspects of QoL, i.e. stigma, social support, cognition and communication (472-475). Especially, stigma was not included in any of the algorithms.

Table 4-6: Characteristics of studies mapping PDQ-39/PDQ-8 scores to EQ-5D-3L utility values

Author, year	Year	Country	Sample size for		Mapping		Mapping models used	Measure of model performance[a]
			Estimation of the algorithm	Validation of the algorithm	From PDQ-39 / 8	To EQ-5D-3L		
Borchani et al. (472)	2012	Spain	448	-	PDQ-39 each question	Each dimension	Markov blanket-based approach using Multi-dimensional Bayesian network classifiers (MBC), class-bridge decomposable MBC, independent Marokov blankets, and independent PC Bayesian networks, back propagation for multi-label learning (BP-MLL), multi-label k-nearest neighbor (ML-kNN), multinomial logistic regression (MNL), ordinary least squares (OLS), and censored least absolute deviations (CLAD)	MSE, MAE, R square, AbsDiff
Cheung et al. (473)	2008	Singapore	162	162	PDQ-8 each question	Overall score	OLS, censored least absolute deviations method	R square, MAE
Dams et al. (474)	2013	Germany	121	Overall number not reported[b]	PDQ-8 each question	Overall score	OLS, fractional polynomial regression; logarithmic function	R square, RMSE, Pregibon link test, BIC
Kent et al. (475)	2015	UK	9,123 pairs from 2043 patients	719 pairs from 352 patients	PDQ-39 each dimension	a. each dimension b. overall score	OLS, 2-part Beta Regression, Finite Mixture Models, Mixtures of linear regressions, mixture of beta regressions, multinomial logistic regression	ME, MAE, MSE.
Young et al. (476)	2013	Austria	80	16	PDQ-39 Each dimension	Each dimension	Ordinal regression with the Cauchit link function,	MAE,RMSE.

Abbreviations: OLS – ordinary least square

a ME - Mean error, calculated as the average difference between observed and predicted utilities; MAE – mean absolute error, calculated as the average of the absolute differences between observed and predicted utilities; MSE – mean square error, calculated as the average of squared differences between observed and predicted utilities; RMSE – root mean square error, calculated as the root square of MSE; AbsDiff – the absolute difference, calculated as the absolute difference between the true and predicted EQ-5D utility mean scores; BIC – Bayesian information criterion; PwP – people with Parkinson's.

b. Data come from three datasets: 1). authors' own unpublished data; 2).Siderowf et al.(Germany, 97PwP) (458); 3). Schrag et al. (124 PwP, UK) (485)

Table 4-7: Summary of PDQ dimensions included in each mapping algorithm

Author, year	PDQ1-Mobility	PDQ2-Activities of daily living	PDQ3-Emotional wellbeing	PDQ4-Stigma	PDQ5-Social support	PDQ6-Cognition	PDQ7-Communication	PDQ8-Bodily discomfort	Other variables in the model	Algorithm of EQ-5D utility
Borchani et al. (472)	✓	✓	✓	-	-	-	-	✓	None	Not reported.
Cheung et al. (473)	✓	✓	✓	-	-	-	-	✓	None	Utility=1 if at least seven responses are "never", otherwise Utility=1-0.135-0.052*PDQ1-0.0034*PDQ2-0.031*PDQ3-0.030*PDQ7 R^2=52.1%.
Dams et al. (474)	✓	✓	-	-	-	-	-	✓	None	Utility=0.9298-0.00004*$PDQ1^2$-0.00002*$PDQ2^2$-0.00004*$PDQ8^2$ R^2=60.34%.
Kent et al. (475)	✓	✓	✓	-	-	-	-	✓	Age, sex	Not reported.
Young et al. (476)	✓	✓	✓	-	✓	✓	✓	✓	None	Overall EQ-5D utility function is not reported, utility function to each EQ-5D dimension is available.

4.5 Summary of results

This chapter systematically reviewed the use of the PbQoL measures in people with Parkinson’s, assessed the construct validity and responsiveness of PbQoL measures, and summarized the mapping algorithms from non-preference based measures to EQ-5D-3L. The EQ-5D-3L was found to be predominantly used as the PbQoL measure in Parkinson’s while the PDQ-39 was the most widely used Parkinson’s-specific QoL measure among included studies.

EQ-5D-3L did achieve statistically significant differences between the known groups divided based on clinical characteristics in most studies, but it may have limited sensitivity to detect differences in QoL among patients with mild Parkinson’s as evidenced by the subgroup analysis in an included study (458). Good evidence of known-group validity has also been demonstrated in the HUI-3, EQ-5D-5L, 15D, HUI-2, and DDI despite limited evidence being available to allow the assessment. HUI-2 may be less sensitive among patients with mild Parkinson’s as there is no difference in the mean utility score between patient groups with first and second quartile UPDRS scores (458).

In terms of convergent validity, overall moderate to strong correlations were shown between the PbQoL measures (EQ-5D-3L, EQ-5D-5L, 15D, DDI, and HUI-II) and Parkinson’s-specific QoL measures/clinical measures. It was found that the EQ-5D-3L, DDI, and HUI-II all correlated most strongly with the physical attributes (i.e., mobility and ADL) of PDQ-39 and least strongly with mental and wellbeing attributes (i.e., social support and stigma).

For responsiveness, most evidence was found for the EQ-5D-3L. The agreement between EQ-5D-3L and the Parkinson’s-specific QoL/clinical measures in regards to the change over time varied across studies. Half of the studies showed that EQ-5D-3L scores reflected changes in clinical status over time as shown on the reference measures, while the other half failed to reach consistent conclusions between the measures. Concerns are raised in the responsiveness of the PbQoL measures especially the EQ-5D-3L to the changes over time that are specific to disease progression in the Parkinson’s population.

Through a summary of the identified mapping algorithms, it is found that four of the five mapping algorithms did not include half of the PDQ dimensions, i.e. stigma, social support, cognition and communication (472-475). Using these mapping algorithms to generate EQ-5D-3L values may be problematic as these algorithms neglected to some extent the impact on mental and overall wellbeing aspects of Parkinson's.

4.6 Discussion

There is evidence from this review that the mental/wellbeing attributes of PDQ-39 may not be fully captured by the EQ-5D instrument. Parkinson's is a chronic, progressive condition which has been shown to affect mental/wellbeing aspects of QoL and as such it is important to include appropriate valuations for improvements in such attributes within priority setting decisions. The importance of these mental/wellbeing aspects is demonstrated by consistent presence of such attributes within Parkinson's-specific QoL measures and by previous literature examining the effect of the mental and wellbeing aspects on Parkinson's patients' QoL (95, 486). With approximately half of the domains in PDQ-39/PDQ-8, PDQUALIF, and PDQL relating to aspects other than physical health, such domains, e.g., social communication, stigma/self-image, emotional functioning, cognition, and outlook, are highly likely to have a substantial impact on patients' QoL. A recent systematic review found that depression was the most frequently identified determinant of HrQoL in people with Parkinson's among all the demographic and clinical factors (487). Therefore, sufficient incorporation of valuations for these broader attributes is crucial when measuring PbQoL in Parkinson's.

The utilities from the PbQoL measures generally discriminated well between groups and correlated well with Parkinson's clinical and QoL measures. However, the inconsistency in findings of responsiveness between those measures cautioned that the change shown on clinical measures may not necessarily lead to the same change in QoL scores. Reuther et al. (251) assumed that there might be other undetected factors leading to the opposite change of QoL scores to the clinical measures. One reason might be the fact that clinical measures such as H&Y and UPDRS focus mostly on the physical symptoms of Parkinson's while QoL measures are subjective to individuals and based on overall experience of health and wellbeing. This may also help explain our finding that the PbQoL measures that focused on physical health should be theoretically able to discriminate between groups defined by clinical factors. Besides this, as clinical status or objective health status is usually one of the primary predictors of QoL, it is reasonable to expect that PbQoL measures would display discriminant and convergent validity.

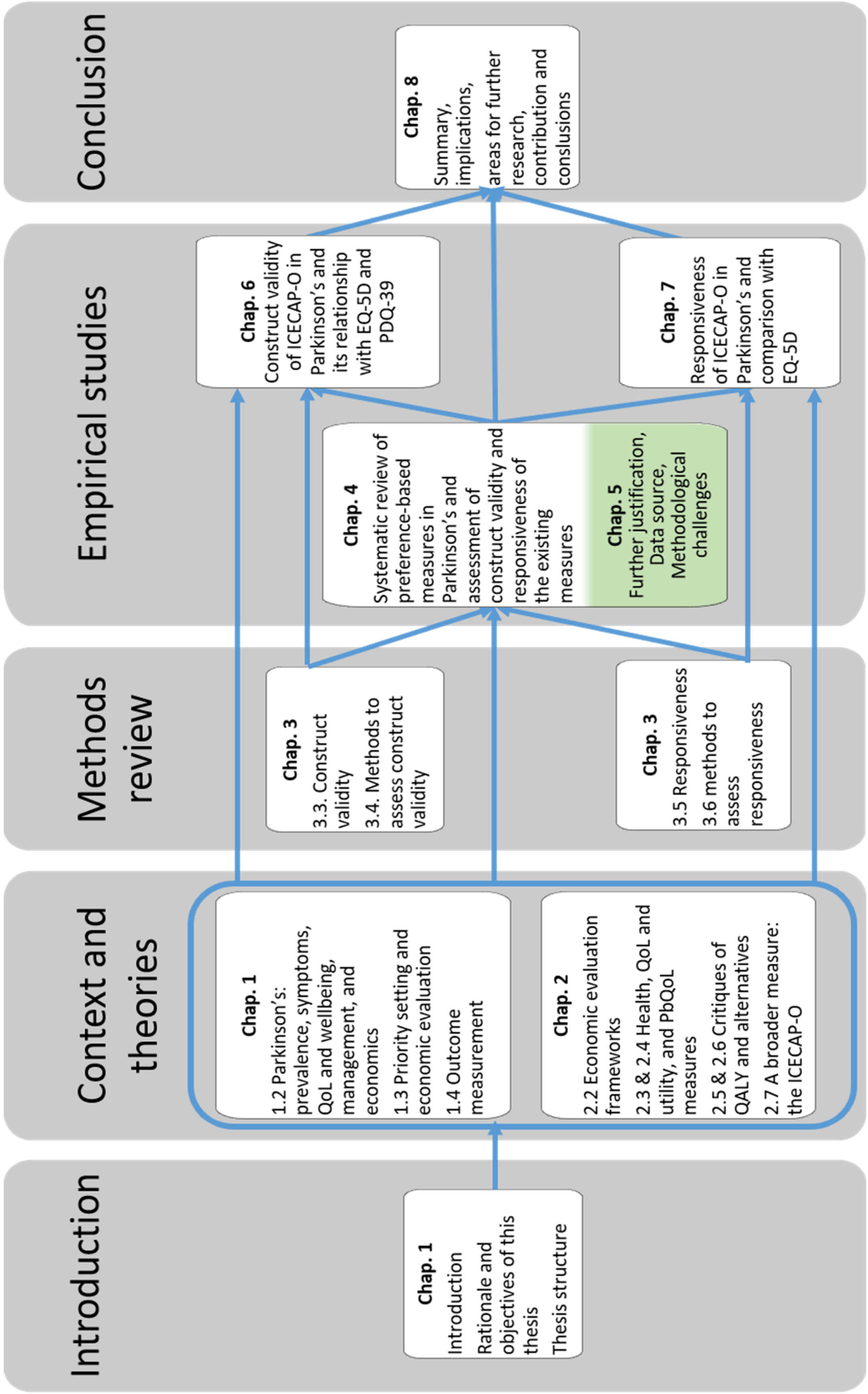
Responsiveness of PbQoL measures is crucial to economic evaluations. In a bid to measure resource use and QALYs, economic evaluations often need to be carried out longitudinally over an appropriate and meaningful time horizon depending upon the intervention being assessed. Previous studies have suggested that the results of economic evaluations are sensitive to the change of utility values when chronic conditions or long-term sequelae are involved (488); Parkinson's is one of those conditions. Therefore, lack of definite evidence of responsiveness may critically undermine the results of CUA analysis in Parkinson's and thus decision-making as QALY gains may differ depending on the derivation of utility values.

4.7 Chapter summary

This chapter reported the objectives, search methods and eligibility criteria of a systematic review of the use of PbQoL measures in Parkinson's, as well as the methods for an assessment of their construct validity and responsiveness in the included studies. In addition, this chapter also summarized the published mapping algorithms identified from the search that could be used to map from the non-preference based measures in Parkinson's to EQ-5D-3L. Results from the search, assessment and the summary of mapping algorithms were reported. The evidence for construct validity of the PbQoL measures identified in this review was generally positive except for in people with milder Parkinson's, nevertheless, there were concerns regarding their responsiveness to the change in QoL over time. The substantial lack of mental and social wellbeing dimensions in the mapping algorithms revealed a concern in EQ-5D-3L's inability to reflect these impact in the Parkinson's population.

Psychometric validation is an iterative process especially when the existing evidence is inconsistent. This chapter demonstrates a need to further explore the construct validity and responsiveness of the PbQoL measures in this population. In particular, exploring the ability of other PbQoL measures to capture wider benefits that are underestimated by the NICE recommended EQ-5D-3L may represent a valuable research development in this area. Therefore, Chapter 6 and 7 will empirically explore the construct validity and responsiveness of the broadly defined preference-based measure, ICECAP-O, in comparison to the EQ-5D-3L using a large longitudinal dataset in Parkinson's. Prior to this, Chapter 5 will provide the information on the data used and discuss the methodological challenges that are to be explored for the case studies in Chapter 6 and 7.

Chapter 5 Case studies: justification, data source and methodological challenges


Introduction
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Context and theories
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Methods review
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Empirical studies
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Conclusion
Chap. 8
Summary, implications, areas for further research, contribution and conslusions

5.1 Introduction

Following the results of the systematic review reported in Chapter 4, which revealed a clear need to further explore the construct validity and responsiveness of the PbQoL measures in the Parkinson's population this chapter will provide a brief digest of rationale for the following two empirical chapters. In addition, as discussed earlier in Section 3.3.1 and 3.5.2 respectively, construct validity and responsiveness are context-specific. This is especially important in the assessment of generic measures as those measures are designed to be generic whereas in practice applied to specific populations with varied characteristics. Tests of measurement properties of generic measures usually requires assessment in different populations to demonstrate its usefulness in each of the specific contexts, thereby a full understanding of where the data come from is essential. Both Chapter 6 and 7 used data from one of the largest trials of medication in Parkinson's (the PD MED RCT), this chapter will provide an overview of the trial and the key outcomes data collected. Given the importance of PDQ-39 in the case studies, a detailed introduction of the PDQ-39 will be provided.

Chapter 3 discussed a number of challenges when applying the classic psychometric testing methods to answer the research question of the case studies. This chapter will provide a summary of these conceptual and practical challenges. They include: (a) How to validate a measure of QoL and wellbeing concept when there is no gold standard measure and no consensus on the concept? (b) Is it appropriate to validate a 'preference-based' measure using non-preference based measures as gold standard? (c) How to set up hypotheses and expectations when there is no prior information regarding capabilities measured by ICECAP-O in Parkinson's? (d) How the variability within the sample would affect the validation of construct validity and responsiveness? These challenges have important implications for the methods chosen and interpretation of results in Chapter 6 and 7.

5.2 Justification for the case studies

In Chapter 1, Section 1.2 described the breadth of Parkinson's motor and non-symptoms, their broad impact on people's health and wellbeing, and the corresponding wide scope of interventions to manage Parkinson's. Chapter 4 revealed evidence that this wide impact, however, was not found to be sufficiently captured by the EQ-5D-3L measure. In particular, concern was raised in EQ-5D-3L's ability to measure and value the mental and social wellbeing dimensions associated with Parkinson's. This raises concerns about whether the EQ-5D captures the full benefit of interventions, especially the interventions that are substantially associated with patients' broader wellbeing. Examples of such interventions may include speech and language therapy (mentioned in 1.2.4.3) that improves patients' communication, and glycopyrronium bromide (a drug) that manages drooling of saliva, both of which may have wide impact on their family relationship, social wellbeing and stigma.

The results from Chapter 4 point to a valuable research direction which is to explore the ability of other PbQoL measures to capture the wider benefits of interventions in Parkinson's. As mentioned in Section 2.7, the ICECAP-O was developed with a view to expand the evaluative space and measure 'capability' wellbeing in older people in response to the need for a broader PbQoL measure (489). It could potentially be used in economic evaluations in older people across health and social areas in which a broader set of outcomes is considered (311, 489). Indeed, NICE recommends the use of ICECAP-O where outcomes in terms of capabilities are considered relevant to the intended effects of social care interventions and programmes. In addition, despite having been developed for only less than 10 years, the ICECAP-O has been found to be the most widely applied older people specific instrument in both community and residential aged care among all the generic preference-based instrument in a recent (2015) systematic review in aged care (20). This review further recommended the use of the EQ-5D to obtain QALYs in combination with a broader QoL measure such as ICECAP-O to facilitate the measurement and valuation of broader QoL benefits as defined by older people (20).

Parkinson's mainly affects elderly people aged over 60 (490) and its impact on people's QoL and social wellbeing is extensive, as demonstrated in Chapter 1 (Section 1.2.3). The validity of the ICECAP-O instrument has not been tested in a Parkinson's population yet and therefore the next two chapters will endeavour to answer the second research question, namely: is the ICECAP-O appropriate to capture the wellbeing impact of interventions in Parkinson's, and is it sufficiently sensitive in this population?

5.3 Case studies: data source

Psychometric assessment relies on the assumption surrounding the relationship between the test measure and the distinguishing characteristics of the population and as such the assessment result is specific to these characteristics. Accordingly, a full understanding of where the data come from and the characteristics of the underlying population is therefore essential to inform the psychometric validation in terms of the design of methods, and interpretation and generalisation of the results (491, 492).

Data for the empirical psychometric testing works in Chapter 6 and 7 were collected from the participants in the PD MED RCT. The PD MED is a large-scale, simple, long-term and 'real-life' study that aims to compare different classes of drugs in terms of their effectiveness, safety and cost-effectiveness, for patients with both early and later stages of Parkinson's (Registration number: ISRCTN69812316). The primary objective was to compare QoL between the different classes of drugs. The PD MED trials included up to three QoL measures – the two most commonly used QoL and wellbeing measures in Parkinson's, EQ-5D-3L and PDQ-39, and a new measure, ICECAP-O. This lays a rich data foundation for this thesis. The development, valuation, and validation of ICECAP-O and EQ-5D-3L have been reviewed and critiqued in Section 2.7.2 and 2.4.4, respectively. This section describes the design, inclusion and exclusion criteria, and the key outcome measures collected from the trial, and introduce in depth the PDQ-39 measure.

5.3.1 Trial design

The PD MED (75, 493) study contains two RCTs, the Early trial and the Later trial. The Early trial is for patients diagnosed with early stages of idiopathic Parkinson's – those just initiated on treatment, while the Later trial is in patients with later stage of idiopathic Parkinson's whose symptoms can no longer be controlled well by the initial therapy. All the drugs in each arm are available in clinical practice and had been tested previously, nevertheless, there is uncertainty around their relative effectiveness and cost-effectiveness due to the small sample size, short-term follow-up and lack of proper QoL measures in previous studies (494, 495). The trials are co-ordinated by the University of Birmingham Clinical Trials Unit. The recruitment started from 2001 until 2009 and participants are being followed up for ten years until 2019.

Whilst maintaining its robustness as a RCT through the randomisation process, PD MED is designed in a pragmatic and more ethically acceptable way to reach the recruitment target number and maximize the relevance of trial finding to clinical practice. The eligibility is not based on rigid entry criteria but on a real-life approach, which is, the 'uncertainty principle'. It allows the clinicians to consider if there is a definite indication for, or a definite contraindication against, a class of drug. In the former case, the patient is not eligible for randomisation; in the latter case, the patients could still be randomised to any of the other two arms. A patient is eligible for the three-arm randomisation only when there is uncertainty regarding which class of drugs should be offered. In addition, to reflect the normal clinical practice, the clinicians could decide the specific drug within each class that they prefer, and vary the dose as they see fit within the bounds of the manufacturer instructions. If patients' symptoms are not adequately controlled by the assigned class of drugs, or adverse effects are observed, adding or switching to a new drug from another drug class is permissible. This pragmatic approach should make the trial participants representative sample of the overall Parkinson's population, and subsequently enhances the generalisability of the results produced from the data to the wider Parkinson's population.

5.3.2 Inclusion/exclusion criteria and recruitment

Patients who met the eligibility criteria were recruited from over 80 neurology & care of the elderly units throughout the UK. The eligibility criteria for the Early trial were: 1) recently diagnosed with idiopathic Parkinson's by movement disorder specialists using UK Brain Bank diagnostic criteria and; 2) previously untreated for Parkinson's and therapeutic intervention was considered appropriate, or the patient had previously been treated with dopaminergic medication for less than 6 months, and there was uncertainty as to which class of drug to use. For the Later trial, the patients were eligible if they developed motor complications that were uncontrolled by levodopa (LD) (alone or in combination with either dopamine agonists (DA) or monoamine oxidase type B inhibitors (MAOBI)), and hence required the addition of another class of drug. For both Early and Later trial, patients were not eligible for the randomisation if they had dementia or unable to give informed consent. If the patient developed dementia during the trial, they can stay in the trial. Patients who had been randomised into the Early trial were re-randomised into the later disease randomisation if motor complications developed that were uncontrolled by the classes of drugs offered in the Early trial.

5.3.3 Outcome measures

The primary outcomes of the PD MED trials were the patient self-reported functional status on the mobility subscale of the PDQ-39 questionnaire and the CUA outcomes (i.e. the EQ-5D-3L, and the QALYs). The secondary outcomes included the other subscales of PDQ-39 questionnaire and the overall score, cognitive function assessed by Mini-Mental State Examination (MMSE), wellbeing of the carers assessed by SF-36 and Carer Experience Scale (CES), resource usage, toxicity and side-effects including mortality rates. In addition, time to onset of motor complications was assessed in the Early trial, and the time to surgical intervention or start of apomorphine was assessed in the Later trial.

To reach the recruitment target and keep the patients in the trial, the extra workload of assessment for the patients was kept to a minimum. The QoL, side-effects and resource usage questionnaires were completed by the patients via postal questionnaires. Meanwhile, MMSE and annual follow-up forms were completed by clinicians and the carer wellbeing forms were completed by carers. All assessments were completed annually after the first year until the end of the ten-year follow-up apart from MMSE, which is measured at baseline and at every subsequent five years. Table 5-1 illustrates the assessment of each of the questionnaires. The ICECAP-O capability measure (140, 141), which was developed in 2006 (140) and valued in 2008 (141), was added to the trial follow-up in 2010 and has since been being collected annually until the end of trial, December 2019.

Table 5-1: Baseline and follow-up assessments of the PD MED trial outcomes

Domains	Outcome measure	Completed by	At Entry	6 months	1st - 10th years, annually	5th, and 10th
Functional status / Quality of Life	PDQ-39, EQ-5D, ICECAP-O*	Patient	✓	✓	✓	
Side effects	Side effect form	Patient		✓	✓	
Health Economics	Resource usage	Patient			✓	
Carer wellbeing	SF-36, CES	Carer	✓	✓	✓	
Cognitive function	MMSE	Clinician	✓			✓
Disease status	Follow-up form, including current H&Y, and complications	Clinician	Rand, notepad		✓	

Note: This table is adapted from PD MED trial protocol (493).
* ICECAP-O was added to the trial since November 2010.
Abbreviation: MMSE – Mini Mental State Examination; CES – Carer Experience Scale; H&Y – Hoehn & Yahr staging scale

5.3.4 The PDQ-39 instrument

The PDQ-39 (131) is the most commonly used condition-specific QoL measure in Parkinson's and is judged to be the most thoroughly tested questionnaire in Parkinson's (433, 496-499). It assesses the effect of Parkinson's on QoL, and is sensitive to changes regarded as important to patients, but not identified by clinical rating scales (62, 131). The PDQ-39 was developed by Crispin Jenkinson, Ray Fitzpatrick and Viv Peto and published in 1997 (131). Aspects of health status were identified through in-depth interviews with 20 people with Parkinson's attending a neurology outpatient clinic, which generated a large number of possible items (500). After scrutinizing, a 65-item questionnaire was developed and piloted to test acceptability and comprehension in 359 individuals. The number of items was further reduced to 39-items with eight dimensions through factor analyses.

The PDQ-39's test-retest reliability was assessed by using Cronbach's alpha statistics with data from two time-point postal surveys with 3-6 days apart. A correlation coefficient value above 0.5 is judged to be acceptable and higher than 0.7 is good. The Cronbach's alpha was found to be good (above 0.7) for all dimensions (500), except for the social support dimension which is 0.66 at time 1 (62, 500). Its construct validity was tested by correlating scale scores with relevant SF-36 scores (62) and UPDRS score in a Spanish study (501).

As seen previously in Chapter 4 Table 4-2, PDQ-39 has 39 questions in total addressing eight domains of functioning and wellbeing in Parkinson's: mobility, ADL, emotional wellbeing, stigma, social support, cognition, communication and bodily discomfort. There are five levels for each attribute: never, occasionally, sometimes, often, always, with scoring being 0, 1, 2, 3 and 4 respectively. The score is calculated by averaging the levels for all the questions within each attribute and then standardizing the 'averaged level' to a scale of 0-100. The summary index (SI) of the PDQ-39, PDQ-39-SI, is the average of the eight attribute scores. The PDQ-39 has a short form version, the PDQ-8, comprising eight of the original 39 items of PDQ-39, with one item selected from each of the eight attributes, and thus the response level for each of the questions in PDQ-8 represents the score for each attribute after standardization (502).

Despite accurately measuring the key condition attributes in Parkinson's, its unweighted scoring system brings limitations for use in CUA. With the summary score being formed without weighting across dimensions and items within each dimension, it is unclear what the combined scores represent and thus hampers their interpretation (503). This instrument cannot be used directly in CUA due to the lack of valuation of attributes. Without valuation of the health states against length of life, or using monetary vehicles, no information is obtained on how its score could be interacted with length of life and how much society would be willing to pay for improvements in scores.

5.4 Methodological challenges of the assessment of the construct validity and responsiveness of ICECAP-O in Parkinson's

5.4.1 No criterion

No 'criterion' exists for a QoL measure. Due to the mixed views on the definition of health and QoL as discussed in Section 2.3, measures are established on different theoretical basis leading to different sets of descriptive system. This creates challenges in identifying the most appropriate anchor for testing responsiveness and the criteria for known-group validity. Using clinical measures it is possible to test whether the test measure is responsive to the change in a particular clinical aspect, however it requires an implicit assumption that the change in the clinical aspects will lead to the change in the test measure. Using another QoL instrument could also be problematic as any negative result can be attributed to the fact that the other QoL measure does not have the same construct as the test measure. As a result, choice of anchor or criteria would affect the interpretation of the validation results.

Chapter 6 and 7 will test the appropriateness of the ICECAP-O capability measure. Then what should be the proper anchor? The optimum anchor for a capability measure is another capability measure with the same construct - a ‘duplicate’ that is almost impossible to have in a trial setting for ethical considerations (i.e. not adding patients’ burden). Using instruments that are not measuring capability as anchors cannot provide evidence regarding how the ICECAP-O capability instrument measures true capability. The HrQoL measures are expected to be correlated with the health measures but this may not be the same for a capability measure with a much broader construct. Some aspects of capability, such as the attachment attribute in ICECAP-O about family and friendship may be affected by health like mobility issues, but not as strongly as other attributes such as control or fulfilment of role.

Although the limitation of no gold standard is not likely to be completely solved, consideration on this can be reflected in the hypotheses drawn through properly specifying the expected strength and correlation between the anchor measure and the test measure. The expectations are on the basis of theoretical understanding and other evidence prior to the testing and thus interpretation of the results should be treated with caution.

5.4.2 How to validate ‘preferences’?

The next challenge is related to the validation of the value set of a preference-based measure since none of psychometric validation approaches are developed considering for the assessment of the PbQoL measures. The essential part of preference-based measures is the value set which incorporates people's preference. As introduced in Section 2.4, preference represents individual's or group's relative desirability of the different outcomes, and hence the more desirable (more preferred) health states receive greater weight or utility score (97). Only by incorporating the preferences against 0 (death) and 1 (full health), the score of a measure can be viewed as weight for length of life in the QALY calculation. This is because the preferences are elicited using 'trade-off' exercises involving risk of death as a gamble or length of life as a trade. Details of this have been provided earlier in Chapter 2 Section 2.4.

Owing to the special purpose of preference-based measures with their score to be combined with length of life, interpretation of the instrument should be based on the overall index value as a whole, which is about the values of a state rather than the state itself (148, 504, 505). Accordingly, lack of construct validity or responsiveness may be due to the inappropriateness in the descriptive system, or could also because of the inappropriate values attached to the states. When it is unknown to what degree the change on the anchored measure is preferred by the patients, the testing methods have to depend on the arbitrary assumptions of people's preferences towards the change on each health state.

Potential ways to validate stated preferences are perhaps by examining revealed preferences or direct elicitation using TTO or SG (118). However these cannot be achieved because it is impractical to measure revealed preference of health state given there is no direct choice of health states. Also, it will be problematic if validating the preference from a generic valued measure by comparing with the preferences from direct elicitation, since studies have shown that the values elicited from direct methods are not interchangeable with that from the indirect methods (118, 506). Arnold et al. (2009) systematically reviewed the studies that compared utilities obtained directly (TTO or SG) (Section 2.4.2 for details) or indirectly (EQ-5D, SF-6D and HUI) (Section 2.4.3 for details) from the same patients (506). They found that direct methods of obtaining utilities yielded systematically higher scores than the indirect methods; mean utility values were 0.81 (SG) and 0.77 (TTO), in comparison to the indirect instruments, 0.59 (EQ-5D), 0.63 (SF-6D), 0.75 (HUI-2) and 0.68 (HUI-3). This means on average the respondents would accept a 19% reduction in lifespan to avoid the condition based on results of SG, and the amount of reduction increased to 41% in lifespan when considering the result from EQ-5D. This massive difference demonstrates that it is not feasible to validate preferences obtained from the off-the-shelf instruments using direct methods. What is also worth mentioning here is that this difference has significant implications for resource allocation decisions as when using a mixture of methods for different decisions is allowed, a motivated choice of method may distort the outcome in a preferred direction (506). This also highlights the importance of carefully scrutinizing the source of how the preferences were elicited before applying it in economic evaluations.

5.4.3 No prior information

Another challenge encountered in the design of the case studies in next two chapters is that there is often a lack of prior information for setting up hypotheses. Capability is an under-defined theory with many unsolved questions (305) when applying it to health economics research and there is no previous information on the possible magnitude of ICECAP-O change in the Parkinson's population. However, hypotheses testing requires forming reasonable hypotheses based on theoretical understanding of the test measure and the other measures and previous evidence. To my knowledge, this is the first study to test the ICECAP-O in people with Parkinson's, and thus no evidence is available to serve as a basis for us to speculate an 'expected effect size' or 'expected correlation coefficients' with which to benchmark the degree of responsiveness and construct validity of this measure in this population. For example, it is difficult to stipulate how much change is expected to happen on ICECAP-O for each degree of clinical improvement on the H&Y clinical scale. A compromise solution is to use another commonly used measure, for instance, EQ-5D-3L, if data are available, as a reference point to set up the hypotheses based on previous studies on their relationships. This partly explains why our case studies employed EQ-5D-3L as a reference measure in addition to aiding the interpretation of the results of the validity of ICECAP-O and add the relevance of the results to current practice.

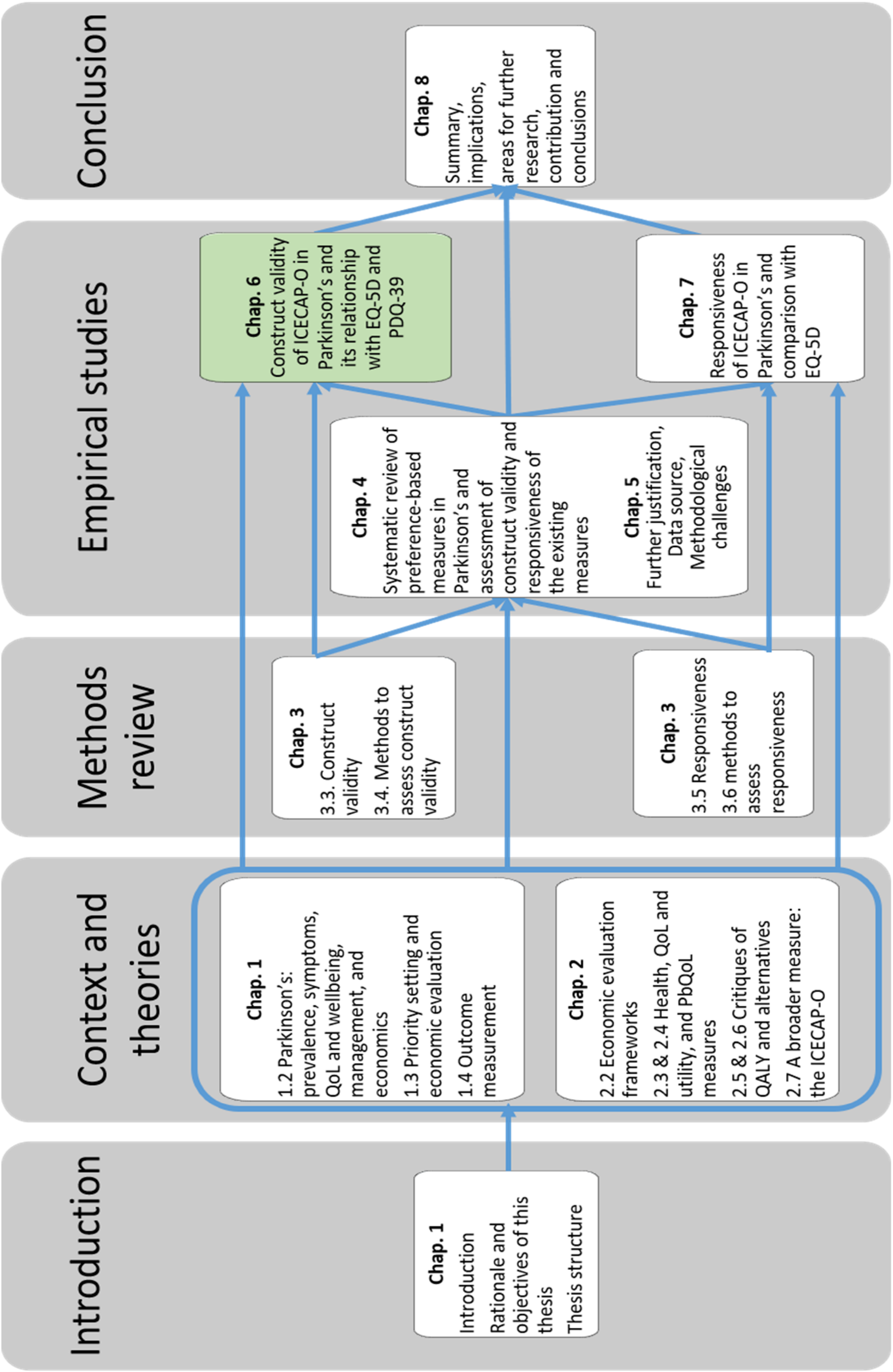
5.4.4 Heterogeneity within the sample

The last challenge comes from the complex nature of Parkinson's which leads to extensive heterogeneity across individuals. Given the large number of influences on QoL that may vary across and within patients, large variability of QoL trajectories over time with disease progression were found in previous studies (507, 508). As Parkinson's UK stated: "each person will have a completely different experience of the condition and their Parkinson's will progress at different rates." (18) The large heterogeneity within the sample may increase the 'noise' in the validation test. For example, in the responsiveness analysis, even the patients are assigned to the 'improved' group based on one certain anchor, it does not mean that all the aspects of Parkinson's are improved. In other words, getting worse in one aspect does not necessarily suggest overall deterioration of the health and QoL. Likewise, getting better in one aspect does not necessarily translate into the overall improvement of the health or QoL. This also affects the assessment of the known-group validity if the grouping criteria predict the test measure in varied ways across the population. In addition, Parkinson's is a chronic progressive disorder and most patient symptoms are in the trend of deterioration. The improvement in one dimension may not suggest the overall health status has not progressed over time and therefore the effect size and the standard response mean for the 'improved' group may be quite small if the anchor is a weak determinant of the test instrument. Acknowledging these issues would help avoid miss-interpretation of the validation results.

5.5 Chapter summary

This chapter started with providing a further justification for the two case studies of exploring the broadly defined measure ICECAP-O in people with Parkinson's by reflecting on the results from last chapter. This is followed by a description of the data source for the case studies, detailed introduction of PDQ-39 instrument and the conceptual and practical challenges related to the assessment of these properties in the case studies presented in Chapter 6 and 7. The next chapter will report the first case study, a cross-sectional assessment of the construct validity of ICECAP-O, in comparison with the EQ-5D-3L and the PDQ-39 in the Parkinson's population.

Chapter 6 Testing the construct validity of the ICECAP-O instrument and exploring its relationship with the EQ-5D-3L and the PDQ-39


Introduction
Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure
Context and theories
Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement
Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O
Methods review
Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity
Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness
Empirical studies
Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures
Chap. 5
Further justification, Data source, Methodological challenges
Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39
Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D
Conclusion
Chap. 8
Summary, implications, areas for further research, contribution and conclusions

6.1 Introduction

Following the justification provided in last chapter for exploring the performance of ICECAP-O instrument in Parkinson's, this chapter and the next chapter will empirically explore the construct validity and responsiveness of the ICECAP-O and how it compares with the EQ-5D-3L in people with Parkinson's, using data collected from the PD MED trials. As introduced in last chapter Section 5.3, the PD MED is the first study to collect the ICECAP-O data in the Parkinson's population hence this thesis will, for the first time provide evidence on the suitability of capability wellbeing as measured by the ICECAP-O in this population, and how its validity compares with existing measures in this context.

As described in Chapter 3 Section 3.3, construct validity represents the ability of an instrument to measure the underlying concept it intends to measure (332, 353). This chapter will start by introducing the objectives of the construct validation. Definition of construct validity and methods of construct validation have been reviewed and discussed in greater depth in Chapter 3 (Section 3.3, 3.4). Specific hypotheses are described regarding the group comparison and convergent validity, and rationale for each of the hypotheses as well as the specific statistical methods are provided. Results of the assessment of each hypothesis are then provided, followed by a summary of results and discussion.

6.2 Aims and objectives

This chapter aims to explore the construct validity of the ICECAP-O instrument in a large-scale RCT of different classes of drugs in Parkinson's (75). Specifically, there are three objectives for this chapter:

1) To explore the impact of Parkinson's on capability-wellbeing;

2) To assess the construct validity of ICECAP-O in people with Parkinson's in terms of its 'known-group' validity and convergent validity with measures with similar construct, i.e., the EQ-5D-3L, and the Parkinson's specific QoL measure, the PDQ-39; and

3) To contrast the construct validity of ICECAP-O versus EQ-5D-3L in this population.

In Chapter 1, Section 1.2.3 detailed the comprehensive impact of Parkinson's on health, QoL and social wellbeing. Therefore, the first objective of this chapter is to use ICECAP-O to provide a broad picture of the capability-wellbeing impact in the population affected by Parkinson's.

As discussed in depth in Chapter 3 (Section 3.3), construct validity is a key property for a newly-developed measure to have established before it can be put into routine use in clinical or research settings. The second objective is to test the construct validity of the ICECAP-O in people with Parkinson's. Establishing 'known-group' validity is a prerequisite for the ICECAP-O to be used in health and wellbeing assessment studies to compare the impact from different interventions. In health economics, the ability of a measure to differentiate between varied health states is the premise for its application In economic evaluations. For example, the index score for the varied health states are directly used as parameters in economic models to generate QALYs. Testing convergent validity would aid the understanding of the underlying construct of ICECAP-O regarding its relationship with the validated measures, especially the existing commonly used measures.

The third objective of this chapter is to test the construct validity of the current recommended measure by NICE, the EQ-5D-3L, as a reference point to help the interpretation of the ICECAP-O results. As discovered in Chapter 4, the EQ-5D-3L was generally able to distinguish between the groups defined by the clinical symptoms of Parkinson's, such as apathy, falling, freezing, visual hallucinations and depression, although evidence was not consistent when distinguishing between groups with or without dyskinesia, and the presence of 'wearing out'. By contrasting with the EQ-5D-3L, this research would be able to relate the results of construct validation of ICECAP-O to current decision-making and generate valuable insights which would help inform the choice of outcome measures in future research.

6.3 Data for this chapter

Data for this chapter were collected from the PD MED RCTs. Details regarding the PD MED RCTs have been provided in last chapter Section 5.3. Cross-sectional data extracted from the overall RCT panel data were used for the analyses in this chapter. For the socio-demographic (i.e., age, sex, presence of a regular carer) and clinical characteristics (i.e., H&Y stages, duration with Parkinson's) variables, they were extracted from the baseline assessment after each participant was recruited into the trial. For the QoL and wellbeing measures, including PDQ-39, ICECAP-O and EQ-5D-3L, responses were extracted only for the year when the participants completed their first ICECAP-O assessment. This analysis did not take the wave data (i.e., equal length of time staying in the RCT for each participant) from the RCT since the recruitment takes 10 years between 2001 and 2009 and thus by the year (2009/10) that the ICECAP-O added into the trial the participants were at varied wave of the trial. This ensures maximisation of sample size for this analysis, as well as generating a broad distribution of participants' characteristics in the dataset, which improves the generalisability of the result.

6.4 Methods

As described in Chapter 3 Section 3.4.1, a five-step hypothesis testing model was used for construct validation. These steps are: theory specification, hypothesis derivation, research design, empirical observation, and revision of the theory and constructs. This study will follow this five-step model by applying data from the PD MED trials to conduct a construct validation of the ICECAP-O in people with Parkinson's. This methods section will start by describing the hypotheses and justifications, followed by describing the statistical methods used to test these hypotheses. Missing data handling strategies are described at the end.

6.4.1 Hypotheses

In Chapter 3, Section 3.4.2 introduced two commonly used approaches to empirically test the construct validity of PbQoL measures: 'known group' method and convergent validity (440). Specific hypotheses were proposed a priori regarding the group differences in ICECAP-O responses (Early versus. Later) and correlations between the three measures, as recommended by Brazier's and COSMIN guideline (326, 327, 440). The detailed description, rationale and statistical methods for each of the hypotheses are presented in Table 6-1. They are summarised as below.

Hypothesis 1: Early versus Later group comparison. This hypothesis contains two sub-hypotheses:

(a) It was expected that the mean ICECAP-O index score in the Early group would be greater than that in the Later group (i.e. greater QoL/capability in the Early group).

(b) It was expected that three ICECAP-O attributes, security, role and independence would be scored more highly (i.e. more capable) in the Early group than in the Later group.

As mentioned previously in Section 3.4.2.2, the choice of criteria is important to the interpretation of the results in that the stronger correlation between the criteria and the test measure, the more favourable the 'known-group' results will be to the test measure. The criteria should be relevant to the use of the measure in practice. The criteria of Early versus Later group was chosen here because there was a clear difference regarding clinical management as distinguished by the PD MED trials. The patients were recruited to the Later trial or progressed from the Early to the Later groups when they developed motor complications or when their symptoms were not controlled by the drugs used in the Early trial. Therefore, as the first study with no previous information regarding the ICECAP-O capability-wellbeing in this population, grouping by the most pragmatic criteria for making decisions about clinical management should provide valuable implications on the validity of ICECAP-O and its relevance to clinical practice in Parkinson's.

As a reference point to aid the interpretation of the result of the ICECAP-O, the responses for the EQ-5D-3L and PDQ-39 were also analysed and compared between the patient group in the Early and Later trial.

Hypothesis 2: convergent validity between instruments. It was expected that there would exist moderate correlation between the ICECAP-O and EQ-5D-3L and between the ICECAP-O and PDQ-39. The correlation coefficient for the ICECAP-O and PDQ-39 was expected to be larger than that of the ICECAP-O and EQ-5D-3L.

This latter hypothesis was developed based on the differing 'constructs' between the two 'criterion' measures; the construct of the PDQ-39 includes broader aspects (e.g. social support, communication) than the EQ-5D-3L whose construct focuses on health-related QoL especially physical aspects. An important terminology note here is that 'criterion' does not refer to the gold standard construct that ICECAP-O must conform to, or must only show high correlation to demonstrate its construct validity. Rather, the term 'criterion' here refers to a widely validated measure whose construct comprises a shared related theory with ICECAP-O in its design to some degree such that a correlation between them is expected (418). In this case, the shared related theory is that health is one of the key determinants of QoL and wellbeing. In addition, more specific hypotheses regarding the correlations between individual attributes of ICECAP-O and PDQ-39/EQ-5D-3L were also proposed, which are presented in Table 6-1.

Table 6-1: Hypotheses, rationale and testing methods

1. Group comparison: groups classified by severity of Parkinson's (Early vs. Later)				
Dependent variable(s)	**Independent variable(s)**		**Hypotheses No. - Expected relationship(s)**	**Testing method**
ICECAP-O total score	Early - Later	1.	The Early group was expected to have lower capability than the Later group.	OLS univariable regression
ICECAP-O total score	Early - Later Age Sex	2.	The mean difference of ICECAP-O total score between the Early and Later group was expected to attenuate after adjusting for sex and age. Age is a potential confounding factor which was found previously to be negatively correlated with capability (312, 313), and those in the advanced group are older on average than those in the Early group. No relationships between capability and sex were found in both of above studies.	OLS multivariable regression
ICECAP-O security	Early - Later Age Sex	3.	The Later group was expected to respond lower level in the 'security' attribute (i.e. more concerns towards future) than the Early group. Parkinson's is a neurodegenerative disease with no cure and hence concerns towards future are common. Depression affects between 20-45% PwP(509) and has been found to be the most frequently identified determinant of HrQoL in PwP(487).	Proportional odds model

ICECAP-O Role	Early - Later Age Sex	4.	The Later group was expected to report lower level in the ‘role’ attribute (i.e. doing things making them feel valued) than the Early group. PwP’s productivity and performance towards their role can be affected by Parkinson’s symptoms. PwP have been found much higher chance to fall than general elderly population; e.g. a meta-analysis has shown 46% of PwP’s experience falls within a three month period(510). Furthermore, the hospital admission rate and length of stay has been found1.44 times more and 1.19 times longer than the age-matched controls(511).	Proportional odds model
ICECAP-O Control	Early - Later Age Sex	5.	The Later group was expected to report lower level in ‘control’ (i.e. be able to be independent) attribute than the Early group. Around 60% PwP have a regular carer in the PD MED trial(75), which suggests that their ability of self-care is limited. In addition, numerous studies have found Parkinson’s affect PwP’s ability of daily living.	Proportional odds model
2. Convergent validity: relationship between ICECAP-O, EQ-5D-3L and PDQ-39				
Variables involved			**Hypotheses -** **Expected relationship(s)**	**Testing method**
ICECAP-O total score	EQ-5D-3L total score, PDQ-39 total score	6. 7.	Medium correlation was expected to show between ICECAP-O and EQ-5D-3L and between ICECAP-O and PDQ-39 as they are measures with related construct but not same. It was also expected that ICECAP-O correlated more strongly with PDQ-39 than EQ-5D-3L because PDQ-39 contains wellbeing attributes as well as HrQoL attributes. Davis et al. analysed the data from the seniors attending fall clinic in Vancouver and found the EQ-5D-3L values in the middle range were consistently lower than the ICECAP-O values but the two measures were consistent in the higher and lower range values (512).	Pearson correlation if the relationship is linear, otherwise Spearman correlation Cocor R package is applied to test the significance of the comparison
ICECAP-O Attachment	PDQ-39 social support	8.	Strong correlation was expected between ICECAP-O attachment attribute and PDQ-39 social support attribute. ICECAP-O attachment attribute asks people how often they feel the love and friendship. This is very similar to the items in the ‘social support’ attribute in the PDQ-39, which asks: “Had problems with your close personal relationships?”, “Lacked support in the ways you need from your spouse or partner?” and “Lacked support in the ways you need from your family or close friends?”	Pearson correlation if the relationship is linear, otherwise Spearman correlation

ICECAP-O Security	PDQ-39 emotional wellbeing, EQ-5D-3L anxiety/depression	9. Strong correlation was expected between ICECAP-O security attribute and PDQ-39 emotional wellbeing attribute. ICECAP-O security attribute asks people to what degree they feel concern about future. Similarly, item 22 under the emotional wellbeing attribute in PDQ-39 asks: “felt worried about your future?” 10. Strong correlation was expected between the security in ICECAP-O and the anxiety in EQ-5D-3L. ‘Feel concerns about future’ and ‘feel anxious and/or depressed’ are associated. This association was found in Coast et al. in the general UK older population (312).	Pearson correlation if the relationship is linear, otherwise Spearman correlation
ICECAP-O role	PDQ-39 stigma EQ-5D-3L usual activities	11. Medium correlation was expected between ICECAP-O role attribute and PDQ-39 stigma attribute. PDQ-39 item36 (stigma) asks “felt ignored by people?” which should be related to the role attribute in ICECAP-O, “doing things make you feel valued”. 12. Medium correlation was expected between the ICECAP-O role attribute and EQ-5D-3L usual activities attribute. Makal et al. found there was medium association between ICECAP-O role and EQ-5D-3L usual activities (r=-0.47), the strongest correlation coefficient in their matrix of ICECAP-O attributes and EQ-5D-3L attributes (313). This was also tested in Davis et al.’s study using contingency table and Wilcoxon test, which, however, showed a few discrepancies between the two attributes (512).	Pearson correlation if the relationship is linear, otherwise Spearman correlation
ICECAP-O enjoyment	PDQ-39 Cognition and communication	13. Medium correlation was expected between ICECAP-O enjoyment attribute and PDQ-39 cognition and communication attributes. The items under these two attributes in PDQ-39 are very similar to the question of ‘enjoyment’ in ICECAP-O: PDQ-39 item-31 (cognition): “had problems with your concentration. E.g. when reading or watching TV?” and PDQ-39 Item-35: “felt unable to communicate with people properly?”	Pearson correlation if the relationship is linear, otherwise Spearman correlation
ICECAP-O control	EQ-5D-3L self-care	14. Medium correlation was expected between ICECAP-O control attribute and EQ-5D-3L self-care attribute. Davis et al. hypothesized there were agreement between EQ-5D-3L self-care and ICECAP-O control in the Vancouver post fall senior population however their Wilcoxon test showed there were significant differences between the two attributes (512).	Pearson correlation if the relationship is linear, otherwise Spearman correlation

Abbreviation: PwP – people with Parkinson’s

6.4.2 Statistical analysis

The statistical method used for each of the hypothesis is summarized in Table 6-1 (Column: 'testing method').

6.4.2.1 'Known-group' analysis

To test if the index scores of measures could differentiate between the Early and Later group (Hypothesis 1a), univariable regression was conducted to obtain the unadjusted mean difference between the groups, followed by multivariable regression which provided an adjusted mean difference after controlling for age and sex. Good evidence of construct validity was demonstrated by statistically significant difference between the groups. Given that statistical significance is dependent on sample size, weak evidence of construct validity was also considered if a statistically significant difference was nearly shown ($p<0.1$).

For hypothesis 1b, the responses for each of the five attributes of ICECAP-O between the Early and Later groups were compared in five ordinal logistic regression models, adjusting for age and sex. Proportional odds model (POM) is a widely used ordinal logistic regression model but the POM's validity relies on its underlying assumption, named the proportional odds assumption or parallel regression assumption (513). POM assumes that the coefficients which describe the relationship between the levels are the same (514). This is reflected in the output of POM that only one coefficient is reported for each variable. This assumption should be tested as it is unknown if the coefficient could be assumed to be the same between levels of ICECAP-O responses. The intervals of preference values between each level are varied, in other words, the increased preference for capability from level 1 (being unable) to 2 (being able in a little area) is different from the increase from level 2 to level 3 (being able in a lot of area). Therefore, the Brant test was conducted to test the proportional odds assumption (515). It allows testing the assumption for each independent variable separately, which is useful since only the variable that indicating the group assignment (Early vs. Later) was the interest of this study. If the assumption was violated for the group indicator variable, a partial proportional odds model would be fitted. This partial proportional odds model could provide a fixed coefficient for the independent variables that meet the assumption and a series of varied coefficients for the independent variables that violate the assumption.

In the proportional odds models, the reference group was the Early group for the models predicting the ICECAP-O attributes and Later group for the models predicting EQ-5D-3L attributes. This is because in EQ-5D-3L, level 1 is the best state (no problem) and level 3 (extreme problems) is the worst state, which is opposite to ICECAP-O in which level 1 is the worst state (least capable level) and level 4 is the best state (most capable level). This discrepancy led to the interpretation of the OR inversely if both of them use Early group as reference group. For ICECAP-O attributes, OR indicates the likelihood of responding to an increased level (more capability) in the Later group compared to Early group, which was expected to be less than 1 due to worsened health state of the Later group, whereas for EQ-5D-3L OR indicates the likelihood of responding to an increased level (worse health state) in the Later group compared to the Early group, which is expected to be larger than 1. Therefore, the reference group was switched to make the OR comparable between EQ-5D-3L and ICECAP-O.

6.4.2.2 Convergent validity

To test the convergent validity between instruments (Hypothesis 2), scatter plots were produced to enable visual examination of the relationship between the three sets of measures in the comparison, i.e. ICECAP-O & EQ-5D-3L, ICECAP-O & PDQ-39, EQ-5D-3L & PDQ-39. Pearson correlations were then conducted to test when a linear relationship was observed; otherwise, Spearman's rank correlation was used. Correlations above 0.6 are determined as strong, between 0.4 and 0.6 as moderate, and below 0.4 as weak. Where the magnitude of two correlations were compared, the web interface of the R statistical package *cocor* was implemented which offers ten statistical tests including Pearson and Filon's z (1898) test and Zou's confidence interval (2007) for testing the significance of the difference between correlations (516).

6.4.2.3 Missing data and imputation

Data were missing or incomplete if patients failed to return a questionnaire, provided an incomplete questionnaire or were lost to follow-up. Depending on the pattern and reason, the missingness mechanism could be missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The data are MCAR only if the missing values are like a random sample of all values so that missingness is not correlated with any variable, observed or unobserved (517, 518). Only under MCAR, which is very rare, a complete-case analysis is unbiased; otherwise, a complete-case analysis may be biased as it omits every observation (row of data) that has a missing value for any of the model variables (i.e. observed data from patients for variables other than the one that is missing, is discarded). In addition, complete-case analysis shrinks the sample size substantially, increasing the uncertainty of the estimates.

In contrast to MCAR, the data can be MAR or MNAR; the difference lies in whether the missingness depends only on information that are already observed. The data are MAR when the missingness is only correlated with observed variables and remains independent of the unobserved variables, so the observed variables are sufficient for predicting missingness (518). When there are unobserved factors that may strongly predict the missingness, the data are MNAR. Notably, it is unlikely to differentiate firmly between MAR or MNAR since proving MNAR requires us to observe, paradoxically, the 'unobserved' information. Nevertheless, it has been argued that the MAR assumption is more likely to be acceptable when adding more relevant variables in the imputation model (519).

Multiple imputation with chained equations (MICE) was conducted which allows the imputation of multiple variables simultaneously. Missing responses for each question of EQ-5D-3L and ICECAP-O and missing subscores of PDQ-39 attributes were imputed. Ologit model was specified for the ordinal EQ-5D-3L and ICECAP-O responses and OLS regression model was specified for the PDQ-39 subscores. Besides the outcomes, Age, sex, duration with Parkinson's and the baseline H&Y scale indicating severity of Parkinson's were also included in the imputation models. Ten imputation datasets were generated and the results were combined using Rubin's rule (520). The statistical analyses were conducted in STATA® 14 (StataCorp. 2015) (521). The imputation code in STATA is shown in Appendix B.

6.5 Results

The results section begins with a description of participant characteristics. Comparison between the Early and Later group are provided in terms of the summary scores of the ICECAP-O, EQ-5D-3L and the PDQ-39 measure, and each of the dimensions for these measures. This is followed by the scatter plot and correlation coefficients between the index scores of the three measures. The correlation coefficients between each dimension of the three measures are provided in the end.

6.5.1 Sample description

Responses from 1,010 participants in the Early group and 227 participants in the Later group were included in the analyses. Sample characteristics and percentage of missing values are presented in Table 6-2. Proportions of missing data were less than 5% for ICECAP-O index score and EQ-5D-3L index score, and around 20% for PDQ-39-SI. The difference in mean age of participants between the Early and later group was approximately two years (73.56 vs. 75.22), whereas the mean duration since diagnosis of Parkinson's differed by four years (5.48 vs. 9.60). Consistent with the longer duration with Parkinson's in the Later group was the increased H&Y stage in this group, which had a median of 2.5 compared to 1.5 in the Early group. The two groups were similar in the gender distribution with around 65% male participants, and approximately 65% participants in both groups had a regular carer.

Besides age, duration with Parkinson's, Parkinson's severity (baseline H&Y scale), expected differences were observed between the Early and Later groups for the ICECAP-O score, EQ-5D-3L score, and PDQ-39-SI (Table 6-2). Distributions for the three measures are presented in Figure 1. The distribution for the Early and Later group are overlaid to enable contrasting between the groups. It shows that the Later group distribution (hollow bins) for all the three measures are consistently shifting to the lower QoL from the Early group distribution (filled with colour), i.e. ICECAP-O to the lower value on the left representing lower capability, EQ-5D-3L to the lower value on the left representing lower health-related QoL, and PDQ-39 to the higher value on the right representing lower QoL and wellbeing. The mean

ICECAP-O score was 0.76 (SD 0.17) for the Early group and 0.69 (SD 0.18) for the Later group, EQ-5D-3L score 0.54 (SD 0.29) for the Early and 0.45 (SD 0.29) for the Later, and PDQ-39-SI 29.99 (SD 17.14) for the Early and 37.87 (SD 17.75) for the Later group.

Table 6-2: Sample characteristics (complete case)

	Early group				Later group			
	Mean	**SD**	**N**	**Missing (%)**	**Mean**	**SD**	**N**	**Missing (%)**
Age	73.56	8.37	1010	0	75.22	8.36	227	0
Duration with Parkinson's	5.48	2.67	1010	0	9.60	4.48	227	0
	Median	IQR	N		Median	IQR	N	
Baseline H&Y	1.5	1-2	1010	0	2.5	2-3	227	0
	Freq.	**%**			**Freq.**	**%**		
Sex				0.4				0
Male	653	64.65			147	64.76		
Female	353	34.95			80	35.24		
with Regular carer				0				0
Yes	671	66.44			149	65.64		
No	339	33.56			78	34.36		
	Mean	**SD**	**N**		**Mean**	**SD**	**N**	
ICECAP-O index score	0.76	0.17	994	1.6	0.69	0.18	225	0.9
EQ-5D-3L index score	0.54	0.29	998	1.2	0.45	0.29	218	4.0
PDQ-39-SI	29.99	17.14	864	14.5	37.87	17.75	178	21.6

Note: score range: H&Y (Hoehn and Yahr scale: a Parkinson's severity measure): 1 (mildest) – 5 (most severe). ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0(death) ~ 1(full health). PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

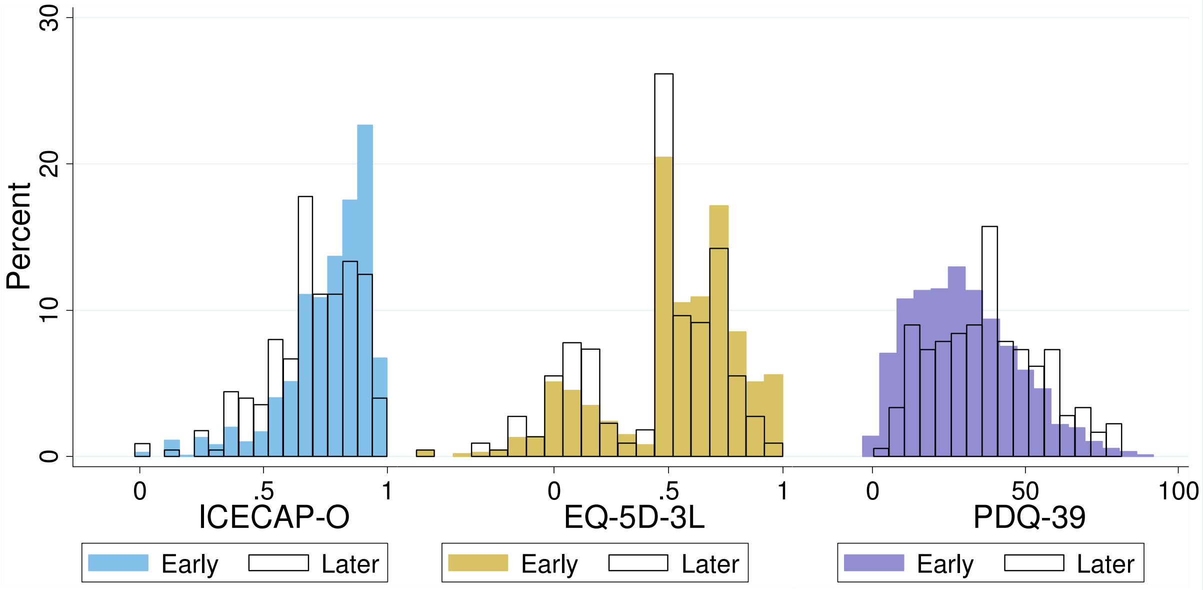


Figure 6-1: Distribution of ICECAP-O, EQ-5D-3L and PDQ-39 in the Early group and the Later group.

Note: score range: ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0 (death) ~ 1(full health). PDQ-39: 0 (least severe) ~ 100 (most severe).

6.5.2 Known-group comparison

6.5.2.1 Summary scores of three measures

Table 6-3 shows the unadjusted and adjusted mean difference of ICECAP-O score, as well as EQ-5D-3L score and PDQ-39-SI. The Later group had a statistically significant lower (p<0.001) health-related QoL and wellbeing as shown by all three measures, in both unadjusted and adjusted analysis. The unadjusted mean difference in the ICECAP-O value between the Early and Later participants (Early - Later) was 0.070 (95%CI 0.044, 0.096), which was slightly attenuated to 0.067 (95%CI 0.041, 0.093) after adjusting for age and sex. The unadjusted and adjusted mean difference in the EQ-5D-3L value had a larger magnitude than the ICECAP-O. This could be due to larger difference in HrQoL between the groups than in capability, or due to the larger possible score range of EQ-5D-3L which is approximately 1.5 times larger than ICECAP-O. Another reference measure, PDQ-39, also showed a statistically significant difference between the two groups with 8.30 (95%CI 5.72, 10.88) points difference after adjusting for age and sex.

Table 6-3: Early versus Later group: mean differences of ICECAP-O, EQ-5D-3L and the PDQ-39

	Unadjusted			Adjusted for age and sex		
Measures	**Mean estimate**	**95% CI**	**p value**	**Mean estimat e**	**95% CI**	**p value**
ICECAP-O	-0.070	-0.096, -0.044	0.000	-0.067	-0.093, - 0.041	0.000
EQ-5D-3L	-0.095	-0.137, -0.052	0.000	-0.091	-0.133, - 0.049	0.000
PDQ-39	8.39	5.81, 10.97	0.000	8.30	5.72, 10.88	0.000

Note: score range: H&Y (Hoehn and Yahr scale: a Parkinson's severity measure): 1 (mildest) – 5 (most severe). ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0(death) ~ 1(full health). PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

6.5.2.2 ICECAP-O dimension responses

The distribution of responses for the ICECAP-O's five attributes are shown in Table 6-4 and Figure 6-2. For attachment, 55.5% and 47.6% of the participants in the Early and Later groups, respectively, reported the highest level of capability (level 4) - "I can have all of the love and friendship that I want". For the other four attributes, security, role, enjoyment and control, the most frequently reported level in the Early group was level 3 (capable in many things), compared to the Later group which was level 2 (capable in a few things). The proportion of participants reporting the least capable level of security was the largest as compared to that of other attributes: 20% of Later participants and 14% of Early participants reported the least capable level (level 1) - 'I can only think about future with a lot of concern'.

Table 6-4: Responses[a] on ICECAP-O for the Early group and Later group

ICECAP-O[b]	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
	Attachment		**Security**		**Role**		**Enjoyment**		**Control**	
Early group										
Levels										
1	23	2.28	141	13.96	75	7.43	42	4.16	74	7.33
2	123	12.18	379	37.52	366	36.24	320	31.68	285	28.22
3	291	28.81	380	37.62	407	40.3	489	48.42	465	46.04
4	561	55.54	105	10.4	159	15.74	156	15.45	184	18.22
Missing	12	1.19	5	0.5	3	0.3	3	0.3	2	0.2
Total	1,010	100	1,010	100	1,010	100	1,010	100	1,010	100
Later group										
Levels										
1	4	1.76	46	20.26	26	11.45	15	6.61	32	14.1
2	41	18.06	107	47.14	113	49.78	107	47.14	89	39.21
3	73	32.16	58	25.55	67	29.52	88	38.77	91	40.09
4	108	47.58	15	6.61	21	9.25	17	7.49	15	6.61
Missing	1	0.44	1	0.44	0	0	0	0	0	0
Total	227	100	227	100	227	100	227	100	227	100

Note: a. The responses are the original observed data from the 1010 participants in this analysis at their first ICECAP-O assessments, including missing values, before multiple imputation.
b. For ICECAP-O, level 1 is the least capable level while level 4 is the most capable level.

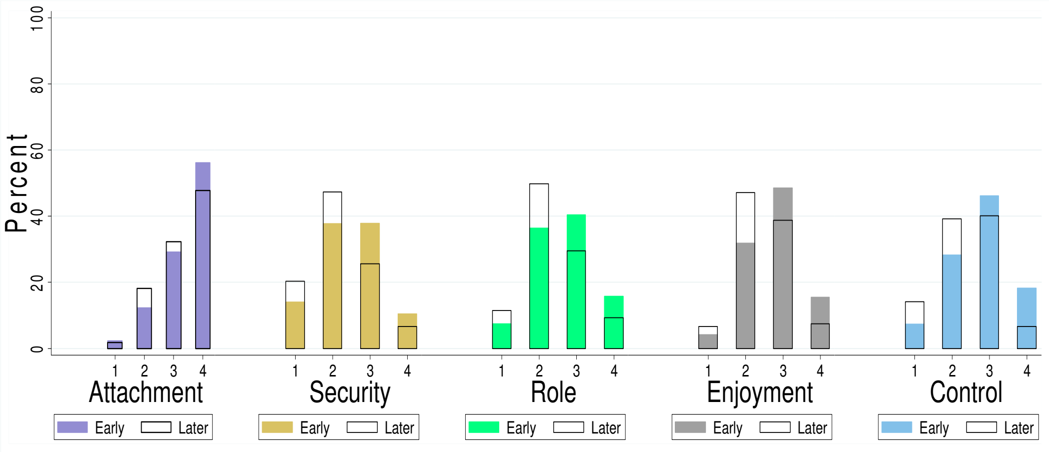


Figure 6-2: Response on ICECAP-O for the Early group and Later group.

Note: For ICECAP-O, level 1 is the least capable level while level 4 is the most capable level

The results of the proportional odds models are presented in Table 6-5. The variable which indicates the group (Later/Early) met the proportional odds assumption for all models, hence, only one odds ratio (OR) was reported for each attribute. The unadjusted ORs between the two groups (Later/Early) were 0.71 (95%CI 0.54, 0.93) for attachment, 0.56 (95%CI 0.43, 0.73) for security, 0.52 (95%CI 0.40, 0.69) for role, 0.49 (95%CI 0.38, 0.65) for enjoyment, and 0.46 (95%CI 0.35, 0.59) for control. The ORs became smaller after adjusting for age and sex for attachment and security attribute, and larger after adjustment for role, enjoyment, and control; but the difference was minimal in both direction. As was theorised, all the ORs are less than one, indicating that participants in the Later group have reduced odds of reporting an increased level of capability in comparison to the participants in the Early group. For example, for the 'role' attribute, the odds that a participant in the Later group answers an increased level (more capability), is approximately half (0.52) the corresponding odds for a participant in the Early group, which means that the later stage group was approximately twice as more likely to respond with a lower level of capability than the early group for the 'role' attribute.

Table 6-5: Odds ratio (OR) between the levels of ICECAP-O attributes (Later group / Early group)

ICECAP-O Attribute	Unadjusted		Adjusted for age and sex	
	OR (95%CI)	P value	OR (95%CI)	P value
Attachment	0.711 (0.541, 0.934)	0.014	0.706 (0.537, 0.928)	0.012
Security	0.558 (0.428, 0.728)	0.000	0.533 (0.408, 0.697)	0.000
Role	0.524 (0.400, 0.686)	0.000	0.548 (0.418, 0.718)	0.000
Enjoyment	0.493 (0.376, 0.647)	0.000	0.506 (0.385, 0.665)	0.000
Control	0.456 (0.349, 0.595)	0.000	0.484 (0.370, 0.633)	0.000

Note for 2b: For ICECAP-O, level 1 is the least capable level while level 4 is the most capable level. For EQ-5D-3L, Level 1 is "no problem", level 2 "some problems", and level 3 "extreme problems".

6.5.2.3 EQ-5D-3L dimension responses

Responses obtained for the EQ-5D-3L are presented in Table 6-6 and visualized in Figure 6-3. For the EQ-5D-3L, the most frequently reported level in both Early and Later group for all attributes was level 2, 'some problems'. There was an obvious proportion shifting from level 1 'no problem' to level 2 'some problems' for mobility and anxiety, and from level 1 ‘no problem’ to level 2 and 3 ‘some and a lot of problems’ for self-care, usual activities and pain/discomfort.

Table 6-6: Responses[a] on EQ-5D-3L for the Early group and Later group

EQ-5D-3L[b]	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
	Mobility		Self-care		Usual activities		Pain/ discomfort		Anxiety/ depression	
Early group										
Levels										
1	234	23.17	429	42.48	202	20.00	238	23.56	435	43.07
2	754	74.65	521	51.58	675	66.83	666	65.94	518	51.29
3	14	1.39	51	5.05	124	12.28	98	9.70	47	4.65
Missing	8	0.79	9	0.89	9	0.89	8	0.79	10	0.99
Total	1010	100	1010	100	1010	100	1010	100	1010	100
Later group										
Levels										
1	22	9.69	62	26.87	21	9.25	34	14.98	76	33.48
2	194	85.46	136	59.91	155	68.28	160	70.48	135	59.47
3	4	1.76	24	10.57	44	19.38	26	11.45	10	4.41
Missing	7	3.08	6	2.64	7	3.08	7	3.08	6	2.64
Total	227	100	227	100	227	100	227	100	227	100

Note: a. The responses are the original observed data from the 1010 participants in this analysis at their first ICECAP-O assessments, including missing values, before multiple imputation.
b. For EQ-5D-3L, Level 1 is “no problem”, level 2 “some problems”, and level 3 “extreme problems”.

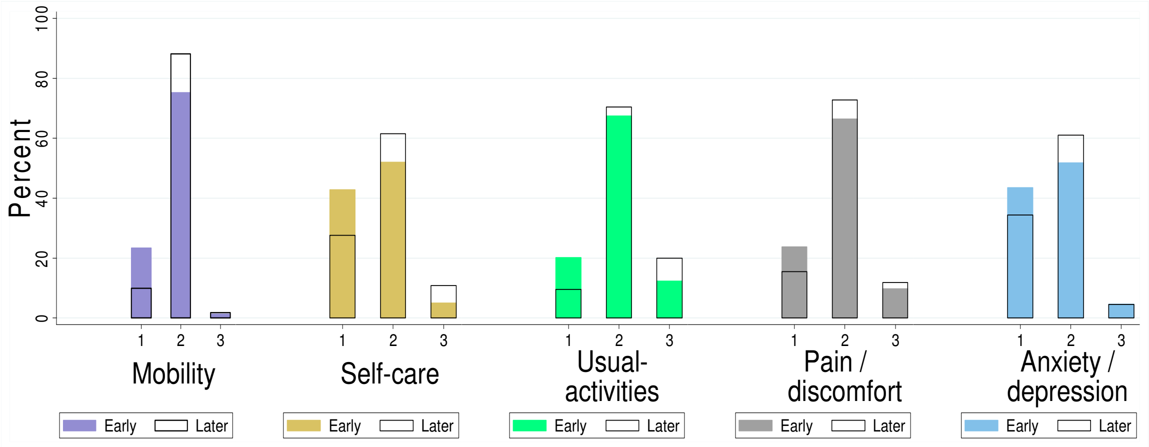


Figure 6-3: Response on EQ-5D-3L for the Early group and Later group.

Note: For EQ-5D-3L, Level 1 is “no problem”, level 2 “some problems”, and level 3 “extreme problems”.

The results of the proportional odds models for the EQ-5D-3L attributes are shown in Table 6-7. Similar to the ICECAP-O attributes, the group variable (Later/Early) met the proportional odds assumption for all five regressions. The ORs were all less than one as expected, indicating that participants in the Early group have reduced odds of reporting an increasing level (a worse health state) compared to the Later group. The group difference can be ranked from the largest to smallest based on the adjusted OR (Early/Later) as mobility (0.43, 95%CI 0.29 - 0.65, the furthest OR from 1), self-care (0.50, 95%CI 0.37 - 0.67), usual activities (0.52, 95%CI 0.38 - 0.71), pain and discomfort (0.67 (95%CI 0.49 - 0.90), and lastly, anxiety and depression (0.70, 95%CI 0.52 - 0.93).

Table 6-7: Odds ratio (OR) between the levels of EQ-5D-3L attributes (Early group / Later group)

EQ-5D-3L attribute[a]	Unadjusted		Adjusted for age and sex	
	OR (95%CI)	P value	OR (95%CI)	P value
Mobility	0.417 (0.277, 0.627)	0.000	0.434 (0.288, 0.653)	0.000
Self-care	0.480 (0.357, 0.645)	0.000	0.497 (0.369, 0.669)	0.000
Usual activities	0.492 (0.362, 0.670)	0.000	0.519 (0.380, 0.707)	0.000
Pain and discomfort	0.668 (0.490, 0.910)	0.000	0.663 (0.486, 0.904)	0.000
Anxiety and depression	0.710 (0.533, 0.945)	0.019	0.698 (0.523, 0.931)	0.014

Note: a. For EQ-5D-3L, level 1 is “no problem”, level 2 “some problems”, and level 3 “extreme problems”. This is opposite to ICECAP-O in which level 1 is the least capable level and level 4 is the most capable level. To make the OR comparable between EQ-5D-3L and ICECAP-O, the comparison was reversed to ‘Early/Later’ with Later group as reference group.

6.5.2.4 PDQ-39 dimensions

Mean scores of the PDQ-39 eight dimensions for the Early and Later groups are presented in Table 6-8 and visualised in Figure 6-4. Based on the magnitude of the score, the aspects that Parkinson’s affected most were mobility, ADL, bodily discomfort, and cognition. Table 6-9 shows the unadjusted and adjusted mean difference (Later – Early) of the PDQ-39 attributes. For all of the eight attributes, the scores in the Later group were statistically significantly higher than the Early group. The magnitude of the mean difference between the Early and the Later group was largest in mobility (15.45, 95%CI 11.14 – 19.76), ADL (12.33, 8.49 – 16.17), and communication (8.82, 95%CI 5.39 – 12.25), which indicates these attributes deteriorated most with the progression of Parkinson’s. These were followed by stigma (7.12, 95%CI 3.58 – 10.65), bodily discomfort (6.56, 95%CI 2.84

- 10.29), emotional wellbeing (6.46, 95%CI 3.19 - 9.72), cognition (5.53, 95%CI 2.33 - 8.74), and social support (4.14, 95%CI 1.54 - 6.74).

Table 6-8: Mean score of PDQ-39 eight dimensions for the Early group and Later group (complete case[a])

PDQ-39 attributes[b]	Early group				Later group			
	Mean	SD	N	Missing (%)	Mean	SD	N	Missing (%)
Mobility	45.84	30.98	969	4.1	63.54	29.70	212	6.6
ADL	39.57	25.77	958	5.1	51.46	25.35	206	9.3
Emotional	27.82	22.33	936	7.3	33.77	22.37	200	11.9
Stigma	21.56	23.13	965	4.5	27.51	26.13	207	8.8
Social support	11.16	17.15	944	6.5	14.20	18.44	203	10.6
Cognition	34.81	21.52	962	4.6	40.15	22.59	205	9.7
Communication	24.17	22.85	965	4.5	32.61	24.92	208	8.4
Bodily discomfort	39.75	25.20	959	5.0	45.73	26.78	207	8.8

Note: a. The data are the original observed data from the 1010 participants in this analysis at their first ICECAP-O assessments, including missing values, before multiple imputation.
b. score range: PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

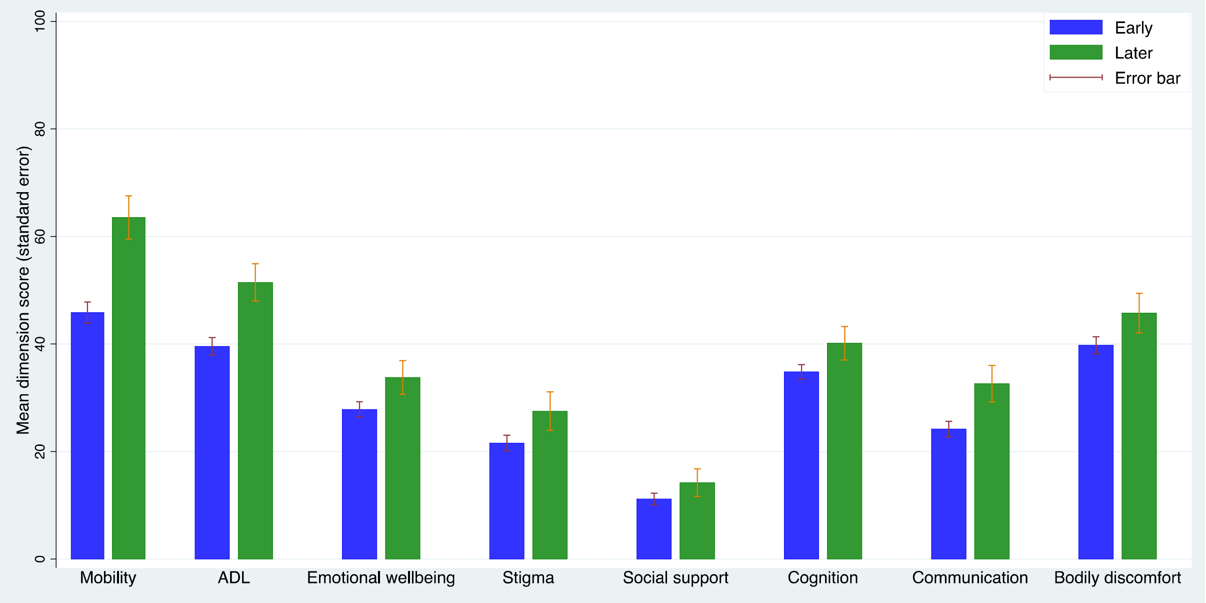


Figure 6-4: Responses on PDQ-39 eight attributes for the Early group and Later group.

Note: score range: PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

Table 6-9: Mean difference (Later - Early) of the PDQ-39 attributes

PDQ-39 attributes	Unadjusted			Adjusted for age and sex		
	Mean difference	95%CI	P value	Mean difference	95%CI	p value
Mobility	16.81	12.33, 21.29	0.000	15.45	11.14, 19.76	0.000
ADL	12.96	9.10, 16.81	0.000	12.33	8.49, 16.17	0.000
Emotional	6.18	2.90, 9.47	0.000	6.46	3.19, 9.72	0.000
Stigma	6.52	2.96, 10.07	0.000	7.12	3.58, 10.65	0.000
Social support	3.64	1.02, 6.26	0.006	4.14	1.54, 6.74	0.002
Cognition	6.21	2.98, 9.43	0.000	5.53	2.33, 8.74	0.001
Communication	8.80	5.35, 12.25	0.000	8.82	5.39, 12.25	0.000
Bodily discomfort	6.00	2.24, 9.77	0.002	6.56	2.84, 10.29	0.001

Note: score range: PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

6.5.3 Convergent validity

Figure 6-5 shows a linear relationship between the ICECAP-O value and the PDQ-39 summary score, the EQ-5D-3L utility values and the PDQ-39 summary score, and the ICECAP-O score and EQ-5D-3L utility values. The ICECAP-O index score was found to be highly correlated with the EQ-5D-3L index value (r = 0.65; p<0.001) and PDQ-39-SI (r = 0.73; p<0.001) (

Table 6-10). The null hypothesis that these two correlation coefficients are equal was rejected for all of the statistical tests (one-sided p value < 0.0001 for all) (Please see Appendix C for the full results). The confidence interval for the 'additional' correlation coefficients for the ICECAP-O - PDQ-39 compared to ICECAP-O - EQ-5D was tested to be 0.0481 and 0.1109. Therefore, it shows that the ICECAP-O index score was more strongly correlated with PDQ-39 score than with the EQ-5D-3L index score (hypothesis 7 in Table 6-1).

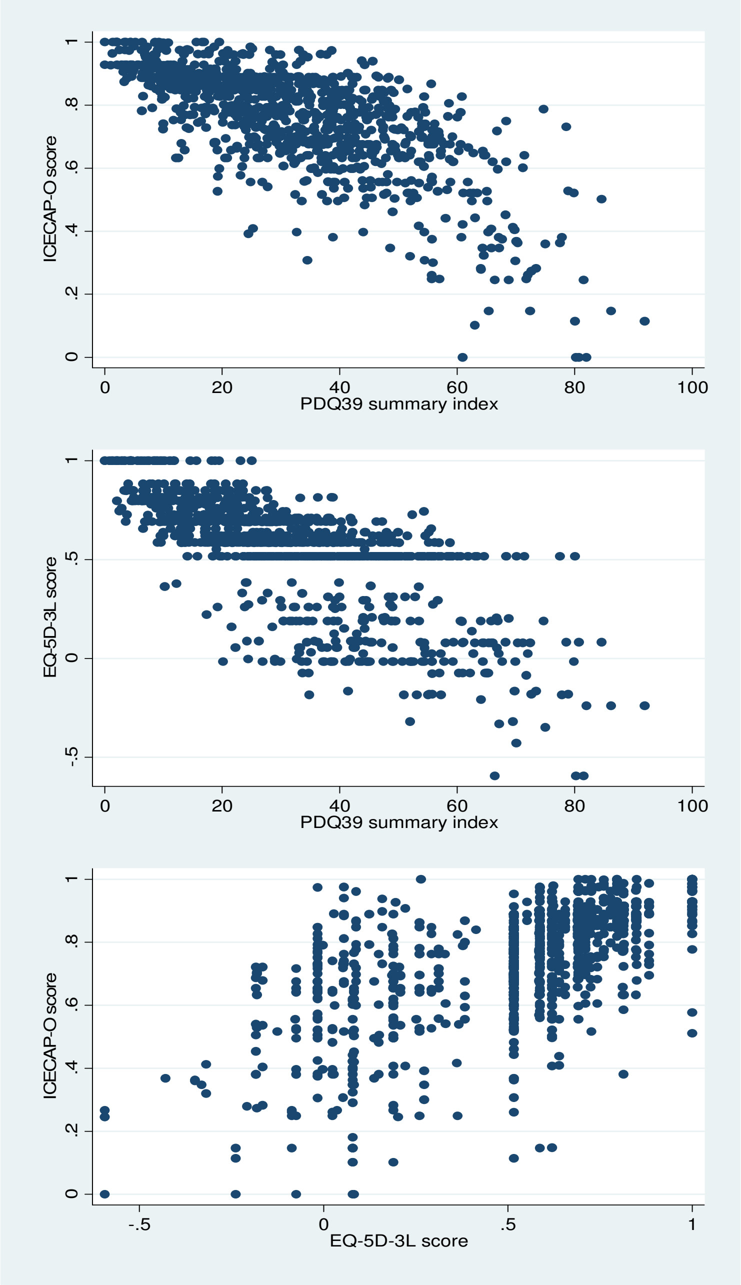


Figure 6-5: Scatter plots of ICECAP-O - PDQ-39, EQ-5D-3L - PDQ-39 and ICECAP-O - EQ-5D-3L.

Note: score range: ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0(death) ~ 1(full health). PDQ-39: 0 (least severe) ~ 100 (most severe).

Table 6-10: Pearson correlations between overall score of ICECAP-O, EQ-5D-3L and PDQ-39

	ICECAP-O index score	EQ-5D-3L index score	PDQ-39-SI
ICECAP-O index score	1		
EQ-5D-3L index score	0.654	1	
PDQ-39-SI	0.733	0.724	1

Note: score range: ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0(death) ~ 1(full health). PDQ-39: 0 (least severe) ~ 100 (most severe).

A correlation matrix between all the attributes of ICECAP-O, PDQ-39 and EQ-5D-3L is presented in Table 6-11. For each of the ICECAP-O attributes, the most correlated attributes from PDQ-39 and EQ-5D-3L respectively were: ICECAP-O ‘attachment’ - PDQ-39 ‘social support’ (r=0.50) and EQ-5D-3L ‘anxiety’ (r=0.26); ICECAP-O ‘security’ - PDQ-39 ‘emotional wellbeing’ (r=0.58) and EQ-5D-3L ‘anxiety’ (r=0.50); ICECAP-O ‘role’ – PDQ-39 ‘mobility’ (r=0.63) and EQ-5D-3L ‘usual activities’ (r=0.56); ICECAP-O ‘enjoyment’ – PDQ-39 ‘mobility’ (r=0.55) and EQ-5D-3L ‘usual activities’ (r=0.51); and ICECAP-O ‘control’ – PDQ-39 ‘mobility’ (r=0.72) and EQ-5D-3L ‘usual activities’ (r=0.62). In addition, the ICECAP-O attributes ‘role’ and ‘enjoyment’ correlated most strongly with each other (r=0.64), as well as ‘role’ and ‘control’ (r=0.63).

Table 6-11: Pearson correlation coefficients matrix - attributes of ICECAP-O, PDQ-39 and EQ-5D-3L

		ICECAP-O					EQ-5D-3L					PDQ-39							
Attributes		Attachment	Security	Role	Enjoyment	Control	Mobility	Self-care	Usual activities	Pain	Anxiety	Mobility	ADL	Emotional Wellbeing	Stigma	Social support	Cognition	Communication	Bodily discomfort
ICECAP-O	Attachment	1																	
	Security	0.302	1																
	Role	0.333	0.469	1															
	Enjoyment	0.458	0.457	0.638	1														
	Control	0.255	0.382	0.627	0.588	1													
EQ-5D-3L	Mobility	0.142	0.253	0.402	0.365	0.484	1												
	Self-care	0.191	0.294	0.511	0.464	0.613	0.432	1											
	Usual activities	0.222	0.322	0.562	0.513	0.624	0.510	0.574	1										
	Pain	0.146	0.239	0.282	0.304	0.333	0.341	0.243	0.313	1									
	Anxiety	0.261	0.495	0.412	0.413	0.389	0.264	0.299	0.330	0.309	1								
PDQ-39	Mobility	0.226	0.369	0.629	0.554	0.724	0.601	0.604	0.651	0.395	0.411	1							
	ADL	0.221	0.331	0.581	0.496	0.681	0.477	0.737	0.609	0.298	0.367	0.745	1						
	Emotional Wellbeing	0.330	0.580	0.522	0.503	0.524	0.340	0.440	0.422	0.348	0.710	0.577	0.552	1					
	Stigma	0.265	0.393	0.399	0.392	0.394	0.224	0.317	0.320	0.244	0.429	0.409	0.465	0.598	1				
	Social support	0.503	0.395	0.352	0.411	0.352	0.198	0.284	0.263	0.237	0.408	0.326	0.360	0.558	0.499	1			
	Cognition	0.194	0.285	0.459	0.429	0.497	0.333	0.444	0.390	0.285	0.390	0.493	0.578	0.534	0.369	0.348	1		
	Communication	0.278	0.337	0.462	0.422	0.486	0.353	0.428	0.408	0.202	0.366	0.436	0.569	0.526	0.465	0.492	0.549	1	
	Bodily discomfort	0.150	0.340	0.364	0.352	0.383	0.362	0.330	0.324	0.565	0.387	0.453	0.452	0.519	0.384	0.335	0.476	0.365	1

6.5.4 Hypotheses tests

The hypotheses (Section 6.4.1) testing results are summarized in Table 6-12.

Table 6-12 Results of hypotheses testing

Known-group hypotheses	Testing result
1. The Early group was expected to have lower capability than the Later group.	Yes
2. The mean difference of ICECAP-O total score between the Early and Later group was expected to attenuate after adjusting for sex and age.	Yes
3. The Later group was expected to respond lower level in the 'security' attribute (i.e. more concerns towards future) than the Early group.	Yes
4. The Later group was expected to report lower level in the 'role' attribute (i.e. doing things making them feel valued) than the Early group.	Yes
5. The Later group was expected to report lower level in 'control' (i.e. be able to be independent) attribute than the Early group.	Yes
Correlation hypotheses	
6. Medium correlation was expected to show between ICECAP-O and EQ-5D-3L and between ICECAP-O and PDQ-39	Yes
7. It was also expected that ICECAP-O correlated more strongly with PDQ-39 than EQ-5D-3L.	Yes
8. Strong correlation was expected between ICECAP-O attachment attribute and PDQ-39 social support attribute.	No, r=0.503
9. Strong correlation was expected between ICECAP-O security attribute and PDQ-39 emotional wellbeing attribute.	No, r=0.580
10. Strong correlation was expected between the security in ICECAP-O and the anxiety in EQ-5D-3L.	No, r=0.495
11. Medium correlation was expected between ICECAP-O role attribute and PDQ-39 stigma attribute	Yes
12. Medium correlation was expected between the ICECAP-O role attribute and EQ-5D-3L usual activities attribute.	Yes
13. Medium correlation was expected between ICECAP-O enjoyment attribute and PDQ-39 cognition and communication attributes.	Yes
14. Medium correlation was expected between ICECAP-O control attribute and EQ-5D-3L self-care attribute.	No, r=0.613

6.6 Summary of results

This study is the first to test the construct validity of the ICECAP-O in the Parkinson's population. Groups with Early and Later stages of Parkinson's in the PD MED trials were compared in terms of the ICECAP-O index score and the responses for each attribute. The EQ-5D-3L and PDQ-39 were also compared between groups as a reference. The Later group was found to have statistically significant lower capability wellbeing in comparison to the Early group. Specifically, for the ICECAP-O attributes ‘security’, ‘role’, ‘enjoyment’ and ‘control’, the Later group was approximately twice as likely to respond with a lower level of capability than the Early group. This demonstrates the substantial impact on those aspects of QoL and wellbeing with the progression of Parkinson’s from the Early stage when symptoms can be controlled by drugs to the Later stage with motor complications developed.

In addition, the ICECAP-O value was found to be highly correlated with the EQ-5D-3L score and PDQ-39-SI, both of which are commonly used and widely validated QoL measures in Parkinson’s. PDQ-39 attributes, ‘social support’, ‘emotional wellbeing’ and ‘mobility’, and EQ-5D-3L attributes, ‘anxiety’ and ‘usual activities’, were found to correlate most strongly with ICECAP-O attributes. In summary, the ICECAP-O’s construct validity was established in terms of discriminating between groups with Early and Later stages of Parkinson’s and association with EQ-5D-3L and the Parkinson’s specific QoL and wellbeing measure PDQ-39. The ICECAP-O shows potential as a preference-based approach to the measurement of capability wellbeing in the Parkinson’s population.

6.7 Results in the context of existing evidence

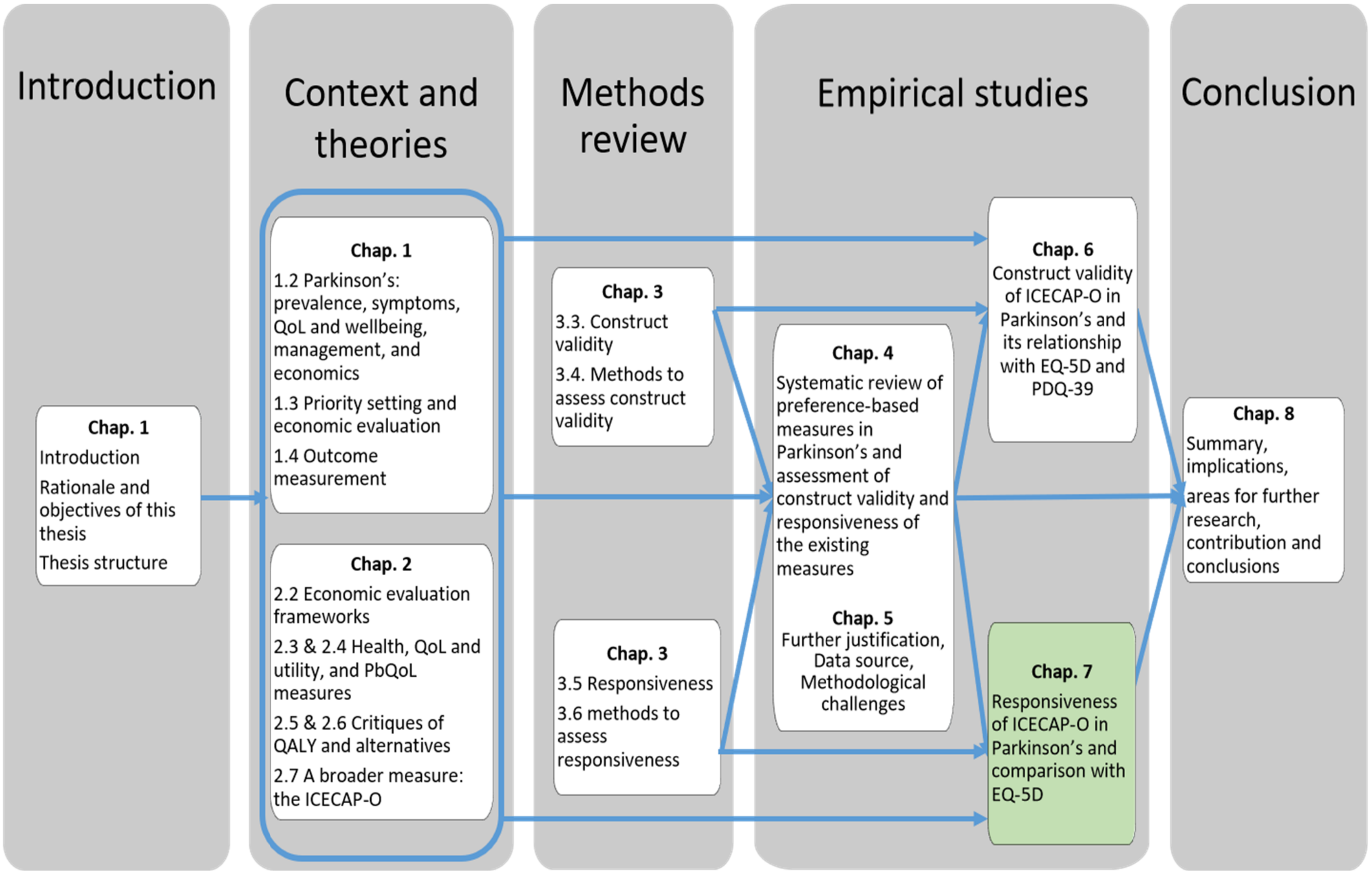
This study found that the mean ICECAP-O value in the Early Parkinson’s group was lower than that in the Later group, both of which (0.76 for the Early group and 0.69 for the Later group) were lower than the reported mean ICECAP-O score (0.83 (SD 0.12)) in the UK general population aged ≥ 65 years assessed by Flynn et al (522). It was also lower than the population in previous validation studies, including in frail older adults (mean 0.78, SD 0.16) (319), post-hospitalized older people overall (mean 0.84, SD 0.14) (313), osteoarthritis patients requiring joint replacement (mean 0.772, SD 0.17) (523), and in older adults with mobility impairment (mean 0.815, SD 0.177) (316). In the study among the post-hospitalized older people, the groups that had an ICECAP-O score lower than 0.8 were the subgroup that were divorced (mean 0.76), or in a nursing home (mean 0.78), limited social activity (mean 0.77) and depressed (mean 0.73) (SD not provided for the subgroups). The above comparisons indicate that the capability impact of Parkinson’s in the affected population is substantial compared to people with other conditions.

It was found that PDQ-39 ‘mobility’ and EQ-5D-3L ‘usual activities’ were the most highly correlated attributes to the ICECAP-O’s attributes ‘role’, ‘enjoyment’ and ‘control’. This finding agrees with a study by Coast et al. which assessed the ICECAP-O in individuals aged 65 and over in the general population of the UK and reported strong relationships between physical measures of health and the above three attributes of ICECAP-O (312). Moreover, PDQ-39’s attributes ‘social support’ and ‘emotional wellbeing’ and EQ-5D-3L’s ‘anxiety’ were the highest correlated attributes with ICECAP-O’s ‘attachment’ and ‘enjoyment’ attributes, which also agrees with findings by Coast et al that mental health measures strongly relates to those two attributes (312). Both being generic preference-based measures, notwithstanding the moderate to strong correlation between the EQ-5D-3L and ICECAP-O, Davis et al. suggested that the EQ-5D-3L and ICECAP-O are complementary measures rather than substitutes as there are several differences between them (512). They conducted exploratory factor analysis on the responses of the two measures among the population that attended a fall clinic and found that the EQ-5D-3L attributes appear to represent a single factor, ‘physical functioning’ and in contrast, ICECAP-O attributes represent an “overall reflection of participants perceived capacity for QoL and wellbeing – ‘psychosocial wellbeing’”.

6.8 Chapter summary

This chapter explored the capability wellbeing in people with Parkinson’s using the PD MED data, and tested the ‘known-group’ validity and convergent validity of the ICECAP-O capability measure in this population. In conclusion, this demonstrated the construct validity of the ICECAP-O in Parkinson’s. Besides construct validity, another measurement property that is important to be tested for ICECAP-O is responsiveness, which will be reported in the next chapter.

Chapter 7 Testing the responsiveness of the ICECAP-O and comparison with the EQ-5D-3L



Introduction

Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure

Context and theories

Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement

Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O

Methods review

Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity

Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness

Empirical studies

Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures

Chap. 5
Further justification, Data source, Methodological challenges

Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39

Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D

Conclusion

Chap. 8
Summary, implications, areas for further research, contribution and conclusions

7.1 Introduction

Chapter 6 demonstrated the construct validity of the ICECAP-O instrument, cross-sectionally, revealing there are important, measurable differences in the ICECAP-O score in patients with varied severity of Parkinson's. Built on the findings in Chapter 6, this chapter will continue addressing the second research question (is the ICECAP-O appropriate to capture the wellbeing impact in Parkinson's, and is it sensitive in this population? (Chapter 1, Section 1.6)) by assessing the ICECAP-O capability measure in a longitudinal way, i.e. responsiveness; that is, the ability to measure a meaningful or clinically important change in various aspects of Parkinson's patients' QoL and wellbeing over time. This is the first time that the responsiveness of ICECAP-O has been assessed and compared with the EQ-5D-3L in this population.

As discussed in Chapter 3 Section 3.5.2, responsiveness refers to the ability of an instrument to detect important change over time in the construct to be measured (322). Responsiveness of PbQoL measures to patients' preference is critical, as it will affect the magnitude of the QALY. The impact is especially fundamental when QALY gain is small, which may result in dramatic change of funding recommendations of new interventions. In other words, the new intervention appearing to be not cost effective may be attributed to the non-responsiveness of the instrument to its intended benefit, rather than lack of benefit. Limited instrument responsiveness will hinder the use of preference-based measures to accurately measure the change in health utilities and jeopardize the rigor and usefulness of the cost-effectiveness result.

This chapter starts with the aims and specific objectives for this assessment of responsiveness, followed by the methods, results and discussion. Definition of responsiveness, along with theoretical basis and statistical methods (e.g. anchor based methods, choice of anchor, group formation, effect size, correlations) to assess the responsiveness have been reviewed and discussed in greater depth in Chapter 3, Section 3.5 and 3.6. This chapter describes the methods in this specific context of Parkinson's population with data from the PD MED trials in terms of the choice of anchors, the grouping information, the statistical methods for the assessment and the approach to handling missing data.

7.2 Aims and objectives

This chapter aims to explore the responsiveness of the ICECAP-O in people with Parkinson's and compare this with the current most commonly used measure in economic evaluations in Parkinson's, the EQ-5D-3L.

Specifically, there are five objectives for this chapter. They are:

1) To explore the impact of Parkinson's progression on individual's capability-wellbeing;

2) To assess the responsiveness of ICECAP-O in people with Parkinson's to the change of patients' overall and eight specific aspects of QoL (i.e. mobility, ADL, emotional wellbeing, stigma, social support, cognitions, communication, and pain, as measured by PDQ-39), and clinical health status (as measured by H&Y);

3) To assess the responsiveness of the EQ-5D-3L and compare with that of the ICECAP-O to the change of the various health and wellbeing aspects as outlined in the second objective above;

4) To explore how to adapt the psychometric methods for the assessment of responsiveness to the assessment of PbQoL measures; and

5) To investigate the impact of missing data on the result of assessment of responsiveness.

The last chapter has demonstrated cross-sectionally that there was statistically significant and large difference in capability wellbeing between the groups with Early and Later stage of Parkinson's. There is no doubt this finding could be interpreted as the substantial impact of Parkinson's progression on people's capability since all the patients will progress from Early to Later stage. Nevertheless, the nature of analysis in last chapter is a cross-sectional group comparison in that it compares between groups that formed by different individuals. This chapter will therefore assess the impact of progression from another perspective, by examining the actual progression of each individual patient over time. This is the first objective of this chapter.

The second objective is to jointly address the overall research question of this thesis with the assessment of construct validity in last chapter. The feature of ‘being broad’ for a QoL measure brings positive voices but also doubts. The ICECAP-O measure is comprised of five broad dimensions of wellbeing: attachment, security, role, enjoyment and control, each with four levels. It was developed based on Sen’s capability approach and in response to the criticism over the application of extra-welfarism (Section 2.7) to broaden the evaluative scope. ICECAP-O is designed to capture the change of QoL in a broader sense and as such the full picture of a Parkinson’s patients’ life could be incorporated into decision-making. Nonetheless, health-care interventions in many circumstances focus on health aspects of QoL (as opposed to the spill-over broader influence), with a core goal to treat diseases. The broad scope of the ICECAP-O instrument and deliberate divergence from ‘health-specific’ attributes to ‘capability wellbeing’ attributes could arguably generate concerns over its sensitivity to capture specific health changes in a narrower context focused on health. This may affect its use in evaluating typical health-care interventions and as such necessitates the comprehensive assessment of the tool’s responsiveness to change in health status. This leads to the second objective of this chapter.

The review in Chapter 4 showed that there are some concerns regarding the responsiveness of the EQ-5D-3L as the agreement over the longitudinal change between the EQ-5D-3L and the Parkinson’s-specific QoL/clinical measures varied across studies. Half of the included studies showed that the EQ-5D-3L scores reflected changes in clinical status over time as shown on the reference measures, while the other half failed to reach consistent conclusions between the measures. The third objective is thus to test the responsiveness of the EQ-5D-3L in the PD MED data and compare with the ICECAP-O results to illustrate which measure is more responsive to the change of which aspects.

A series of approaches are proposed in the literature to assess responsiveness (397, 403) (please see section 3.6). However, none of them was developed specifically for the assessment of the ‘preference-based’ measures. As discussed in Chapter 5 (Section 5.4), there raised methodological challenges in the methods and interpretation of the results - the testing methods may have to depend on the arbitrary assumptions of people's preferences towards the change on each aspect of health. Therefore, the fourth objective of this chapter is to describe how these challenges are tackled in the methods when applying the traditional psychometric approaches to the test of the PbQoL measures in the PD MED context.

In Chapter 6, the imputation strategy was straight-forward as the data structure was cross-sectional and contained low percentage of missing data (<5%). In contrast, repeated measurements are used in this chapter, which contains up to 70% of missing data for some waves, creating much tougher challenges for the imputation. Simply dropping the whole observation that contains missing data may lead to biased result. Little information about how to address the missing data issue was provided in previous studies of assessment of responsiveness. Therefore, the last objective of this chapter is to fill this gap by examining how different missing data handling strategies would affect the result of assessment of responsiveness.

7.3 Methods

This section will introduce the data used for this analysis, choice of anchor measures, criteria for grouping the participants to different 'change groups' for each type of anchor, how to address missing data, and statistical methods for assessing responsiveness.

7.3.1 Data

The data for these analyses were obtained from the PD MED trials (75, 493). Details for the PD MED trial has been provided in Section 5.3. The data used for this chapter include: socio-demographic characteristics collected at randomisation and the annually collected PDQ-39, ICECAP-O, EQ-5D-3L, and follow-up questionnaires collected since 2010 when the ICECAP-O was added to the trial. The time horizon for this analysis was originally chosen to be four years between 2011 and 2015 but changed to two years given the large amount of missing data after the trial has been running for more than ten years since 2001. However, the four-year analysis was conducted as part of sensitivity analysis.

7.3.2 The anchor based method

As discussed in greater depth in Chapter 3 (Section 3.6), anchor-based approaches examine the relationship between the change in scores of the test measure and the independent anchor or external criterion (404, 409). The anchor(s) may be a clinical objective measure, or a subjective measure such as patient self-reported QoL measures (16), the selection of which depends on the correlations between the anchor and the test instrument, and whether it would increase understanding and of importance to the researchers (14, 15).

The literature strongly recommends the use of multiple anchors, each of which is deliberately constructed with different focus, to examine responsiveness of a multi-dimensional measure from different aspects (403, 408, 524). This is especially of importance in this case of assessing the ICECAP-O capability measure and EQ-5D-3L, the QoL measures. As described in Chapter 2 Section 2.3, QoL has been defined as a concept encompassing a broad range of physical and psychological characteristics and limitations, which describe an individual's ability to function and derive satisfaction from doing so (525). Similarly, capability focuses on the freedom of choosing among a combination of functionings which allows for measuring an even broader set of dimensions of wellbeing (135, 526, 527). This wide definition of QoL and capability wellbeing requires that the choice of anchors for this assessment of responsiveness should cover a wide range of related concepts.

Ten anchors were therefore identified as most appropriate allowing a comprehensive assessment of the impact of Parkinson's. They measure health and QoL aspects that could be specifically affected by Parkinson's to make this assessment context-specific and result relevant to Parkinson's population. The ten anchors are: the most commonly used clinical measure in Parkinson's, the Hoehn and Yahr staging scale (H&Y), and the most commonly used QoL measure in Parkinson's, PDQ-39-SI (SI: summary index) along with its eight dimensions. The justifications for the choice of anchors are provided in the following section.

7.3.2.1 The modified Hoehn and Yahr staging scale

The modified H&Y staging scale consists of seven stages (1.0, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0) of the typical progression of motor function from the mildest stage 1.0 ‘unilateral involvement only’ to the most severe, stage 5.0 ‘wheelchair bound or bedridden unless aided’ (528). It is objectively assessed by clinicians. Using H&Y as an anchor will inform whether the QoL measures are responsive to the clinical change in motor symptoms over time. However, it is not comprehensive in its component for impairment and disability by Parkinson’s, and there is no information concerning non-motor and mental aspects impaired by Parkinson’s. In line with aforementioned methodological considerations concerning the responsiveness test in Chapter 5 (Section 5.4), the result of responsiveness when using H&Y as an anchor should be interpreted with caution since H&Y and the ICECAP-O capability measure are essentially measuring different albeit related concept. Moreover, the broad definition of the staging may cause the insensitivity of H&Y to the slow progression of Parkinson’s, thereby bringing the concern that the anchor itself cannot act perfectly as a criterion for the test measure in terms of responsiveness to the change. Finally, the non-linear relationship between each H&Y stage creates challenges in the anchor group formation by H&Y staging which will be discussed later in the section 7.3.3.1.

7.3.2.2 The PDQ-39 QoL measure

In light of the above limitations with the use of a clinical measure as an anchor, choosing a condition-specific QoL instrument appears to be an optimal addition. The summary score of PDQ-39 QoL measure and its eight dimension scores were identified as relevant anchors to assess the responsiveness of the ICECAP-O. As discussed in greater depth in Section 5.3.4, the PDQ-39 is designed to measure QoL specifically in people with Parkinson’s (62, 131). It assesses impact of Parkinson’s on eight aspects of patients’ QoL, including mobility, ADL, emotional wellbeing, stigma, social support, cognitions, communication and bodily discomfort. Its validity, responsiveness and reliability have been well demonstrated in previous studies (503, 529, 530). Theoretically, all eight dimensions should influence patients’ overall wellbeing and therefore the change (improvement/deterioration) on any of the eight dimensions is expected to affect the overall perception of capability wellbeing as measured by ICECAP-O accordingly. The choice of the eight dimensions also has the benefit of covering a wide scope of physical, mental and social wellbeing domains, which will broaden the understanding of the ICECAP-O capability measure in terms of its relationships with the wide range of aspects affected in Parkinson's patients lives.

It is recommended to use correlations to aid the justification for the choice of the anchor (403, 404) when data are available. A change correlation of 0.3 between the test measure and the anchor is regarded as the minimum threshold to warrant the choice of the anchor (403, 404). Chapter 6 found the correlations between the five dimensions of ICECAP-O and the eight dimensions of PDQ-39 to be 'moderate to strong' (Section 6.5.3, Table 6-10). The cross-sectional summary scores of the ICECAP-O and PDQ-39 were found to be highly correlated with the correlation coefficient larger than 0.7. Therefore, PDQ-39-SI and its eight dimensions were sufficiently correlated with ICECAP-O to be selected as anchors.

7.3.3 Change group formation

7.3.3.1 The modified Hoehn and Yahn scale

Five change groups were defined according to the change of the staging on the modified H&Y scale (Table 7-1). The modified H&Y scale is an ordinal measure with each of the stages presenting distinct classification/clinical states from each other, which facilitates the group stratification. Every change to the next up staging of H&Y is associated with some extent of decreased level of quality of life according to previous studies (458, 531). This has also served as basis for previous decision-analytic modelling studies which constructed the health states based on H&Y staging (144, 145, 532) and thus this was used as grouping criteria for this study. Based on the H&Y, the 'largely improved' (or 'largely deteriorated') group is comprised of the participants with a decrease (or increase) in H&Y by more than one stage (e.g. 'largely deteriorated' if changing from stage 1 to 2.5). The 'slightly improved' (or 'slightly deteriorated') group is comprised of participants who had a change equal or less than one (e.g. 'slight improved' if changing from stage 2.5 to 2 or 1.5). The 'no change' group if the H&Y stage kept the same between the two assessment points.

Table 7-1: Criteria for forming change groups in the base case analysis: H&Y and PDQ-39-SI and eight dimension scores

Anchor measure	From baseline to Follow-up			Score range
	largely Improved/ Deteriorated	A little improved/ Deteriorated	No change group	
H&Y	|change| > 1	0 < |change| ≤ 1	|change| = 0	1 (mildest) ~ 5 (confined to bed)
PDQ-39-SI and dimensions	|change| ≥ 5 *MID	MID ≤ |change|< 5*MID	|change| < MID	0 (best) ~ 100 (worst)

Note: MID – minimally important difference

7.3.3.2 The PDQ-39 QoL measure

MID in the PDQ-39-SI and its eight dimensions were used as the threshold to classify the participants to the change groups (Table 7-1). As described in Section 3.6.2.4, grouping based on MID is the preferred method for continuous measures to assess preference-based measures as it implicitly incorporates the 'importance' in its survey question when asking the participants to determine if they had experienced a little better or worse in their health states.

Understanding the process of how the MID was obtained for the anchor measure is essential for judging its appropriateness and identifying limitations of its use. The MID for PDQ-39 were investigated and reported in Peto, Jenkinson and Fitzpatrick's study (533). They conducted a postal survey by sending PDQ-39 questionnaires to randomly selected members of Parkinson's Disease Society members on two occasions, six-month apart. Additional transition questions were asked at the second time to indicate how much change ('a lot better', 'a little better', 'the same', 'a little worse' or 'a lot worse') the participants had experienced since baseline in overall health and in each of the eight domains of the questionnaire. 728 participants completed the questionnaires at both time points. The mean change in scores for the summary index and each dimension of the PDQ-39 were calculated and compared with the responses to the relevant transition question. The authors argued that the MID could be the mean change in either the 'a little better' group, or the 'a little worse' group, as supported by previous evidence (45) that small positive or negative change from baseline can be treated as equivalent for measuring MID. Although this assumption of equal MID in the better and worse groups was not always supported by literature (46), given the proportion of the sample who reported ‘a little better’ (between 3.7-13.0%) was much less than that who reported ‘a little worse’ (> 25%) for most of the dimensions, it is reasonable that the authors decided to use the mean change score in the ‘a little worse’ sample to calculate MIDs for each dimension and summary score. The limitation of this assumption will be discussed later in this Chapter (Section 7.6.3.2).

Peto et al.’s study (533) showed that the MID (with SD) varies among dimensions: 3.2 (13.26) for mobility, 4.4 (16.56) for ADL, 4.2 (17.09) for emotional wellbeing, 5.6 (22.98) for stigma, 11.4 (23.28) for social support, 1.8 (15.56) for cognition, 4.2 (18.74) for communication, 2.1 (18.68) for pain, and 1.6 (8.89) for overall score. Statistically significant differences on the change scores were found for all except for the cognition and pain dimension. Among those, the sample size for the social support dimension was small as only 33 participants reported ‘a little worse’ while the majority of participants (n=547) reported ‘about the same’. The varied size of MID for each of the PDQ-39 dimensions is a reflection of the different weights that patients put on each dimension when generating their overall perception about their health and wellbeing. This addresses to some extent the issue of lacking a preference-based measure as anchor in this study.

Based on MID, the ‘improved’ (or ‘deteriorated’) group is comprised of the participants with positive (or ‘negative’) change equal or larger than the size of MID; and the no change group is defined as the change is smaller than the size of MID. However, as mentioned in section 3.6.2.4, a limitation of this MID approach is that it does not provide guidance on the plausible cut-offs between the ‘a little change’ and ‘a lot of change’. Peto et al.’s study (533) did not report the average mean change in the ‘a lot better’ and ‘a lot worse’ groups providing its dedicated objective to determine the MID. Given no information is available from previous literature on this issue, sensitivity analysis was conducted to explore the impact of the varied threshold on the result – 5 * (multiplied by) MID was used in the base case analysis and 3 * (multiplied by) MID in sensitivity analysis. 5*MID was preferred over 3*MID as it is a more conservative estimate which is expected to overcome the ‘noise’ brought by the large SD of the MID estimates from Peto et al.’s study.

In addition, sensitivity analysis was also conducted by collapsing the ‘a little’ and ‘a lot’ categories so that the five categories were reduced to three. This method has been used in a previous study (534).

7.3.4 Missing data and multiple imputation

7.3.4.1 Missingness diagnostics

In Chapter 6, Section 6.4.2.3 has detailed the patterns of missing data (MCAR, MAR and MNAR) and the importance of imputation. Therefore prior to evaluating responsiveness, patterns of missing data were explored and logistic regression was conducted to investigate whether the probability of missing was associated with any patient characteristics. A binary variable was created which representing the missingness status (0: missing; 1: non-missing) of PDQ-39-SI. The MCAR assumption could be easily rejected as long as any link between the variable of interest and the probability of missing is identified. The variables of interest were selected from the pool of outcome measures and patient characteristics which may influence the missingness, including age, sex, duration with Parkinson’s, H&Y staging, EQ-5D-3L five dimensions, ICECAP-O five dimensions, number of hospitalisation days for treatment, whether or not having dementia, and whether or not the patient has a caregiver.

If the logistic regression provides the evidence against the MCAR assumption, multiple imputation (MI) should be conducted to impute the missing data (535). MI has to meet a less restrictive assumption, the ‘missing at random’ (MAR). MAR allows missingness to be correlated with observed variables so long as it remains conditionally independent of the unobserved values (518), so the observed variables must suffice for predicting missingness.

7.3.4.2 Imputation strategy

The optimal imputation model should incorporate the correlation between the repeated assessments and allow the maximum inclusion of the variables which can predict missingness. However, it could not be executed successfully after many attempts of adjustment which might be because of the huge imputation burden from the multiplied number of the variables and covariates in the imputation model and the large proportion of missing data. The details will be further discussed later in this Chapter (Section 7.6.3.4).

Compromising imputation strategy was then applied. The time horizon was adjusted to two years (wave 2 to wave 4) from the original four years (wave 1 to wave 5) due to the lower proportion of missing data at wave 2-4. The scores of ICECAP-O, EQ-5D-3L and PDQ-39 were imputed along with the status of dementia and the H&Y scale indicating severity of Parkinson's (19). Predictive mean matching was specified for the continuous ICECAP-O, EQ-5D-3L and PDQ-39 scores, ologit model was specified for the H&Y scale and the logit model was specified for the status of dementia. Age, sex, duration with Parkinson's and the baseline H&Y status were included in the imputation models. Thirty imputed datasets were generated as the overall missing percentage is around 10-30% for each variable (shown in the result section, Table 7-7). The results were combined using Rubin's rule (35). The STATA code for this imputation is provided in Appendix B.

In addition, another two strategies were conducted as sensitivity analysis to the current method. The first used a multi-variate latent normal model with the Realcom Impute software (536), which could handle the two-level (multiple waves within each patient) data structure. However, a weakness was revealed in its formula in that it only used the variables that did not contain any missing data to predict the value of missing variables. This limits its prediction ability as the missingness may be most accurately predicted by the non-missing values of the same variable at other waves.

Another imputation strategy in the sensitivity analysis which treated each wave from the same patient as independent. This was technically practical as it did not multiply the number of variables in the imputation model, and also enabled the missing values to be predicted by both the non-missing values from other waves and the non-missing variables. However, this method neglects the multi-level structure of the data.

After the completion of imputation, the statistical analyses were conducted based on Rubin's rules (520). The missing whole-waves were, despite being imputed, excluded from the analysis (537). The performance of the different imputation strategies was compared with imputation diagnostics strategies. The effect size statistics were calculated in the dataset without imputation and with each method of imputation.

7.3.5 Statistical analysis

The statistical analyses conducted included: 1) a scatter plot using the change of the ICECAP-O index scores / EQ-5D-3L index scores against the change of the PDQ-39-SI, and the correlations between them; 2) effect size and the standardized response mean of the change of ICECAP-O index scores /EQ-5D-3L index scores in each of the change groups formed based on the selected anchors; 3) regression analyses to explore the factors associated with the changes of ICECAP-O index score / EQ-5D-3L index score. These statistical methods have been described in depth in Chapter 3 (please see section 3.6.3). This section will describe the application of these methods with the PD MED data and the modified statistical approaches with the imputed datasets. The statistical analyses were conducted in STATA® 14 (StataCorp. 2015) (521).

7.3.5.1 Scatter plot

To visualize the relationship between the test measure and the anchor, two scatter plots were produced using 1) the change of ICECAP-O score 2) the EQ-5D-3L score, against the change of the PDQ-39-SI for each patient. Fitted linear regression lines predicting the ICECAP-O/EQ-5D-3L change score using PDQ-39-SI change score were added to the plots with 95% confidence intervals. The change was averaged over the imputed datasets. Given the different scales for the three QoL measures, the change scores were standardized using the following formula before plotting to facilitate visual comparisons of the distributions:

standardized change = (change - mean (change)) / SD (change).

7.3.5.2 Correlation

Correlations of the change of ICECAP-O / EQ-5D-3L index value against the change of PDQ-39-SI score were examined. This provides an indication of the extent to which the change score of the anchor and test measure (ICECAP/EQ-5D-3L) are associated; a stronger correlation typically means a stronger responsiveness of the test measure to the anchor (415). The correlation coefficient describes both the strength and direction of the relationship.

Linear regressions of the ICECAP-O / EQ-5D-3L against PDQ-39 were conducted and the normality of the residuals were tested to choose between the Pearson's correlation coefficient and a Spearman's rank correlation coefficient. To enable the normality testing in the multiply imputed datasets, the approach proposed by White et al. (538) was adopted. It recommends calculating fitted values and residuals for each imputed dataset first, and then, for each imputed dataset, plot these residuals against the fitted values. Following White's recommendation, the kernel density plot was conducted to allow a visual comparison of the distribution of the residuals against an overlaid normal distribution (427, 428). The Shapiro-Wilk W test was then performed for significance testing of the assumption that the distribution was normal (427, 429). Given the number of imputed datasets was relatively large (n=30) in this study, five out of the 30 datasets were randomly selected for the normality checking, i.e., the 1st, 3rd, 10th, 20th and 30th imputed dataset. Pearson's correlation coefficient was used when the linear relationship was demonstrated and when it was not met, Spearman's rank correlation coefficient was used (430). As with the correlation comparison in Chapter 6 (please see Section 6.4.2.2), the correlation coefficients of the change scores of ICECAP-O and PDQ-39 were compared with EQ-5D and PDQ-39 using the R statistical package *cocor* which examines the significance of the difference of correlation coefficients using a range of statistical tests.

The Pearson correlation coefficient of the imputed datasets was calculated with Harel's Fisher's r to z transformation utilizing the normal distribution of z (539). This was realized with the user-written STATA command '*mibeta*' with the option '*fisherz*' (539). Detail of this transformation is provided in Appendix D. Compared to the Pearson correlation coefficient, Herel's transformation method cannot be applied to Spearman rank correlation since the linear relationship assumption does not meet. Neither does a user-written command exist to calculate Spearman rank correlation in imputed datasets and hence a compromise approach was adopted. The average of the standardized score change among the 30 imputed datasets was calculated (i.e. not combined using Rubin's rule) and then the routine procedure (STATA command *spearman*) for Spearsman's rank correlation was performed.

7.3.5.3 Effect sizes and standardised response means

Effect size (ES) (i.e. mean change divided by the standard deviation (SD) of the baseline score) and the standardised response mean (SRM) (i.e. mean change divided by the SD of the change score) as introduced in Chapter 3 (Section 3.6.3.1) were calculated for each of the groups classified by the ten anchors. The standard error and confidence intervals for the estimates of SRM and ES were generated through bootstrap with 1000 replicates.

Without imputation, SD is calculated by multiplying SE with square root of the sample size, whereas after the imputation, the sample size of each change groups varies across the imputed datasets depending on the imputed values. To address this, a transformation was performed. The ES/SRM was calculated individually for each patient using the SD for each imputed dataset. Then the means of the individual ES/SRM in each imputed dataset were combined using Rubin's rule. This process can be summarized with the formula below:

For each of the five change groups defined within imputed dataset (i=1,2,…30)), let the change of the measure (i.e. ICECAP-O or EQ-5D-3L) be noted as xic (x1c, x2c, x3c…x30c), and the baseline measurement be noted as xib (x1b, x2b, x3b…x30b).

$$\text{ES} = \frac{\mu\,(\text{xic})}{\text{SD}\,(\text{xib})} = \mu\,(\frac{xic}{SD\,(xib)})\ \text{(i=1,2,...30)}$$

$$\text{SRM} = \frac{\mu\,(\text{xic})}{\text{SD}\,(\text{xic})} = \mu\,(\frac{xic}{SD\,(xic)})\ \text{(i=1,2,…30)}$$

This transformation bypasses the difficulty with calculating SD, and allowed the pooled estimates of ES/SRM applying Rubin's rule.

According to Cohen's rule of thumb (419), an ES below 0.2 represents 'very small' effect, 0.2 to 0.5 – 'small', 0.5 to 0.8 – 'medium', and a ES score higher than 0.8 represents a 'large' ES (421). It was inconsistent in the literature how to interpret the SRM with some suggesting SRM may be interpreted with the same rule (540) while other suggested it cannot be directly interpreted (422). However, as mentioned in Chapter 3, the COSMIN checklist manual gave a warning on using this rule to assess the responsiveness of a measure. Cohen's rule applies in situation where effect size is used as an indicator for treatment effect, rather than for testing the responsiveness of an instrument to a change (327), and therefore simply examining the size of ES/SRM is inappropriate. They proposed that the judgement on the responsiveness should be established on whether the magnitude of ES/SRM meet the expectations on the relationship between the test measure and the anchors. This has been discussed in more detail in Chapter 3 (Section 3.6.3.1).

In addition, this study used EQ-5D-3L as reference, which facilitates the interpretation of the ES/SRM statistics through the comparison between ICECAP-O and EQ-5D-3L. In this study, the expectations of the magnitude of ES/SRM are indicated by the change (direction and magnitude) of the anchor measure in each of the five change groups. For example, the participants in the 'largely improved group' anchored by PDQ-39 are expected to have medium to large ES/SRM in ICECAP-O. In this way, a measure is judged to be more responsive when it meets the expectation of each of the change group to a larger degree than the other measure.

7.3.5.4 Paired t-test

The paired t-test (415, 425) was used to assess change of the ICECAP-O / EQ-5D-3L between baseline and follow-up in each of the four change groups (i.e. except for the 'no change' groups). Due to its high dependence on the sample size, the result from the paired t-test only plays a supportive role in the interpretation of the responsiveness.

7.3.5.5 Regression

Two multivariate linear regressions were conducted to further examine the extent to which the difference in the change of the ICECAP-O / EQ-5D-3L index score was influenced / explained by the change of eight QoL aspects as measured by the PDQ-39 eight dimensions. This method has been commonly used in previous validation of preference-based instrument (384, 541, 542). For example, in addition to assessing validity and responsiveness of the EORTC-8D relative to the EQ-5D-3L, Lorgelly (2017) employed regression analysis to understand the determinants of the difference in QALYs generated independently by EORTC-8D and EQ-5D-3L (384). This aids the understanding of the interaction between the test measure and the anchors, and interpretation of the effect size results. A small effect size is due to the relative magnitude of ‘noise’ and in this context, the ‘noise’ could come from the fact that patients value the change in other aspects of life more than the aspects that is being examined. A multivariate regression analysis will help the understanding of to what degree the change of Parkinson’s specific QoL aspects would lead to the change of the ICECAP-O after adjusting for each other. It would also inform which of these QoL aspects are deemed most relevant for a patient’s capability wellbeing.

7.3.5.6 Sensitivity analysis

Three sensitivity analyses were carried out to explore the impact of different methodologies on the responsiveness result. These sensitivity analysis have been mentioned in the methods above while describing the main analysis and this section provides a brief summary of them. First, a two-level imputation model provided by the Realcom Impute software and imputation in long-form were carried out using the four year data as well as the two year data in addition to the imputation strategy used for the main analysis. Second, 3 * MID (in contrast to 5 * MID in the main analysis) was used as the cut-off for PDQ-39-SI between a small change group and a large change group given it represents the midpoint between 1 and 5. Lastly, the number of change groups reduced from five to three by collapsing the ‘a little’ and ‘a lot’ categories to become ‘worse’, ‘no change’ and ‘improved’. For all the sensitivity analyses, effect size results with PDQ-39-SI as anchor were estimated and compared with the main analysis results.

7.4 Results

The results section begins with a description of participant characteristics. Two scatter plots are provided to illustrate the relationships between the change on the ICECAP-O/EQ-5D-3L and the change on PDQ-39, which are explained by the correlations between the summary scores of the three measures. The grouping is summarized, followed by the effect size statistics of each group for each chosen anchor. The result of the regression analysis is presented at the end.

7.4.1 Patient characteristics

Characteristics of the PD MED study participants used for this responsiveness analysis are presented in Table 7-2. A total of 1,238 participants were included for this analysis. Among them, 1,023 participants were eligible for the assessment of responsiveness who did not have unit-nonresponse (i.e. provided some responses to the assessment) at the two assessment points[8]. The mean age of participants at baseline for this analysis is approximately 74 years, with the mean duration since diagnosis of Parkinson's 6.5 years. 65 % of the participants are male, which is in line with a previous finding that the relative risk of the incident rate of Parkinson's in men is 1.5 times greater than women (543).

[8] We included the participants that only had one ICECAP-O assessment to maximize information for the multiple imputation. However, they were excluded for analysis after imputation completed as change score requires two waves however in these participants another wave was purely imputed which may introduce overfitting of the estimation model (537).

Table 7-2: Participant characteristics, at baseline and two-year follow-up (complete case)

Characteristics	Baseline						Two years later					
	Mean	SD	Median	IQR	Sample size	Missing (%)	Mean	SD	Median	IQR	Sample size	Missing (%)
Age	74.15	8.25	75.34	69.42, 79.78	1238	0						
Duration with PD	6.54	3.44	6.03	4.07,8.31	1238	0						
H&Y			2.5	2, 3	1060	14.38			3	2, 3	806	34.89
ICECAP-O index[1]	0.75	0.18	0.79	0.67, 0.89	1056	14.70	0.72	0.19	0.77	0.63, 0.87	879	29.00
EQ-5D-3L index[1]	0.52	0.29	0.59	0.62, 0.69	1128	8.89	0.46	0.32	0.52	0.19, 0.66	882	28.76
PDQ-39-SI[1]	31.72	17.41	29.87	19.09, 42.97	972	21.49	34.30	17.83	32.86	21.38, 45.16	704	56.87
	Freq	% of the observed	Missing %				Freq	% of the observed	Missing %			
Sex (male)	805	65.02	0									
Dementia			13.65						33.52			
With	82	7.67					126	15.31				
without	987	92.33					697	84.69				
Motor complication[2]			49.60						46.20			
With	364	58.33					283	42.49				
Without	260	41.67					383	57.51				

1 Note: score range: ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0 (death) ~ 1(full health). PDQ-39: 0 (least severe) ~ 100 (most severe).

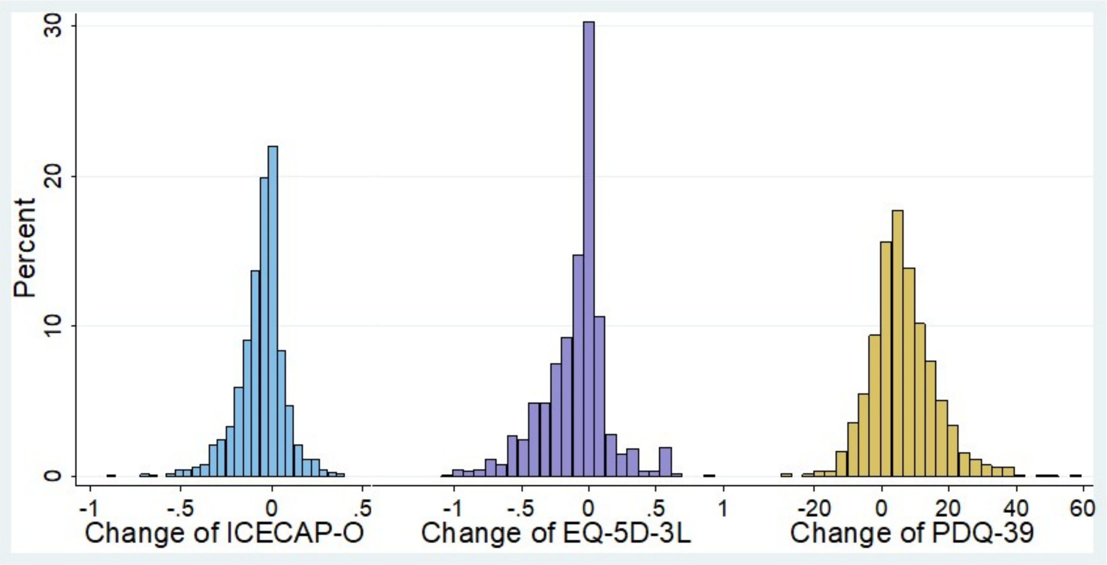
2 Motor complication includes motor fluctuation and dyskinesia. Motor fluctuation refers to the situation that patients oscillate between 'on', during which the patient experiences a positive response to medication and 'off' state, during which the symptoms cannot be controlled by the medication (544). Dyskinesia are involuntary movements in the muscles, often include twitches, jerks, twisting or writhing movements (545).

An overall trend of deterioration in clinical scores, health related quality of life and capability wellbeing over the two years was observed. The median H&Y stage increased (i.e a deterioration in health) from 2.5 (IQR: 2-3) to 3 (IQR: 2-3). Table 7-3 shows specifically the distribution of the H&Y staging at baseline and two-year follow-up. There is a decrease in the percentage of the patients at the milder stages, i.e., stage 1-2.5, and an increase in the percentage of the patients at the more severe stages, stage 3-5. In terms of having dementia, 6.6% of the participants reported having dementia at baseline, which increased to 10.2% at endpoint, although the percentage of missing data also increased from 13.65% to 33.52% (Table 7-2). 29.4% patients reported they had motor complication at baseline, which, unexpectedly, decreased to 22.86% two years later. This might be due to the very large proportion of missing data presented for this variable, i.e. patients that developed motor complications might take other treatment options such as deep brain stimulation surgery and thus may be more likely to drop out from the trial.

Table 7-3: Distribution of H&Y scores, at baseline and two-year follow-up (complete case)

Characteristic	Baseline			Two years later		
	Freq	% of the observed	Missing (%)	Freq	% of the observed	Missing (%)
H&Y staging			14.38			34.89
1	67	6.32		15	1.86	
1.5	90	8.49		36	4.47	
2	276	26.04		167	20.72	
2.5	199	18.77		123	15.26	
3	301	28.4		293	36.35	
4	96	9.06		122	15.14	
5	31	2.92		50	6.2	

All three QoL measures showed overall deterioration in QoL (Table 7-2, Figure 7-1). Figure 7-1 shows the distribution of the change score of the three measures with the imputed data. A change score being 0 indicates that there was no change over the two years. For EQ-5D-3L and ICECAP-O, a positive change score (>0) indicates the improvement of health status while a negative change score (<0) indicates deterioration. This trend is the opposite for PDQ-39 where a positive change score indicates a worse health status and negative indicates a better health. Figure 7-1 shows that larger area of distribution is on the worsened health status side for all the three measures. In particular, a spike at 0 was observed for EQ-5D-3L, which was partly attributed to the high percentage of no change in score for approximately 23.3% (238/1023) of patients. The mean change for ICECAP-O, EQ-5D-3L and PDQ-39 were -0.057 (95%CI -0.066, -0.046), -0.095 (95%CI -0.11, -0.077), and 6.36 (95%CI 5.36, 7.36) respectively.



Mean change (95% CI)	-0.057 (-0.066, -0.046)	-0.095 (-0.11, -0.077)	6.36 (5.36, 7.36)

Figure 7-1: The distribution of the change of ICECAP-O, EQ-5D-3L and PDQ-39

Note: score range: ICECAP-O: 0 (no capability) ~ 1 (full capability). EQ-5D-3L: -0.59 (worse than death) ~0 (death) ~ 1(full health). PDQ-39: 0 (least severe) ~ 100 (most severe).

In terms of the responses to each dimension of ICECAP-O, Table 7-4 and Figure 7-2 shows that there was a higher proportion of lower levels (Level 1 and 2) of capability at two years later compared to baseline and decreased percentages of responses for all of the more capable levels (level 3 and 4) over time. Specifically, the proportion of level 1 (lowest level of capability) increased and other levels dropped for security dimension, level 1 & 2 increased and 3 & 4 dropped for role, enjoyment and control, and level 1 & 2 & 3 increased and 4 dropped for attachment. This suggests that the main deteriorating aspects were control, followed by security, role, enjoyment, and lastly, attachment.

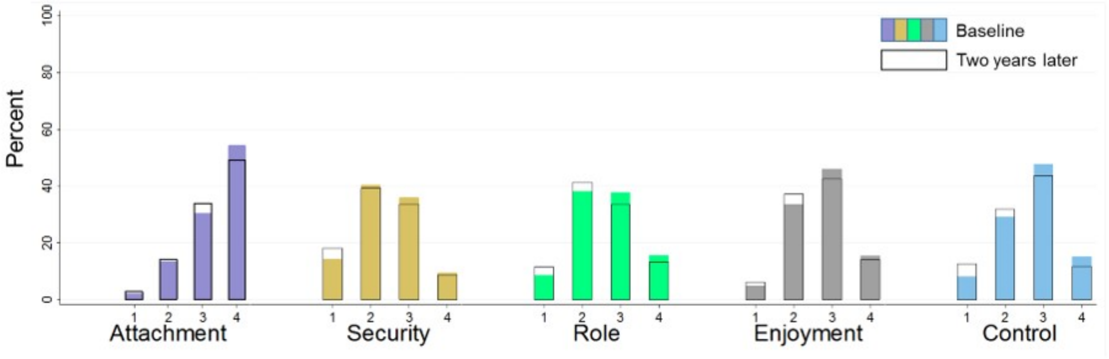


Figure 7-2: Change of the distribution of responses for each of the ICECAP-O attributes over the two years (complete case)

Note: For ICECAP-O, level 1 is the least capable level while level 4 is the most capable level.

Table 7-4: Responses (percentages) for each of the ICECAP-O attributes, at baseline and two-year follow-up (complete case)

ICECAP-O attributes [a,b](%)	Baseline					Two years later				
	Level 1 No capability	Level 2 A little capability	Level 3 Some capability	Level 4 A lot of capability	Missing	Level 1 No capability	Level 2 A little capability	Level 3 Some capability	Level 4 A lot of capability	Missing
Attachment	2.07	13.37	30.32	54.24	14.22	2.83	14.16	33.97	49.04	28.68
Security	14.33	40.44	35.86	9.37	13.73	18.25	39.3	33.67	8.78	28.27
Role	8.61	38.01	37.73	15.64	13.73	11.6	41.45	33.67	13.29	28.27
Enjoyment	4.88	33.58	46.06	15.48	13.89	6.18	37.2	42.47	14.16	28.11
Control	8.06	29.14	47.6	15.19	13.81	12.6	31.95	43.75	11.7	28.19

Note: a. the proportion for level 1-4 did not account for missing values (i.e. with total observed as denominator). b. For ICECAP-O, level 1 is the least capable level while level 4 is the most capable level.

Table 7-5: Responses (percentages) for each of the EQ-5D-3L attributes, at baseline and two-year follow-up (complete case)

EQ-5D-3L Attributes [a,b] (%)	Baseline[a]				Two years later[a]			
	Level 1 No problem	Level 2 Some problems	Level 3 A lot of problems	Missing	Level 1 No problem	Level 2 Some problems	Level 3 A lot of problems	Missing
Mobility	19.63	78.88	1.49	8.24	15.99	80.75	3.26	28.27
Self-care	37.97	56.21	5.81	8.32	30.10	59.64	10.26	28.35
Usual activities	17.02	69.49	13.49	8.40	15.45	64.82	19.73	28.35
Pain/discomfort	20.56	68.58	10.86	8.48	19.41	68.17	12.42	28.43
Anxiety/depression	42.33	53.88	3.79	8.41	34.08	60.27	5.65	28.44

Note: a. the proportion for level 1-3 did not account for missing values (i.e. with total observed as denominator). b. For EQ-5D-3L, Level 1 is “no problem”, level 2 “some problems”, and level 3 “extreme problems”.

The proportion of responses for each level of EQ-5D-3L five dimensions at baseline and two years later are shown in Table 7-5 and Figure 7-3. Approximately 8% of the data were missing at baseline and this increased to around 28% two years later. Despite the missing data, the deterioration trend appears clear for all the five dimensions. This trend is particularly noticeable for self-care, depression/anxiety dimension and usual activities; 20% less in level 1 (no problem) and a corresponding increase in level 2 and 3 were observed for the former two, and a decrease in level 1 (no problem) & 2 and 30% more in 3 (a lot of problems) for the latter.

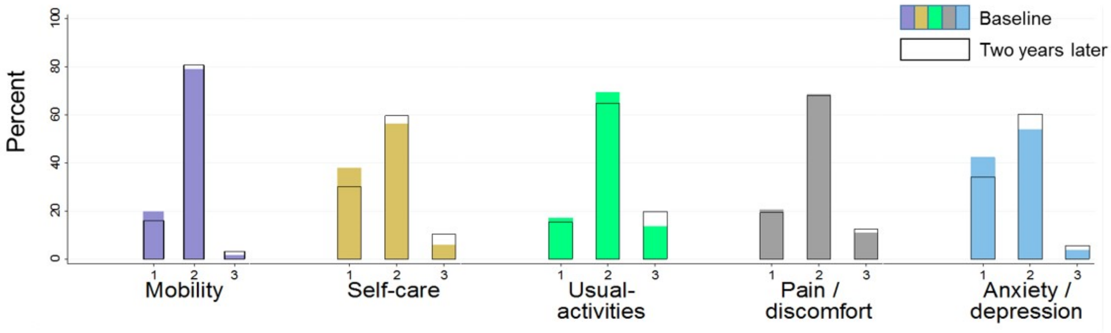


Figure 7-3: Change of the distribution of responses for each of the EQ-5D-3L attributes over the two years (complete case)

Note: For EQ-5D-3L, Level 1 is “no problem”, level 2 “some problems”, and level 3 “extreme problems”.

This deterioration trend was also demonstrated in the Parkinson’s specific QoL measure PDQ-39 with the increased overall score (Table 7-2) and the increased score for all the eight dimensions (Table 7-6 and Figure 7-4) over time. Deterioration of the dimensions over the two years can be ranked according to the size of the mean change (largest to smallest), as mobility (mean 10.54 (SD 18.25)), ADL (7.84 (17.60)), communication (5.54 (16.60), cognition (5.14 (15.90)), emotional wellbeing (4.47 (17.07)), bodily discomfort (3.79 (20.14)), social support (3.21 (14.47), and stigma (1.78 (17.40)).

Table 7-6: PDQ-39 eight dimension scores, at baseline and two-year follow-up (complete case)

PDQ-39 eight dimensions	Baseline				Two years later				Difference over the two years			
	Mean	SD	Sample size	Missing (%)	Mean	SD	Sample size	Missing (%)	Mean	SD	Sample size	Missing (%)
Mobility	49.53	30.77	1102	10.99	56.06	30.65	863	30.29	10.54	18.25	801	35.30
ADL	41.90	25.83	1080	12.76	46.19	26.25	799	35.46	7.84	17.60	755	39.01
Emotional wellbeing	28.79	22.41	1062	14.22	31.16	23.19	785	36.59	4.47	17.07	733	40.79
Stigma	22.15	23.56	1082	12.60	22.68	22.87	807	34.81	1.78	17.40	762	38.45
Social support	11.98	17.73	1055	14.78	13.89	18.65	783	36.75	3.21	14.67	731	40.95
Cognition	35.77	21.86	1082	12.60	37.70	22.02	801	35.30	5.14	15.90	758	38.77
Communication	26.08	22.80	1087	12.20	29.23	23.62	802	35.22	5.54	16.60	760	38.61
Bodily discomfort	41.00	24.98	1080	12.76	43.73	25.39	798	35.54	3.79	20.14	756	38.93

Note: score range: PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

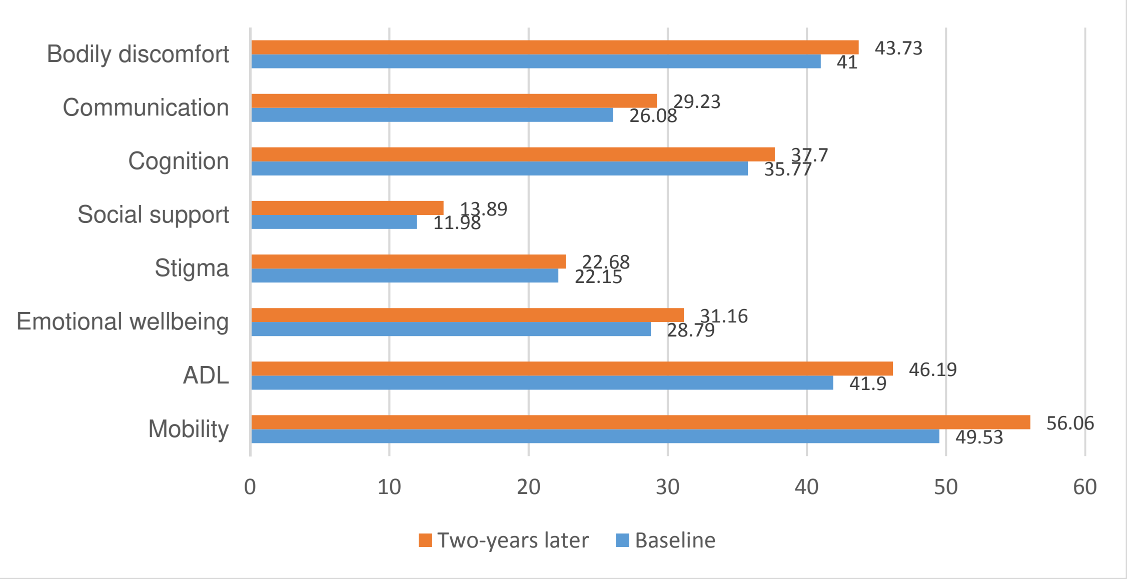


Figure 7-4 : PDQ-39 eight dimension scores, at baseline and two-year follow-up (complete case)

Note: score range: PDQ-39-SI and each of the attributes: 0 (least severe) ~ 100 (most severe).

7.4.2 Missingness prediction

The result for the logistic regression for explaining missingness is shown in Table 7-7. Four independent variables were found to be associated with missingness of PDQ-39: age, ICECAP-O attachment, number of days treated in hospital, and lastly, whether the patient has dementia. Among them, the presence of dementia strongly predicted the missing observations – the odds of missing when patients have dementia is around ten times than the patients do not have dementia (OR=0.109, p<0.0001). The pseudo R square of the logistic regression was tested to be 0.153 and the p value for the whole model was <0.0001. The C-statistic (area under the ROC curve) was 0.728. This demonstrated that the MCAR assumption was not met and provided justification for the multiple imputation approach.

Table 7-7: Predictors for the probability of missing values (0: missing, 1: non-missing) for PDQ-39-SI

Independent variables	Odds Ratio	Standar d error	P value	Lower 95% CI	Upper 95% CI	Note for coding
Age (years)	**0.978**	0.008	**0.0080**	0.963	0.994	Continuous
Sex	0.895	0.112	0.3760	0.701	1.144	Binary: 1- male; 2-female
Duration with Parkinson's (years)	0.972	0.017	0.1060	0.940	1.006	Continuous
H&Y stage (1-5)	1.011	0.084	0.8980	0.858	1.190	Ordinal, 1-mildest, 5-most severe
EQ-5D-3L						
Mobility	0.859	0.155	0.4020	0.603	1.225	Ordinal (1,2,3), Level 1: 'no problem'; level 3: 'quite a lot problems'
Self-care	1.148	0.158	0.3150	0.877	1.504	
Usual activities	0.856	0.129	0.3010	0.637	1.149	
Pain/discomfort	1.078	0.129	0.5280	0.853	1.364	
Anxiety/depression	1.014	0.132	0.9170	0.786	1.308	
ICECAP-O						
Attachment	**1.211**	0.102	**0.0230**	1.027	1.427	Ordinal (1,2,3,4), Level 1: 'no capability'; level 4: 'full capability'
Security	1.030	0.085	0.7250	0.875	1.211	
Role	1.053	0.109	0.6180	0.859	1.291	
Enjoyment	0.971	0.111	0.7940	0.776	1.214	
Control	1.178	0.129	0.1350	0.950	1.460	
Days treated in Hospital	**0.984**	0.008	**0.0470**	0.969	1.000	Continuous
Dementia	**0.109**	0.018	**0.0000**	0.079	0.151	Binary, 0-no dementia;1-have dementia,
Carer	0.902	0.131	0.4800	0.678	1.200	Binary: 0–no carer; 1- has carer
Constant	20.213	19.847	0.0020	2.950	138.494	

7.4.3 Scatter plot of the change in ICECAP-O/EQ-5D-3L

Figure 7-5 and Figure 7-6 shows the scatter plot of the standardized change of ICECAP-O / EQ-5D-3L against the standardized change of PDQ-39-SI. The fitted red lines in the plots are the predicted ICECAP-O/EQ-5D-3L change score from a linear regression of ICECAP-O/EQ-5D-3L change score on PDQ-39-SI change score, with 95% confidence interval highlighted in the grey area. In the plots, each matched dyad (pairing of change score on ICECAP-O and change score on PDQ-39-SI or pairing of change score on EQ-5D-3L and change score on PDQ-39-SI) represents one patient. A dyad on x-axis (y=0) indicates there was no change of ICECAP-O/EQ-5D-3L over the two years, and similarly a dyad on y-axis (x=0) means no change of PDQ-39-SI happened over the two years. A dyad on the origin means there was no change for both measures over the two years.

There is no substantial difference in terms of the pattern shown in the two plots. The dyads in Figure 7-5 appear more concentrated around the predicted linear regression line compared to the dyad in Figure 7-6, indicating a slightly stronger correlation between ICECAP-O change score and PDQ-39-SI change score. The size of the standardized EQ-5D-3L change score is larger than that of the ICECAP-O change score which indicated a larger mean of EQ-5D-3L change score and larger SD.

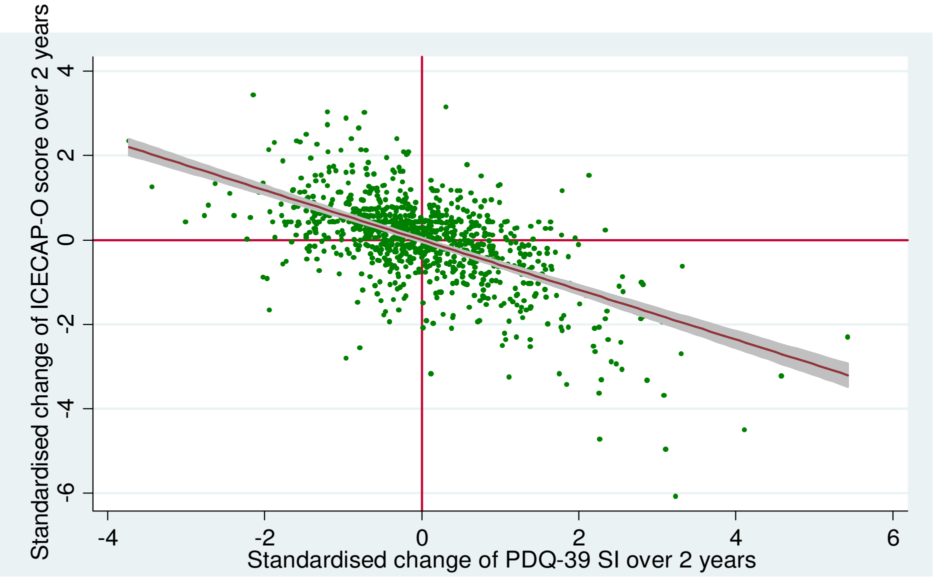


Figure 7-5: Scatter plot of the change of ICECAP-O and PDQ-39-SI

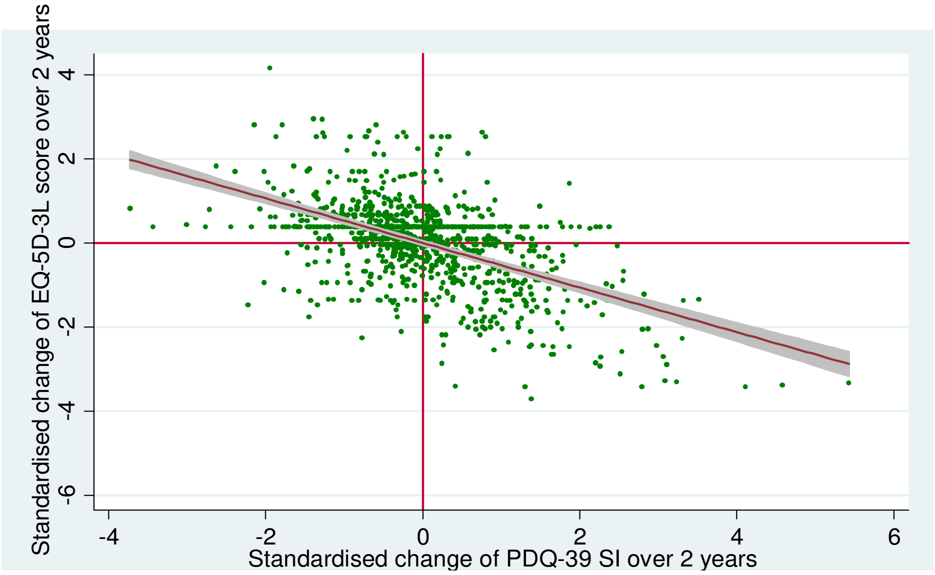


Figure 7-6: Scatter plot of the change of EQ-5D-3L and PDQ-39SI

7.4.4 Correlation

Two regressions were conducted predicting ICECAP-O / EQ-5D-3L index value with PDQ-39-SI respectively and the normality of the residual of the regression were checked. For both of the regressions, the Shapiro-Wilk W test rejected the assumption that the residuals have a normal distribution with a p value <0.00001 for all of the randomly selected imputed datasets (n=1, 3, 10, 20 and 30). The kernel density graphs of the residuals had a sharp bulge shape above the normal density reference line. Figure 7-7 shows an example of the kernel density graph.

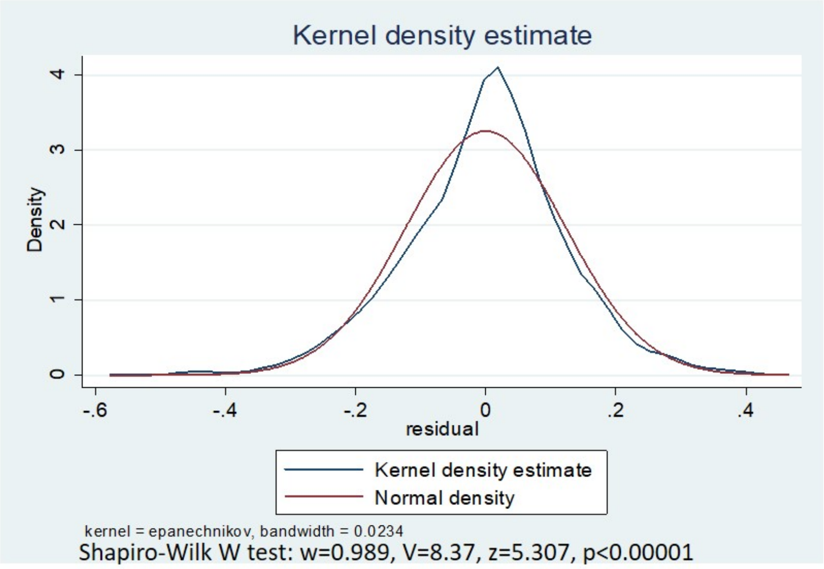


Figure 7-7: An example for checking the normality of the regression residual to choose between Pearson correlation and Spearman correlation: ICECAP-O change score vs. PDQ-39 change score (m=10)

Therefore, a linear relationship between the ICECAP-O/EQ-5D-3L and PDQ-39 cannot be met and accordingly Spearman correlation coefficient was determined to be more appropriate than Pearson correlation coefficient. The detailed result of the residuals normality checking is shown in the Appendix E.

Table 7-8 shows the correlation coefficients between the change of the three measures. The Pearson correlation coefficients are also presented here in addition to Spearman correlation coefficients as additional information. The correlation coefficients agreed with the scatter plot that the change of ICECAP-O index value was more strongly correlated with the change of the PDQ-39-SI (r = -0.526) than the change of EQ-5D-3L index value (r=-0.483) (one sided p value = 0.07 based on Pearson and Filon’s z statistics, 95% confidence interval: -0.0145, 0.1006). Full results of the statistical tests of the difference of correlations are provided in Appendix C.

Table 7-8: Correlation coefficients between the change scores of the ICECAP-O, EQ-5D-3L and PDQ-39

Change of the summary scores / index scores	ICECAP-O vs. PDQ-39	EQ-5D-3L vs. PDQ-39	ICECAP-O vs. EQ-5D-3L
Spearman correlation coefficient	-0.526	-0.483	0.401
Pearson correlation coefficient	-0.536	-0.482	0.425

7.4.5 Effect size statistics

7.4.5.1 Anchor by PDQ-39-SI

A total of 933 patients were eligible for this analysis after excluding the patients that had missing whole-wave assessments for either baseline or follow-up assessment at two-years later for this analysis. Table 7-9 shows the grouping by PDQ-39-SI using the MID information. The sample size column depicts the median and average number of patients in each group across the 30 imputations. Nearly half of the participants are in the 'largely deteriorated group' for PDQ-39-SI. This may be because the MID for PDQ-39-SI is 1.6, and thus the 5*MID threshold (=8) between the slight and large change is still relatively small compared to the overall score range (0-100). This also explains the relatively small number of participants in the 'no change group'. Despite the low threshold, a large overall deterioration is still observed in this 'largely deteriorated group' with a mean change score of PDQ-39-SI being 17.09 (95%CI 15.82, 18.36). From the 'largely improved group' to the 'largely deteriorated group', the mean change of the PDQ-39-SI increased from -13.58 (95%CI -15.11, -12.05) to 17.26 (95%CI 16.22, 18.29).

Table 7-9: Change group formation: PDQ-39-SI in each change group anchored by PDQ-39-SI

Stata wide form (yr2, yr4) Anchor: PDQ-39-SI	Sample size*		Change of the anchor measure		
	Median	Average	Mean	SE	95% CI
Largely Improved group	85	86.5	-13.577	0.771	-15.106, -12.049
Slightly improved group	144	144.3	-4.734	0.192	-5.113,-4.356
No change group	128.5	128.3	0.050	0.093	-0.134,0.233
Slightly deteriorated	277.5	277.1	4.412	0.120	4.177,4.648
Largely deteriorated group	387	386.8	17.257	0.525	16.222,18.292

* The sample size column depicts the median and average number of patients in each group across the 30 imputations

Table 7-10 shows the change of ICECAP-O and EQ-5D-3L in each change group defined by the change of PDQ-39-SI. In line with the trend of the PDQ-39-SI, from the ‘largely improved group’ to the ‘largely deteriorated group’, the mean change of ICECAP-O index value in each group decreased from the best 0.050 (95%CI 0.011, 0.088) (p=0.011) to the worst -0.125 (95%CI -0.144, -0.105) (p<0.0001), and the mean change of the EQ-5D-3L index value decreased from the best 0.059 (95%CI -0.009, 0.126) (p=0.086) to the worst -0.210 (95%CI -0.244, -0.176) (p<0.0001).

The ES and SRM results of ICECAP-O and EQ-5D-3L with the anchor PDQ-39-SI are described in Table 7-11 and visualised in Figure 7-8. For the change groups, the larger ES or SRM means the more responsiveness of the measure to the change of PDQ-39-SI. Although EQ-5D-3L had a larger mean change than ICECAP-O in each group, it also had a larger size of SD for its means, leading to a smaller ES and SRM of EQ-5D-3L compared to ICECAP-O. The confidence intervals for the SRM presented as error bars in Figure 7-8 were overlapping, indicating that there was no statistically significantly difference between the responsiveness of EQ-5D-3L and ICECAP-O to the change of the PDQ-39-SI.

Table 7-10: Change score of the ICECAP-O and EQ-5D-3L over the two years for each of the five change groups (anchored by PDQ-39-SI)

Stata wide form (yr2,yr4) Anchor: PDQ-39-SI	Change in ICECAP-O score				Change in EQ-5D-3L score			
	Mean	SE	95% CI	p value	Mean	SE	95% CI	p value
Largely Improved group	0.050	0.019	0.011,0.088	0.0107	0.059	0.034	-0.009,0.126	0.0858
Slightly improved group	0.009	0.012	-0.015,0.032	0.4835	0.002	0.024	-0.044,0.049	0.9326
No change group	-0.016	0.012	-0.04,0.009	0.2095	-0.016	0.023	-0.061,0.029	0.4901
Slightly deteriorated	-0.043	0.008	-0.059,-0.028	0.0000	-0.067	0.015	-0.095,-0.038	0.0000
Largely deteriorated group	-0.125	0.010	-0.144,-0.105	0.0000	-0.210	0.017	-0.244,-0.176	0.0000

Table 7-11: ES and SRM statistics of ICECAP-O and EQ-5D-3L with the anchor PDQ-39-SI

Stata wide form (yr2,yr4) Anchor: PDQ-39-SI	Effect size (ES)						Standardised response mean (SRM)					
	ICECAP-O			EQ-5D			ICECAP-O			EQ-5D		
	Mean	SE	95% CI	Mean	SE	95% CI	Mean	SE	95% CI	Mean	SE	95% CI
Largely Improved group	0.256	0.062	0.130,0.375	0.194	0.072	0.046,0.328	0.353	0.090	0.160,0.513	0.228	0.085	0.047,0.380
Slightly improved group	0.052	0.048	-0.034,0.155	0.006	0.053	-0.101,0.107	0.073	0.068	-0.050,0.215	0.007	0.068	-0.131,0.134
No change group	-0.105	0.048	-0.223,-0.037	-0.057	0.051	-0.160,0.042	-0.145	0.065	-0.311,-0.056	-0.080	0.072	-0.223,0.059
Slightly deteriorated	-0.277	0.042	-0.353,-0.189	-0.261	0.039	-0.332,-0.180	-0.388	0.052	-0.486,-0.281	-0.331	0.048	-0.419,-0.232
Largely deteriorated group	-0.778	0.070	-0.910,-0.637	-0.791	0.065	-0.913,-0.660	-0.754	0.044	-0.837,-0.663	-0.700	0.049	-0.796,-0.605

Note: an effective size statistic below 0.2 represents ‘very small’ effect, 0.2 to 0.5 – ‘small’, 0.5 to 0.8 – ‘medium’, and a ES score higher than 0.8 represents a ‘large’ ES (377). But this should be interpreted together with the expected ES/SRM as categorised in each of the change group. A measure is judged to be more responsive when it meets the expectation of the assignment of the change group to a larger degree than the other measure. Please see Section 7.3.5.3 for details.

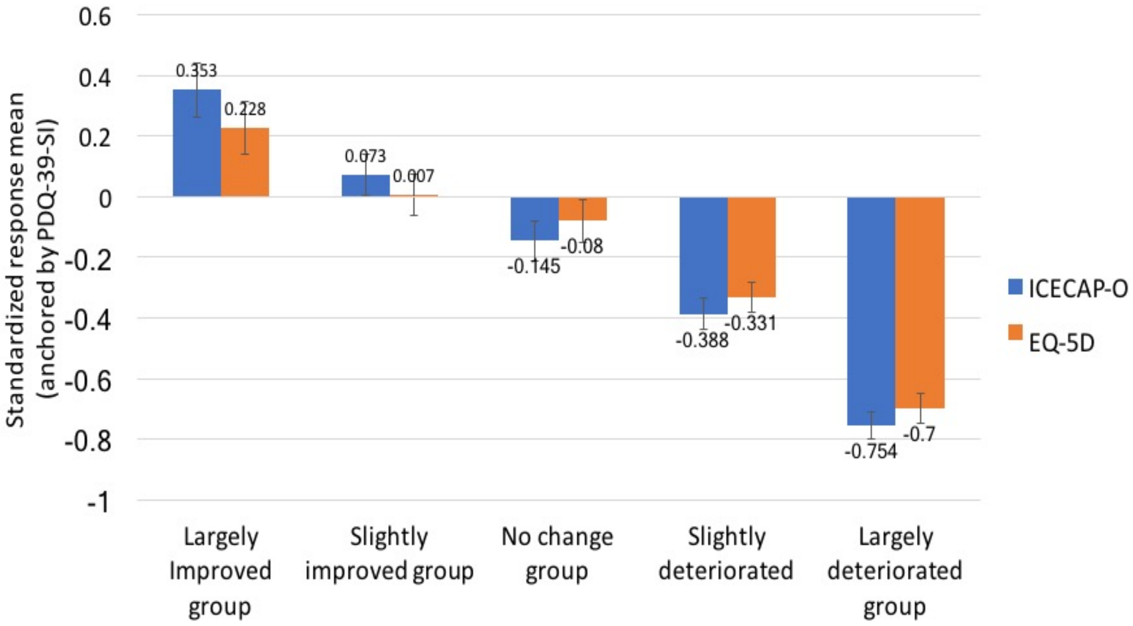


Figure 7-8: SRM (SE) of ICECAP-O and EQ-5D-3L in the five change groups anchored by PDQ-39-SI

Note: an effective size statistic below 0.2 represents 'very small' effect, 0.2 to 0.5 – 'small', 0.5 to 0.8 – 'medium', and a ES score higher than 0.8 represents a 'large' ES (377). But this should be interpreted together with the expected ES/SRM as categorised in each of the change group. A measure is judged to be more responsive when it meets the expectation of the assignment of the change group to a larger degree than the other measure. Please see Section 7.3.5.3 for details.

7.4.5.2 Anchor by H&Y

Five change groups were formed based on the change of H&Y over the two years (Table 7-12). There were large numbers of patients assigned to 'no change' group (median of sample size = 407) or the 'slightly deteriorated' group (median of sample size = 400), while only roughly[9] five patients were in the 'largely improved' group.

[9] 'roughly' is because sample size may be different for each imputed dataset.

Table 7-12: Change group formation: H&Y in each change group anchored by H&Y

Stata wide form,(yr2,yr4) Anchor H&Y	Sample size		H&Y at year 2		H&Y at year 4	
	Median	Average	Mean	SE	Mean	SE
Largely Improved group	5	5.3	3.174	0.406	1.593	0.325
Slightly improved group	109.5	110.9	2.828	0.085	2.149	0.074
No change group	407	406.9	2.725	0.042	2.725	0.042
Slightly deteriorated	399	399.9	2.222	0.040	2.985	0.045
Largely deteriorated group	100.5	100.0	2.043	0.079	3.908	0.095

Table 7-13 shows the change score of the ICECAP-O and EQ-5D-3L over two years in the five change groups. In contrast to anchoring by PDQ-39-SI, with which the two test measures were changing in the same direction as the anchor, the ICECAP-O and EQ-5D-3L both showed a decreased QoL in the improved groups defined by H&Y, being -0.085 (-0.268, 0.098) and -0.146 (-0.377, 0.085) respectively for the 'largely improved' group, and -0.026 (-0.056, 0.004) and -0.025 (0.078, 0.029) respectively for the 'slightly improved' group. For the 'no change' and 'deteriorated' groups, both the EQ-5D-3L and the ICECAP-O showed a statistically significant deterioration with an expected increasing size of mean change from 'no change' group to the 'largely deteriorated' group.

Table 7-13 shows the ES and SRM of the ICECAP-O and EQ-5D-3L in the five change groups. The mean SRM of EQ-5D-3L was larger than that of the ICECAP-O (-0.857 and -0.800) in the largely deteriorated group but overall the mean of the ES and SRM for the two measures were comparable with an overlapping confidence intervals (Figure 7-9).

Table 7-13: Change score of the ICECAP-O and EQ-5D-3L over two years (anchored by H&Y)

Stata wide form,(yr2,yr4) Anchor H&Y	Change of ICECAP-O				Change of EQ-5D-3L score			
	Mean	SE	95% CI	p value	Mean	SE	95% CI	p value
Largely Improved group	-0.085	0.093	-0.268,0.098	0.3666	-0.146	0.117	-0.377,0.085	0.2139
Slightly improved group	-0.026	0.015	-0.056,0.004	0.0845	-0.025	0.027	-0.078,0.029	0.3727
No change group	-0.038	0.008	-0.053,-0.022	0.0000	-0.052	0.013	-0.078,-0.025	0.0001
Slightly deteriorated	-0.064	0.009	-0.082,-0.045	0.0000	-0.114	0.016	-0.145,-0.083	0.0000
Largely deteriorated group	-0.125	0.026	-0.177,-0.072	0.0000	-0.264	0.040	-0.343,-0.184	0.0000

Table 7-14: ES and SRM statistics of ICECAP-O and EQ-5D-3L with the anchor H&Y

	Effect size (ES)						Standardise response mean (SRM)					
Stata wide form (yr2,yr4) Anchor: PDQ-39-SI	ICECAP-O			EQ-5D-3L			ICECAP-O			EQ-5D-3L		
	Mean	SE	95% CI	Mean	SE	95% CI	Mean	SE	95% CI	Mean	SE	95% CI
Largely Improved group	-1.007	1.644	-4.327,2.118	-0.733	0.313	-1.332,-0.104	-0.549	0.272	-1.025,0.042	-1.149	0.483	-2.107,-0.215
Slightly improved group	-0.168	0.071	-0.299,-0.021	-0.113	0.078	-0.239,0.067	-0.183	0.077	-0.34,-0.038	-0.125	0.083	-0.258,0.067
No change group	-0.222	0.037	-0.300,-0.156	-0.166	0.042	-0.254,-0.09	-0.262	0.041	-0.376,-0.214	-0.203	0.047	-0.304,-0.121
Slightly deteriorated	-0.403	0.049	-0.469,-0.275	-0.421	0.050	-0.513,-0.317	-0.418	0.040	-0.474,-0.315	-0.415	0.044	-0.498,-0.325
Largely deteriorated group	-0.686	0.147	-1.034,-0.457	-0.992	0.147	-1.317,-0.739	-0.800	0.093	-0.883,-0.518	-0.857	0.084	-1.04,-0.712

Note: an effective size statistic below 0.2 represents ‘very small’ effect, 0.2 to 0.5 – ‘small’, 0.5 to 0.8 – ‘medium’, and a ES score higher than 0.8 represents a ‘large’ ES (377). But this should be interpreted together with the expected ES/SRM as categorised in each of the change group. A measure is judged to be more responsive when it meets the expectation of the assignment of the change group to a larger degree than the other measure. Please see Section 7.3.5.3 for details.

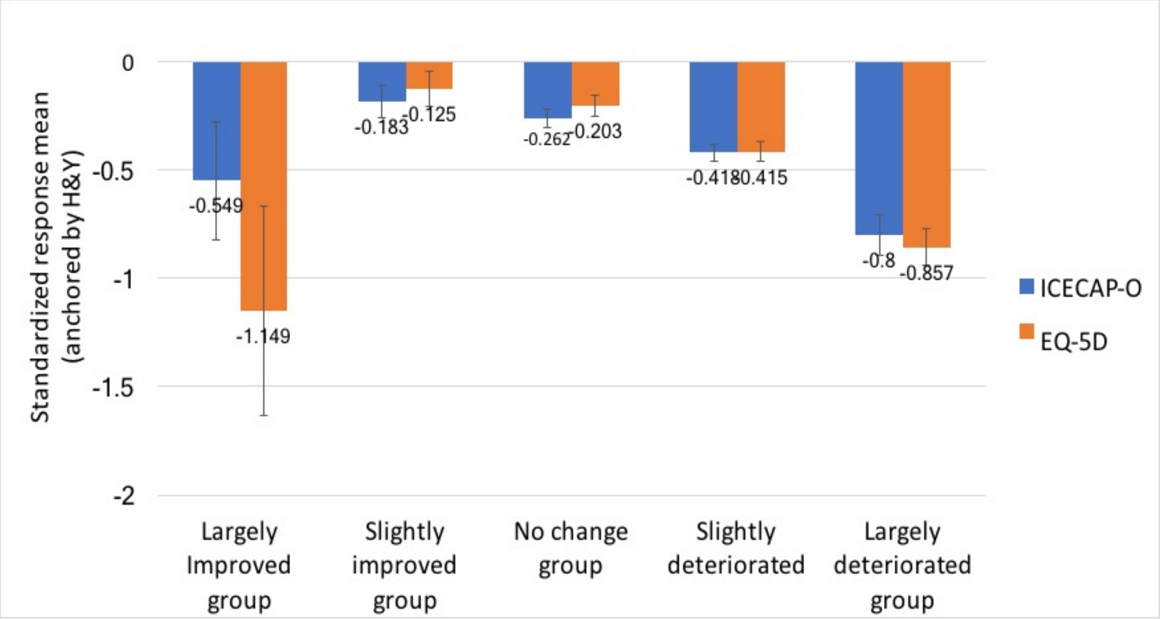


Figure 7-9: SRM (SE) of ICECAP-O and EQ-5D-3L in the five change groups anchored by H&Y

Note: an effective size statistic below 0.2 represents 'very small' effect, 0.2 to 0.5 – 'small', 0.5 to 0.8 – 'medium', and a ES score higher than 0.8 represents a 'large' ES (377). But this should be interpreted together with the expected ES/SRM as categorised in each of the change group. A measure is judged to be more responsive when it meets the expectation of the assignment of the change group to a larger degree than the other measure. Please see Section 7.3.5.3 for details.

7.4.5.3 Anchor by PDQ-39 sub-dimensions

Five change groups were formed based on the change score of each of the PDQ-39 eight dimensions with the complete data at year 2 and year 4 (Table 7-15). The distribution of the sample size and the mean change in the five change groups varied across the sub-dimensions due to the different sizes of the MID. The larger size of the MID, the more likely that a patient was grouped into a no change or small change group but not a large change group. Among all the dimensions, the sample size for 'no change' group anchored by the social support dimension was the largest (n=459) since it has the largest MID, i.e., 11.4 points (533). For the dimension cognition and bodily discomfort, there were no observations / patients in the small change groups due to the small MID and the relatively small number of questions within the dimension. The possible scores for dimensions are discrete due to the limited number of questions and levels and there may be few scores that falls between one MID and 5 * MID, resulting in no observations in the small change group.

Table 7-15: Change groups formed by PDQ-39 dimensions: PDQ-39 dimension score in each change group anchored by that dimension

Groups defined by the change of anchor measure (in bold) (MID in bracket)	**Sample size**		**Change of the anchor measure**		
	ICECAP-O	**EQ-5D-3L**	**Mean**	**SE**	**95% CI**
Mobility (3.2)					
Largely Improved group	82	88	-16.733	0.764	-18.232, -15.234
Slightly improved group	44	43	-5.795	0.178	-6.144, -5.447
No change group	167	183	0.189	0.133	-0.073, 0.451
Slightly deteriorated	96	98	6.262	0.125	6.017, 6.508
Largely deteriorated group	339	375	24.804	0.760	23.313, 26.295
Activities of daily living (4.4)					
Largely Improved group	62	67	-22.699	0.818	-24.305, -21.093
Slightly improved group	63	66	-10.075	0.253	-10.571, -9.578
No change group	231	248	0.753	0.209	0.343, 1.163
Slightly deteriorated	131	134	10.205	0.177	9.858, 10.553
Largely deteriorated group	205	231	27.849	0.813	26.253, 29.445
Emotional wellbeing (4.2)					
Largely Improved group	66	70	-21.488	0.876	-23.207, -19.769
Slightly improved group	92	98	-10.333	0.209	-10.744, -9.923
No change group	268	286	0.345	0.185	-0.019, 0.709
Slightly deteriorated	111	119	10.434	0.190	10.061, 10.807
Largely deteriorated group	132	149	29.276	1.214	26.892, 31.661
Stigma (5.6)					
Largely Improved group	82	85	-28.338	1.138	-30.572, -26.104
Slightly improved group	124	136	-8.560	0.258	-9.066, -8.054
No change group	231	240	0.000	0.000	--
Slightly deteriorated	161	179	8.924	0.231	8.47, 9.377
Largely deteriorated group	97	111	30.580	1.274	28.079, 33.082
Social support (11.4)					
Largely Improved group	5	7	-46.429	3.454	-53.209, -39.648
Slightly improved group	70	69	-18.403	0.599	-19.578, -17.228
No change group	459	493	0.434	0.182	0.076, 0.792
Slightly deteriorated	118	127	20.410	0.554	19.322, 21.498
Largely deteriorated group	20	24	46.833	1.527	43.836, 49.831
Cognition (1.8)					
Largely Improved group	214	222	-12.416	0.516	-13.43, -11.403
Slightly improved group	0	0	--	--	--
No change group	114	125	0.000	0.000	--
Slightly deteriorated	0	0	--	--	--
Largely deteriorated group	368	400	16.441	0.571	15.319, 17.563
Communication (4.2)					
Largely Improved group	71	75	-22.478	0.987	-24.417, -20.54
Slightly improved group	87	95	-8.333	0.000	-8.333, -8.333
No change group	226	236	0.000	0.000	--
Slightly deteriorated	121	131	8.333	0.000	8.333, 8.333
Largely deteriorated group	189	212	25.845	0.768	24.338, 27.352
Bodily discomfort (2.1)					
Largely Improved group	213	229	-18.678	0.762	-20.175, -17.182
Slightly improved group	0	0	--	--	--
No change group	152	165	0	0	-
Slightly deteriorated	0	0	--	--	--
Largely deteriorated group	327	351	20.282	0.686	18.935, 21.628

Note: complete case analysis. Change was from year 2 to year 4. 3*MID is used as the cut-off between slight change and large change.

Overall, the responsiveness of the ICECAP-O and EQ-5D-3L to the eight dimensions were similar, nonetheless, there were some differences to a small degree in responsiveness between the two measures to all the dimensions except for cognition and communication (Table 7-16). The trend of the SRM and ES were similar to that when anchored by other measures - the statistics decreased from the largely improved group to the largely deteriorated group.

For mobility, stigma and social support, the ICECAP-O met the expectations of the assignment of the change group to a slightly larger degree than the EQ-5D-3L, indicated by a larger size of SRM in the change groups with an expected sign (positive mean change in the improved groups, and negative mean change in the deteriorated groups). Likewise, EQ-5D-3L met the expectations to a slightly larger degree with the dimension ADL, emotional wellbeing, and bodily discomfort.

Due to the overall deterioration trend, investigation was focused on the mean change and SRM in the improved groups. The larger size with a positive sign of the SRM for a measure, indicates the more consistent a measure with the anchor. Across the anchors of mobility, ADL and emotional wellbeing, both ICECAP-O and EQ-5D-3L had an increased mean, although not statistically significant, in the largely improved group. The SRM of ICECAP-O in the largely improved group anchored by social support is 1.041, yet this is more likely by chance given the sample size is very small (n=5).

Table 7-16: ES and SRM statistics of ICECAP-O and EQ-5D-3L, anchored by the PDQ-39 eight dimensions

Groups defined by the change of anchor (MID in bracket)	**Change of ICECAP-O score**							**Change of EQ-5D-3L score**							**Which more responsive?[a]**
	N	**Mean**	**SE**	**95% CI**	**p value**	**ES**	**SRM**	**N**	**Mean**	**SE**	**95% CI**	**p value**	**ES**	**SRM**	
Mobility (3.2)															
Largely Improved group	82	0.011	0.011	-0.010,0.033	0.2988	0.090	**0.116**	88	0.016	0.028	-0.038,0.070	0.5775	0.067	0.061	
Slightly improved group	44	-0.007	0.019	-0.045,0.031	0.7223	-0.047	-0.056	43	0.002	0.039	-0.076,0.079	0.9715	0.006	0.006	
No change group	167	-0.024	0.009	-0.042,-0.005	0.0110	-0.124	-0.197	183	-0.019	0.014	-0.046,0.009	0.1792	-0.053	-0.100	**ICECAP-O**
Slightly deteriorated	96	-0.032	0.013	-0.057,-0.007	0.0116	-0.230	-0.258	98	-0.037	0.022	-0.079,0.006	0.0878	-0.135	-0.172	
Largely deteriorated group	339	-0.089	0.008	-0.105,-0.073	0.0000	-0.669	**-0.602**	375	-0.154	0.014	-0.181,-0.126	0.0000	-0.650	-0.568	
Activities of daily living (4.4)															
Largely Improved group	62	0.016	0.016	-0.015,0.047	0.3208	0.094	0.127	67	0.040	0.027	-0.013,0.094	0.1404	0.158	**0.180**	
Slightly improved group	63	-0.004	0.013	-0.028,0.021	0.7901	-0.032	-0.036	66	-0.029	0.020	-0.069,0.011	0.1528	-0.108	-0.176	
No change group	231	-0.018	0.007	-0.030,-0.005	0.0076	-0.121	-0.176	248	-0.009	0.013	-0.035,0.016	0.4788	-0.031	**-0.046**	EQ-5D-3L
Slightly deteriorated	131	-0.047	0.009	-0.064,-0.029	0.0000	-0.362	-0.460	134	-0.064	0.019	-0.101,-0.027	0.0007	-0.241	-0.293	
Largely deteriorated group	205	-0.093	0.011	-0.114,-0.072	0.0000	-0.661	-0.608	231	-0.181	0.019	-0.218,-0.144	0.0000	-0.767	**-0.630**	
Emotional wellbeing (4.2)															
Largely Improved group	66	0.015	0.017	-0.018,0.048	0.3701	0.097	0.112	70	0.048	0.031	-0.013,0.109	0.1255	0.171	**0.183**	
Slightly improved group	92	-0.015	0.011	-0.037,0.006	0.1664	-0.116	-0.145	98	-0.042	0.023	-0.087,0.003	0.0673	-0.165	-0.185	
No change group	268	-0.031	0.006	-0.042,-0.019	0.0000	-0.238	-0.318	286	-0.038	0.012	-0.063,-0.014	0.0018	-0.147	-0.185	EQ-5D-3L
Slightly deteriorated	111	-0.043	0.010	-0.063,-0.023	0.0000	-0.307	-0.400	119	-0.081	0.019	-0.119,-0.044	0.0000	-0.277	-0.393	
Largely deteriorated group	132	-0.105	0.015	-0.134,-0.076	0.0000	-0.666	-0.619	149	-0.189	0.024	-0.236,-0.143	0.0000	-0.752	**-0.653**	
Stigma (5.6)															
Largely Improved group	82	-0.021	0.017	-0.053,0.012	0.2145	-0.144	-0.138	85	-0.024	0.025	-0.073,0.025	0.3512	-0.074	-0.102	
Slightly improved group	124	-0.016	0.011	-0.038,0.005	0.1363	-0.114	-0.134	136	-0.049	0.020	-0.087,-0.010	0.0128	-0.208	-0.213	
No change group	231	-0.036	0.006	-0.049,-0.024	0.0000	-0.292	-0.369	240	-0.043	0.014	-0.070,-0.016	0.0021	-0.174	-0.199	ICECAP-O
Slightly deteriorated	161	-0.035	0.009	-0.052,-0.018	0.0001	-0.279	-0.311	179	-0.086	0.020	-0.125,-0.047	0.0000	-0.326	-0.321	
Largely deteriorated group	97	-0.114	0.016	-0.145,-0.083	0.0000	-0.665	**-0.738**	111	-0.144	0.027	-0.198,-0.09	0.0000	-0.482	-0.496	
Social support (11.4)															
Largely Improved group	5	**0.080**	0.034	0.013,0.147	**0.0197**	0.564	**1.041**	7	-0.049	0.195	-0.433,0.335	0.8138	-0.130	-0.095	ICECAP-O

Groups defined by the change of anchor (MID in bracket)	Change of ICECAP-O score							Change of EQ-5D-3L score							Which more responsive?[a]
	N	Mean	SE	95% CI	p value	ES	SRM	N	Mean	SE	95% CI	p value	ES	SRM	
Slightly improved group	70	-0.017	0.018	-0.053,0.019	0.3551	-0.112	-0.112	69	-0.060	0.030	-0.119,-0.002	0.0409	-0.225	-0.246	
No change group	459	-0.027	0.005	-0.036,-0.017	0.0000	-0.208	-0.253	493	-0.057	0.010	-0.076,-0.038	0.0000	-0.226	-0.266	
Slightly deteriorated	118	-0.074	0.012	-0.098,-0.050	0.0000	-0.579	-0.557	127	-0.085	0.026	-0.135,-0.034	0.0011	-0.314	-0.291	
Largely deteriorated group	20	-0.169	0.048	-0.264,-0.074	0.0005	-0.756	**-0.783**	24	-0.246	0.070	-0.384,-0.108	0.0005	-0.772	-0.714	
Cognition (1.8)															
Largely Improved group	214	-0.015	0.009	-0.032,0.003	0.1057	-0.102	-0.111	222	-0.013	0.015	-0.043,0.017	0.4012	-0.045	-0.057	ICECAP-O/EQ-5D-3L
Slightly improved group	0	--	--	--	--	--	--	0	--	--	--	--	--	--	
No change group	114	-0.029	0.008	-0.045,-0.013	0.0003	-0.258	-0.339	125	-0.027	0.017	-0.060,0.005	0.0984	-0.110	-0.148	
Slightly deteriorated	0	--	--	--	--	--	--	0	--	--	--	--	--	--	
Largely deteriorated group	368	-0.061	0.007	-0.074,-0.047	0.0000	-0.413	-0.465	400	-0.110	0.013	-0.136,-0.084	0.0000	-0.420	-0.411	
Communication (4.2)															
Largely Improved group	71	-0.005	0.017	-0.038,0.028	0.7918	-0.030	-0.033	75	-0.010	0.031	-0.070,0.050	0.7579	-0.032	-0.038	ICECAP-O/EQ-5D-3L
Slightly improved group	87	-0.025	0.014	-0.052,0.002	0.0689	-0.160	-0.195	95	-0.031	0.026	-0.082,0.020	0.2292	-0.127	-0.124	
No change group	226	-0.027	0.006	-0.039,-0.015	0.0000	-0.238	-0.286	236	-0.025	0.013	-0.051,0.002	0.0657	-0.091	-0.120	
Slightly deteriorated	121	-0.027	0.010	-0.047,-0.006	0.0103	-0.221	-0.233	131	-0.060	0.018	-0.095,-0.024	0.0010	-0.257	-0.288	
Largely deteriorated group	189	-0.088	0.011	-0.109,-0.067	0.0000	-0.576	-0.602	212	-0.159	0.019	-0.196,-0.122	0.0000	-0.582	-0.573	
Bodily discomfort (2.1)															
Largely Improved group	213	-0.037	0.008	-0.053,-0.021	0.0000	-0.257	-0.311	229	-0.036	0.017	-0.069,-0.002	0.0362	-0.130	-**0.138**	EQ-5D-3L
Slightly improved group	0	--	--	--	--	--	--	0	--	--	--	--	--	--	
No change group	152	-0.030	0.009	-0.048,-0.012	0.0010	-0.211	-0.268	165	-0.030	0.017	-0.063,0.003	0.0725	-0.098	-0.140	
Slightly deteriorated	0	--	--	--	--	--	--	0	--	--	--	--	--	--	
Largely deteriorated group	327	-0.049	0.008	-0.064,-0.034	0.0000	-0.354	-0.360	351	-0.108	0.013	-0.133,-0.082	0.0000	-0.431	-**0.441**	

Note: a. an effective size statistic below 0.2 represents 'very small' effect, 0.2 to 0.5 – 'small', 0.5 to 0.8 – 'medium', and a ES score higher than 0.8 represents a 'large' ES (377). A measure is judged to be more responsive when it meets the expectation of the assignment of the change group to a larger degree than the other measure. Please see Section 7.3.5.3 for details.

7.4.6 Regression analysis

Four PDQ-39 dimensions showed statistical significance when predicting the change of ICECAP-O scores (Table 7-17). They are: (in the order of the size of coefficient) change in social support (-0.00168, p<0.001), change in emotional wellbeing (-0.00153, p=0.001), change in mobility (-0.00135, p<0.001), and the change in ADL (-0.00124, p=0.002). This result is slightly different from the complete case analysis, in which the determinants that showed statistical significance were only two dimensions: PDQ-39 social support (-0.00245, p=0.006), and PDQ-39 emotional (-0.00223, p=0.016) (result presented in Appendix F). The size of coefficient became smaller after imputation. The p value also became smaller which may be due to the substantial increase of sample size. The median of the adjusted R squared among the imputed datasets was 0.383 (range: 0.340, 0.421).

Table 7-17: Regression analysis to predict the change of ICECAP-O index score from the change of the PDQ-39 eight dimensions

Change of PDQ-39 dimensions	Coefficient	SE.	P	95% CI
Mobility	**-0.00135**	0.00032	**0.000**	-0.00198, -0.00072
ADL	**-0.00124**	0.00038	**0.002**	-0.00200, -0.00047
Emotional wellbeing	**-0.00153**	0.00043	**0.001**	-0.00239, -0.00068
Stigma	-0.00059	0.00031	0.058	-0.00121, 0.00002
Social support	**-0.00168**	0.00039	**0.000**	-0.00245, -0.00090
Cognition	-0.00020	0.00035	0.557	-0.00089, 0.00048
Communication	-0.00021	0.00035	0.547	-0.00092, 0.00049
Bodily discomfort	0.00034	0.00028	0.222	-0.00021, 0.00089
Constant	-0.02614	0.00761	0.001	-0.04113, -0.01115

Table 7-18 showed that there were four dimensions that showed a statistically significant difference when predicting the change of EQ-5D-3L: (in the order of the size of coefficient) change in ADL (-0.00309, p<0.001), change in emotional (-0.00276, p<0.001), change in mobility (-0.00273, p<0.001), and change in bodily discomfort (-0.00187, p<0.001) (Table 7-18). The median of the adjusted R squared was 0.409 (range: 0.382, 0.447).

The result was also slight different from the complete case analysis, in which the variables that showed statistical significance were emotional wellbeing (-0.00364, p=0.001), mobility (-0,00247, p=0.009), and ADL (-0.00197, p=0.036). The size of coefficient of ADL increased after the imputation, while the size of coefficient of emotional wellbeing decreased.

Table 7-18: Regression analysis to predict the change of EQ-5D-3L index score from the change of the PDQ-39 eight dimensions

Change of PDQ-39 dimensions	Coefficient	SE	P	95% CI
Mobility	**-0.00273**	0.00052	**0.000**	-0.00375, -0.00172
ADL	**-0.00309**	0.00060	**0.000**	-0.00428, -0.00190
Emotional wellbeing	**-0.00276**	0.00073	**0.000**	-0.00422, -0.00131
Stigma	-0.00044	0.00054	0.412	-0.00151, 0.00062
Social support	-0.00013	0.00070	0.855	-0.00153, 0.00127
Cognition	0.00039	0.00064	0.545	-0.00089, 0.00167
Communication	0.00031	0.00051	0.543	-0.00069, 0.00131
Bodily discomfort	**-0.00187**	0.00043	**0.000**	-0.00273, -0.00101
Constant	-0.03168	0.01317	0.017	-0.05762, -0.00573

7.4.7 Sensitivity analysis

Four missing data handling strategies including two imputation strategies and two complete case analyses were compared with the primary strategy and the results were summarized in Table 7-19. Except for imputation with long form, the effect size results using other strategies all agreed that the ICECAP-O was more responsive to the change of PDQ-39-SI than the EQ-5D-3L. When classifying the participants to three change groups (no change, improved/deteriorated group), the result agreed with the five groups that ICECAP-O is slightly more responsive to the change of the PDQ-39-SI than the EQ-5D-3L (Table 7-20). The conclusion remains the same when the cut-off between the largely change group and the small change group changed to 3 * MID (Table 7-20).

Table 7-19: Sensitivity analysis - comparing the impact of different imputation strategies to the result of responsiveness

Groups defined by the change of PDQ-39-SI	Sample size*		Change of ICECAP-O score					Change of EQ-5D score					Which more responsive? (marginal)
	Average		Mean	SE	p value	ES	SRM	Mean	SE	p value	ES	SRM	
Stata long form, (yr1, yr5)													
Largely Improved group	125.2		0.046	0.020	0.0188	0.204	0.209	0.126	0.030	0.0000	0.353	0.373	EQ-5D/ ICECAP-O
Slighly improved group	83.1		-0.004	0.019	0.8477	-0.020	-0.023	-0.017	0.038	0.6769	-0.046	-0.048	
No change group	73.4		-0.045	0.021	0.0374	-0.248	-0.243	0.012	0.039	0.7727	0.034	0.036	
Slightly deteriorated	141.4		-0.063	0.015	0.0000	-0.359	-0.355	-0.101	0.028	0.0003	-0.306	-0.305	
Largely deteriorated group	509.9		-0.179	0.010	0.0000	-1.287	-0.798	-0.312	0.018	0.0000	-1.148	-0.783	
Realcom Impute, wide form, (yr1, yr5)													
Largely Improved group	36.6		0.067	0.029	0.0225	0.285	0.380	0.023	0.042	0.5991	0.075	0.091	ICECAP-O
Slighly improved group	70.2		0.010	0.017	0.5793	0.047	0.068	-0.045	0.041	0.2742	-0.114	-0.131	
No change group	80.1		-0.027	0.017	0.1128	-0.148	-0.178	-0.004	0.035	0.9104	-0.012	-0.014	
Slightly deteriorated	207.7		-0.052	0.010	0.0000	-0.272	-0.350	-0.106	0.021	0.0000	-0.281	-0.348	
Largely deteriorated group	538.5		-0.154	0.008	0.0000	-0.830	-0.785	-0.248	0.015	0.0000	-0.854	-0.719	
Stata wide form, (yr2, yr4)													
Largely Improved group	76.8		0.055	0.018	0.0017	0.322	0.455	0.084	0.032	0.0086	0.316	0.387	ICECAP-O
Slighly improved group	143.1		0.011	0.011	0.3465	0.062	0.102	0.007	0.023	0.7830	0.054	0.070	
No change group	123.3		-0.015	0.012	0.2060	-0.086	-0.120	-0.011	0.022	0.6355	-0.019	-0.026	
Slightly deteriorated	271.5		-0.041	0.008	0.0000	-0.259	-0.354	-0.061	0.013	0.0000	-0.231	-0.313	
Largely deteriorated group	408.3		-0.124	0.010	0.0000	-0.766	-0.817	-0.211	0.017	0.0000	-0.808	-0.743	
Complete case, (yr1, yr5)	**ICECAP**	**EQ-5D**											
Largely Improved group	5	17	0.035	0.059	0.5604	0.432	0.268	0.035	0.060	0.5693	0.187	0.142	ICECAP-O
Slighly improved group	10	23	0.021	0.018	0.2422	0.130	0.372	0.011	0.051	0.8440	0.032	0.044	
No change group	18	37	-0.012	0.024	0.6207	-0.119	-0.121	0.052	0.036	0.1455	0.178	0.240	
Slightly deteriorated	37	61	-0.070	0.021	0.0010	-0.615	-0.544	-0.113	0.027	0.0000	-0.400	-0.532	
Largely deteriorated group	54	128	-0.142	0.024	0.0000	-1.389	-0.806	-0.216	0.026	0.0000	-0.894	-0.732	
Complete case, (yr2, yr4)	**ICECAP**	**EQ-5D**											
Largely Improved group	41	41	0.034	0.018	0.0619	0.245	0.291	0.041	0.031	0.1830	0.153	0.208	ICECAP-O
Slighly improved group	85	91	0.010	0.011	0.3910	0.075	0.094	-0.001	0.022	0.9756	-0.003	-0.004	
No change group	80	87	-0.010	0.011	0.3707	-0.074	-0.101	-0.004	0.019	0.8353	-0.015	-0.024	
Slightly deteriorated	185	187	-0.039	0.007	0.0000	-0.307	-0.400	-0.048	0.012	0.0001	-0.203	-0.297	
Largely deteriorated group	185	209	-0.087	0.011	0.0000	-0.598	-0.603	-0.162	0.020	0.0000	-0.622	-0.565	

Table 7-20: Sensitivity analysis - results of three change groups (as versus five change groups in the base case) and 3*MID (as versus 5*MID in the base case)

Groups defined by the change of PDQ-39-SI	Sample size*	Change of ICECAP-O score					Change of EQ-5D score					Which more responsive? (marginal)
	Average	Mean	SE	p value	ES	SRM	Mean	SE	p value	ES	SRM	
Three change groups												
Largely Improved group	219.9	0.026	0.010	0.0064	0.154	0.220	0.034	0.018	0.0635	0.144	0.184	ICECAP-O
No change group	123.3	-0.015	0.012	0.2060	-0.086	-0.126	-0.011	0.022	0.6355	-0.019	-0.026	
Largely deteriorated group	679.8	-0.091	0.007	0.0000	-0.568	-0.631	-0.151	0.012	0.0000	-0.576	-0.578	
3 * MID												
Largely Improved group	149.4	0.039	0.013	0.0019	0.234	0.340	0.049	0.022	0.0241	0.201	0.251	ICECAP-O
Slightly improved group	70.5	-0.001	0.018	0.9782	-0.027	-0.037	0.001	0.034	0.9784	0.019	0.026	
No change group	123.3	-0.015	0.012	0.2060	-0.086	-0.126	-0.011	0.022	0.6355	-0.019	-0.026	
Slightly deteriorated	159.7	-0.034	0.011	0.0015	-0.223	-0.302	-0.053	0.018	0.0032	-0.214	-0.281	
Largely deteriorated group	520.1	-0.108	0.008	0.0000	-0.672	-0.729	-0.181	0.014	0.0000	-0.682	-0.662	

7.5 Summary of results

This chapter has, for the first time, empirically tested the responsiveness of ICECAP-O in comparison to EQ-5D-3L to the change of a variety of health, QoL and wellbeing aspects in people with Parkinson's. An overall deterioration of health status, HrQoL as well as wellbeing was found as expected in this cohort. The change of ICECAP-O index score and the change of EQ-5D-3L index score were moderately correlated and there was no substantial difference in the pattern shown in the scatter plots with PDQ-39-SI and in the responsiveness performance to all the aspects. Nevertheless, the ICECAP-O was slightly more responsive to the change in general health and QoL as measured by PDQ-39-SI, and in contrast, EQ-5D-3L was slightly more responsive to the change of motor symptoms as measured by the clinical scale H&Y staging. Results remain the same in the sensitivity analyses when varying the imputation strategies (except for imputation with long form), dividing to three change groups, and using 3 * MID as the cut-off between 'largely' and 'slightly' change for PDQ-39-SI.

For PDQ-39 dimensions, the ICECAP-O was shown to be marginally more responsive to mobility, stigma and social support dimensions than the EQ-5D-3L, but less responsive to ADL, emotional wellbeing, and bodily discomfort. Again, the difference in their responsiveness to the PDQ-39 dimensions was minimal. The spike at 0 in the histogram for EQ-5D-3L adds to the existing concern that its three levels may lack sensitivity to small changes (e.g. patients may stay in level 2, 'some problems' and the change was not big enough for patients to answer 'no problem' or 'a lot of problems').

The regression analysis showed that the change in mobility, ADL, and emotional dimensions could predict the change of both ICECAP-O and EQ-5D-3L, however, the size of change that these dimensions could predict was smaller in ICECAP-O than in EQ-5D-3L. Besides the shared predictors, it was found that the change of social support strongly predicted the change in ICECAP-O value, while the change of bodily discomfort predicted the change in EQ-5D-3L value.

As discussed in Chapter 5 (Section 5.4), a variety of methodological challenges were met in this assessment of responsiveness, including the selection of the optimal anchor, use of the MID method to form change groups anchored by PDQ-39-SI and its dimensions, addressing the missing data problem, and applying the statistical methods in the imputed datasets. Assumptions were made to tackle these problems throughout the methods in this case study. These challenges will be further discussed in Section 7.6.3.

7.6 Discussion

7.6.1 Interpretation of results

This study shows that the progression of Parkinson’s over time leads to a noticeable deterioration in health status, HrQoL and capability wellbeing. This is in line with the finding in last chapter that there is an important difference in all of these aspects between the Early and Later group. The result that ICECAP-O was more responsive than the EQ-5D to the change of PDQ-39 overall score while EQ-5D was more responsive than ICECAP-O to the change of H&Y is as expected, although the difference was small and did not cross the categories of the Cohen’s interpretation of the effect size statistics.

The results of PDQ-39 dimensions are mixed. The PDQ-39 dimensions are not mutually exclusive, for example, stigma and emotional wellbeing are conceptually related, however stigma is more sensitive to be measured by ICECAP-O while emotional wellbeing prefers EQ-5D-3L. Therefore these results should be treated with caution. One reason for the mixed results may be due to the fact that they are analysed using complete case analysis which risks selection bias. Also, both QoL and capability are multi-faceted concepts that have numerous determinants, while each of eight dimensions only represent one aspect and thus the ‘noise’ (the impact of other factors contributing to HrQoL or capability) is large. For these reasons the analysis anchored by the dimension scores can only play a supplementary role in concluding the research.

One potential reason to explain the smaller sizes of change coefficient of ICECAP-O than EQ-5D as predicted by mobility, ADL and emotional wellbeing dimensions is due to the difference in scoring range between the measures. EQ-5D-3L has wider score range than the ICECAP-O. Another contributing factor might be the deliberately broad nature of the ICECAP-O dimensions which could be affected by many determinants, thereby certain change in a few dimensions may not have as large an impact as the health-focused EQ-5D-3L.

Besides the shared predictors, this study found that the change of social support strongly predicted the change in ICECAP-O value, while the change of bodily discomfort predicted the change in EQ-5D-3L value. This meets the expectation, which could be explained by the similar questions in the ICECAP-O or EQ-5D-3L as those in the PDQ-39. The attachment dimension in ICECAP-O asks “are you able to feel the love and friendship?”, which is similar to the PDQ-39 social support dimension which asks “had problems with your close relationships?”, “lacked support in the ways you need from your spouse or partner?”, “lacked support in the ways you need from your family or close friends?”. In addition, the ‘attachment’ has the highest set of weights among the five dimensions of the ICECAP-O and thus determines more strongly the total capability score than the other dimensions, which adds to the explanation for the significance of the social support dimension which is similar to ‘attachment’ to predict the total capability score. For the EQ-5D-3L, the pain/discomfort dimension asks “have problems in pain/discomfort?”, which is similar to the PDQ-39 ‘bodily discomfort’, which asks ‘had painful muscle cramps or spasms?”, “had aches and pains in your joints or body?”, and “felt unpleasantly hot or cold?”. This explains the stronger relationship between these two dimensions in EQ-5D-3L and the PDQ-39.

7.6.2 Critique of previous studies assessing responsiveness of ICECAP-O

Three studies were identified from literature which, respectively, assessed responsiveness of ICECAP-O in frail older adults in the Netherlands (319), among older adults (aged ≥ 70) at risk of mobility impairment in Vancouver (546), and in a cohort of patients with a hip fracture in the UK (547). These studies generated mixed conclusions regarding the comparison of EQ-5D and ICECAP-O in each population. Two out of the three studies (546, 547), unfortunately, failed to interpret appropriately from their results and thus their conclusions should be treated with caution. As it is the interest of this chapter to critique the assessment of methods of responsiveness, this discussion subsection will provide a brief critique of the methods used in these studies.

The earliest study of the three, Parsons et al. (2014), measured EQ-5D, ICECAP-O, and the Oxford Hip Score (a hip specific measure, OHS) in the patients that had hip operation following a hip fracture at baseline (pre-operation), 4 weeks, 4 months post operatively (547). They concluded that ICECAP-O was not responsive to the change for patients recovering from hip fracture whereas EQ-5D could be used to measure outcome for patients recovering from hip fracture. This conclusion is based on two statistics: a. effect size as used in this chapter (change scores divided by SD of baseline scores); b. correlations between the change score of clinical measure OHS and the EQ-5D/ICECAP-O at each time point. The effect size of ICECAP-O at 4-month was found to be much smaller than that of EQ-5D, and the correlation between ICECAP-O and OHS at each assessment point was found to be smaller than the correlation between EQ-5D and OHS, based on which the conclusion was generated favouring the responsiveness of EQ-5D. However, there are two crucial methodological flaws of this study.

Firstly, for the effect size statistics, as mentioned in Chapter 3 (Section 3.6.3.1) and this chapter Section 7.3.5.3, effect size statistics were initially invented to detect the size of clinical change by an intervention, rather than determining the responsiveness of measures. The difference between the former and the latter is that in the case of former, the responsiveness of the measure is proved, based on which the responsive measure is used to test whether the intervention is effective or not. In contrast, in the case of latter, it would require the effectiveness of the intervention is proved and known, based on which we could then determine whether the measure is responsive.

This means that when the effectiveness of an intervention is unknown, comparison of the magnitude of the effect size statistic is futile to determining the responsiveness of measures (327, 415). In Parson et al.'s study, the effectiveness of hip operation on the change of capability wellbeing is unknown, and no assumptions were made regarding this expected change, therefore this study design essentially failed to measure responsiveness.

Second, for the correlation comparison, although the authors did not mention it, they used OHS as a surrogate to confirm the effectiveness of the intervention, thereby a high correlation of EQ-5D/ICECAP- with OHS was judged in this study as representing high responsiveness. However, according to this study, the OHS measure is a clinical measure “which quantify disability secondary to hip osteoarthrosis”, that gives an overall score for hip function from 0 to 48, where “0 indicates excellent hip function and 48 indicates very poor hip function.”. It is a typical clinical measure focusing specifically on hip function and thus a high correlation with OHS can only mean the construct of the measure is highly related with the hip function. In other words, the OHS was used as an ‘anchor’ in this study which, however, may not be an ideal anchor as it measures related but still quite different concepts from capability-wellbeing in general. This is similar to using H&Y clinical scale in this thesis which can only provide information regarding whether ICECAP-O is responsive to one aspect of clinical change, but not on ‘whether it is responsive to the important changes in the concept that the measure is constructed to measure’, as per the definition of responsiveness (see Section 3.5.2 for details)

The second study, by Davis et al. (2017), had a similar issue as it compared the magnitude of mean change and SD between the measures in a population that did not have any external criteria to confirm what the change was expected to be (546). This study followed 359 patients who had experienced a minimum of one minimal displacement non-syncopal fall and attended the Vancouver Falls Prevention Clinic in the past 12 months for one year. It compared the responsiveness of EQ-5D-3L and ICECAP-O in terms of their mean change and SD; the larger the mean change relative to SD, the more responsiveness of the measure. Based on this, it concluded that EQ-5D-3L was more responsive than the ICECAP-O. However, this conclusion was an incorrect interpretation of the results, because there was no external criteria to confirm what should be the real change for HrQoL and capability. External criteria and assumptions are key for testing responsiveness (327, 415); if in fact there is no real change for the patients over the year, then the measure that had a no mean change should be the most responsive one, and if in fact there is a small change for the patients over the year, then the measure that had a small mean change should be the most responsive, and so on. Therefore, it is incorrect to associate ‘more responsiveness’ to a higher correlation coefficient, without any external information regarding what the actual correlation is expected to be.

The last study, by van Leeuwen et al. (2015), compared responsiveness of EQ-5D-3L, ICECAP-O and ASCOT based on the correlation coefficient between each of these measures and with eight external measures (319). The methods used in this study were appropriate as these external measures covered a broad range of health, HrQoL and wellbeing aspects, which provided a comprehensive concept examination of the three preference-based measures. The external measures included health GRS (Global Rating Scale), ADL limitations, impact of physical limitations (SF-12 physical), impact of emotional influences (SF-12 mental), QoL GRS, mastery (through Pearlin Mastery Scale, reflects the extent to which a person perceives himself or herself to be in control of events and ongoing situations), and client-centeredness (Client-Centered Care Questionnaire, reflects the extent to which respondents feel recognized and respected by nurses and to which they experience autonomy with respect to the way in which care is delivered).

In addition, ex ante hypotheses were provided in van Leeuwen et al's study with reasons regarding the comparative strength of correlations between the measures, for example, hypothesis 3a stated that "ICECAP-O change scores are less strongly correlated to ADL limitations than the EQ-5D-3L change scores" (319). A correlation table was provided which listed all the coefficients between the change score of each of the three measures in addition to the summary results of the hypothesis. Among the eight broad aspects, EQ-5D-3L change score was shown to be correlated strongest with the impact of physical limitations (SF-12 physical, r=0.23), whereas ICECAP-O was shown to be correlated strongest with ADL limitations (r=0.26), ASCOT (See Section 2.6.4) (r=0.31), and impact of emotional influences (SF-12 mental, r=0.22). This result, as concluded by the authors, is in line with the findings of this thesis which support the adoption of ICECAP-O as outcome measures in economic evaluations of care interventions for older adults that have a broader aim than HrQoL.

7.6.3 Methodological considerations

As the fourth and fifth objectives of this chapter are related to the methods (i.e. (4) to explore how to adapt the psychometric methods for the assessment of responsiveness to the assessment of PbQoL measures; (5) to investigate the impact of missing data on the result of assessment of responsiveness), this section will therefore discuss the methodological considerations of the methods used in this chapter.

7.6.3.1 Choice of anchors

Ten anchors were chosen in this study to ensure the comprehensiveness of our assessments for a broadly defined QoL measure. However, the interpretation of the results with each anchor should be treated with caution since they are dependent on the underlying assumptions. As mentioned earlier, preference-based measures are designed to be responsive to the change of overall utility rather than a particular aspect of clinical change. Therefore, the justification for using the PDQ-39 dimensions and the motor symptom scale H&Y would be that the improvement shown on these anchor measures are expected to lead to improvement of the overall QoL utilities. This study showed that the ES and SRM results anchored by the PDQ-39 eight dimensions were with smaller size than that when anchored by PDQ-39-SI. This is as expected because the change in only one dimension may be too small to affect the overall QoL, or it may be offset by the change in other areas which may not be captured by that specific dimension, or the change in one dimension may not affect the overall preferences. For this exact reason, the PDQ-39-SI is judged to be the more suitable anchor compared to the others since it measures similar broad QoL and wellbeing concept and thus the closest to the gold standard when testing the psychometric properties of ICECAP-O in Parkinson's population.

7.6.3.2 The MID method

Despite this, PDQ-39 is still not perfect as it does not measure preferences - to what degree the patients would judge the change to be 'important' so that they would like to trade a larger amount of length of life or accept a higher risk of death for it. To minimize this concern, this study used the MID to form the change groups by PDQ-39-SI as by definition MID is the smallest change that the patient considers 'important' and the 'important' should implicitly incorporate patients' general perception about their health. Although 'important' is an inherently different concept from 'preference', this MID approach is considered the best approximation given the limited information. In the future, valuation of PDQ-39 will facilitate the type of study which would provide the preference information for PDQ-39. This will be further discussed in Chapter 8 Section 8.6.2.

In addition to the reason above, to facilitate the interpretation of 'preference', the MID method also presents a feasible and justified avenue to form groups based on the continuous scoring measure, PDQ-39. Chapter 3 (Section 3.6.2) outlined four approaches to forming the change groups, including the Global rating scale approach, using intervention groups with confirmed effectiveness, using anchor measures which contain ordinal distinctive objective measures or MID for a continuous anchor measure. As the PD MED trials do not contain any GRS, and its interventions have uncertainties in their comparative effectiveness, the anchor measure method was used to classify the groups. This avoids the recall bias of the GRS approach (See Section 3.6.2.1), and the uncertainty in the outcome of the interventions, although this anchor approach must meet the prerequisite of the required correlation strength between the test measure and the condition-specific anchor measure (548).

The grouping by PDQ-39-SI and its eight dimensions applies the MID method. Unlike the H&Y scale which can be easily used to divide the groups due to its discrete staging system, the PDQ-39 generates a continuous score ranging from 0, no impairment of QoL, to 100, worst state that every aspect of QoL is substantially affected (131). It would be arbitrary to simply put, say a one-point change, or five-point change, as the criteria for the 'change' of the health status. It is also inappropriate to apply the same amount of points change for all the dimension scores and the overall score as they would fundamentally mean different health states. For example, a five-point change in the mobility dimension score may indicate a two level change (5*4*10/100, the scoring system is explained previously in Chapter 5 Section 5.3.4) in one of the mobility questions, e.g. changing from 'occasionally' to 'often' for the question, 'needed someone else to accompany you when you went out'. Meanwhile, five-point change in the emotional wellbeing dimension may indicate roughly one level change (5*4*6/100) in one of the questions of the emotional dimension, e.g. changing from 'sometimes' to 'often' in the question 'felt isolated and lonely?'. The actual change in these two situations is different to some degree, which however, translate to the same amount (five-point) of point change in their dimension score. Therefore, to avoid misjudgement on the meaning of the size of the change, the published MIDs (533) in the PDQ-39-SI and each of its dimension scores were determined to be the best available evidence to define groups that had changed (improved or worsened) by equal to or greater than the MID.

Although the MID estimates come from a published source (533) as described earlier in this chapter (Section 7.3.3.2), the limitation of the original study should be noted. The MID was calculated from the mean difference in the 'about the same' and 'a little worse' group (533). An underlying assumption is that the MID gained for the improved group is the same from the MID in the deteriorated group. If this assumption is not valid, using a constant MID for the both the improved and deteriorated groups may cause bias. Walters and Brazier (2005) compared the MID generated from the 'somewhat better' group and 'somewhat worse' group (400) in eight disease areas. They showed that those who improved and those who deteriorated have different MID although these difference was not statistical significant for most disease areas.

In terms of EQ-5D, the difference was 0.089 (p=0.42) for leg ulcer, 0.12 (p=0.14) for early rheumatoid arthritis, 0.072 (p=0.57) for limb reconstruction, 0.099 (p=0.15) irritable bowel syndrome, 0.006 (p=0.93) for acute myocardial infarction, 0.020 (p=0.83) and 0.166 (p=0.05) respectively for the two studies with patients of chronic obstructive pulmonary disease. This not statistically significant difference could be either due to the homogeneous nature of the MID for the improved group and the deteriorated group, or due to the small sample size in most of the studies (<30 in most groups). The comparison of whose MID was higher, those who improved or those who deteriorated, is not consistent across diseases. The exceptions that have a statistically significantly different MID was SF-6D in patients with back pain, and the MID of EQ-5D-3L in patients with Osteoarthritis of the knee. For the former, MID from the patients that answered 'somewhat better' was 0.115 which became 0.035 in people who answered 'somewhat worse', p=0.02; for the latter, MID from those improved was 0.261, which became much smaller in those deteriorated, i.e., 0.014, p=0.001.

Furthermore, the magnitude of MID may depend on the average health status of the participants, in other words, the transferability of the MID results in other studies. In Peto's study, the mean age of sample (N=728) was 70.4 years and 58.9% were men, with an average of 8.6 years of diagnosis of Parkinson's (533). These characteristics are similar to the PD MED cohort used in this thesis, and therefore it was judged that the MID results from Peto's study should be able to generalized to this analysis with low risk of bias.

7.6.3.3 Missing data

Another challenge of this analysis comes from the large amount of missing data in this dataset. This phenomenon is not rare for patient self-reported data. Previous studies (549) reported that the problem of missing patient reported outcomes is common in clinical trials (550-552). This creates challenges for data analysis, and may mislead the result and compromise the interpretability and credibility of the findings (553-555).

There are various reasons for the missing data in the PD MED trials. First, with the progression of the disease, the physical symptoms may impair the patients' ability of holding a pen and completing the questionnaires. Parkinson's may also affect patients' cognitive function and thus the patients may find the questionnaires difficult to grasp or answer. Sometimes the onsite guidance would help clarify the questions but this was unable to achieve in this trial as all the questionnaires were sent through post. Missing data were therefore produced, item missing in particular where only some items within the questionnaires were missing. Besides, numerous reasons may cause the whole follow-up assessment to be missing, such as patients drop out, loss of contact, hospitalisation, severe deterioration of disease or death. The loss of follow-up seems inevitable in the PD MED trial as it has been on-going for 15 years and circumstances of patients may change.

Given the large amount of missing data in this dataset, simply analysing the participants that have complete data may lead to biased result as they may not be a proper representative sample of the whole group (518, 555, 556). This chapter tested the mechanism of missing data and the result showed that the missingness was associated with age, ICECAP-O attachment dimension, number of days treated in hospital, and whether the patient has dementia. This demonstrated that the complete-case analysis may be biased and necessitated the implementation of imputation method to the data.

Sensitivity analysis of this study compared different missing data handling strategies (Table 7-19). Although the complete case analysis generated the same conclusion regarding the comparative responsiveness between ICECAP-O and EQ-5D-3L, the magnitude of effect size of the complete case analysis was smaller for all change groups than the imputed datasets. The imputation almost doubled the sample size for all groups and made use of all of the information available for the analysis, which enhanced the confidence in the study results.

In the future, measures should be undertaken to improve the completeness of the questionnaire (549). On-site assistance could be provided if the patients may have problems in the interpretation of the questions . But this should be treated with caveats to avoid leaving the answer to be 'contaminated' by the varied perception of the question of staff for a patient-self-reported questionnaire. Computerized questionnaire distribution would also prevent the item-non-response by providing notices, explanations to the question, and completeness checking. It would also prevent whole-wave missingness when the participants change contact address, or the correspondence lost in the mail system.

7.6.3.4 Compromised multiple imputation strategy

MI was used to impute the missing data since the MCAR assumption was not met (535). It is recommended to use MI in wide form in STATA for the data with panel structure with repeated measurements (537). Incorporation of auxiliary variables which influence the probability of missing values was also found to improve the imputation model (537). MI in wide form has the advantage of handling the correlation between the data from each wave for the same patient in the imputation model. On the other hand, it requires computational power as it multiples the number of the variables in the imputation model. The original model specification was comprised of 19 variables with missing data to impute, plus four baseline variables without missing data (age, sex, H&Y stage at baseline, and years since diagnosis of Parkinson's). The 19 variables included the five dimensions of ICECAP-O, five dimensions of EQ-5D, eight dimension of PDQ-39, H&Y staging, and whether the patient has dimension. After transforming the five waves to wide form, there were a total of 95 variables to impute in the imputation model. The imputation was initially carried out on the dimension level rather than the overall score level because this analysis need correlations between the change score of each dimensions and also dimension scores were used as anchors for ES statistics. However, despite its theoretical appropriateness, the imputation model cannot be fulfilled by STATA as it cannot converge due to the large proportion of missing data.

Several compromise measures were then undertaken after the failure of the optimal model. First, effort was taken by adding a small group of variables in the model gradually and it was found that the model seemed to be reaching a 'saturation' state when the number of variables to impute increased to around 20 and adding any extra variable will initiate the convergence failure. Therefore, the number of variables to impute in the model had to be reduced. After many attempts, the time horizon was adjusted to two years (wave 2 to wave 4) from the original four years (wave 1 to wave 5) due to the lower proportion of missing data at wave 2-4. The number of variables to impute was also reduced from 19 to 4 by imputing the summary score of the QoL measures rather than the dimensions. In this way, an imputation in wide form for wave 2 and 4 was applied in the primary analysis.

7.6.3.5 The effect size statistics

All the effect size methods are based on mean and SDs, which has an implicit assumption that the data follows a normal distribution. However, our results showed that the distribution of the change of the PDQ-39 and the EQ-5D did not meet the normal distribution assumption. In this case, Fayer (2007) suggests that the medians and interquartile ranges may replace means and SDs however little work has been carried out in this area (410).

Calculation of the effect size formula requires SD and mean change, however the SD cannot be calculated directly with STATA post estimation command after imputation. The SE (standard error) of the pooled mean result was provided in the result, which is the SD of the sampling distribution. SE describes how much the sample mean will vary from the mean for the whole population (557). A large SE means a wider sampling distribution and the mean from one sample may have a higher chance to be different from the whole population. The SE can be calculated by dividing the SD with the square root of the sample size (SE= SD/ square root(sample size)), and therefore, when the SE is known from the MI output using Rubin's rule, the SD can be calculated by multiplying the SE with square root of the sample size (557).

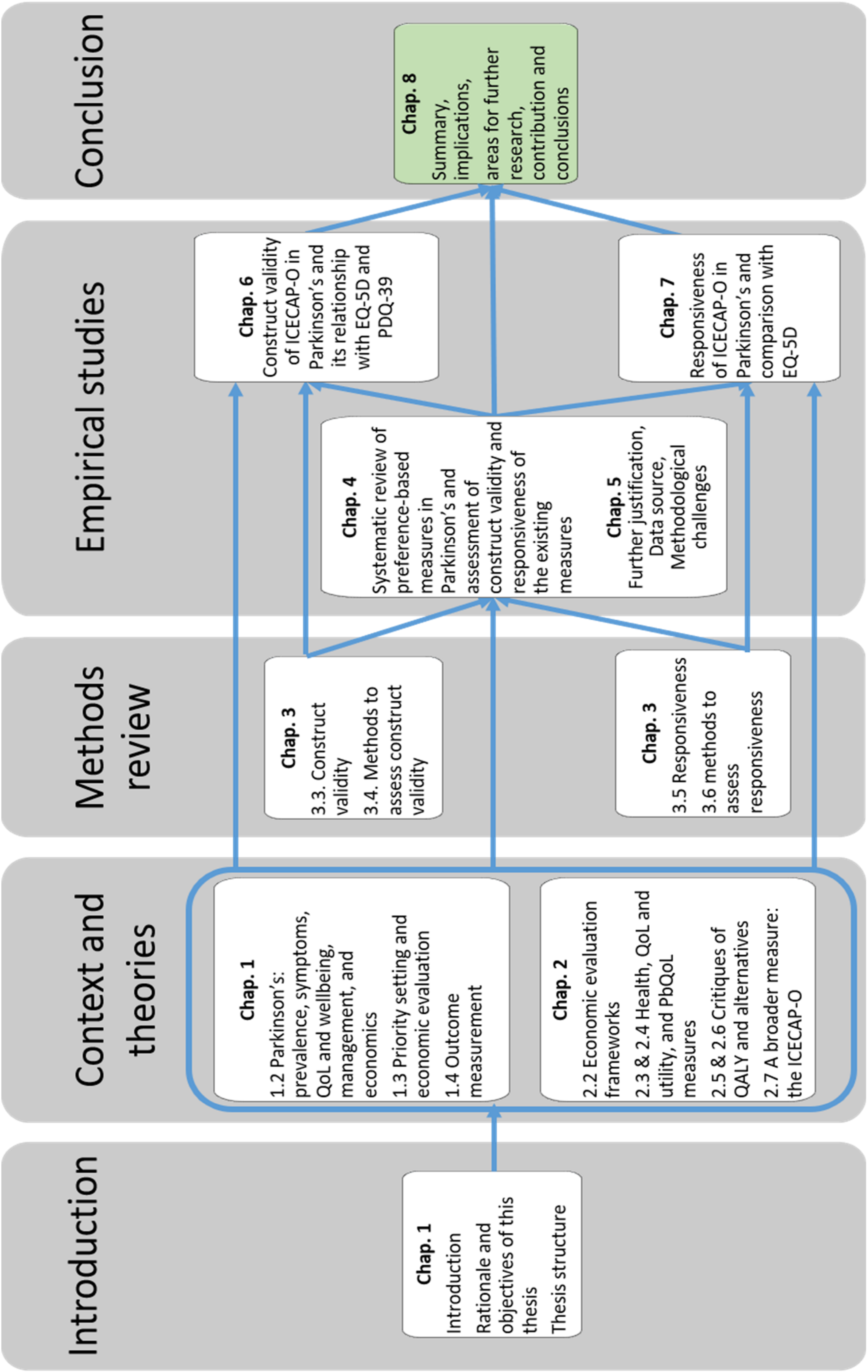
This brought another issue which is that the calculation of sample size is complicated due to the imputation. When the anchor variable contains missing data, the values imputed may be different for each imputed dataset. The different values may cause the patients to be categorised to different change groups for each imputation datasets, leading to unequal sample size for each of the five change groups between the imputation datasets. Therefore, the transformation was used to enable the calculation of ES/SRM bypassing the difficulty with calculating standard deviation, and allow the pooled estimates of ES/SRM applying Rubin's rule (Section 7.3.5.3).

7.7 Chapter summary

Following the gap identified in Chapter 4 regarding the limited responsiveness of EQ-5D-3L, this chapter reported the objectives, methods, and results of an empirical assessment of responsiveness of the capability wellbeing ICECAP-O instrument in comparison with EQ-5D-3L in people with Parkinson's. It found the ICECAP-O was more responsive to changes in overall QoL and wellbeing, and the EQ-5D-3L was more responsive to change in motor symptoms. Close relationships (social support) and bodily discomfort from the PDQ-39 were identified as the unique predictors for the change score of ICECAP-O and EQ-5D respectively, which reflects the difference in their construct and could be highlighted as their respective unique advantages in future studies. Discussion surrounding the results and methodologies were provided, and previous studies that tested the responsiveness of ICECAP-O in other populations were critiqued.

Although the differences in the estimates were small and not statistically significant, the results eliminate the concern that the sensitivity of a broad scoped measure to the specific health change might be inferior compared to a health-related QoL measure. It shows that the ICECAP-O instrument was able to provide rich information on capability wellbeing in the Parkinson's population without compromising its sensitivity to the clinical and specific QoL change in this patient group. This is a key point for its longitudinal use to measure change in a broader way than current practice and established a foundation of its use in economic evaluations of interventions in Parkinson's population. Implications of these results will be further discussed in the next chapter.

Chapter 8 Discussion and conclusion



Introduction

Chap. 1
Introduction
Rationale and objectives of this thesis
Thesis structure

Context and theories

Chap. 1
1.2 Parkinson's: prevalence, symptoms, QoL and wellbeing, management, and economics
1.3 Priority setting and economic evaluation
1.4 Outcome measurement

Chap. 2
2.2 Economic evaluation frameworks
2.3 & 2.4 Health, QoL and utility, and PbQoL measures
2.5 & 2.6 Critiques of QALY and alternatives
2.7 A broader measure: the ICECAP-O

Methods review

Chap. 3
3.3. Construct validity
3.4. Methods to assess construct validity

Chap. 3
3.5 Responsiveness
3.6 methods to assess responsiveness

Empirical studies

Chap. 4
Systematic review of preference-based measures in Parkinson's and assessment of construct validity and responsiveness of the existing measures

Chap. 5
Further justification, Data source, Methodological challenges

Chap. 6
Construct validity of ICECAP-O in Parkinson's and its relationship with EQ-5D and PDQ-39

Chap. 7
Responsiveness of ICECAP-O in Parkinson's and comparison with EQ-5D

Conclusion

Chap. 8
Summary, implications, areas for further research, contribution and conclusions

8.1 Introduction

Parkinson's is a neurodegenerative condition which is associated with lifelong disability in many aspects of body function, significantly affecting patients' lives. Interventions in Parkinson's may benefit patients in a wide range of ways related to patients' health, QoL and wellbeing. In an economic evaluation context, these benefits require sufficiently accurate measurement and valuation by preference-based instruments in order to inform decision-making.

The aim of this thesis was to examine the performance of the existing preference-based outcome measures in people with Parkinson's, and evaluate the potential of using a generic preference-based capability-wellbeing measure, ICECAP-O, to incorporate broader aspects affected by Parkinson's in economic evaluations. ICECAP-O had not been used in the Parkinson's population before and therefore it was unknown how this population would score using this measure, and to what degree it is valid compared to existing health-related QoL measures. There were two overarching research questions of this thesis:

1) Are the existing PbQoL measures appropriate to be used in the Parkinson's population? In other words, do existing preference-based generic measures capture all important aspects of QoL in People with Parkinson's?

and,

2) Is the ICECAP-O capability wellbeing measure appropriate to capture the wellbeing impact of interventions in Parkinson's, and is it sensitive in this population?

This thesis addressed these questions through three empirical works. It firstly explored the use of current PbQoL measures in people with Parkinson's and assessed existing evidence of the construct validity and responsiveness of these measures in this population through a systematic review. This was followed by two studies evaluating the construct validity, and responsiveness, respectively of the ICECAP-O, in comparison to the EQ-5D-3L measure in people with Parkinson's. This chapter will provide an overall discussion of the main findings from the empirical works making reference to the two research questions above, and summarise the implications to policy making and research. Strengths and limitations will be provided, as well as a summary of the contributions and conclusions.

8.2 Summary of the findings

8.2.1 Evaluating the performance of existing PbQoL measures in Parkinson’s

Chapter 4 explored the use, the construct validity and responsiveness of PbQoL measures in Parkinson’s population. Not surprisingly, the EQ-5D-3L instrument was predominantly used as the PbQoL measure in Parkinson’s. Furthermore, PDQ-39 was the most widely used Parkinson’s specific QoL measure among the included studies, which further justified the choice of PDQ-39 as a Parkinson’s-specific ‘gold standard’ for the case studies of validation of ICECAP-O in the Parkinson’s population.

8.2.1.1 Construct validity

Overall, the EQ-5D-3L and the other identified PbQoL measures were found to be able to differentiate between patients with different characteristics, although the grouping criteria in the included studies were favourable for a difference to be found, e.g. general population vs. people with Parkinson’s. Despite the overall positive evidence, EQ-5D-3L and HUI-2’s ability to differentiate patients with mild Parkinson’s disease (UPDRS first quantile vs. second quantile) was found to be limited. This raised a query about whether EQ-5D-3L and HUI-2 would be sensitive to detect the benefit of interventions that may slow down the progression from a very early stage of Parkinson’s to a moderate stage. Many patients at early stage of Parkinson’s may still be working or leisurely active. Early interventions that can impede the progression of disease may not only keep them in reasonable functioning status but also may prevent them from retiring from work early (i.e. saving productivity cost) or keep the level of their leisure activities, and also reduce cost of caring. Failing to accurately value the benefit of these early interventions by the PbQoL measures may lead to a poorly informed resource allocation decision.

Convergent validity analysis showed that the EQ-5D-3L, DDI and HUI-II all correlated most strongly with the physical attributes (i.e., mobility and ADL) of the PDQ-39 measure and most weakly with mental and wellbeing attributes (i.e., social support and stigma). This finding echoed the concern that these PbQoL measures may not be scoped broadly enough to capture the full 'valued' impact of disease and the associated benefit of potential interventions that may improve those aspects.

8.2.1.2 Responsiveness

For responsiveness, agreement was mixed between EQ-5D-3L and the Parkinson's-specific QoL/clinical measures in regards to the change over time across studies. The inconsistency in findings of responsiveness between these measures cautioned that the change shown on clinical measures may not necessarily translate to the same change in PbQoL scores. This may be potentially explained by either of the following two reasons. The aspects of change on the clinical measures were not 'important' to patients (i.e. patients would not trade any length of life for the improvement), thereby no difference was apparent in PbQoL scores. However, if the patients do consider the improved clinical aspects to be 'important', it could be due to the fact that the PbQoL measure does not sufficiently reflect the patients' true preferences.

8.2.1.3 Mapping

The summary of mapping algorithms from PDQ-39/8 to EQ-5D-3L utility values found that half of the PDQ-39/8 eight dimensions, which are mainly related to mental health and wellbeing, i.e. stigma, social support, cognition and communication (472-475) were not included in four out of five algorithms. This, from an alternative perspective, suggests that these important aspects to patients with Parkinson's may not be sufficiently counted in the EQ-5D.

8.2.1.4 Summary

In brief, the first empirical work of this thesis detailed the gap between 'what is counted' and 'what counts' through assessing the evidence identified in a systematic review. The gap could be summarized as follows:

- Patients value changes in their progression of Parkinson's from very mild to moderate state, however PbQoL measures may not adequately measure and value a change in their scores; what counts (the deterioration in health) is not sufficiently counted;

- In two out of ten intervention studies, the Parkinson's specific QoL measures showed a statistically significant improvement in the overall score and majority of dimensions, however, the EQ-5D did not find a difference or even showed an opposite direction of change; what counts (the benefit of some types of intervention) is not sufficiently counted;

- When the EQ-5D is obtained through mapping from the PDQ-39, the four dimensions that are related to mental health and wellbeing, no matter what levels they are at, do not affect the value of the EQ-5D score and consequent decision-making; what counts (mental health and wellbeing) is not sufficiently counted.

These gaps in PbQoL measures may have profound impact on resource allocation decisions related to the cost-effectiveness of interventions that have an impact on these aspects of people's lives. This highlights a need to seek a broadly scoped mental health and wellbeing inclusive measure to incorporate such aspects in economic evaluations.

8.2.2 The performance of the ICECAP-O instrument in Parkinson's

Following the gaps identified in Chapter 4, Chapter 6 and 7 explored if the broadly defined capability wellbeing measure, the ICECAP-O, is an appropriate measure to reflect the broader aspects of Parkinson's that count to the patients, and if it is sensitive to use in the specific disease context (i.e. as additional to the social care and public health context which have been recommended by NICE (117, 142)).

8.2.2.1 Construct validity

The capability wellbeing of the group with later stage Parkinson's was found to be statistically significantly lower than the group with early stage Parkinson's. For the ICECAP-O attributes of 'security', 'role', 'enjoyment' and 'control', the later stage group was approximately twice as likely to respond with a lower level of capability than the early group. This demonstrated that beyond the impact on health and daily functioning, the expected broader impacts of Parkinson's were well captured by the ICECAP-O instrument. The ICECAP-O index value was found to be correlated slightly more strongly with PDQ-39-SI (r=-0.73) than with the EQ-5D-3L index score (r=-0.65). The PDQ-39 attributes 'social support', 'emotional wellbeing' and 'mobility', and EQ-5D-3L attributes, 'anxiety' and 'usual activities', were found to correlate most strongly with ICECAP-O attributes.

8.2.2.2 Responsiveness

There was a statistically significant decline in capability wellbeing (mean change=-0.057) as measured by the ICECAP-O over the two-year time period. Similar to the construct validity result, a slightly stronger correlation between the ICECAP-O and PDQ-39 change scores (r=-0.54) was found than between the EQ-5D-3L and PDQ-39 change scores (r=-0.48), although differences were small. Results further showed that the ICECAP-O was more responsive to the change of overall QoL and wellbeing as confirmed by the PDQ-39, while in contrast, the EQ-5D-3L was more responsive to the change of motor symptoms. The differences in responsiveness were not statistically significant, which may be due to the fact that EQ-5D and ICECAP-O are moderately correlated both conceptually and statistically (correlation coefficient = 0.43). Nevertheless, the 'incremental' correlation between ICECAP-O and PDQ-39, compared with EQ-5D-3L and PDQ-39, was consistent with the effect size result when anchored by PDQ-39, both of which demonstrated that the ICECAP-O may be a more appropriate measure than EQ-5D for measuring the full impact of Parkinson's on patients QoL and wellbeing.

8.2.2.3 Summary

In summary, this thesis contributed to the literature by demonstrating, for the first time, the construct validity and responsiveness of the ICECAP-O in people with Parkinson's. As a subjective wellbeing measure, its sensitivity to the change in the 'gold standard' PDQ-39 in measuring the impact of Parkinson's was not inferior to the EQ-5D, and, indeed, may even surpass EQ-5D. This thesis provided initial evidence to support the continued development of the ICECAP-O as a preference-based instrument in the Parkinson's population. To realise the potential use of ICECAP-O in decision making future research will need to address other barriers such as the lack of a threshold related to ICECAP-O derived benefits (Section 8.4.3 for further details).

8.3 Strengths and limitations of this thesis

8.3.1 Strengths

This thesis has a number of strengths. The systematic review in Chapter 4 comprehensively identified studies that reported PbQoL measures which facilitated a robust summary of the use of PbQoL measures in people with Parkinson's and an assessment of the existing evidence on construct validity and responsiveness in these measures. The large sample size of the data rigorously collected through the 'pragmatic' PD MED RCT (please see details discussed in Section 5.3.1), containing patients with a wide variation of severity of Parkinson's, provided robust estimation and generalizability of metrics for the assessment of construct validity and responsiveness of the ICECAP-O instrument.

This thesis also demonstrated good practice in the assessment of construct validity and responsiveness of preference-based multi-dimensional measures. The issues regarding applying the classic psychometric methods to the assessment of preference-based QoL/capability measures were discussed in depth in Chapter 3 and 5. The assumptions were extensively discussed in the assessment of responsiveness and tested in sensitivity analysis. The missing data in the dataset were investigated and appropriately handled with imputation strategies, recognising that missing data were unlikely to be missing completely at random. Several missing data handling strategies were compared to investigate the impact of the strategies on the assessment result in the sensitivity analysis which was not been explored before to my knowledge.

8.3.2 Limitations

There are, however, also a few limitations noted with this research. Previous studies have argued that given no 'gold standard' has been established for measuring PbQoL, the test of validity can only provide a reference of a measure's performance rather than leading to a rigorous conclusion (432, 558). The ICECAP-O is a capability wellbeing measure which should be tested on wellbeing attributes in a broader sense than health only. The other capability measures could potentially be options, such as the OCAP-18 which was developed for use in public health interventions assessing central human capabilities, or the ASCOT instrument as mentioned earlier in Chapter 2 (Section 2.6.4) which purport to be based on Sen's capability approach as the ICECAP. However, given the family of capability measures are all quite new, the validity of the other capability measures is not quite clear yet, thereby the other capability measures cannot be considered as 'gold standard' if used in the assessment of ICECAP-O. Besides, the capability measures are developed for different purposes, such as ASCOT developed for use in social care thereby it including social care specific dimensions, it may be controversial to use it to validate ICECAP.

Typically, clinical trials that compare pharmacological treatments are not designed to collect data on attributes of wellbeing defined in a broader sense, rather than solely health attributes which affect QoL. Hence, there are a lack of variables to test the aspects of ICECAP-O that go beyond health. Future validation studies may consider collecting primary survey data to expand the validation angle for ICECAP in this population. The case studies used the PDQ-39 measure as a 'gold standard' as the PDQ-39 contains a wide range of attributes which covers both HrQoL and more general wellbeing, and it was designed specifically for Parkinson's and hence it should be the most relevant measure for Parkinson's.

Despite its relevance and broad coverage, the PDQ-39 has not been valued which means the attributes have not been weighted against length of life or risk of death and thus the meaning of their scores is not as clear as the preference-based measures. Although correlating the PbQoL against another non-preference QoL measure is arguably not the best test of convergent validity, as both instruments were designed to measure QoL, the trend of the scores (i.e., higher value represents better QoL) should be similar and hence the validity of the test should still provide useful information. Besides, the research of responsiveness used MID to inform grouping based on PDQ-39, which implicitly incorporates the 'importance' in its survey question when asking the participants to determine if they had experienced a little better or worse in their health states, even if this cannot replace the valuation exercise.

Besides, the known-group construct validity testing did not use H&Y staging as grouping criteria, in addition to the existing Early vs. Advanced known-group testing. Using H&Y staging to classify the participants to seven groups (Stage 1, 1.5, 2, 2.5, 3, 4, 5) will detect whether ICECAP-O is able to differentiate between the H&Y stages. This will potentially enable a comparison of the known-group results with the responsiveness results anchoring with H&Y in Chapter 7. This may also facilitate the potential future modelling studies that use H&Y states for defining the health states.

Another limitation to note is that floor and ceiling effects were not assessed in this research. It was found in Chapter 4 that the EQ-5D-3L and HUI-2 have limited ability to discriminate between patients with varied levels of mild Parkinson's. Some may argue that this may be related to the ceiling effect of the EQ-5D-3L and HUI-2 as found in other studies (270, 559-561). However, ceiling effect usually exists in a healthy general population whereby 'no problem' for all dimensions is likely to be answered, whereas Parkinson's patients have been found to have poor QoL - as mentioned in Chapter 1 (Section 1.2.3), the EQ-5D-3L value of Parkinson's patients was the lowest scores among 29 conditions, which was 0.440, compared to 0.835 among the general population in Finland (42). In this Finnish study, 6% of Parkinson's patients reported no problems on all EQ-5D-3L dimension (i.e. scoring 1), whereas typically it would need at least over 10% of respondents score to raise attention of ceiling effect.

Finally, this thesis was restricted to measurement instruments and the data that were available in the PD MED trials and thus was limited to assessment using quantitative evidence. NICE recommends the use of qualitative research to explore the content validity of EQ-5D in specific populations when its appropriateness is in doubt. Qualitative research would facilitate the justifications for the quantitative psychometric assessment. For instance, in this Parkinson's case study, 'what counts is not counted?' can be initially explored using qualitative research. However, despite the lack of qualitative investigation, this thesis drew upon the views from patients' group and the difference of descriptive systems between the generic PbQoL and the Parkinson's specific QoL measures, which are still considered to be sufficient to raise the concern that 'what counts is not counted'. The value of a qualitative research will be further discussed in 8.6 Areas for further research.

8.4 Implications of research for policy making

8.4.1 Limitations of the use of EQ-5D in Parkinson's

The findings from this thesis have significant implications for policy making. Many HTA agencies across the world are using QALYs in its decision making, such as the Netherlands, France, Sweden, Australia, Canada, New Zealand and China (562-564). Even in the US, a country that by statute prohibits the use of the cost-per-QALY approach as a basis for HTA (565), the American College of Physicians issued a position paper (2016) explicitly calling for use of QALY approach to compare the value of interventions and control drug cost (566). Among all the HTA agencies, NICE is undoubtedly the agency that most strongly relying on CUA evidence in addition to effectiveness evidence to inform decision making. NICE has explicitly stated that EQ-5D is the preferred measure for measuring outcomes for all types of guidance (117, 122, 142). However, the narrow scope that is covered by EQ-5D and the significance of EQ-5D in decision making may create some major issues.

8.4.1.1 From a decision-making perspective

For decision-making, as discussed throughout this thesis, limitation of EQ-5D would affect the estimation of the 'true effects' of interventions, and lead to inaccurate ICER calculation which may affect funding decisions or, at least endangering the role of economic evaluations in decision making. For example, DBS is a surgery option to improve motor symptoms of Parkinson's (as introduced in Chapter 1 Section 1.2.4.1) which has been shown to significantly improve patients' motor symptoms and QoL (74, 86, 87), but it is expensive as well. The PD SURG trial demonstrated that DBS is effective compared to best medical therapy (BMT) in improving patients QoL as evidenced by the significant decrease (i.e. meaning get better) on PDQ-39-SI and five out of eight dimensions of PDQ-39. These differences were not only statistical significant but with noticeable large magnititude (all much larger than MID[10]): -5.6 (95%CI -8.9 to -2.4, p=0.0008) score difference on PDQ-39-SI, -12.0 (95%CI -17.5 to -6.6, p<0.0001) on mobility,

[10] The MID (with SD) for each of the PDQ-39 eight dimensions and summary index are: 3.2 (13.26) for mobility, 4.4 (16.56) for ADL, 4.2 (17.09) for emotional wellbeing, 5.6 (22.98) for stigma, 11.4 (23.28) for social support, 1.8 (15.56) for cognition, 4.2 (18.74) for communication, 2.1 (18.68) for pain, and 1.6 (8.89) for overall score (533).

-14.0 (95% -18.7 to -9.3, p<0.0001) on ADL, stigma -9.5 (95%CI -14.9 to -4.1, p=0.0006) and -10.9 (95%CI -16.1 to -5.7, p<0.0001) on bodily discomfort (74). The EQ-5D-3L, however, only had a small and not statistically significant difference, 0.05 (-0.01, 0.11), at one year between the DBS and the BMT group (567). The modest improvement on EQ-5D-3L partly led to the very large ICER estimate of £468,528 per QALY gained over one-year time horizon of the trial. Even after extrapolation of the time horizon to five years, the ICER was still estimated to be higher than the threshold at £45,180 (567). However, given the considerable difference shown on PDQ-39, the effect is likely to be underestimated by the EQ-5D measurement and the ICER is expected to be much smaller than the current estimates if broader benefits were taken into account. If decision-making was based on ICECAP, it is likely that the considerable difference shown on PDQ-39 could be better captured and the benefit of the intervention may be more sufficiently considered in decision making.

To provide specific guidance for people with advanced Parkinson's, NICE did not calculate the utility directly measured from this subgroup from the trial, instead, the EQ-5D data were remodelled as a function of the UPDRS, off-time, and PDQ-39 variables (568). This generates a much larger EQ-5D difference between the DBS and BMT, i.e., 0.12 (95%CI 0.02, 0.22) at one year, and the lifetime ICER is £34,524. Even though this is still higher than the threshold, given the considerable improvement in PDQ-39 and some considerations on the opportunity cost of local NHS commissioning bodies, NICE's 2017 latest updated guideline recommended that DBS could be considered for people with advanced Parkinson's whose symptoms are not adequately controlled by BMT (76, 568).

This example demonstrated firstly EQ-5D is limited in capturing all the benefit of the intervention. Notably, of the four dimensions that showed improvement with certainty, except for stigma, the other three dimensions (i.e., mobility, ADL and pain/discomfort) are all related directly with EQ-5D attributes, it is surprising that the large difference in those aspects of PDQ-39 did not reflect on EQ-5D. Secondly, the very small utility gain leads to the very large ICER which provides a negative recommendation to the funding decisions, however, the reliability of this result has been questioned due to the discrepancy in effectiveness based on the evidence provided from the disease specific measure and the EQ-5D. The remodelling approach demonstrated a case of taking disease specific QoL measures into consideration in NICE's decision making when the benefit of intervention is suspected to be not fully captured by EQ-5D. However, the remodelling procedure was ad-hoc and suffers from the limitations noted in relation to mapping (Section 2.6.1). This case study highlights a perceived deficiency of directly observed EQ-5D data by the relevant decision makers. Potentially, addition of a broader outcome measure such as ICECAP-O in the setting could have provided a more responsive measurement to the improvement of PDQ-39.

8.4.1.2 From a manufacturer perspective

For the manufacturers, the substantial role of EQ-5D for NICE's decision may distort R&D decisions to focus on improvement of these aspects that covered by EQ-5D, especially the aspects that are attached with greater weights in its value sets, i.e. pain and discomfort. On the other hand, it will disincentivise the development of interventions that have little effect on the direct health related aspects but more broader benefits in general wellbeing. This may even move the disease of Parkinson's to lower level of prioritisation. This is because Parkinson's is not life threatening and thus extension of life expectancy is usually not the aim of interventions, but rather improvement of QoL. Its full mechanism for the wide range of symptoms is not yet clear and therefore developing interventions that would make a dramatic improvement is not likely at current stage of medicical innovation. Resources for R&D are limited and pharmaceutical companies aim for profit. Undervaluing the benefit of interventions in Parkinson's disease by EQ-5D would lead to those interventions that may potentially benefit patients' wellbeing being unfavourable in the decision-making process, causing equity issues across disease areas. Although NICE stated clearly in its guideline that a QALY has the same weight across all population groups and equity weighting is not included in economic evaluations (11), its recommendation of EQ-5D may cause inequity issues across population groups and disease sectors.

8.4.2 Sensitivity of EQ-5D levels

Chapter 7 identified a spike at zero in the histogram of the change of the EQ-5D-3L, which adds to the existing criticism around lack of sensitivity of the EQ-5D-3L, especially for chronic incurable diseases like Parkinson's. This insensitivity in the three levels, however, may be addressed by the newly developed EQ-5D-5L measure. However, two issues should be noted. Firstly, in NICE's latest position statement, it recommends the use of a mapping algorithm to map the answers from EQ-5D-5L to EQ-5D-3L before the validity of the newly developed EQ-5D-5L value set is tested (259). As indicated by summarizing the mapping algorithms in Chapter 4, mapping algorithms varied from each other in the characteristics of the original population where the statistical relationship was generated and in the statistical approaches taken. It is not known to what degree this NICE recommended mapping algorithm is valid; if repeating the original mapping study, whether the resulted algorithm would be the same, or (more practically) varied within an acceptable range. Furthermore, it is in doubt that the intended increased sensitivity of EQ-5D-5L could still maintain after this mapping. The second issue is that although EQ-5D-5L was developed in a bid to improve the sensitivity and indeed by judging its new five levels intuitively it will, the concern over the limited ability of EQ-5D still exists since the additional levels do not change the measuring scope. EQ-5D-5L remains a health-related QoL measure and its ability may remain insufficient to capture the broader impact of interventions.

8.4.3 Use of ICECAP-O instrument for decision-making

8.4.3.1 Role of capability approach and ICECAP-O for technology appraisal

NICE currently recommends the use of the ICECAP-O in its social care and public health guideline to capture capability wellbeing (117, 142). This thesis demonstrated that the ICECAP-O has the potential to fill the gap of health focused QoL measure by providing valid information regarding broader impact of disease or intervention. Given the comprehensive impact of Parkinson's disease and the subsequent potential broad benefit of interventions, it could be argued that NICE's recommendation of the use of ICECAP-O to measure capability may be generalized to any interventions and populations when sufficient justification is provided that the intervention may significantly impact on patients' capability wellbeing. For example, dance has been suggested as an alternative to traditional exercises for addressing the difficulties with gait and balance among people with Parkinson's; notably, of all the benefits such as activation of some parts brain that controlling motor areas, improvement of movement and endurance, the benefits of enhancing social support networks, community involvement and self-expression should not be ignored in their contribution to improve QoL (569). This important aspect of benefits due to its social nature may be more appropriately measured and valued by the capability wellbeing measures such as the ICECAP-O instrument.

Lorgelly (2015) argued that the application of capability approach may not be confined to public health and social care, but could be of benefit to evaluations of pharmacotherapies and other technologies, i.e. technology appraisal (125). NICE's 'Methods for Technology Appraisal' (119) guides the assessment of drugs and devices for conditions or diseases where economic evaluation is primarily used and has an important role to assist decision-making. Arguments for a broader measure in the disease / health domain have been arisen especially in the recent decade with the increasing awareness of the complex nature of some diseases and/or the corresponding complex outcome of interventions (125, 570).

Payne et al. (2013) (570, 571) used the genetic diagnostic services and tests as an example for complex interventions that have broader objectives than health gain only, as the information provided may influence behaviour and decisions of the parents, affecting their wellbeing. Therefore they suggested the non-health benefit "may usefully be measured using the concept related to capability, which we called 'empowerment'" (570). Lorgelly (2015) (125) gave a further example from Australia to argue for the importance of capturing non-health benefit in technology appraisals of pharmaceuticals. To manage the attention-deficit/hyperactivity disorder (ADHD), a drug named lisdexamfetamine is listed positively by the Australisan Pharmaceutical Benefits Advisory Committee (PBAC), based on the evidence on its health benefit of improving the symptoms of ADHD. However, it is argued that the lisdexamfetamine could have considerable non-health benefit as it will improve the impact of ADHD on children's behaviour including poor education performance and increased criminal activity. These non-health benefits should be considered in decision making, which however cannot be captured by EQ-5D, where capability instruments would have a role.

In addition, given the growing awareness of diseases' impact on carer's burden and the associated invisible and substantial caring cost, incorporation of informal care into decision making for resource allocation purposes has attracted increasing amount of attention in research. However, a review published in 2012 found that there is a huge heterogeneity in terms of the methods applied to measure and value informal care (572). Specifically, of the 17 studies that incorporated outcomes for carers, a mixture of outcomes were used including the generic health-related measures, EQ-5D-3L, HUI, SF-6D, and well as several carer specific measures. The issue of using generic health-related instruments to measure carer's outcome is that the majority of the items on these instruments are unrelated to the carer's burden; actually probably all but the psychological items such as anxiety/depression for EQ-5D. Accordingly using these generic health-related measures is not ideal. On the other hand, the carer specific measures are specific to carers but not patients and thus inevitably creating an important issue regarding how to combine the carer's outcome and the patient's outcome and incorporate these together into economic evaluation. In this situation, ICECAP may potentially provide an alternative feasible avenue. It is a measure that is proposed to have an important role in the context of integration of health care and social care (141). Its broad capability wellbeing attributes (i.e. for ICECAP-O, attachment, security, role, enjoyment and control) are relevant to both patients and carers. As such ICECAP may have a role in the work of incorporating informal care into economic evaluation for decision making which will enable the measurement, valuation and combination of patients and carers' outcomes.

8.4.3.2 Issues of the use of ICECAP-O

Chapter 2 (Section 2.7.2) showed that the ICECAP-O was valued using a best worst scaling approach, where individuals simply choose the best and worst attribute level, as opposed to the trade-off methods, where the individuals have to scarifice A for B. The authors argued that best worst scaling approach is appropriate as it reflects a value judgement which is more aligned with Sen's capability approach (141). The valuation used 'no capability' as anchor rather than 'death', and as such when combing ICECAP value with length of life would not generate QALY and cannot apply the same decision rule (£20,000 – 30,000 per QALY gained). Details of this have been provided in Chapter 2, Section 2.7.3 'use of ICECAP-O for decision making'.

There are, however, other outstanding issues when applying the capability approach in decision making as outlined by Lorgelly (2015) (125). Besides the ICECAP, other instruments have been developed purporting to measure capability, such as ASCOT and the OCAP, and for all the three measures many versions have been developed to suit different populations. Just like the evidence showing that EQ-5D, HUI, and SF-6D index values are not interchangeable (127, 267-270), these measures have important difference between their descriptive systems and valuation methods, thereby selection of one or another as a capability wellbeing measure to inform decision making is challenging.

However, a bigger picture discussion is when the scope is broadened to wellbeing, whether the scope for cost should also be broadened beyond the health budget (125). Subjective wellbeing is about life satisfaction in general and the wellbeing attributes in ICECAP such as attachment or role are a result of multiple sectors not confined to health. When the interventions provided by health sector can have benefit beyond health sector, the budget may need to be redistributed in a way that is multisectoral to match the scope of benefit. Remme et al. (2017) proposed a 'cofinancing' approach, in which the other sectors could contribute towards a health intervention which would achieve non-health benefit in other sectors, and vice versa. Research on this in health economics is still at initial stage and certainly more research is required to explore such bigger issues of broadening evaluative scope.

8.5 Implications and recommendations for future research

8.5.1 Choice of preference-based outcomes

Awareness should be raised in Parkinson's research regarding the potential limited scope of the EQ-5D in this population and it is advised that researchers should be mindful of the advantages and limitations of each outcome measure before using it to measure the effect of different types of intervention. It is recommended that when the main objective of an intervention is to improve specific health aspects that are covered by EQ-5D, the addition of other measures may not be necessary. When the intervention is expected to have broader impact beyond health alone, the addition of ICECAP-O which could capture the broader impact and incorporate it into economic evaluations is recommended. This should be particularly considered if there is a doubt that the bespoke benefit of the intervention is not captured by the EQ-5D questionnaire. This thesis provided a reference point to future studies regarding the capability level in the early and later stages of Parkinson’s and the change of capability over time, which will facilitate future studies to set up their hypotheses, especially the intervention studies regarding the expected change in capability within a certain time frame.

8.5.2 The necessity of setting up hypotheses

For future studies to assess responsiveness, it is recommended that a priori hypotheses / expectations must be set based on the intended construct of the test measure. This has not been rigorously applied by many studies as shown in Chapter 3 and 7. Simply giving conclusions based on the strength of correlation / size of ES statistics without prior expectations will lead to incorrect interpretation of the assessment result. Although setting up a hypothesis is crucial to the interpretation of results, it is usually difficult to determine precisely how strongly the generic QoL measures is expected to correlate with the clinical measures. There is very limited guidance on what constitutes reasonable correlation, or effect size, in the test of responsiveness of a multi-dimensional concept. Despite the COSMIN checklist providing a standard guidance on how the properties should be assessed, consensus was not reached on the criteria of adequacy of measurement properties (326). This will be further discussed in Section 8.6.3 for future research.

8.5.3 Outcome measures for defining health state in decision-analytic modelling

Chapter 7 found that when H&Y stage changes, EQ-5D utility values do not always change with the same direction, especially in the group that have improved H&Y stage and no change H&Y stage. This raised a caveat in decision-analytic modelling studies when simply using H&Y to define health states in the models. Although there are good reasons to predict utility values based on H&Y stage, this prediction is not perfect since utility value, again, is obtained from a multi-dimensional concept, of which motor complication is only one (albeit important) of the many dimensions. Decision-analytic modellers in Parkinson’s are advised to consider the scope of the targeted effect of interventions and the available outcome measures in studies, and where appropriate, to use a combination of measures to enable a broader scope and account for the progression of the disease.

8.6 Areas for further research

8.6.1 Comparison between EQ-5D-3L and ICECAP-O measure in economic evaluation in Parkinson's population

PbQoL measures are developed for use in economic evaluations, and economic evaluations are conducted for making recommendations for resource allocation between interventions. This thesis provided evidence on the use and relevance of the EQ-5D-3L in Parkinson's population particularly in relation to the scope and sensitivity between milder disease states. This thesis demonstrated that the ICECAP-O, the broadly scoped measure is responsive and exhibit construct validity to be used in this population. This naturally leads to the next important question – whether an economic evaluation using ICECAP-O or EQ-5D from the same study would provide different recommendations of a health care intervention to the decision makers? The varied constructs and purposes of the two measures mean that if the recommendations are different, then the discrepancy is most likely to happen for the interventions that have a focus on broader wellbeing. This however, requires further research to test, and most crucially, a decision rule for using ICECAP-O. A health QALY and a capability QALY are not comparable; the current NICE decision rule, £20,000 – 30,000 per QALY gained only applies to the health QALY but not the capability QALY. At the moment, research led by Dr. Philip Kinghorn funded by the Medical Research Council is ongoing which is looking at establishing the social willingness to pay for gains in capability outcomes up to the sufficient level of capability given the consideration of equity (i.e. YSC, please see Chapter 2, Section 2.7.3 for details) (320, 573). This research will have vital methodological and practical significance in broadening the evaluative space of health care decision-making in the UK.

8.6.2 Validation of 'preference'

As mentioned earlier in the limitation section (8.3.2), one limitation of this thesis is using a non-preference based measure as a 'gold standard' to validate a preference-based measure. Some may argue that the meaning of the change on the non-preference based measure is not clear and thus the assumption for the validation studies is not valid. Although the MID method was used, with the assumption that ‘importance’ is similar to ‘preference’, in fact these two terms have conceptual difference from each other. When patients answered ‘a little better’ in MID research, there is no ‘trade off’ in this preference expression, whereas when patients answered ‘preferred’ in preference elicitation exercise, it means they would trade more length of life or accept a higher risk of death for it.

This conceptual difference should be reflected in the interpretation of results. Papaioannou, Brazier, and Parry (2011) previously pointed out that “where health dimensions and changes appear to have been missed by preference-based HRQoL measures, these may not actually be important to patients or valued by the general population’ thus it cannot be determined as a weakness of the measure” (246). This is indeed the truth. However, it could be argued that this thesis assumed PDQ-39 as ‘gold standard’ because its dimensions were developed based on extensive qualitative and quantitative research and therefore all these dimensions are deemed important to patients to some degree. One avenue to test this assumption is to conduct a valuation exercise of the PDQ-39 measure, i.e. generate preference values for this condition-specific measure. This would strengthen the fundamental assumption of Chapter 6 and 7 of this thesis – PDQ-39 is the current best available instrument to measure ‘what counts’ for people with Parkinson’s, and would add weight to the validation of generic preference-based measures in the context of Parkinson’s population in the future. In addition, this valuation should be conducted in both a patient population and among the general public to identify the differences, if any, between the elicited values.

8.6.3 Developing robust methods of demonstrating psychometric properties for generic PbQoL measures

NICE methods guideline recommends the assessment of content validity, construct validity, and responsiveness to make a case that the EQ-5D is inappropriate. This requires a standard and robust method to implement. Many studies have tested the construct validity and responsiveness of the EQ-5D but may generate different conclusions depending on the patient population, the hypotheses tested and quality of the methods. There is no standard for how hypotheses should be set, what hypothesis is important to a PbQoL measure, and what size of the statistics should be expected.

For example, for known-group validity, it is not difficult for the measure to differentiate between the groups with later stage Parkinson's and early Parkinson's given there are a lot of patients characteristics that are expected to differ, however a large difference does not automatically mean a higher construct validity – it all depends on how it matches the expectation as stated by outlined hypothesis. Therefore, how to determine what the expectation should be is quite crucial, which requires future research. Similarly, for responsiveness, larger size of effect size does not necessarily mean a measure being more responsive – again, it is dependent upon the expectations. The COSMIN checklist and previous studies (246, 327) recommended to be explicit in the hypothesis about the expected mean difference or correlation however this is challenging to implement in practice, especially for a measure used in a population for the first time.

Modelling methods may be one solution to help set up expectations and resolve hypotheses of a multi-dimensional 'no agreement' concept like QoL/capability. When there is gold standard, the test measure is expected to make exactly the same judgement as the gold standard, e.g. a diagnostic test. However, when there is no gold standard such as QoL/capability, it is difficult to set up expectations on the strength of the correlation or the ES statistics. This is because beside what is measured by the anchor, there may be other underlying factors that also contributing to the score of the test measure, especially when the anchor measure only measures one dimension of the multi-dimensional concept. For example, Chapter 6 found that when using the disease severity measure, H&Y as an anchor, there were good reasons to predict that very strong correlations might exist between H&Y and ICECAP-O. The patients with more advanced stage of H&Y are more likely to have poor independence, feel worried about their future, less ability to do activities with family and friend or social. However, in our results, this relationship was not strong and there are other factors contributing to ICECAP-O beyond H&Y. In these circumstances, potentially, a modelling method that can adjust for other factors affecting the size of the measure would help to set up a hypothesis or expectations for the relationship between the test measure and the anchor measure. There is little guidance in the literature regarding how to use modelling methods to adjust for other factors so this could be an important direction for future research in the area of psychometrics testing.

Overall, a comprehensive assessment of PbQoL measures would also require the investigation of the routine criteria, as with other non-preference-based measures, i.e., practicality, reliability and validity, but with an additional layer of complexity due to the nature of no gold standard, self-reported responses, and a scoring algorithm based on preferences elicited from general public (118). The COSMIN checklist as previously mentioned in Chapter 3 (326) and Brazier et al. (2017) (118) provide the current best available guidance on the assessment. In particular, built on the earlier checklist of psychometric criteria written by Brazier and Deverill (1999) (148), the book by Brazier et al. (2017) focused on the assessment of preference based measures and provides an updated checklist of criteria that should be assessed for this special type of measures. These criteria include: practicality (i.e., acceptability to respondents, and burden of administration and completion), reliability (i.e., test-retest, inter-rater, and between methods of administration), and validity. Within validity, the tests are categorized to three aspects: (a) the assessment of the description, including content validity and face validity; (b) assessment of the valuation including whose preference was elicited, technique of valuation, and the quality of data, and; (c) empirical validity which relies on empirical data to test which includes the testing of stated preference, and hypothesized preferences including the test of construct validity and responsiveness using empirical data. Future research of validation of ICECAP may assess other criteria contained in this checklist and thus provides together with this thesis a more comprehensive picture of the merit of ICECAP.

8.6.4 Value of qualitative research

One of the ideas behind this thesis stemmed from the expressed concern from the Parkinson’s UK and the large difference between the Parkinson’s specific QoL measures and the EQ-5D. Hence, qualitative research on content validity to explore the relevance of EQ-5D to the Parkinson’s disease would be a useful addition to the literature. Interviews with patients as well as clinicians, researchers, physiotherapists and other stakeholders, would provide an in-depth insight into the gap between ‘what counts’ and ‘what is counted’. In addition, an interview with Parkinson’s patients, experts and manufacturers regarding the comparison between ICECAP-O and EQ-5D would also provide important information from their perspectives to what extent QoL and wellbeing should be incorporated into decision-making. Psychometric testing is an iterative process. Qualitative interviews based on the finding of this research will further inform setting up hypotheses on the relationships tested in the quantitative studies in the future.

8.7 Contribution of the thesis

The contribution of this thesis can be summarized as follows:

- This is the first study to provide information regarding the capability wellbeing of Parkinson's patients. This extends our knowledge of capability wellbeing of different patient groups.

- This is the first study to test construct validity and responsiveness of the ICECAP-O in Parkinson's population. This thesis revealed that the use of the ICECAP-O in Parkinson's population is warranted.

- This thesis provided evidence to eliminate the concern that a broadly defined subjective measure would not be sensitive to specific health aspects; sensitivity of ICECAP-O was not inferior to the HrQoL measure EQ-5D-3L.

- This thesis provided guidance on the choice of measure in the studies of Parkinson's – when the benefit of the intervention is projected /suspected to be broad and beyond health, ICECAP-O is recommended in addition to the EQ-5D (3L/5L) instrument required by NICE.

- This thesis demonstrated good practice in the assessment of construct validity and responsiveness of a preference-based multi-dimensional measure. This was achieved through the application of rigorous methodological critique, statistical testing, multiple imputation and extensive sensitivity analysis surrounding the methodological assumptions.

- This thesis provided a reference point to future studies regarding the capability level in early and later stages of Parkinson's and the change of capability over time and the natural history of the disease in this group of patients. This will facilitate future studies to set up their hypothesis regarding the expectations on capability in this population.

8.8 Conclusions

This thesis identified the gap between ‘what counts’ to people with Parkinson’s and ‘what is counted’ in the current PbQoL measures, with a focus on the NICE recommended PbQoL measure EQ-5D, and demonstrated the appropriateness of using ICECAP-O in the Parkinson’s population, which may potentially fill this gap in decision-making process.

Through a systematic review, this thesis has detailed the limited ability of the existing preference-based measures in people with Parkinson’s and suggested that EQ-5D may underestimate the value placed on the mental and social wellbeing aspects in Parkinson’s disease. This highlights the need to seek a broadly scoped mental health and wellbeing inclusive measure to incorporate such aspects in economic evaluations. This thesis established the construct validity and responsiveness of the ICECAP-O in Parkinson’s and demonstrated that there are valued capability wellbeing attributes in Parkinson’s beyond those reflected by the EQ-5D instrument. This thesis contributes to understanding the use of broadly scoped outcome measure for economic evaluations in Parkinson’s by showing that the ICECAP-O instrument was able to provide rich information on these under-represented aspects in the Parkinson’s population, without compromising its sensitivity to the clinical and specific physical QoL dimensions in this patient group. It should be also noted that choosing one or the other measure does not simply depend on whichever performed better in the psychometric test, it can also depend on many other factors, such as: which one is more acceptable to decision makers, which one uses an evaluative space that best matches with the aim of NHS, which one has a decision rule (threshold) that can be applied, which one the patients more willing to answer and the clinical researchers prefer, and so on. This thesis, therefore, achieved progress within a bigger picture towards the use of ICECAP-O in economic evaluations to measure the impact of interventions in a broader way than current practice in Parkinson’s populations.

This thesis has contributed to the goal of making what counts to patients with Parkinson’s become what is counted in the decision-making process - letting decision-making meet patients’ need. Although there are differences in preferences from patients’ and decision makers’ perspectives, there is one fundamental principle in the NHS since it was launched in 1948 – ‘it meets the needs of everyone’. Making what counts counted is a necessary step to make the needs met. What is measured and valued by the PbQoL measures in the health care decision-making process determines what is valued by the health care system, and jointly determines with cost what is prioritised for NHS funding. Scoping issues are unavoidable, due to the broad ranging disease symptoms, their impact on life, and the subsequent benefit of interventions. This thesis demonstrated that the broadly scoped capability/wellbeing measure, ICECAP-O, should be considered as an instrument for evaluations of interventions in Parkinson’s, looking beyond health gains to produce an overall valuation of their benefit.

References

1. Cameron WB. Informal sociology: a casual introduction to sociological thinking. New York: Random House; 1963.

2. Quote Investigator. (2010). Not Everything That Counts Can Be Counted. https://quoteinvestigator.com/2010/05/26/everything-counts-einstein/. Accessed 15 November 2017.

3. Nussbaum RL, Ellis CE. Alzheimer's disease and Parkinson's disease. *N Engl J Med*. 2003;348(14):1356-64.

4. Saarni SI, Harkanen T, Sintonen H, Suvisaari J, Koskinen S, Aromaa A, et al. The impact of 29 chronic conditions on health-related quality of life: a general population survey in Finland using 15D and EQ-5D. *Qual Life Res*. 2006;15(8):1403-14.

5. Winter Y, von Campenhausen S, Popov G, Reese JP, Balzer-Geldsetzer M, Kukshina A, et al. Social and clinical determinants of quality of life in Parkinson's disease in a Russian cohort study. *Parkinsonism Relat Disord*. 2010;16(4):243-8.

6. DeMaagd G, Philip A. Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis. *Pharm Ther*. 2015;40(8):504-32.

7. Todorova A, Jenner P, Ray Chaudhuri K. Non-motor Parkinson's: integral to motor Parkinson's, yet often neglected. *Pract Neurol*. 2014;14(5):310-22.

8. Bovolenta TM, de Azevedo Silva SMC, Arb Saba R, Borges V, Ferraz HB, Felicio AC. Systematic review and critical analysis of cost studies associated with Parkinson's Disease. *Parkinsons Dis*. 2017;2017:3410946.

9. Gumber A, Ramaswamy B, Ibbotson R, Ismail M, Thongchundee O, Harrop D, et al. Economic, Social and Financial Cost of Parkinson's on Individuals, Carers and their Families in the UK. Project report. Centre for Health and Social Care Research, Sheffield Hallam University; 2017.

10. Findley LJ. The economic impact of Parkinson's disease. *Parkinsonism Relat Disord*. 2007;13 Suppl:S8-S12.

11. National Institute for Health and Care Excellence. (2014). Developing NICE guidelines: the manual [PMG20]. https://www.nice.org.uk/process/pmg20/chapter/incorporating-economic-evaluation#the-reference-case. Accessed 3 January 2018.

12. Scottish Medicines Consortium. Guidance to manufacturers Question and answer document on economic submissions to the Scottish Medicines Consortium. http://www.scottishmedicines.org.uk/Submission_Process/Submission_guidance_and_forms/Templates-Guidance-for-Submission/Economic_Question_and_Answer_Document. Accessed 9 August 2017.

13. National Institute for Health and Care Excellence. Glossary. https://www.nice.org.uk/glossary?letter=h. Accessed 05 Feb 2018.

14. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16(3):199-208.

15. National Institute for Clinical Excellence. (2004). Guide to the Methods of Technology Appraisal.

16. Karimi M, Brazier J. Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? *Pharmacoeconomics*. 2016;34(7):645-9.

17. Centers for Disease Control and Prevention (CDC). Well-Being Concepts. https://www.cdc.gov/hrqol/wellbeing.htm#three. Accessed 05 Feb 2018.

18. National Institute of Health and Care Excellence. (2013). Methods of Technology Appraisal Consultation. Responses to TA Methods Addendum Public Consultation between 27th March 2014 and 1st July 2014. https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/VBA-Consultation-Comments.pdf. Accessed 16 November 2017.

19. Finch AP, Brazier JE, Mukuria C. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *Eur J Health Econ*. 2017.

20. Bulamu NB, Kaambwa B, Ratcliffe J. A systematic review of instruments for measuring outcomes in economic evaluation within aged care. *Health Qual Life Outcomes*. 2015;13:23.

21. Yang YL, Brazier J, Longworth L. EQ-5D in skin conditions: an assessment of validity and responsiveness. *Eur J Health Econ*. 2015;16(9):927-39.

22. Yang Y, Brazier J, Tsuchiya A. Effect of adding a sleep dimension to the EQ-5D descriptive system: a "bolt-on" experiment. *Med Decis Making*. 2014;34(1):42-53.

23. Kuspinar A, Mayo NE. Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? *Health Qual Life Outcomes*. 2013;11(1):71.

24. Monsy K, N. (2015). Movement is their Mantra. https://www.khaleejtimes.com/wknd/movement-is-their-mantra. Accessed 15 November 2017.

25. Fernandez M. (2012). Elder George Bush is hospitalized, but condition is stable. https://thecaucus.blogs.nytimes.com/2012/11/29/former-president-george-h-w-bush-is-hospitalized-but-in-stable-condition/. Accessed 15 November 2017.

26. Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov Disord*. 2014;29:1583-90.

27. Parkinson J. An essay on the shaking palsy. 1817. *J Neuropsychiatry Clin Neurosci*. 2002;14(2):223-36; discussion 2.

28. Jankovic J, Tolosa E. Parkinson's disease & movement disorders. Sixth ed. Philadelphia: Lippincott Williams & Wilkins; 2015.

29. Hou J-GG, Lai EC. Non-motor Symptoms of Parkinson's Disease. *Int J Gerontol*. 2007;1(2):53-64.

30. Lumen - Boundless Anatomy and Physiology. Motor Pathways. https://courses.lumenlearning.com/boundless-ap/chapter/motor-pathways/. Accessed 1 Feb 2018.

31. Jellinger KA. Neuropathology of sporadic Parkinson's disease: evaluation and changes of concepts. *Mov Disord*. 2012;27(1):8-30.

32. Baumann CR. Epidemiology, diagnosis and differential diagnosis in Parkinson's disease tremor. *Parkinsonism Relat Disord*. 2012;18(Suppl 1):S90-2.

33. Jankovic J. Parkinson's disease: clinical features and diagnosis. *J Neurol Neurosurg Psychiatry*. 2008;79(4):368-76.

34. Suchowersky O, Reich S, Perlmutter J, Zesiewicz T, Gronseth G, Weiner WJ. Practice Parameter: diagnosis and prognosis of new onset Parkinson disease (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*. 2006;66(7):968-75.

35. Xia R, Mao ZH. Progression of motor symptoms in Parkinson's disease. *Neurosci Bull*. 2012;28(1):39-48.

36. Jimenez MC, Vingerhoets FJ. Tremor revisited: treatment of PD tremor. *Parkinsonism Relat Disord*. 2012;18(Suppl 1):S93-5.

37. Berardelli A, Rothwell JC, Thompson PD, Hallett M. Pathophysiology of bradykinesia in Parkinson's disease. *Brain*. 2001;124(Pt 11):2131-46.

38. Grabli D, Karachi C, Welter ML, Lau B, Hirsch EC, Vidailhet M, et al. Normal and pathological gait: what we learn from Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 2012;83(10):979-85.

39. Garcia Ruiz PJ, Catalan MJ, Fernandez Carril JM. Initial motor symptoms of Parkinson disease. *Neurologist*. 2011;17(6 Suppl 1):S18-20.

40. Hallett M. Parkinson's disease tremor: pathophysiology. *Parkinsonism Relat Disord*. 2012;18 Suppl 1:S85-6.

41. Moustafa AA, Chakravarthy S, Phillips JR, Gupta A, Keri S, Polner B, et al. Motor symptoms in Parkinson's disease: A unified framework. *Neurosci Biobehav Rev*. 2016;68:727-40.

42. Barone P, Antonini A, Colosimo C, Marconi R, Morgante L, Avarello TP, et al. The Priamo Study: A Multicenter Assessment of Nonmotor Symptoms and Their Impact on Quality of Life in Parkinson's Disease. *Mov Disord*. 2009;24(11):1641-9.

43. Martinez-Martin P, Rodriguez-Blazquez C, Kurtis MM, Chaudhuri KR. The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease. *Mov Disord*. 2011;26(3):399-406.

44. Salawu FK, Danburam A, Olokoba AB. Non-motor symptoms of Parkinson's disease: diagnosis and management. *Niger J Med*. 2010;19(2):126-31.

45. Parkinson's UK. (2017). Explore information and support. https://www.parkinsons.org.uk/information-and-support/symptoms. Accessed 13 November 2017.

46. Poewe W. Non-motor symptoms in Parkinson's disease. *Eur J Neurol*. 2008;15 Suppl 1:14-20.

47. Kurtis M, Logishetty K, Martinez-Martin P. An in-Depth Look at the Non Motor Symptom Scale. In: Chaudhuri KR, Martinez-Martin P, Odin P, Antonini A, editors. Handbook of Non-Motor Symptoms in Parkinson's Disease. Heidelberg: Springer Healthcare UK; 2011. p. 37-43.

48. Kano O, Ikeda K, Cridebring D, Takazawa T, Yoshii Y, Iwasaki Y. Neurobiology of depression and anxiety in Parkinson's disease. *Parkinsons Dis*. 2011;2011:143547.

49. Szatmari S, Illigens BM-W, Siepmann T, Pinter A, Takats A, Bereczki D. Neuropsychiatric symptoms in untreated Parkinson’s disease. *Neuropsychiatr Dis Treat*. 2017;13:815-26.

50. Aarsland D, Marsh L, Schrag A. Neuropsychiatric symptoms in Parkinson's disease. *Mov Disord*. 2009;24(15):2175-86.

51. Paulus W, Jellinger K. The neuropathologic basis of different clinical subgroups of Parkinson's disease. *J Neuropathol Exp Neurol*. 1991;50(6):743-55.

52. Braak H, Ghebremedhin E, Rub U, Bratzke H, Del Tredici K. Stages in the development of Parkinson's disease-related pathology. *Cell Tissue Res*. 2004;318(1):121-34.

53. Das P, Naylor C, Majeed A. Bringing together physical and mental health within primary care: a new frontier for integrated care. *J R Soc Med*. 2016;109(10):364-6.

54. Mental health foundation. (2017). Physical health and mental health. https://www.mentalhealth.org.uk/a-to-z/p/physical-health-and-mental-health. Accessed 13 November 2017.

55. Ravina B, Camicioli R, Como PG, Marsh L, Jankovic J, Weintraub D, et al. The impact of depressive symptoms in early Parkinson disease. *Neurology*. 2007;69(4):342-7.

56. Starkstein SE, Mayberg HS, Leiguarda R, Preziosi TJ, Robinson RG. A prospective longitudinal study of depression, cognitive decline, and physical impairments in patients with Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 1992;55(5):377-82.

57. Hughes TA, Ross HF, Mindham RH, Spokes EG. Mortality in Parkinson's disease and its association with dementia and depression. *Acta Neurol Scand*. 2004;110(2):118-23.

58. Reijnders JS, Ehrt U, Weber WE, Aarsland D, Leentjens AF. A systematic review of prevalence studies of depression in Parkinson's disease. *Mov Disord*. 2008;23(2):183-9; quiz 313.

59. Aarsland D, Larsen JP, Lim NG, Janvin C, Karlsen K, Tandberg E, et al. Range of neuropsychiatric disturbances in patients with Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 1999;67(4):492-6.

60. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. *Ann Med*. 2001;33(5):328-36.

61. Nijhof G. Uncertainty and lack of trust with Parkinson's disease. *Eur J Public Health*. 1996;6(1):58-63.

62. Jenkinson C, Fitzpatrick R, Peto V, Dummett S, Morley D, Saunders P. The Parkinson's Disease Questionnaires - User Manual. Third ed: Oxford University Innovation Limited; 2012.

63. Smith LJ, Shaw RL. Learning to live with Parkinson's disease in the family unit: an interpretative phenomenological analysis of well-being. *Med Health Care Philos*. 2017;20(1):13-21.

64. Tanji H, Anderson KE, Gruber-Baldini AL, Fishman PS, Reich SG, Weiner WJ, et al. Mutuality of the marital relationship in Parkinson's disease. *Mov Disord*. 2008;23(13):1843-9.

65. Carter JH, Stewart BJ, Archbold PG, Inoue I, Jaglin J, Lannon M, et al. Living with a person who has Parkinson's disease: the spouse's perspective by stage of disease. Parkinson's Study Group. *Mov Disord*. 1998;13(1):20-8.

66. Maffoni M, Giardini A, Pierobon A, Ferrazzoli D, Frazzitta G. Stigma Experienced by Parkinson's Disease Patients: A Descriptive Review of Qualitative Studies. *Parkinsons Dis*. 2017;2017:7203259.

67. Nijhof G. Parkinson's Disease as a problem of shame in public appearance. *Sociol Health Illn*. 1995;17(2):193-205.

68. Hellqvist C, Bertero C. Support supplied by Parkinson's disease specialist nurses to Parkinson's disease patients and their spouses. *Appl Nurs Res*. 2015;28(2):86-91.

69. Karlsen KH, Tandberg E, Arsland D, Larsen JP. Health related quality of life in Parkinson's disease: a prospective longitudinal study. *J Neurol Neurosurg Psychiatry*. 2000;69(5):584-9.

70. Hartelius L, Svensson P. Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. *Folia Phoniatr Logop*. 1994;46(1):9-17.

71. Miller N, Deane KH, Jones D, Noble E, Gibb C. National survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: therapists' practices. *Int J Lang Commun Disord*. 2011;46(2):189-201.

72. Miller N, Noble E, Jones D, Burn D. Life with communication changes in Parkinson's disease. *Age Ageing*. 2006;35(3):235-9.

73. Soleimani MA, Negarandeh R, Bastani F, Greysen R. Disrupted social connectedness in people with Parkinson's disease. *Br J Community Nurs*. 2014;19(3):136-41.

74. Williams A, Gill S, Varma T, Jenkinson C, Quinn N, Mitchell R, et al. Deep brain stimulation plus best medical therapy versus best medical therapy alone for advanced Parkinson's disease (PD SURG trial): a randomised, open-label trial. *Lancet Neurol*. 2010;9(6):581-91.

75. Gray R, Ives N, Rick C, Patel S, Gray A, Jenkinson C, et al. Long-term effectiveness of dopamine agonists and monoamine oxidase B inhibitors compared with levodopa as initial treatment for Parkinson's disease (PD MED): a large, open-label, pragmatic randomised trial. *Lancet*. 2014;384(9949):1196-205.

76. National Institute for Health and Care Excellence. (2017). Parkinson’s disease in adults. NICE guideline [NG71]. https://www.nice.org.uk/guidance/ng71. Accessed 15 November 2017.

77. Poewe W. Treatments for Parkinson disease--past achievements and current clinical needs. *Neurology*. 2009;72(7 Suppl):S65-73.

78. Lees AJ. The on-off phenomenon. *J Neurol Neurosurg Psychiatry*. 1989;Suppl:29-37.

79. Schrag A, Quinn N. Dyskinesias and motor fluctuations in Parkinson's disease. A community-based study. *Brain*. 2000;123 (Pt 11):2297-305.

80. Voon V, Mehta AR, Hallett M. Impulse control disorders in Parkinson's disease: recent advances. *Curr Opin Neurol*. 2011;24(4):324-30.

81. Weintraub D, Koester J, Potenza MN, Siderowf AD, Stacy M, Voon V, et al. Impulse control disorders in Parkinson disease: a cross-sectional study of 3090 patients. *Arch Neurol*. 2010;67(5):589-95.

82. Avanzi M, Baratti M, Cabrini S, Uber E, Brighetti G, Bonfa F. Prevalence of pathological gambling in patients with Parkinson's disease. *Mov Disord*. 2006;21(12):2068-72.

83. Cilia R, Cho SS, van Eimeren T, Marotta G, Siri C, Ko JH, et al. Pathological gambling in patients with Parkinson's disease is associated with fronto-striatal disconnection: a path modeling analysis. *Mov Disord*. 2011;26(2):225-33.

84. Stacy M, Bowron A, Guttman M, Hauser R, Hughes K, Larsen JP, et al. Identification of motor and nonmotor wearing-off in Parkinson's disease: comparison of a patient questionnaire versus a clinician assessment. *Mov Disord*. 2005;20(6):726-33.

85. Groiss SJ, Wojtecki L, Sudmeyer M, Schnitzler A. Deep brain stimulation in Parkinson's disease. *Ther Adv Neurol Disord*. 2009;2(6):20-8.

86. Krack P, Batir A, Van Blercom N, Chabardes S, Fraix V, Ardouin C, et al. Five-year follow-up of bilateral stimulation of the subthalamic nucleus in advanced Parkinson's disease. *N Engl J Med*. 2003;349(20):1925-34.

87. Rodriguez-Oroz MC, Obeso JA, Lang AE, Houeto JL, Pollak P, Rehncrona S, et al. Bilateral deep brain stimulation in Parkinson's disease: a multicentre study with 4 years follow-up. *Brain*. 2005;128(Pt 10):2240-9.

88. Deuschl G, Schade-Brittinger C, Krack P, Volkmann J, Schafer H, Botzel K, et al. A randomized trial of deep-brain stimulation for Parkinson's disease. *N Engl J Med*. 2006;355(9):896-908.

89. Guehl D, Cuny E, Benazzouz A, Rougier A, Tison F, Machado S, et al. Side-effects of subthalamic stimulation in Parkinson's disease: clinical evolution and predictive factors. *Eur J Neurol*. 2006;13(9):963-71.

90. Volkmann J, Moro E, Pahwa R. Basic algorithms for the programming of deep brain stimulation in Parkinson's disease. *Mov Disord*. 2006;21 Suppl 14:S284-9.

91. Kurtis MM, Rajah T, Delgado LF, Dafsari HS. The effect of deep brain stimulation on the non-motor symptoms of Parkinson's disease: a critical review of the current evidence. *NPJ Parkinsons Dis*. 2017;3:16024.

92. Shulman LM, Taback RL, Rabinstein AA, Weiner WJ. Non-recognition of depression and other non-motor symptoms in Parkinson's disease. *Parkinsonism Relat Disord*. 2002;8(3):193-7.

93. Chaudhuri KR, Prieto-Jurcynska C, Naidu Y, Mitra T, Frades-Payo B, Tluk S, et al. The nondeclaration of nonmotor symptoms of Parkinson's disease to health care professionals: an international study using the nonmotor symptoms questionnaire. *Mov Disord*. 2010;25(6):704-9.

94. Hiseman JP, Fackrell R. Caregiver Burden and the Nonmotor Symptoms of Parkinson's Disease. *Int Rev Neurobiol*. 2017;133:479-97.

95. Martinez-Martin P, Rodriguez-Blazquez C, Abe K, Bhattacharyya KB, Bloem BR, Carod-Artal FJ, et al. International study on the psychometric attributes of the non-motor symptoms scale in Parkinson disease. *Neurology*. 2009;73(19):1584-91.

96. Knapp M, Mangalore R, Simon J. The global costs of schizophrenia. *Schizophr Bull*. 2004;30(2):279-93.

97. Drummond M, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW, Askews, et al. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2015.

98. Cunningham SJ. Economic evaluation of healthcare - is it important to us? *Br Dent J*. 2000;188(5):250-4.

99. Office for National Statistics. (2017). UK health Accounts: 2015. Healthcare expenditure statistics, produced to the international definitions of the System of Health Accounts 2011. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/bulletins/ukhealthaccounts/2015. Accessed 13 June 2017.

100. Office for National Statistics. (2015). Healthcare spending in the UK, 2013. http://webarchive.nationalarchives.gov.uk/20160105160709/http:/www.ons.gov.uk/ons/rel/psa/expenditure-on-healthcare-in-the-uk/2013/info-healthcare-spending-in-the-uk--2013.html. Accessed 14 June 2017.

101. Mooney G, Russell EM, Weir RD. Choices for health care: a practical introduction to the economics of health provision. 2nd ed. London: Macmillan; 1986.

102. Palmer S, Raftery J. Opportunity cost. *BMJ*. 1999;318(7197):1551-2.

103. Debreu G. Theory of value: an axiomatic analysis of economic equilibrium. New York: Wiley; 1959.

104. Sloman J. Economics. 6th ed. Harlow: Financial Times Prentice Hall; 2006.

105. Drummond MF. Output measurement for resource allocation decisions in health care. *Oxf Rev Econ Policy*. 1989;5(1):59-74.

106. Palmer GR, Ho MT. Health economics: a critical and global analysis. Basingstoke: Palgrave Macmillan; 2008.

107. Shiell A, Donaldson C, Mitton C, Currie G. Health economic evaluation. *J Epidemiol Community Health*. 2002;56(2):85.

108. Office of Health Economics. (2007). The economics of health care. https://www.ohe.org/sites/default/files/TheEconomicsofHeathCare2007.pdf. Accessed 03 Jan 2018.

109. Bator FM. The Anatomy of Market Failure. *Q J Econ*. 1958;72(3):351-79.

110. Arrow KJ. Uncertainty and the welfare economics of medical care. 1963. *Bull World Health Organ*. 2004;82(2):141-9.

111. Elliott R, Payne K. Essentials of Economic Evaluation in Healthcare. London, UK: Pharmaceutical Press; 2005.

112. Donaldson C, Gerard K. Market Failure in Health Care. Economics of Health Care Financing: The Visible Hand. London: Macmillan Education UK; 1993. p. 26-48.

113. Mwachofi A, Al-Assaf AF. Health Care Market Deviations from the Ideal Market. *Sultan Qaboos Univ Med J*. 2011;11(3):328-37.

114. Drummond MF. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford: Oxford University Press; 2005.

115. World Health Organisation. (2017). Technology, Health. http://www.who.int/topics/technology_medical/en/. Accessed 9 August 2017.

116. Kobelt G. Health economics: an introduction to economic evaluation: Office of Health Economics; 2013.

117. National Institute for Health and Care Excellence. (2012). Methods for the development of NICE public health guidance (third edition) [PMG4]. https://www.nice.org.uk/process/pmg4/chapter/introduction. Accessed 3 January 2018.

118. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press Inc.; 2017.

119. National Institute for Health and Care Excellence. (2013). Guide to the methods of technology appraisal [PMG9]. https://www.nice.org.uk/process/pmg9/chapter/the-reference-case. Accessed 9 August 2017.

120. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. *Value Health*. 2009;12 Suppl 1:S5-9.

121. Brooks R. (2015). 28 Years of the EuroQol Group: An Overview.

122. National Institute of Health and Care Excellence. (2013). 5.3. Measuring and valuing health effects - 5.3.1. Guide to the methods of technology appraisal 2013 [PMG9]. https://www.nice.org.uk/process/pmg9/chapter/the-reference-case#measuring-and-valuing-health-effects. Accessed 3 August 2017.

123. Devlin NJ, Brooks R. EQ-5D and the EuroQol Group: Past, Present and Future. *Appl Health Econ Health Policy*. 2017;15(2):127-37.

124. Robinson R. Cost-effectiveness analysis. *BMJ*. 1993;307(6907):793-5.

125. Lorgelly PK. Choice of Outcome Measure in an Economic Evaluation: A Potential Role for the Capability Approach. *Pharmacoeconomics*. 2015;33(8):849-55.

126. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*. 2001;33:337-43.

127. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ*. 2004;13(9):873-84.

128. Mulhern B, Mukuria C, Barkham M, Knapp M, Byford S, Soeteman D, et al. Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D. *Br J Psychiatry*. 2014;205(3):236-43.

129. McIntosh E, Clarke PM, Frew EJ, Louviere JJ. Applied methods of cost-benefit analysis in health care. Oxford: Oxford University Press; 2010.

130. Fox M. (2012). 21 motivational Michael J Fox quotes on living with Parkinson's disease. http://parkinsonslife.eu/21-motivational-michael-j-fox-quotes-on-living-with-parkinsons-disease/. Accessed 3 January 2018.

131. Jenkinson C, Fitzpatrick R, Peto V, Greenhall R, Hyman N. The Parkinson's Disease Questionnaire (PDQ-39): development and validation of a Parkinson's disease summary index score. *Age Ageing*. 1997;26(5):353-7.

132. Welsh M, McDermott MP, Holloway RG, Plumb S, Pfeiffer R, Hubble J. Development and testing of the Parkinson's disease quality of life scale. *Mov Disord*. 2003;18(6):637-45.

133. de Boer AG, Wijker W, Speelman JD, de Haes JC. Quality of life in patients with Parkinson's disease: development of a questionnaire. *J Neurol Neurosurg Psychiatry*. 1996;61(1):70-4.

134. Calne S, Schulzer M, Mak E, Guyette C, Rohs G, Hatchard S, et al. Validating a quality of life rating scale for idiopathic parkinsonism: Parkinson's Impact Scale (PIMS). *Parkinsonism Relat Disord*. 1996;2:55-61.

135. Sen A. Commodities and capabilities. Delhi: Oxford University Press; 1999.

136. Sen A. Capability and well-being. In M. C. Nussbaum (Ed.), The quality of life. Oxford: Clarendon Press. 1993.

137. Coast J, Smith R, Lorgelly P. Should the capability approach be applied in health economics? *Health Econ*. 2008;17(6):667-70.

138. Coast J, Smith RD, Lorgelly P. Welfarism, extra-welfarism and capability: the spread of ideas in health economics. *Soc Sci Med*. 2008;67(7):1190-8.

139. Sen A. Inequality reexamined. New York: Russell Sage Foundation; Clarendon Press; 1992.

140. Grewal I, Lewis J, Flynn T, Brown J, Bond J, Coast J. Developing attributes for a generic quality of life measure for older people: preferences or capabilities? *Soc Sci Med*. 2006;62(8):1891-901.

141. Coast J, Flynn TN, Natarajan L, Sproston K, Lewis J, Louviere JJ, et al. Valuing the ICECAP capability index for older people. *Soc Sci Med*. 2008;67(5):874-82.

142. National Institute for Health and Care Excellence. (2013). The social care guidance manual [PMG10]. 7 Incorporating economic evaluation. https://www.nice.org.uk/process/pmg10/chapter/incorporating-economic-evaluation. Accessed 3 January 2018.

143. Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework; Committee on National Statistics; Division on Behavioral and Social Sciences and Education; National Research Council. 4.3. sensitivity of ExWB measures to changing conditions. In: AA S, Mackie C, editors. Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience. Washington (DC): National Academies Press (US); 2013.

144. Eggington S, Valldeoriola F, Chaudhuri KR, Ashkan K, Annoni E, Deuschl G. The cost-effectiveness of deep brain stimulation in combination with best medical therapy, versus best medical therapy alone, in advanced Parkinson's disease. *J Neurol*. 2014;261(1):106-16.

145. Dams J, Siebert U, Bornschein B, Volkmann J, Deuschl G, Oertel WH, et al. Cost-effectiveness of deep brain stimulation in patients with Parkinson's disease. *Mov Disord*. 2013;28(6):763-71.

146. Walter E, Odin P. Cost-effectiveness of continuous subcutaneous Apomorphine in the treatment of Parkinson s Disease in the UK and Germany. *J Med Econ*. 2014:1-36.

147. Tomaszewski KJ, Holloway RG. Deep brain stimulation in the treatment of Parkinson's disease: a cost-effectiveness analysis. *Neurology*. 2001;57(4):663-71.

148. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics. *Health Econ*. 1999;8(1):41-51.

149. Quercioli C, Messina G, Barbini E, Carriero G, Fani M, Nante N. Importance of sociodemographic and morbidity aspects in measuring health-related quality of life: performances of three tools: comparison of three questionnaire scores. *Eur J Health Econ*. 2009;10(4):389-97.

150. Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. *Qual Life Res*. 2005;14(7):1651-68.

151. Joore M, Brunenberg D, Nelemans P, Wouters E, Kuijpers P, Honig A, et al. The Impact of Differences in EQ-5D and SF-6D Utility Scores on the Acceptability of Cost–Utility Ratios: Results across Five Trial-Based Cost–Utility Studies. *Value Health*. 2010;13(2):222-9.

152. Sach TH, Barton GR, Jenkinson C, Doherty M, Avery AJ, Muir KR. Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter? *Med Care*. 2009;47(8):889-94.

153. Brooks R. EuroQol: the current state of play. *Health Policy*. 1996;37(1):53-72.

154. Sculpher M, Claxton K. Sins of omission and obfuscation: IQWIG's guidelines on economic evaluation methods. *Health Econ*. 2010;19(10):1132-6.

155. Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making*. 1989;9(2):142-9.

156. Garau M, Shah KK, Mason AR, Wang Q, Towse A, Drummond MF. Using QALYs in cancer: a review of the methodological limitations. *Pharmacoeconomics*. 2011;29(8):673-85.

157. Knapp M, Mangalore R. "The trouble with QALYs...". *Epidemiol Psichiatr Soc*. 2007;16(4):289-93.

158. Kind P, Lafata JE, Matuszewski K, Raisch D. The use of QALYs in clinical and patient decision-making: issues and prospects. *Value Health*. 2009;12.

159. Woods B, Revill P, Sculpher M, Claxton K. Country-Level Cost-Effectiveness Thresholds: Initial Estimates and the Need for Further Research. *Value Health*. 2016;19(8):929-35.

160. National Institute of Health and Care Excellence. (2013). 6.3 Decision-making. Guide to the methods of technology appraisal 2013 [PMG9]. https://www.nice.org.uk/process/pmg9/chapter/the-appraisal-of-the-evidence-and-structured-decision-making#decision-making. Accessed 3 August 2017.

161. BBC. (2015). NICE 'sets price too high for NHS medicines'. http://www.bbc.co.uk/news/health-31507861. Accessed 7 Feb 2018.

162. The Guardian. (2015). Patients suffer when NHS buys expensive new drugs, says report. https://www.theguardian.com/society/2015/feb/19/nhs-buys-expensive-new-drugs-nice-york-karl-claxton-nice. Accessed 7 Feb 2018.

163. Office of Health Economics. (2015). OHE Occasional Paper Critiques the Claxton et al. £13,000 per QALY Estimate. https://www.ohe.org/news/ohe-occasional-paper-critiques-claxton-et-al-%C2%A313000-qaly-estimate. Accessed 7 Feb 2018.

164. Barnsley P, Towse A, Karlsberg Schaffer S, Sussex J. (2013). Critique of CHE Research Paper 81: Methods for the Estimation of the NICE Cost Effectiveness Threshold.

165. Dillon A. (2015). Carrying NICE over the threshold Sir Andrew Dillon. https://www.nice.org.uk/news/blog/carrying-nice-over-the-threshold. Accessed 7 Feb 2018.

166. Frew EJ, Whynes DK, Wolstenholme JL. Eliciting willingness to pay: comparing closed-ended with open-ended and payment scale formats. *Med Decis Making*. 2003;23(2):150-9.

167. Carson RT, Flores NE, Martin KM, Wright JL. Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods. *Land Economics*. 1996;72(1):80-99.

168. McIntosh E, Donaldson C, Ryan M. Recent advances in the methods of cost-benefit analysis in healthcare. Matching the art to the science. *Pharmacoeconomics*. 1999;15(4):357-67.

169. Brazier J, Tsuchiya A. Improving Cross-Sector Comparisons: Going Beyond the Health-Related QALY. *Appl Health Econ Health Policy*. 2015;13(6):557-65.

170. Olsen JA, Smith RD. Theory versus practice: a review of 'willingness-to-pay' in health and health care. *Health Econ*. 2001;10(1):39-52.

171. Lorgelly PK, Lawson KD, Fenwick EA, Briggs AH. Outcome measurement in economic evaluations of public health interventions: a role for the capability approach? *Int J Environ Res Public Health*. 2010;7(5):2274-89.

172. Huber M, Knottnerus JA, Green L, van der Horst H, Jadad AR, Kromhout D, et al. How should we define health? *BMJ*. 2011;343:d4163.

173. Huber M. Towards a new, dynamic concept of Health. Its operationalisation and use in public health and healthcare, and in evaluating health effects of food. ISBN 978-94-6259-471-5. 2014.

174. Osborne R. The History Written on the Classical Greek Body. Cambridge, New York: Cambridge University Press; 2011.

175. Shorter E. Doctors and their patients, a social history. New Brunswick, New Jersey: Transaction Publishers; 1991.

176. World Health Organisation. Constitution of WHO: principles WHO.

177. World Health Organisation. (1986). The Ottawa Charter for Health Promotion. . http://www.who.int/healthpromotion/conferences/previous/ottawa/en/. Accessed 19 July 2017.

178. Larson JS. The World Health Organization's definition of health: Social versus spiritual health. *Soc Indic Res*. 1996;38(2):181-92.

179. Evans RG, Wolfson AD, Economics UoBCDo. Faith, Hope, and Charity: Health Care in the Utility Function: Department of Economics, University of British Columbia; 1980.

180. Evans RG, Wolfson AD. Faith, hope and charity: health care in the utility function. Department of Economics, University of British Columbia and department of Health Administration, University of Toronto, unpublished paper. 1980.

181. Brazier J, Roberts J. Methods for developing preference-based measures of health. In: Jones MA, editor. The Elgar Companion to Health Economics. Glos, UK: Edward Elgar Publishing, Inc.; 2006.

182. Syme SL. Rethinking disease: where do we go from here? *Ann Epidemiol*. 1996;6(5):463-8.

183. Institute of Medicine (US) Committee on Using Performance Monitoring to Improve Community Health. 2 Understanding Health and Its Determinants. In: Durch J, Bailey L, Stoto M, editors. Improving Health in the Community: A Role for Performance Monitoring. Washington (DC): National Academies Press (US); 1997.

184. Donaldson C, Atkinson A, Bond J, Wright K. Should QALYs be programme-specific? *J Health Econ*. 1988;7(3):239-57.

185. Jadad AR, O'Grady L. How should health be defined? *BMJ*. 2008;337.

186. Smith R. (2008). The BMJ Opinion. The end of disease and the beginning of health. http://blogs.bmj.com/bmj/2008/07/08/richard-smith-the-end-of-disease-and-the-beginning-of-health/. Accessed 19 July 2017.

187. World Health Organisation. The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. *Soc Sci Med*. 1995;41:1403–09.

188. Post MW. Definitions of quality of life: what has happened and how to move on. *Top Spinal Cord Inj Rehabil*. 2014;20(3):167-80.

189. Hartge P. A Dictionary of Epidemiology, Sixth Edition Edited by Miquel Porta. *Am J Epidemiol*. 2015;181(8):633-4.

190. Bowling A. Measuring health: a review of quality of life measurement scales. 3rd ed. Maidenhead, Berkshire: Open University Press; 2005.

191. Lam CLK. Subjective Quality of Life Measures – General Principles and Concepts. In: Preedy VR, Watson RR, editors. Handbook of Disease Burdens and Quality of Life Measures. New York: Springer; 2010. p. 381-99.

192. Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. Oxford; New York: Oxford University Press; 1993.

193. Miravitlles M, Ferrer J, Baró E, Lleonart M, Galera J. Differences between physician and patient in the perception of symptoms and their severity in COPD. *Respir Med*. 2013;107(12):1977-85.

194. Sewitch MJ, Abrahamowicz M, Dobkin PL, Tamblyn R. Measuring differences between patients' and physicians' health perceptions: the patient-physician discordance scale. *J Behav Med*. 2003;26(3):245-64.

195. Jachuck SJ, Brierley H, Jachuck S, Willcox PM. The effect of hypotensive drugs on the quality of life. *J R Coll Gen Pract*. 1982;32(235):103-5.

196. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess*. 1998;2(14):i-iv, 1-74.

197. Spitzer WO. State of science 1986: quality of life and functional status as target variables for research. *J Chronic Dis*. 1987;40(6):465-71.

198. Ware JE, Jr. Standards for validating health measures: definition and content. *J Chronic Dis*. 1987;40(6):473-80.

199. Torrance GW. Utility approach to measuring health-related quality of life. *J Chronic Dis*. 1987;40.

200. Ebrahim S. Clinical and public health perspectives and applications of health-related quality of life measurement. *Soc Sci Med*. 1995;41(10):1383-94.

201. Leplege A, Hunt S. The problem of quality of life in medicine. *JAMA*. 1997;278(1):47-50.

202. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med*. 1993;118.

203. Neumann PJ, Ganiats TG, Russell LB, Sanders GD, Siegel JE. Cost-effectiveness in health and medicine. Second ed. New York, NY: Oxford University Press; 2016.

204. Prutkin JM, Feinstein AR. Quality-of-life measurements: origin and pathogenesis. *Yale J Biol Med*. 2002;75(2):79-93.

205. Karnofsky D, A., Burchenal J, H. In: Evaluation of chemotherapeutic agents. MacLeod CM, editor. New York: Columbia University Press; 1949. The clinical evaluation of chemotherapeutic agents in cancer; pp. 191–205.

206. Campbell A, Converse PE, Rodgers WL. The quality of American life: perceptions, evaluations, and satisfactions. New York: Russell Sage Foundation; 1976.

207. Carlens E, Dahlstrom G, Nou E. Comparative measurements of quality of survival of lung cancer patients after diagnosis. *Scand J Respir Dis*. 1970;51(4):268-75.

208. Tofler OB. Life units. A discussion in the Department of Cardiology, Royal Perth Hospital, Australia. *Br Heart J*. 1970;32(6):771-3.

209. Priestman TJ, Baum M. Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet*. 1976;1(7965):899-900.

210. Bowling A. Measuring disease: a review of disease-specific quality of life measurement scales. 2nd ed. Ballmoor, UK.: Open University Press; 2001.

211. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473-83.

212. Rowen D, Brazier J, Young T, Gaugris S, Craig BM, King MT, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value Health*. 2011;14(5):721-31.

213. MacKillop E. (2017). Valuing life: Exploring the history of Quality-Adjusted Life-Years (QALY) 2017. https://remedianetwork.net/2017/05/03/valuing-life-exploring-the-history-of-quality-adjusted-life-years-qaly/. Accessed 29 Jan 2018.

214. Klarman HE, John O'S F, Rosenthal GD. Cost Effectiveness Analysis Applied to the Treatment of Chronic Renal Disease. *Medical Care*. 1968;6(1):48-54.

215. York Health Economics Consortium. (2016). Utility [online]. http://www.yhec.co.uk/glossary/utility/. Accessed.

216. Marshall A. Principles of economics. 8th ed. Basingstoke: Palgrave Macmillan; 2013.

217. the EuroQol Group. About EQ-5D. http://www.euroqol.org/about-eq-5d.html. Accessed 03 March 2017.

218. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull*. 2010;96:5-21.

219. Wisloff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. *Pharmacoeconomics*. 2014;32(4):367-75.

220. Mongin P. (1997). Expected utility theory. https://studies2.hec.fr/jahia/webdav/site/hec/shared/sites/mongin/acces_anonyme/page%20internet/O12.MonginExpectedHbk97.pdf. Accessed 25 July 2017.

221. Glick H, Doshi JA, Sonnad SS, Polsky D. Economic evaluation in clinical trials. Second ed. Oxford: Oxford University Press; 2015.

222. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making*. 2001;21(4):329-34.

223. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. *Med Decis Making*. 1997;17(2):208-16.

224. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. *Med Decis Making*. 2001;21(1):17-27.

225. Parducci A, Wedell DH. The category effect with rating scales: number of categories, number of stimuli, and method of presentation. *J Exp Psychol Hum Percept Perform*. 1986;12(4):496-516.

226. Bell DE, Farquhar PH. Perspectives on Utility Theory. *Operations Research*. 1986;34(1):179-83.

227. Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1944.

228. Torrance GW. Social preferences for health states: An empirical evaluation of three measurement techniques. *Socioecon Plann Sci*. 1976;10(3):129-36.

229. Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care*. 1989;5(4):559-75.

230. Finnell SME, Carroll AE, Downs SM. The Utility Assessment Method Order Influences Measurement of Parents' Risk Attitude. *Value Health*. 2012;15(6):926-32.

231. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. *Health Serv Res*. 1972;7(2):118-33.

232. van der Pol M, Roux L. Time preference bias in time trade-off. *Eur J Health Econ*. 2005;6(2):107-11.

233. Brazier J, Rowen D, Yang Y, Tsuchiya A. (2009). Using rank and and discrete choice data to estimate health state utility values on the QALY scale. HEDS Discussion paper 09/10.

234. Brazier J. Is the EQ-5D fit for purpose in mental health? *Br J Psychiatry*. 2010;197(5):348-9.

235. Longworth L, Yang Y, Young T, Mulhern B, Hernandez Alava M, Mukuria C, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technol Assess*. 2014;18(9):1-224. doi: 10.3310/hta18090.

236. Riepe MW, Mittendorf T, Forstl H, Frolich L, Haupt M, Leidl R, et al. Quality of life as an outcome in Alzheimer's disease and other dementias--obstacles and goals. *BMC Neurology*. 2009;9:47.

237. Brazier J, Rowen D, Yang Y, Tsuchiya A. Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. *Eur J Health Econ*. 2012;13(5):575-87.

238. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ*. 2002;21.

239. The University of Sheffield. (2017). The SF-6D: A new, internationally adopted measure for assessing the cost- effectiveness of health care interventions. https://www.sheffield.ac.uk/economics/research/impact/sf6d. Accessed.

240. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med*. 2001;33(5):375-84.

241. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Qual Life Outcomes*. 2003;1:54.

242. Kind P. The EuroQol Instrument: An Index of Health-Related Quality of Life. In: Spilker B (ed.). *Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd edn Lippincott-Rivera, Philadelphia, PA, pp191-201*. 1996.

243. the EuroQol Group. EQ-5D-3L Self-complete version on paper. http://www.euroqol.org/eq-5d-products/eq-5d-3l/self-complete-version-on-paper.html. Accessed 13 March 2017.

244. Xie F, Gaebel K, Perampaladas K, Doble B, Pullenayegum E. Comparing EQ-5D valuation studies: a systematic review and methodological reporting checklist. *Med Decis Making*. 2014;34(1):8-20.

245. Dolan P. Modeling valuations for EuroQol health states. *Med Care*. 1997;35.

246. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. *Value Health*. 2011;14(6):907-20.

247. Garau M SK, Towse A, Wang Q, Drummond M, Mason A. (2009). Assessment and appraisal of oncology medicines: does NICE's approach include all relevant elements? What can be learnt from international HTA experiences? Report for the Pharmaceutical Oncology Initiative (POI). Office of Health Economics. London. UK. https://www.ohe.org/publications/assessment-and-appraisal-oncology-medicines-nices-approach-and-international-hta. Accessed 16 August 2015.

248. Hounsome N, Orrell M, Edwards RT. EQ-5D as a quality of life measure in people with dementia and their carers: evidence and key issues. *Value Health*. 2011;14(2):390-9.

249. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: Validity, responsiveness and reliability of EuroQol (EQ-5D). *Br J Rheumatol*. 1997;36(5):551-9.

250. Luo N, Low S, Lau PN, Au WL, Tan LC. Is EQ-5D a valid quality of life instrument in patients with Parkinson's disease? A study in Singapore. *Ann Acad Med Singapore*. 2009;38(6):521-8.

251. Reuther M, Spottke EA, Klotsche J, Riedel O, Peter H, Berger K, et al. Assessing health-related quality of life in patients with Parkinson's disease in a prospective longitudinal study. *Parkinsonism Relat Disord*. 2007;13(2):108-14.

252. Feng Y, Devlin N, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? *Health Qual Life Outcomes*. 2015;13(1):171.

253. Ringbaek T, Brondum E, Martinez G, Lange P. EuroQoL in assessment of the effect of pulmonary rehabilitation COPD patients. *Respir Med*. 2008;102(11):1563-7.

254. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. *Health Qual Life Outcomes*. 2014;12:74.

255. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-36.

256. Wailoo A, Davis S, Tosh J. (2010). The incorporation of health benefits in cost utility analysis using the EQ-5D. Report by the decision support unit http://www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D%20final%20report%20-%20submitted.pdf. Accessed 20 July 2015.

257. van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health*. 2012;15(5):708-15.

258. Devlin N, Shah K, Feng Y, Mulhern B, Van Hout B. (Jan 2016). Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England.

259. National Institute for Health and Care Excellence. (2017). Position statement on use of the EQ-5D-5L valuation set. https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf. Accessed 30 November 2017.

260. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res*. 2013;22(7):1717-27.

261. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care*. 2007;45(3):259-63.

262. Lin F, Longworth L, Pickard A. Evaluation of content on EQ-5D as compared to disease-specific utility measures. *Qual Life Res*. 2013;22(4):853-74.

263. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F, Gaeta L, et al. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. *Qual Life Res*. 2013;22(7):1707-16.

264. Rosser RM, Watts VC. The measurement of hospital output. *Int J Epidemiol*. 1972;1(4):361-8.

265. Connell J, Brazier J, O'Cathain A, Lloyd-Jones M, Paisley S. Quality of life of people with mental health problems: a synthesis of qualitative research. *Health Qual Life Outcomes*. 2012;10:138.

266. National Health Executive. (2017). NICE to review quality of life measures across health and social care. http://www.nationalhealthexecutive.com/Health-Care-News/nice-to-review-quality-of-life-measures-across-health-and-social-care. Accessed 3 August 2017.

267. Moock J, Kohlmann T. Comparing preference-based quality-of-life measures: results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. *Qual Life Res*. 2008;17(3):485-95.

268. Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. *Health Econ*. 2008;17(7):815-32.

269. McDonough CM, Grove MR, Tosteson TD, Lurie JD, Hilibrand AS, Tosteson AN. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants. *Qual Life Res*. 2005;14(5):1321-32.

270. Macran S, Weatherly H, Kind P. Measuring population health: a comparison of three generic health status measures. *Med Care*. 2003;41(2):218-31.

271. Richardson J, Khan MA, Iezzi A, Maxwell A. Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Med Decis Making*. 2015;35(3):276-91.

272. Xie F, Li SC, Luo N, Lo NN, Yeo SJ, Yang KY, et al. Comparison of the EuroQol and short form 6D in Singapore multiethnic Asian knee osteoarthritis patients scheduled for total knee replacement. *Arthritis Rheum*. 2007;57(6):1043-9.

273. Wailoo A, Alava MH, Grimm S, Pudney S, Gomes M, Sadique Z, et al. (2017). Comparing the EQ-5D-3L and 5L versions. What are the implications for cost effectiveness estimates? Report by the decision support unit.

274. Mott DJ, Najafzadeh M. Whose preferences should be elicited for use in health-care decision-making? A case study using anticoagulant therapy. *Expert Rev Pharmacoecon Outcomes Res*. 2016;16(1):33-9.

275. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res*. 2003;12(6):599-607.

276. Rowen D, Mulhern B, Banerjee S, Tait R, Watchurst C, Smith SC, et al. Comparison of general population, patient, and carer utility values for dementia health states. *Med Decis Making*. 2015;35(1):68-80.

277. Ratcliffe J, Brazier J, Palfreyman S, Michaels J. A comparison of patient and population values for health states in varicose veins patients. *Health Econ*. 2007;16(4):395-405.

278. De Wit GA, Busschbach JJ, De Charro FT. Sensitivity and perspective in the valuation of health status: whose values count? *Health Econ*. 2000;9(2):109-26.

279. Stamuli E. Health outcomes in economic evaluation: who should value health? *Br Med Bull*. 2011;97:197-210.

280. Insinga RP, Fryback DG. Understanding differences between self-ratings and population ratings for health in the EuroQOL. *Qual Life Res*. 2003;12(6):611-9.

281. Kahneman D. Determinants of health economic decisions in actual practice: the role of behavioral economics. Summary of the presentation given by Professor Daniel Kahneman at the ISPOR 10th Annual International Meeting First Plenary Session, May 16, 2005, Washington, DC, USA. *Value Health*. 2006;9(2):65-7.

282. Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, et al. The MAPS Reporting Statement for Studies Mapping onto Generic Preference-Based Outcome Measures: Explanation and Elaboration. *Pharmacoeconomics*. 2015;33(10):993-1011.

283. National Institute of Health and Care Excellence. (2013). 5.3. Measuring and valuing health effects - 5.3.9. Guide to the methods of technology appraisal 2013 [PMG9]. https://www.nice.org.uk/process/pmg9/chapter/the-reference-case#measuring-and-valuing-health-effects. Accessed 3 August 2017.

284. Rivero-Arias O, Dakin H, Gray A. (2016). Mapping algorithms from non-preference to preference-based outcome measures: do they really work in practice? https://www.ndph.ox.ac.uk/study/dphil-population-health-2016-entry/2016-DPhil-research-projects-list/mapping-algorithms-from-non-preference-to-preference-based-outcome-measures-do-they-really-work-in-practice. Accessed 15March2017.

285. Dakin H, Burns R, Y. Y. (2016). HERC database of mapping studies, Version 5.0 (Last updated: 16th May 2016). http://www.herc.ox.ac.uk/downloads/herc-database-of-mapping-studies. Accessed 15March2017.

286. Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health Qual Life Outcomes*. 2013;11:151.

287. Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, et al. Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement. *J Med Econ*. 2015;18(11):851-7.

288. Yang Y, Brazier JE, Tsuchiya A, Young TA. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. *Med Decis Making*. 2011;31(2):281-91.

289. Versteegh MM, Leunis A, Uyl-de Groot CA, Stolk EA. Condition-specific preference-based measures: benefit or burden? *Value Health*. 2012;15.

290. Brazier J, Rowen D, Tsuchiya A, Yang Y, Young TA. The impact of adding an extra dimension to a preference-based measure. *Soc Sci Med*. 2011;73(2):245-53.

291. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability? *Health Econ*. 2010;19(2):125-9.

292. Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol*. 1999;52(4):293-301.

293. McTaggart-Cowan H. Elicitation of informed general population health state utility values: a review of the literature. *Value Health*. 2011;14(8):1153-7.

294. Dolan P, Kahneman D. Interpretations Of Utility And Their Implications For The Valuation Of Health*. *The Economic Journal*. 2008;118(525):215-34.

295. Ryan RM, Deci EL. On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu Rev Psychol*. 2001;52:141-66.

296. Netten A, Burge P, Malley J, Potoglou D, Towers AM, Brazier J, et al. Outcomes of social care for adults: developing a preference-weighted measure. *Health Technol Assess*. 2012;16(16):1-166.

297. Flynn TN, Huynh E, Peters TJ, Al-Janabi H, Clemens S, Moody A, et al. Scoring the Icecap-a Capability Instrument. Estimation of a UK General Population Tariff. *Health Econ*. 2015;24:258-69.

298. Dolan P, Peasgood T, White M. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *J Econ Psychol*. 2008;29(1):94-122.

299. Smith DM, Brown SL, Ubel PA. Are subjective well-being measures any better than decision utility measures? *Health Econ Policy Law*. 2008;3(Pt 1):85-91.

300. Sampson CJ, Wailoo A, M. H. (2012). Subjective well-being and generic preference-based measures of health: an empirical contribution.

301. Bleichrodt H, Quiggin J. Capabilities as menus: A non-welfarist basis for QALY evaluation. *J Health Econ*. 2013;32(1):128-37.

302. Sugden R. Welfare, Resources, and Capabilities: A Review of Inequality Reexamined by Amartya Sen.31(4):1947-62.

303. Cookson R. QALYs, and the capability approach. *Health Econ*. 2005;14(8):817-29.

304. Karimi M, Brazier J, Basarir H. The Capability Approach: A Critical Review of Its Application in Health Economics. *Value Health*. 2016;19(6):795-9.

305. Keeley T. Capability as an outcome measure in randomised controlled trials. Birmingham: University of Birmingham; 2014.

306. Dang A-T. Amartya Sen's Capability Approach: A Framework for Well-Being Evaluation and Policy Analysis? *Rev Soc Econ*. 2014;72(4):460-84.

307. Sugden R. Welfare, Resources, and Capabilities: A Review of Inequality Reexamined by Amartya Sen. *J Econ Lit*. 1993;31(4):1947-62.

308. Culyer A. The Normative Economics of Health Care Finance and Provision. *Oxf Rev Econ Policy*. 1989;5(1):34-58.

309. Brouwer WBF, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. *J Health Econ*. 2008;27(2):325-38.

310. Birch S, Donaldson C. Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? *Soc Sci Med*. 2003;56(5):1121-33.

311. Coast J. Is economic evaluation in touch with society's health values? *BMJ*. 2004;329(7476):1233-6.

312. Coast J, Peters TJ, Natarajan L, Sproston K, Flynn T. An assessment of the construct validity of the descriptive system for the ICECAP capability measure for older people. *Qual Life Res*. 2008;17(7):967-76.

313. Makai P, Koopmanschap MA, Brouwer WB, Nieboer AA. A validation of the ICECAP-O in a population of post-hospitalized older people in the Netherlands. *Health Qual Life Outcomes*. 2013;11:57.(doi).

314. Couzner L, Ratcliffe J, Lester L, Flynn T, Crotty M. Measuring and valuing quality of life for public health research: application of the ICECAP-O capability index in the Australian general population. *Int J Public Health*. 2012;58:367-76.

315. Couzner L, Crotty M, Norman R, Ratcliffe J. A comparison of the EQ-5D-3L and ICECAP-O in an older post-acute patient population relative to the general population. *Appl Health Econ Health*. 2013;11:415-25.

316. Davis JC, Bryan S, McLeod R, Rogers J, Khan K, Liu-Ambrose T. Exploration of the association between quality of life, assessed by the EQ-5D and ICECAP-O, and falls risk, cognitive function and daily function, in older adults with mobility impairments. *BMC geriatrics*. 2012;12:65.

317. Makai P, Beckebans F, van Exel J, Brouwer WBF. Quality of life of nursing home residents with dementia: validation of the German version of the ICECAP-O. *PloS one*. 2014;9:e92016.

318. Horwood J, Sutton E, Coast J. Evaluating the Face Validity of the ICECAP-O Capabilities Measure: A “Think Aloud” Study with Hip and Knee Arthroplasty Patients. *Appl Res Qual Life*. 2013;9:667-82.

319. van Leeuwen KM, Bosmans JE, Jansen APD, Hoogendijk EO, van Tulder MW, van der Horst HE, et al. Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. *Value Health*. 2015;18:35-43.

320. Kinghorn P. Use of ICECAP in decision making. http://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/HE/ICECAP/decision-making/index.aspx. Accessed 9 August 2017.

321. Mitchell PM, Roberts TE, Barton PM, Coast J. Assessing sufficient capability: A new approach to economic evaluation. *Soc Sci Med*. 2015;139:71-9.

322. Kinghorn P. (2017). Using deliberative methods to establish a sufficient level of capability well-being for use in decision-making in the contexts of public health and social care.

323. National Institute of Health and Care Excellence. (2013). 5.3. Measuring and valuing health effects - 5.3.10. Guide to the methods of technology appraisal 2013 [PMG9]. https://www.nice.org.uk/process/pmg9/chapter/the-reference-case#measuring-and-valuing-health-effects. Accessed 15 August 2017.

324. Carmines EG, Zeller RA. Reliability and Validity Assessment. Beverly Hills, CA: Sage.; 1979.1979.

325. Bowling A, Ebooks Corporation L. Research methods in health: investigating health and health services. Fourth ed. Maidenhead, Berkshire: Open University Press; 2014.

326. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19(4):539-49.

327. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. (2012). COSMIN checklist manual. http://www.cosmin.nl/COSMIN%20checklist.html. Accessed 22 January 2016.

328. Diamantopoulos A, Riefler P, Roth KP. Advancing formative measurement models. *J Bus Res*. 2008;61(12):1203-18.

329. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. *J R Stat Soc Series A Stat Soc*. 2002;165(2):233-53.

330. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *J Pers Assess*. 2003;80(3):217-22.

331. Avila ML, Stinson J, Kiss A, Brandão LR, Uleryk E, Feldman BM. A critical review of scoring options for clinical measurement tools. *BMC Res Notes*. 2015;8:612.

332. Bowling A. Research methods in health: investigating health and health services: McGraw-Hill Education 2009; 2009.

333. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012;22(3):276-82.

334. Hamilton B, Whiteley R, Almusa E, Roger B, Geertsema C, Tol JL. Excellent reliability for MRI grading and prognostic parameters in acute hamstring injuries. *British Journal of Sports Medicine*. 2014;48(18):1385.

335. Fielitz L, Coelho J, Horne T, Brechue W. Inter-Rater Reliability and Intra-Rater Reliability of Assessing the 2-Minute Push-Up Test. *Mil Med*. 2016;181(2):167-72.

336. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press; 2011.

337. Krabbe P. The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. Saint Louis: Elsevier; 2016.

338. Ferguson L. External validity, generalizability, and knowledge utilization. *J Nurs Scholarsh*. 2004;36(1):16-22.

339. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Fifth ed. Oxford: Oxford University Press; 2015.

340. Holden RR. Face Validity. The Corsini Encyclopedia of Psychology: John Wiley & Sons, Inc.; 2010.

341. Jull A. Evaluation of studies of assessment and screening tools, and diagnostic tests. *Evidence Based Nursing*. 2002;5(3):68.

342. Rogelberg SG. Encyclopedia of Industrial and Organizational Psychology: SAGE Publications; 2007.

343. Hanks GWC. Oxford textbook of palliative medicine. 4th ed. Oxford: Oxford University Press; 2010.

344. Uttl B. Measurement of individual differences: lessons from memory assessment in research and clinical practice. *Psychol Sci*. 2005;16(6):460-7. doi: 10.1111/j.0956-7976.2005.01557.x.

345. Martin PR, Bateson PPG. Measuring behaviour: an introductory guide. 3rd ed. Cambridge: Cambridge University Press; 2008.

346. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. *Health Qual Life Outcomes*. 2014;12:74.(doi):10.1186/477-7525-12-74.

347. Bharmal M, Thomas J, 3rd. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value Health*. 2006;9(4):262-71.

348. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ*. 2003;12(12):1061-7.

349. Ferreira LN, Ferreira PL, Pereira LN, Brazier J, Rowen D. A Portuguese Value Set for the SF-6D. *Value Health*. 2010;13(5):624-30.

350. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care*. 2004;42(9):851-9.

351. Jones AM, Ebooks Corporation L. The Elgar companion to health economics. Second ed. Cheltenham: Edward Elgar Publishing Limited; 2012.

352. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006.

353. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52(4):281-302.

354. Kane MT. Current Concerns in Validity Theory. *J Educ Meas*. 2001;38(4):319-42.

355. Messick S. Validity. In: Linn R, editor. Educational measurement. 3 ed. Washington, DC: American Council on Education / Macmillan; 1989. p. 13-103.

356. Zumbo BD, Chan EKH. Setting the Stage for Validity and Validation in Social, Behavioral, and Health Sciences: Trends in Validation Practices. In: Zumbo BD, Chan EKH, editors. Validity and Validation in Social, Behavioral, and Health Sciences. 1 ed: Springer International Publishing; 2014.

357. Cronbach LJ. Test validation. In: Thorndike RL, editor. Educational measurement. Washington, DC: American Council on Education; 1971. p. 221-37.

358. Landy FJ. Stamp Collecting Versus Science: Validation as Hypothesis Testing. *Am Psychol*. 1986;41(11):1183-92.

359. Messick S. Standards of Validity and the Validity of Standards in Performance Asessment. *Educ Meas Issues Pract*. 1995;14(4):5-8.

360. Grimm KJ, Widaman KF. Construct Validity. APA handbook of research methods in psychology. 1. Washington, DC, US.: American Psychological Association.; 2012. p. 621 - 42.

361. American Psychological A, American Educational Research A, Joint Committee on Standards for E, Psychological T, National Council on Measurement in E. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2004.

362. Zumbo BD. Validity as contextualized and pragmatic explanation, and its implications for validation practice. In: R.W. L, editor. The concept of validity: Revisions, new directions and applications. Charlotte, NC: Information Age; 2009. p. 65-82.

363. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J Consult Clin Psychol*. 1978;46(4):806-34.

364. Meehl PE. Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*. 1990;1(2):108-41.

365. Smith GT. On construct validity: issues of method and measurement. *Psychol Assess*. 2005;17(4):396-408.

366. Weimer WB. Notes on the methodology of scientific research. Hillsdale, NJ: Erlbaum; 1979.

367. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*. 1959;56(2):81-105.

368. Hattie J, Cooksey R. Procedures for Assessing the Validities of Tests Using the "Known-Groups" Method. *Applied Psychological Measurement*. 1984;8(3):295-305.

369. O'Leary-Kelly SW, J. Vokurka R. The empirical assessment of construct validity. *J Oper Manag*. 1998;16(4):387-405.

370. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. *Annu Rev Clin Psychol*. 2009;5:1-25.

371. Williams A. (1995). The role of the Euroqol instrument in QALY calculations. Discussion paper 130.

372. Davidson M. Known-Groups Validity. In: Michalos AC, editor. Encyclopedia of Quality of Life and Well-Being Research. Dordrecht: Springer Netherlands; 2014. p. 3481-2.

373. King MT, Costa DSJ, Aaronson NK, Brazier JE, Cella DF, Fayers PM, et al. QLU-C10D: a health state classification system for a multi-attribute utility measure based on the EORTC QLQ-C30. *Qual Life Res*. 2016;25(3):625-36.

374. Shah HA, Dritsaki M, Pink J, Petrou S. Psychometric properties of Patient Reported Outcome Measures (PROMs) in patients diagnosed with Acute Respiratory Distress Syndrome (ARDS). *Health Qual Life Outcomes*. 2016;14:13.

375. Sakthong P, Munpan W. A Head-to-Head Comparison of UK SF-6D and Thai and UK EQ-5D-5L Value Sets in Thai Patients with Chronic Diseases. *Appl Health Econ Health Policy*. 2017;15(5):669-79.

376. Mulhern B, Pink J, Rowen D, Borghs S, Butt T, Hughes D, et al. Comparing Generic and Condition-Specific Preference-Based Measures in Epilepsy: EQ-5D-3L and NEWQOL-6D. *Value Health*. 2017;20(4):687-93.

377. Finch AP, Dritsaki M, Jommi C. Generic Preference-based Measures for Low Back Pain: Which of Them Should Be Used? *Spine*. 2016;41(6):E364-E74.

378. Yang F, Lau T, Lee E, Vathsala A, Chia KS, Luo N. Comparison of the preference-based EQ-5D-5L and SF-6D in patients with end-stage renal disease (ESRD). *Eur J Health Econ*. 2015;16(9):1019-26.

379. Ratcliffe J, Flint T, Easton T, Killington M, Cameron I, Davies O, et al. An Empirical Comparison of the EQ-5D-5L, DEMQOL-U and DEMQOL-Proxy-U in a Post-Hospitalisation Population of Frail Older People Living in Residential Aged Care. *Appl Health Econ Health Policy*. 2017;15(3):399-412.

380. Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, et al. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technol Assess*. 2014;18(34):vii-viii, xiii-xxv, 1-188.

381. Stavem K, Bjornaes H, Lossius MI. Properties of the 15D and EQ-5D utility measures in a community sample of people with epilepsy. *Epilepsy Res*. 2001;44(2-3):179-89.

382. Maddigan SL, Feeny DH, Johnson JA. Construct validity of the RAND-12 and Health Utilities Index Mark 2 and 3 in type 2 diabetes. *Qual Life Res*. 2004;13(2):435-48.

383. Al-Janabi H, Peters TJ, Brazier J, Bryan S, Flynn TN, Clemens S, et al. An investigation of the construct validity of the ICECAP-A capability measure. *Qual Life Res*. 2013;22(7):1831-40.

384. Lorgelly PK, Doble B, Rowen D, Brazier J, Canc I. Condition-specific or generic preference-based measures in oncology? A comparison of the EORTC-8D and the EQ-5D-3L. *Qual Life Res*. 2017;26(5):1163-76.

385. Sarabia-Cobo CM, Paras-Bravo P, Amo-Setien FJ, Alconero-Camarero AR, Saenz-Jalon M, Torres-Manrique B, et al. Validation of the Spanish Version of the ICECAP-O for Nursing Home Residents with Dementia. *Plos One*. 2017;12(1):13.

386. Goranitis I, Coast J, Day E, Copello A, Freemantle N, Seddon J, et al. Measuring Health and Broader Well-Being Benefits in the Context of Opiate Dependence: The Psychometric Performance of the ICECAP-A and the EQ-5D-5L. *Value Health*. 2016;19(6):820-8.

387. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. *Qual Life Res*. 1992;1(1):73-5.

388. Patrick DL, Chiang YP. Measurement of health outcomes in treatment effectiveness evaluations: conceptual and methodological challenges. *Med Care*. 2000;38(9 Suppl):Ii14-25.

389. Reise SP, Waller NG. Item response theory and clinical measurement. *Annu Rev Clin Psychol*. 2009;5:27-48.

390. Linn RL, Slinde JA. The Determination of the Significance of Change between Pre- and Posttesting Periods. *Rev Educ Res*. 1977;47(1):121-50.

391. U.S. National Library of Medicine. What is precision medicine? https://ghr.nlm.nih.gov/primer/precisionmedicine/definition. Accessed 30 Jan 2018.

392. Hodson R. Precision medicine. *Nature*. 2016;537:S49.

393. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987;40(2):171-8.

394. Liang MH. Longitudinal construct validity - Establishment of clinical meaning in patient evaluative instruments. *Medical Care*. 2000;38(9):84-90.

395. Streiner DL, Norman GR, Oxford University P. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford: Oxford University Press; 2008.

396. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol*. 1989;42(5):403-8.

397. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Qual Life Res*. 2003;12(4):349-62.

398. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407-15.

399. Schünemann HJ, Guyatt GH. Commentary—Goodbye M(C)ID! Hello MID, Where Do You Come From? *Health Serv Res*. 2005;40(2):593-7.

400. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res*. 2005;14(6):1523-32.

401. Drummond M. Introducing economic and quality of life measurements into clinical studies. *Ann Med*. 2001;33(5):344-9.

402. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis*. 1985;38(1):27-36.

403. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102-9.

404. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S, Ind Advisory Comm Int Soc Q. Methods for interpreting change over time in patient-reported outcome measures. *Qual Life Res*. 2013;22(3):475-83.

405. Nguyen C, Berezne A, Mestre-Stanislas C, Lefevre-Colau MM, Rannou F, Guillevin L, et al. Changes over Time and Responsiveness of the Cochin Hand Function Scale and Mouth Handicap in Systemic Sclerosis Scale in Patients with Systemic Sclerosis A Prospective Observational Study. *Am J Phys Med Rehabil*. 2016;95(12):E189-E97.

406. de Boer A, van Lanschot JJB, Stalmeier PFM, van Sandick JW, Hulscher JBF, de Haes J, et al. Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Qual Life Res*. 2004;13(2):311-20.

407. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*. 2002;77(4):371-83.

408. Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A, Cella D. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) instrument. *Value Health*. 2005;8(2):117-27.

409. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7(5):541-6.

410. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. 2nd ed. Chichester: John Wiley; 2007.

411. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther*. 2009;17(3):163-70.

412. Lafave MR, Hiemstra L, Kerslake S, Heard M, Buchko G. Validity, Reliability, and Responsiveness of the Anterior Cruciate Ligament Quality of Life Measure: A Continuation of Its Overall Validation. *Clin J Sport Med*. 2017;27(1):57-63.

413. Greco NJ, Anderson AF, Mann BJ, Cole BJ, Farr J, Nissen CW, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. *Am J Sports Med*. 2010;38(5):891-902.

414. Herrmann D. Reporting current, past, and changed health status. What we know about distortion. *Med Care*. 1995;33(4 Suppl):As89-94.

415. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53(5):459-68.

416. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Verstappen SM, Watson K, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. *Qual Life Res*. 2009;18(9):1195-205.

417. Eurich DT, Johnson JA, Reid KJ, Spertus JA. Assessing responsiveness of generic and specific health related quality of life measures in heart failure. *Health Qual Life Outcomes*. 2006;4.

418. Keeley T, Al-Janabi H, Nicholls E, Foster NE, Jowett S, Coast J. A longitudinal assessment of the responsiveness of the ICECAP-A in a randomised controlled trial of a knee pain intervention. *Qual Life Res*. 2015;24(10):2319-31.

419. Cohen J. Statistical power analysis for the behavioral sciences. New York;London;: Psychology Press; 1988.

420. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res*. 1995;4(4):293-307.

421. Brożek JL, Guyatt GH, Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health Qual Life Outcomes*. 2006;4:69-.

422. Sivan M. Interpreting effect size to estimate responsiveness of outcome measures. 2009(1524-4628 (Electronic)).

423. Middel B, Kuipers-Upmeijer H, Bouma J, Staal M, Oenema D, Postma T, et al. Effect of intrathecal baclofen delivered by an implanted programmable pump on health related quality of life in patients with severe spasticity. *J Neurol Neurosurg Psychiatry*. 1997;63(2):204-9.

424. Middel B, van Sonderen E. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *Int J Integr Care*. 2002;2:e15.

425. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12(4 Suppl):142s-58s.

426. Sedgwick P. Pearson's correlation coefficient. *BMJ*. 2012;345.

427. Regression with Stata. Chapter 2 - Regression Diagnostics. UCLA: Statistical Consulting Group. . http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm. Accessed 16 Feburary 2017.

428. Parzen E. On Estimation of a Probability Density Function and Mode. *Ann Stat*. 1962;33(3):1065-76.

429. Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*. 1965;52(3/4):591.

430. Sedgwick P. Spearman's rank correlation coefficient. *BMJ*. 2014;349.

431. Jenkinson C, Gray A, Doll H, Lawrence K, Keoghane S, Layte R. Evaluation of index and profile measures of health status in a randomized controlled trial. Comparison of the Medical Outcomes Study 36-Item Short Form Health Survey, EuroQol, and disease specific measures. *Med Care*. 1997;35(11):1109-18.

432. Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. *Value Health*. 2012;15(1):118-27.

433. Martinez-Martin P, Jeukens-Visser M, Lyons KE, Rodriguez-Blazquez C, Selai C, Siderowf A, et al. Health-related quality-of-life scales in Parkinson's disease: critique and recommendations. *Mov Disord*. 2011;26(13):2371-80.

434. Dodel R, Jonsson B, Reese JP, Winter Y, Martinez-Martin P, Holloway R, et al. Measurement of costs and scales for outcome evaluation in health economic studies of Parkinson's disease. *Mov Disord*. 2014;29(2):169-76.

435. Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. *Expert Rev Pharmacoecon Outcomes Res*. 2001;1.

436. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. *Pharmacoeconomics*. 2000;17(1):13-35.

437. Martinez-Martin P, Gil-Nagel A, Gracia LM, Gomez JB, Martinez-Sarries J, Bermejo F. Unified Parkinson's Disease Rating Scale characteristics and structure. The Cooperative Multicentric Group. *Mov Disord*. 1994;9(1):76-83.

438. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. *Neurology*. 1967;17(5):427-42.

439. Maccorquodale K, Meehl PE. On a distinction between hypothetical constructs and intervening variables. *Psychol Rev*. 1948;55(2):95-107.

440. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess*. 1999;3(9):i-iv, 1-164.

441. Guttman M. Double-blind comparison of pramipexole and bromocriptine treatment with placebo in advanced Parkinson's disease. International Pramipexole-Bromocriptine Study Group. *Neurology*. 1997;49(4):1060-5.

442. Bowling A, Ebrahim S. Handbook of health research methods. Investigation, measurement and analysis. Maidenhead, U.K.: Open University Press; 2005. 625 p.

443. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol*. 2001;54(12):1204-17.

444. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Physical Therapy*. 1998;78(11):1186-96.

445. McGill University Health Center Libraries. MEDLINE VIA PUBMED VS CINAHL. 2013.

446. Rickman K. Embase vs. pubmed & medline: how do they differ.

447. American Psychological Association. PsycINFO. http://www.apa.org/pubs/databases/psycinfo/index.aspx. Accessed 6 September 2017.

448. ASSIA: Applied Social Sciences Index and Abstracts. http://www.proquest.com/products-services/ASSIA-Applied-Social-Sciences-Index-and-Abstracts.html. Accessed 6 September 2017.

449. Social Services Abstracts. http://www.proquest.com/products-services/ssa-set-c.html. Accessed 6 September 2017.

450. (2015). DARE and NHS EED. https://www.crd.york.ac.uk/CRDWeb/AboutPage.asp. Accessed 6 September 2017.

451. Scottish Intercollegiate Guidelines Network. Search filters. http://www.sign.ac.uk/search-filters.html. Accessed 6 September 2017.

452. Benito-Leon J, Cubo E, Coronell C. Impact of apathy on health-related quality of life in recently diagnosed Parkinson's disease: the ANIMO study. *Mov Disord*. 2012;27(2):211-8.

453. Garcia-Gordillo MA, Del Pozo-Cruz B, Adsuar JC, Sanchez-Martinez FI, Abellan-Perpinan JM. Validation and comparison of 15-D and EQ-5D-5L instruments in a Spanish Parkinson's disease population sample. *Qual Life Res*. 2013;23(4):1315-26.

454. Jones CA, Pohar SL, Patten SB. Major depression and health-related quality of life in Parkinson's disease. *Gen Hosp Psychiatry*. 2009;31(4):334-40.

455. Martinez-Martin P, Rodriguez-Blazquez C, Forjaz MJ, Alvarez-Sanchez M, Arakaki T, Bergareche-Yarza A, et al. Relationship between the MDS-UPDRS domains and the health-related quality of life of Parkinson's disease patients. *Eur J Neurol*. 2014;21(3):519-24.

456. Pohar SL, Allyson Jones C. The burden of Parkinson disease (PD) and concomitant comorbidities. *Arch Gerontol Geriatr*. 2009;49(2):317-21.

457. Rodriguez-Blazquez C, Forjaz MJ, Frades-Payo B, de Pedro-Cuesta J, Martinez-Martin P. Independent validation of the scales for outcomes in Parkinson's disease-autonomic (SCOPA-AUT). *Eur J Neurol*. 2010;17(2):194-201.

458. Siderowf A, Ravina B, Glick HA. Preference-based quality-of-life in patients with Parkinson's disease. *Neurology*. 2002;59(1):103-8.

459. Swinn L, Schrag A, Viswanathan R, Bloem BR, Lees A, Quinn N. Sweating dysfunction in Parkinson's disease. *Mov Disord*. 2003;18(12):1459-63.

460. Daley DJ, Deane KH, Gray RJ, Clark AB, Pfeil M, Sabanathan K, et al. Adherence therapy improves medication adherence and quality of life in people with Parkinson's disease: a randomised controlled trial. *Int J Clin Prac*. 2014;68(8):963-71.

461. Ebersbach G, Hahn K, Lorrain M, Storch A. Tolcapone improves sleep in patients with advanced Parkinson's disease (PD). *Arch Gerontol Geriatr*. 2010;51(3):e125-8.

462. Jarman B, Hurwitz B, Cook A, Bajekal M, Lee A. Effects of community based nurses specialising in Parkinson's disease on health outcome and costs: randomised controlled trial. *BMJ*. 2002;324(7345):1072-5.

463. Larisch A, Reuss A, Oertel WH, Eggert K. Does the clinical practice guideline on Parkinson's disease change health outcomes? A cluster randomized controlled trial. *J Neurol*. 2011;258(5):826-34.

464. Luo N, Ng WY, Lau PN, Au WL, Tan LC. Responsiveness of the EQ-5D and 8-item Parkinson's Disease Questionnaire (PDQ-8) in a 4-year follow-up study. *Qual Life Res*. 2010;19(4):565-9.

465. Noyes K, Dick AW, Holloway RG. Pramipexole versus levodopa in patients with early Parkinson's disease: effect on generic and disease-specific quality of life. *Value Health*. 2006;9(1):28-38.

466. Nyholm D, Nilsson Remahl AI, Dizdar N, Constantinescu R, Holmberg B, Jansson R, et al. Duodenal levodopa infusion monotherapy vs oral polypharmacy in advanced Parkinson disease. *Neurology*. 2005;64(2):216-23.

467. Schroder S, Martus P, Odin P, Schaefer M. Impact of community pharmaceutical care on patient health and quality of drug treatment in Parkinson's disease. *Int J Clin Pharm*. 2012;34(5):746-56.

468. Stocchi F, Giorgi L, Hunter B, Schapira AH. PREPARED: Comparison of prolonged and immediate release ropinirole in advanced Parkinson's disease. *Mov Disord*. 2011;26(7):1259-65.

469. Trend P, Kaye J, Gage H, Owen C, Wade D. Short-term effectiveness of intensive multidisciplinary rehabilitation for people with Parkinson's disease and their carers. *Clin Rehabil*. 2002;16(7):717-25.

470. Wade DT, Gage H, Owen C, Trend P, Grossmith C, Kaye J. Multidisciplinary rehabilitation for people with Parkinson's disease: a randomised controlled study. *J Neurol Neurosurg Psychiatry*. 2003;74(2):158-62.

471. Zhu XL, Chan DT, Lau CK, Poon WS, Mok VC, Chan AY, et al. Cost-effectiveness of subthalmic nucleus deep brain stimulation for the treatment of advanced Parkinson disease in Hong Kong: a prospective study. *World Neurosurg*. 2014;82(6):987-93.

472. Borchani H, Bielza C, Marti Nez-Marti NP, Larranaga P. Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: an application to predict the European Quality of Life-5 Dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39). *J Biomed Inform*. 2012;45(6):1175-84.

473. Cheung YB, Tan LC, Lau PN, Au WL, Luo N. Mapping the eight-item Parkinson's Disease Questionnaire (PDQ-8) to the EQ-5D utility index. *Qual Life Res*. 2008;17(9):1173-81.

474. Dams J, Klotsche J, Bornschein B, Reese JP, Balzer-Geldsetzer M, Winter Y, et al. Mapping the EQ-5D index by UPDRS and PDQ-8 in patients with Parkinson's disease. *Health Qual Life Outcomes*. 2013;11(1):35.

475. Kent S, Gray A, Schlackow I, Jenkinson C, McIntosh E. Mapping from the Parkinson's Disease Questionnaire PDQ-39 to the Generic EuroQol EQ-5D-3L: The Value of Mixture Models. *Med Decis Making*. 2015;29:pii: 0272989X15584921.

476. Young MK, Ng SK, Mellick G, Scuffham PA. Mapping of the PDQ-39 to EQ-5D scores in patients with Parkinson's disease. *Qual Life Res*. 2013;22(5):1065-72.

477. Rosser R, Kind P. A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol*. 1978;7(4):347-58.

478. Gudex C, Kind P. (1989). The QALY tool kit - Discussion Paper 38. York: Centre for Health Economics, University of York. http://www.york.ac.uk/che/pdf/dp38.pdf. Accessed 01 April 2016.

479. Kristiansen IS, Bingefors K, Nyholm D, Isacson D. Short-term cost and health consequences of duodenal levodopa infusion in advanced Parkinson's disease in Sweden: an exploratory study. *Appl Health Econ Health Policy*. 2009;7(3):167-80.

480. Lundqvist C, Beiske AG, Reiertsen O, Kristiansen IS. Real life cost and quality of life associated with continuous intraduodenal levodopa infusion compared with oral treatment in Parkinson patients. *J Neurol*. 2014;261(12):2438-45.

481. Noyes K, Dick AW, Holloway RG. Pramipexole and levodopa in early Parkinson's disease: dynamic changes in cost effectiveness. *Pharmacoeconomics*. 2005;23(12):1257-70.

482. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care*. 1996;34.

483. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care*. 2002;40.

484. Pechevis M, Clarke CE, Vieregge P, Khoshnood B, Deschaseaux-Voinet C, Berdeaux G, et al. Effects of dyskinesias in Parkinson's disease on quality of life and health-related costs: a prospective European study. *Eur J Neurol*. 2005;12(12):956-63.

485. Schrag A, Selai C, Jahanshahi M, Quinn NP. The EQ-5D--a generic quality of life measure-is a useful instrument to measure quality of life in patients with Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 2000;69(1):67-73.

486. Martinez-Martin P, Rodriguez-Blazquez C, Kurtis MM, Chaudhuri KR. The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease. *Movement Disorders*. 2011;26(3):399-406.

487. Soh S-E, Morris ME, McGinley JL. Determinants of health-related quality of life in Parkinson's disease: A systematic review. *Parkinsonism Relat Disord*. 2011;17(1):1-9.

488. McDonough CM, Tosteson AN. Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. *Pharmacoeconomics*. 2007;25(2):93-106.

489. Grewal I, Lewis J, Flynn T, Brown J, Bond J, Coast J. Developing attributes for a generic quality of life measure for older people: preferences or capabilities? *Social science & medicine*. 2006;62:1891-901.

490. Reeve A, Simcox E, Turnbull D. Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Res Rev*. 2014;14:19-30.

491. Kukull WA, Ganguli M. Generalizability: the trees, the forest, and the low-hanging fruit. *Neurology*. 2012;78(23):1886-91.

492. Polit DF, Beck CT. Generalization in quantitative and qualitative research: myths and strategies. *Int J Nurs Stud*. 2010;47(11):1451-8.

493. Gray R, Baker M, Fitzpatrick R, Gray A, Greenhall R, Overstall P, et al. (2010). HTA - 98/03/02: A large randomised assessment of the relative cost-effectiveness of different classes of drugs for Parkinson's disease (PD MED). Protocol version 8. http://www.birmingham.ac.uk/Documents/college-mds/trials/bctu/PDMed/Investigators/PD-MED-Protocol-Version-8.pdf. Accessed 2 Feb 2016.

494. Rascol O, Goetz C, Koller W, Poewe W, Sampaio C. Treatment interventions for Parkinson's disease: an evidence based assessment. *Lancet*. 2002;359(9317):1589-98.

495. Wheatley K, Stowe RL, Clarke CE, Hills RK, Williams AC, Gray R. Evaluating drug treatments for Parkinson's disease: how good are the trials? *BMJ*. 2002;324(7352):1508-11.

496. Krygowska-Wajs AT, Gorecka-Mazur A, Tomaszewski KA, Potasz K, Furgala A. Psychometric validation of the Polish version Parkinson's disease questionnaire-39 (PDQ-39) and its short form (PDQ-8). *Mov Disord*. 2015;30:S426-S.

497. Morley D, Dummett S, Kelly L, Dawson J, Jenkinson C. An electronic version of the PDQ-39: acceptability to respondents and assessment of alternative response formats. *Journal of Parkinson's disease*. 2014;4(3):467.

498. Zhang J-L, Chan P. Reliability and validity of PDQ-39: a quality-of-life measure for patients with PD in China. *Qual Life Res*. 2012;21(7):1217-21.

499. Park HJ, Sohng KY, Kim S. Validation of the Korean version of the 39-Item Parkinson's Disease Questionnaire (PDQ-39). *Asian Nurs Res*. 2014;8(1):67-74.

500. Jenkinson C, Fitzpatrick R. 2 - The development and validation of the Parkinson's Disease Questionnaire and related measures. In: Jenkinson C, Peters M, Bromberg B, M., editors. Quality of Life Measurement in Neurodegenerative and Related Conditions. Cambridge: Cambridge University Press; 2011.

501. Martinez-Martin P, Frades Payo B. Quality of life in Parkinson's disease: validation study of the PDQ-39 Spanish version. The Grupo Centro for Study of Movement Disorders. *J Neurol*. 1998;245 Suppl 1:S34-8.

502. Jenkinson C, Fitzpatrick R, Peto V, Greenhall R, Hyman N. The PDQ-8: Development and validation of a short-form parkinson's disease questionnaire. *Psychol Health*. 1997;12:805-14.

503. Hagell P, Nygren C. The 39 item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence based medicine. *J Neurol Neurosurg Psychiatry*. 2007;78(11):1191-8.

504. Lenert L, Kaplan RM. Validity and interpretation of preference-based measures of health-related quality of life. *Med Care*. 2000;38(9 Suppl):II138-50.

505. Malley JN, Towers AM, Netten AP, Brazier JE, Forder JE, Flynn T. An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people. *Health Qual Life Outcomes*. 2012;10:21.(doi):10.1186/477-7525-10-21.

506. Arnold D, Girling A, Stevens A, Lilford R. Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. *BMJ*. 2009;339:b2688.

507. Klotsche J, Reese JP, Winter Y, Oertel WH, Irving H, Wittchen HU, et al. Trajectory classes of decline in health-related quality of life in Parkinson's disease: a pilot study. *Value Health*. 2011;14(2):329-38.

508. Mavandadi S, Nazem S, Ten Have TR, Siderowf AD, Duda JE, Stern MB, et al. Use of latent variable modeling to delineate psychiatric and cognitive profiles in Parkinson disease. *Am J Geriatr Psychiatry*. 2009;17(11):986-95.

509. Rickards H. Depression in neurological disorders: Parkinson's disease, multiple sclerosis, and stroke. *J Neurol Neurosurg Psychiatry*. 2005;76 Suppl 1:i48-52.

510. Pickering RM, Grimbergen YAM, Rigney U, Ashburn A, Mazibrada G, Wood B, et al. A meta-analysis of six prospective studies of falling in Parkinson's disease. *Mov Disord*. 2007;22:1892-900.

511. Guttman M, Slaughter PM, Theriault M-E, DeBoer DP, Naylor CD. Burden of parkinsonism: a population-based study. *Mov Disord*. 2003;18:313-9.

512. Davis JC, Liu-Ambrose T, Richardson CG, Bryan S. A comparison of the ICECAP-O with EQ-5D in a falls prevention clinical setting: are they complements or substitutes? *Qual Life Res*. 2013;22(5):969-77. doi: 10.1007/s11136-012-0225-4. Epub 2012 Jun 22.

513. Das S, Rahman RM. Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. *Nutr J*. 2011;10:124.

514. Long JS, Freese J. Regression models for categorical dependent variables using Stata. Third ed. College Station, Texas: StataCorp LP; 2006. 527 p.

515. Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*. 1990;46(4):1171-8.

516. Diedenhofen B, Musch J. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*. 2015;10(4):e0121945.

517. Eddings W, Marchenko Y. Diagnostics for multiple imputation in Stata. *Stata J*. 2012;12(3):353-67.

518. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken, N.J.: Wiley; 2002. xv, 381 p. p.

519. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157-66.

520. Rubin DB. Multiple imputation for nonresponse in surveys. New York ; Chichester: Wiley; 1987. v. p.

521. Schrag A, Jahanshahi M, Quinn N. What contributes to quality of life in patients with Parkinson's disease? *J Neurol Neurosurg Psychiatry*. 2000;69(3):308-12.

522. Flynn TN, Chan P, Coast J, Peters TJ. Assessing quality of life among British older people using the ICEPOP CAPability (ICECAP-O) measure. *Appl Health Econ Health Policy*. 2011;9:317-29.

523. Mitchell PM, Roberts TE, Barton PM, Pollard BS, Coast J. Predicting the ICECAP-O capability index from the WOMAC osteoarthritis index: is mapping onto capability from condition-specific health status questionnaires feasible? *Med Decis Making*. 2013;33(4):547-57.

524. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res*. 2005;14(2):285-95.

525. Schipper H. Guidelines and caveats for quality of life measurement in clinical practice and research. *Oncology (Williston Park)*. 1990;4(5):51-7; discussion 70.

526. Nussbaum M, Sen A. Capability and well-being. The Quality of Life. Oxford: Oxford University Press; 1993. p. 30-53.

527. Lorgelly PK, Lawson Kd Fau - Fenwick EAL, Fenwick Ea Fau - Briggs AH, Briggs AH. Outcome measurement in economic evaluations of public health interventions: a role for the capability approach? (1660-4601 (Electronic)).

528. Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Counsell C, et al. Movement disorder society task force report on the Hoehn and Yahr staging scale: Status and recommendations. *Mov Disord*. 2004;19(9):1020-8.

529. Marinus J, Ramaker C, van Hilten JJ, Stiggelbout AM. Health related quality of life in Parkinson's disease: a systematic review of disease specific instruments. *J Neurol Neurosurg Psychiatry*. 2002;72(2):241-8.

530. Damiano AM, Snyder C, Strausser B, Willian MK. A review of health-related quality-of-life concepts and measures for Parkinson's disease. *Qual Life Res*. 1999;8(3):235-43.

531. Palmer CS, Schmier JK, Snyder E, Scott B. Patient preferences and utilities for 'off-time' outcomes in the treatment of Parkinson's disease. *Qual Life Res*. 2000;9(7):819-27.

532. Johnson SJ, Diener MD, Kaltenboeck A, Birnbaum HG, Siderowf AD. An economic model of Parkinson's disease: Implications for slowing progression in the United States. *Mov Disord*. 2013;28(3):319-26.

533. Peto V, Jenkinson C, Fitzpatrick R. Determining minimally important differences for the PDQ-39 Parkinson's disease questionnaire. *Age Ageing*. 2001;30(4):299-302.

534. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the patient-reported outcomes measurement information system (PROMIS) 20-item physical functioning short form in a prospective observational study of rheumatoid arthritis. *Ann Rheum Dis*. 2015;74(1):104-7.

535. Buhi ER, Goodson P, Neilands TB. Out of sight, not out of mind: Strategies for handling missing data. *Am J Health Behav*. 2008;32(1):83-92.

536. Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *J Stat Softw*. 2011;45(5):1-14.

537. Young R, Johnson DR. Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. *J Marriage Fam*. 2015;77(1):277-94.

538. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-99.

539. group, USc. How can I estimate R-squared for a model estimated with multiply imputed data? [UCLA Statistical consulting group]. http://www.ats.ucla.edu/stat/stata/faq/mi_r_squared.htm. Accessed 16 Feburary 2017.

540. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes*. 2003;1:4.

541. Abel H, Kephart G, Packer T, Warner G. Discordance in Utility Measurement in Persons with Neurological Conditions: A Comparison of the SF-6D and the HUI3. *Value Health*. 2017;20(8):1157-65.

542. Wu J, Han YR, Zhao FL, Zhou J, Chen ZJ, Sun H. Validation and comparison of EuroQoL-5 dimension (EQ-5D) and Short Form-6 dimension (SF-6D) among stable angina patients. *Health Qual Life Outcomes*. 2014;12:11.

543. Wooten GF, Currie LJ, Bovbjerg VE, Lee JK, Patrie J. Are men at greater risk for Parkinson's disease than women? *J Neurol Neurosurg Psychiatry*. 2004;75(4):637-9.

544. Mouradian MM, Juncos JL, Fabbrini G, Chase TN. Motor fluctuations in Parkinson's disease: pathogenetic and therapeutic studies. *Ann Neurol*. 1987;22(4):475-9.

545. Parkinson's UK. (2014). Wearing off and involuntary movements (dyskinesia). https://www.parkinsons.org.uk/sites/default/files/publications/download/english/fs73_wearingoffandinvoluntarymovements.pdf. Accessed 21 Febuary 2017.

546. Davis JC, Best JR, Dian L, Khan KM, Hsu CL, Chan W, et al. Are the EQ-5D-3L and the ICECAP-O responsive among older adults with impaired mobility? Evidence from the Vancouver Falls Prevention Cohort Study. *Qual Life Res*. 2017;26(3):737-47.

547. Parsons N, Griffin XL, Achten J, Costa ML. Outcome assessment after hip fracture is EQ-5D the answer? *Bone Joint Res*. 2014;3(3):69-75.

548. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol*. 2003;56(5):395-407.

549. Mercieca-Bebber R, Palmer MJ, Brundage M, Calvert M, Stockler MR, King MT. Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review. *BMJ Open*. 2016;6.

550. Bylicki O, Gan HK, Joly F, Maillet D, You B, Peron J. Poor patient-reported outcomes reporting according to CONSORT guidelines in randomized clinical trials evaluating systemic cancer therapy. *Ann Oncol*. 2015;26(1):231-7.

551. Efficace F, Fayers P, Pusic A, Cemal Y, Yanagawa J, Jacobs M, et al. Quality of patient-reported outcome reporting across cancer randomized controlled trials according to the CONSORT patient-reported outcome extension: A pooled analysis of 557 trials. *Cancer*. 2015;121(18):3335-42.

552. Fielding S, Maclennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials*. 2008;9:51.

553. Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Stat Med*. 1998;17(5-7):781-96.

554. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Stat Methods Med Res*. 2014;23(5):440-59.

555. Gomes M, Gutacker N, Bojke C, Street A. Addressing Missing Data in Patient-Reported Outcome Measures (PROMS): Implications for the Use of PROMS for Comparing Provider Performance. *Health Econ*. 2016;25(5):515-28.

556. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*. 2010;63(7):721-7.

557. Altman DG, Bland JM. Statistics notes - Standard deviations and standard errors. *BMJ*. 2005;331(7521):903-.

558. Walters SJ. Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation. Chichester: John Wiley & Sons, Ltd; 2009.

559. Luo N, Johnson JA, Shaw JW, Coons SJ. Relative efficiency of the EQ-5D, HUI2, and HUI3 index scores in measuring health burden of chronic medical conditions in a population health survey in the United States. *Med Care*. 2009;47(1):53-60.

560. McDonough CM, Tosteson TD, Tosteson AN, Jette AM, Grove MR, Weinstein JN. A longitudinal comparison of 5 preference-weighted health state classification systems in persons with intervertebral disk herniation. *Med Decis Making*. 2011;31(2):270-80.

561. Sung L, Greenberg ML, Doyle JJ, Young NL, Ingber S, Rubenstein J, et al. Construct validation of the Health Utilities Index and the Child Health Questionnaire in children undergoing cancer chemotherapy. *Br J Cancer*. 2003;88(8):1185-90.

562. Holmes AM. A QALY-based societal health statistic for Canada, 1985. *Soc Sci Med*. 1995;41(10):1417-27.

563. Angelis A, Lange A, Kanavos P. Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. *Eur J Health Econ*. 2018;19(1):123-52.

564. Rios-Diaz AJ, Lam J, Ramos MS, Moscoso AV, Vaughn P, Zogg CK, et al. Global Patterns of QALY and DALY Use in Surgical Cost-Utility Analyses: A Systematic Review. *PLoS One*. 2016;11(2):e0148304.

565. Devlin NJ, Lorgelly PK. QALYs as a measure of value in cancer. *J Cancer Policy*. 2017;11:19-25.

566. Daniel H. Stemming the Escalating Cost of Prescription Drugs: A Position Paper of the American College of Physicians. *Ann Intern Med*. 2016.

567. McIntosh E, Gray A, Daniels J, Gill S, Ives N, Jenkinson C, et al. Cost-utility analysis of deep brain stimulation surgery plus best medical therapy versus best medical therapy in patients with Parkinson's: Economic evaluation alongside the PD SURG trial. *Mov Disord*. 2016;31(8):1173-82.

568. National Institute for Health and Care Excellence. (2017). Parkinson’s disease in adults: diagnosis and management.Full guideline NICE guideline NG71. Methods, evidence and recommendations.

569. Earhart GM. Dance as Therapy for Individuals with Parkinson Disease. *Eur J Phys Rehabil Med*. 2009;45(2):231-8.

570. Payne K, McAllister M, Davies LM. Valuing the economic benefits of complex interventions: when maximising health is not sufficient. *Health Econ*. 2013;22(3):258-71.

571. Gray E, Eden M, Vass C, McAllister M, Louviere J, Payne K. Valuing Preferences for the Process and Outcomes of Clinical Genetics Services: A Pilot Study. *Patient*. 2016;9(2):135-47.

572. Goodrich K, Kaambwa B, Al-Janabi H. The inclusion of informal care in applied economic evaluation: a review. *Value Health*. 2012;15(6):975-81.

573. Kinghorn P. (2018). Investigating deliberative methods for setting a monetary capability threshold in the context of social care and public health. http://gtr.rcuk.ac.uk/projects?ref=MR/N014790/1. Accessed 1 Feb 2018.

574. Harel O. The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *J Appl Stat*. 2009;36(10):1109-18.

Appendices

Appendix A: Search strategy

Search strategy and number of results in each database

Search first run: inception of each database - 26 November, 2013

Search update: 9 June 2015 (same search strategy, only limit the date, 01/01/2013- present (9 June 2015))

PUBMED

1st Result: 1196

2nd Result: 314

Search Strategy:

Search (((parkinsonian disorders[MeSH Terms]) OR parkinson*[Title/Abstract])) AND ((cost effective*[Title/Abstract]) OR Cost-Benefit Analysis[Mesh]) OR Quality of Wellbeing[Title/Abstract]) OR Quality of Wellbeing[Title/Abstract]) OR QWB[Title/Abstract]) OR Health Utilities Index[Title/Abstract]) OR cost benefit*[Title/Abstract]) OR visual analogue scale[Title/Abstract]) OR time trade off) OR time tradeoff) OR standard gamble) OR discrete choice) OR dce) OR conjoint analysis) OR contingent valuation) OR preference*[Title/Abstract]) OR utility[Title/Abstract]) OR willingness to pay[Title/Abstract]) OR wtp[Title/Abstract]) OR QALY[Title/Abstract]) OR Quality-Adjusted Life Years[MeSH Terms]) OR QALE[Title/Abstract]) OR QALD[Title/Abstract]) OR Qtime[Title/Abstract]) OR quality adjusted life expectancy[Title/Abstract]) OR DALY[Title/Abstract]) OR Disability adjusted life[Title/Abstract]) OR HYE[Title/Abstract]) OR HYEs[Title/Abstract]) OR healthy year equivalent) OR SF-6D[Title/Abstract]) OR SF6D[Title/Abstract]) OR EuroQOL[Title/Abstract]) OR Euro qol[Title/Abstract]) OR EQ-5D[Title/Abstract]) OR HUI[Title/Abstract]) OR EQ5D[Title/Abstract]) OR HUI1[Title/Abstract]) OR HUI2[Title/Abstract]) OR HUI3[Title/Abstract]) Filters: English

Ovid MEDLINE® In-Process & Other Non-Indexed Citations and Ovid MEDLINE® 1946 to Present with Daily & Weekly Update

1st Result: 1202

2nd Result: 300 (01/01/2013-current)

Search Strategy:

1 exp "cost effective"/
2 cost effective*.ab.
3 cost utility*.ab.
4 cost benefit*.ab.
5 cost benefit analysis.sh.
6 visual analogue scale.ab.

7 standard gamble.af.
8 time trade off.af.
9 time tradeoff.mp
10 dce.mp.
11 discrete choice.mp.
12 conjoint analysis.af.
13 willingness to pay.mp.
14 wtp.mp.
15 Patient Preference/ or preference.mp.
16 preference*.ab.
17 contingent valuation.mp.
18 QALY$.mp. or Quality-Adjusted Life Years/
19 QALE$.mp.
20 QALD$.mp.
21 Qtime$.mp.
22 quality adjusted life expectancy.mp.
23 quality adjusted life day$.mp.
24 DALY$.mp.
25 Disability adjusted life.mp.
26 HYE.mp.
27 HYEs.mp.
28 Health$ year$ equivalent$.mp.
29 SF-6D.mp.
30 SF6D.mp.
31 EuroQOL.mp.
32 Euro qol.mp.
33 EQ-5D.mp.
34 EQ5D.mp.
35 HUI.mp.
36 HUI1.mp.
37 HUI2.mp.
38 HUI3.mp.
39 Health Utilities Index.mp.
40 QWB.mp.
41 Quality of Wellbeing.mp.
42 Quality of Wellbeing.mp.
43 utilit$.mutilitp.
44 or/1-43
45 parkinson*.ab.
46 parkinsonian disorders.sh.
47 45 or 46
48 44 and 47
49 limit 48 to english language

Embase

1947 – Present, updated daily

1st Result: 1516

2nd Result: 553 (01/01/2013-current)

Search strategy:

1. Parkinsonian Disorders/
2. parkinson*.ab.
3. Parkinsonian Disorders.sh.
4. 1 or 2 or 3
5. cost effective*.ab.
6. cost utility*.ab.
7. cost benefit*.ab.
8. cost benefit analysis.sh.
9. visual analogue scale.ab.
10. time trade off.af.
11. standard gamble.af.
12. Patient Preference/ or preference.mp.
13. preference*.ab.
14. discrete choice.mp.
15. conjoint analysis.af.
16. utilit$.mp.
17. willingness to pay.mp.
18. wtp.mp.
19. dce.mp.
20. contingent valuation.mp.
21. QALY$.mp. or Quality-Adjusted Life Years/
22. QALE$.mp.
23. QALD$.mp.
24. Qtime$.mp.
25. quality adjusted life expectancy.mp.
26. quality adjusted life day$.mp.
27. DALY$.mp.
28. Disability adjusted life.mp.
29. HYE.mp.
30. HYEs.mp.
31. Health$ year$ equivalent$.mp.
32. SF-6D.mp.
33. SF6D.mp.
34. EuroQOL.mp.
35. Euro qol.mp.
36. EQ-5D.mp.
37. EQ5D.mp.
38. HUI.mp.
39. HUI1.mp.
40. HUI2.mp.
41. HUI3.mp.
42. Health Utilities Index.mp.
43. QWB.mp.
44. Quality of Wellbeing.mp.
45. Quality of Wellbeing.mp.
46. time tradeoff.mp.
47. exp "cost benefit analysis"/ or exp "cost effectiveness analysis"/
48. 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47
49. 4 and 48
50. limit 49 to (human and english language)

CINAHL (EBSCO)

1st Result: 102

2nd Result: 25 (01/01/2013-present)

Search strategy:

S1	MW parkinsonian disorders
S2	AB parkinson*
S3	S1 OR S2
S4	MW quality adjusted life years
S5	AB cost effective*
S6	AB cost utility
S7	MW cost benefit analysis
S8	AB cost benefit
S9	TX visual analogue scale
S10	TX time trade off
S11	TX standard gamble
S12	AB preference*
S13	TX discrete choice
S14	TX conjoint analysis
S15	TX willingness to pay
S16	TX time tradeoff
S17	TX dce
S18	AB utilit*
S19	TX wtp
S20	TX contingent valuation
S21	(MH "Quality-Adjusted Life Years") OR "QALY"
S22	TX hye
S23	TX hyes
S24	TX qaly
S25	TX qale
S26	TX Quality adjusted life expectancy
S27	TX QALD
S28	TX quality adjusted life days
S29	TX DALY
S30	TX disability adjusted life
S31	TX health* year* equivalent
S32	TX eq5d
S33	TX hui1
S34	TX hui2
S35	TX hui3
S36	TX eq-5d
S37	TX euroqol
S38	TX euro qol
S39	TX sf 6d
S40	TX sf6d
S41	TX health utilit* index
S42	TX qwb

S43	TX quality of wellbeing
S44	TX quality of well being
S45	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44
S46	S3 AND S45
S47	S3 AND S45 (Limiters - English Language)

PsycINFO(EBSCO)

1st Result: 440

2nd Result: 213 (01/01/2013-present)

Search strategy:

S1	MM "Parkinsonism"
S2	AB parkinson*
S3	S1 AND S2
S4	MM "Quality of Life" OR MM "Quality of Work Life"
S5	AB cost effective*
S6	AB cost utility
S7	TX visual analogue scale
S8	TX time trade off
S9	TX standard gamble
S10	AB preference*
S11	TX discrete choice
S12	TX conjoint analysis
S13	TX willingness to pay
S14	TX time tradeoff
S15	TX dce
S16	TX wtp
S17	TX contingent valuation
S18	QALY
S19	TX qaly
S20	TX disability adjusted life
S21	TX quality adjusted life year*
S22	TX eq5d
S23	TX EQ-5D
S24	TX hui1
S25	TX hui2
S26	TX hui3
S27	TX euroqol
S28	TX utilit*
S29	TX sf 6d
S30	TX sf6d

S31	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30
S32	S1 OR S2
S33	S31 AND S32

Applied Social Sciences Index and Abstracts (ASSIA) (Proquest)

1st Result: 30

2nd Result: 6 (01/01/2013-present)

Search strategy:

(ab(parkinson*) OR su(parkinsonian disorders)) AND (su(cost effective*) OR ab(cost effective*) OR ab(cost utility) OR ab(cost benefit) OR su(cost - benefit analysis) OR (standard gamble) OR (time trade off) OR (visual analogue scale) OR (discrete choice) OR ab(preference*) OR (conjoint analysis) OR (willingness to pay) OR (contingent valuation) OR (time tradeoff) OR (dce) OR (wtp) OR su(Quality-Adjusted Life Years) OR (QALY) OR (QALE) OR (QALD) OR (Qtime) OR (quality adjusted life expectancy) OR (quality adjusted life day*) OR (DALY) OR (disability adjusted life) OR (hye) OR (hyes) OR (health* year* equivalent*) OR (sf6d) OR (sf 6d) OR (euroqol) OR (euro qol) OR (eq 5d) OR (eq5d) OR (hui) OR (hui1) OR (hui2) OR (hui3) OR (health utilities index) OR (qwb) OR (quality of wellbeing) OR (quality of well being) OR ab(utility*))

SOCIAL service abstract (SSA) (Proquest)

1st Result: 2

2nd Result: 1 (01/01/2013-present)

Search strategy:

(ab(parkinson*) OR su(parkinsonian disorders)) AND (su(cost effective*) OR ab(cost effective*) OR ab(cost utility) OR ab(cost benefit) OR su(cost - benefit analysis) OR (standard gamble) OR (time trade off) OR (visual analogue scale) OR (discrete choice) OR ab(preference*) OR (conjoint analysis) OR (willingness to pay) OR (contingent valuation) OR (time tradeoff) OR (dce) OR (wtp) OR su(Quality-Adjusted Life Years) OR (QALY) OR (QALE) OR (QALD) OR (Qtime) OR (quality adjusted life expectancy) OR (quality adjusted life day*) OR (DALY) OR (disability adjusted life) OR (hye) OR (hyes) OR (health* year* equivalent*) OR (sf6d) OR (sf 6d) OR (euroqol) OR (euro qol) OR (eq 5d) OR (eq5d) OR (hui) OR (hui1) OR (hui2) OR (hui3) OR (health utilities index) OR (qwb) OR (quality of wellbeing) OR (quality of well being) OR ab(utility*))

AgeInfo open search

1st Result: 15

2nd Result: 1 (01/01/2013-present)

Search strategy:

Parkinson* and quality of life

Database of Abstracts of Reviews of Effects (DARE) (CRD York)

1st Result: 51

2nd Result: 8 (01/01/2013-present)

Search strategy:

(Parkinson*) AND (quality of life) OR (utility*) IN DARE

NHS Economic evaluation database (NHS EED)

1st Result: 26

2nd Result: 2 (01/01/2013-present)

Search strategy:

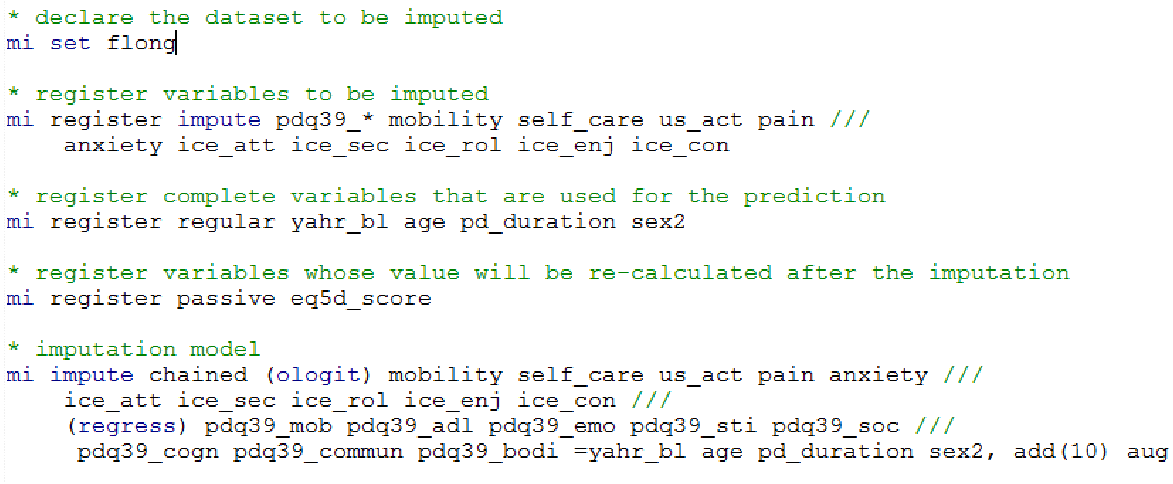
(Parkinson*) AND (quality of life) OR (utility*) IN NHSEED

Appendix B: Stata codes for multiple imputation

This appendix contains the STATA codes for the multiple imputation with chained equations implemented for imputing missing values in Chapter 6 (B-1) and Chapter 7 (B-2).

B-1. STATA codes for multiple imputation in Chapter 6

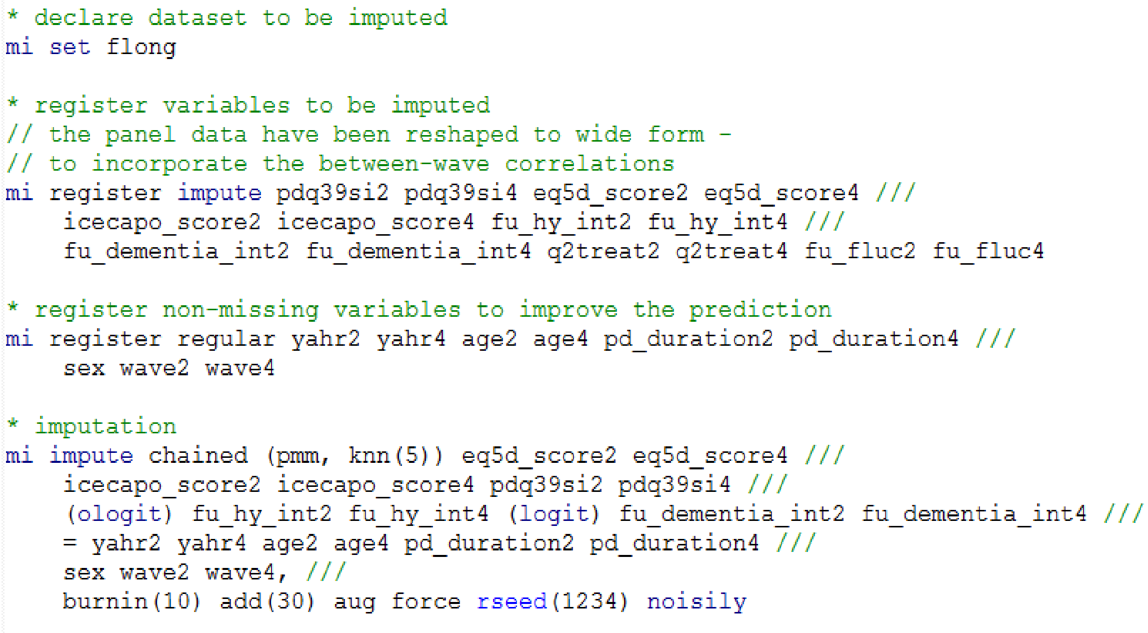
A description for this imputation model is provided in Section 6.4.2.3.



```
* declare the dataset to be imputed
mi set flong

* register variables to be imputed
mi register impute pdq39_* mobility self_care us_act pain ///
    anxiety ice_att ice_sec ice_rol ice_enj ice_con

* register complete variables that are used for the prediction
mi register regular yahr_bl age pd_duration sex2

* register variables whose value will be re-calculated after the imputation
mi register passive eq5d_score

* imputation model
mi impute chained (ologit) mobility self_care us_act pain anxiety ///
    ice_att ice_sec ice_rol ice_enj ice_con ///
    (regress) pdq39_mob pdq39_adl pdq39_emo pdq39_sti pdq39_soc ///
     pdq39_cogn pdq39_commun pdq39_bodi =yahr_bl age pd_duration sex2, add(10) aug
```

B-2. STATA codes for multiple imputation in Chapter 7 .

A description for this imputation model is provided in Section 7.3.4.2. Note: the panel data structure have been reshaped to wide form prior to the beginning of the imputation procedure. This generated the “2” and “4” at the end of each variable name which represents respectively the variable value at year 2 and year 4.



```
* declare dataset to be imputed
mi set flong

* register variables to be imputed
// the panel data have been reshaped to wide form -
// to incorporate the between-wave correlations
mi register impute pdq39si2 pdq39si4 eq5d_score2 eq5d_score4 ///
    icecapo_score2 icecapo_score4 fu_hy_int2 fu_hy_int4 ///
    fu_dementia_int2 fu_dementia_int4 q2treat2 q2treat4 fu_fluc2 fu_fluc4

* register non-missing variables to improve the prediction
mi register regular yahr2 yahr4 age2 age4 pd_duration2 pd_duration4 ///
    sex wave2 wave4

* imputation
mi impute chained (pmm, knn(5)) eq5d_score2 eq5d_score4 ///
    icecapo_score2 icecapo_score4 pdq39si2 pdq39si4 ///
    (ologit) fu_hy_int2 fu_hy_int4 (logit) fu_dementia_int2 fu_dementia_int4 ///
    = yahr2 yahr4 age2 age4 pd_duration2 pd_duration4 ///
    sex wave2 wave4, ///
    burnin(10) add(30) aug force rseed(1234) noisily
```

Appendix C: Results of statistical tests for the comparison of correlation coefficients using the R cocor package.

This appendix provides the results of the statistical tests for comparison of correlation coefficients using the R cocor package in Chapter 6 (C-1) and 7 (C-2).

C-1: hypothesis in Chapter 6 for testing the correlation coefficient using the cross-sectional data: the correlation coefficient for the ICECAP-O and PDQ-39 was expected to be larger than that of the ICECAP-O and EQ-5D-3L. Null hypothesis: The correlation coefficient for the ICECAP-O and PDQ-39 was equal to that of the ICECAP-O and EQ-5D-3L.

Results:

cocor – comparing correlations, 1.1–3, http://comparingcorrelations.org

INPUT:

```
require(cocor) # load package
cocor.dep.groups.overlap(r.jk=0.733, r.jh=0.654, r.kh=0.724, n=1010,
alternative="greater", alpha=0.05, conf.level=0.95, null.value=0)
```

OUTPUT:

```
Results of a comparison of two overlapping correlations based on
dependent groups

Comparison between r.jk = 0.733 and r.jh = 0.654
Difference: r.jk - r.jh = 0.079
Related correlation: r.kh = 0.724
Group size: n = 1010
Null hypothesis: r.jk is equal to r.jh
Alternative hypothesis: r.jk is greater than r.jh (one-sided)
Alpha: 0.05

pearson1898: Pearson and Filon's z (1898)
z = 4.9579, p-value = 0.0000
Null hypothesis rejected

hotelling1940: Hotelling's t (1940)
t = 5.1411, df = 1007, p-value = 0.0000
Null hypothesis rejected

williams1959: Williams' t (1959)
t = 5.0790, df = 1007, p-value = 0.0000
Null hypothesis rejected

olkin1967: Olkin's z (1967)
z = 4.9579, p-value = 0.0000
Null hypothesis rejected

dunn1969: Dunn and Clark's z (1969)
```

```
z = 5.0510, p-value = 0.0000
Null hypothesis rejected

hendrickson1970: Hendrickson, Stanley, and Hills' (1970) modification of
Williams' t (1959)
t = 5.1411, df = 1007, p-value = 0.0000
Null hypothesis rejected

steiger1980: Steiger's (1980) modification of Dunn and Clark's z (1969)
using average correlations
z = 5.0389, p-value = 0.0000
Null hypothesis rejected

meng1992: Meng, Rosenthal, and Rubin's z (1992)
z = 5.0336, p-value = 0.0000
Null hypothesis rejected
95% confidence interval for r.jk - r.jh: 0.0934 0.2125
Null hypothesis rejected (Lower boundary > 0)

hittner2003: Hittner, May, and Silver's (2003) modification of Dunn and
Clark's z (1969) using a backtransformed average Fisher's (1921) Z
procedure
z = 5.0291, p-value = 0.0000
Null hypothesis rejected

zou2007: Zou's (2007) confidence interval
95% confidence interval for r.jk - r.jh: 0.0481 0.1109
Null hypothesis rejected (Lower boundary > 0)
```

C-2: comparing the correlation coefficients between the change scores of ICECAP-O – PDQ-39 and ICECAP-O – EQ-5D. Null hypothesis: the correlation coefficient between the change scores of ICECAP-O and PDQ-39 is equal to that of the ICECAP-O and EQ-5D.

Results:

```
cocor - comparing correlations, 1.1-3, http://comparingcorrelations.org

INPUT:
require(cocor) # load package
cocor.dep.groups.overlap(r.jk=0.526, r.jh=0.483, r.kh=0.401, n=933,
alternative="greater", alpha=0.05, conf.level=0.95, null.value=0)

OUTPUT:
Results of a comparison of two overlapping correlations based on
dependent groups

Comparison between r.jk = 0.526 and r.jh = 0.483
Difference: r.jk - r.jh = 0.043
Related correlation: r.kh = 0.401
Group size: n = 933
Null hypothesis: r.jk is equal to r.jh
Alternative hypothesis: r.jk is greater than r.jh (one-sided)
Alpha: 0.05

pearson1898: Pearson and Filon's z (1898)
```

```
z = 1.4669, p-value = 0.0712
Null hypothesis retained

hotelling1940: Hotelling's t (1940)
t = 1.5033, df = 930, p-value = 0.0665
Null hypothesis retained

williams1959: Williams' t (1959)
t = 1.4663, df = 930, p-value = 0.0715
Null hypothesis retained

olkin1967: Olkin's z (1967)
z = 1.4669, p-value = 0.0712
Null hypothesis retained

dunn1969: Dunn and Clark's z (1969)
z = 1.4655, p-value = 0.0714
Null hypothesis retained

hendrickson1970: Hendrickson, Stanley, and Hills' (1970) modification of
Williams' t (1959)
t = 1.5033, df = 930, p-value = 0.0665
Null hypothesis retained

steiger1980: Steiger's (1980) modification of Dunn and Clark's z (1969)
using average correlations
z = 1.4652, p-value = 0.0714
Null hypothesis retained

meng1992: Meng, Rosenthal, and Rubin's z (1992)
z = 1.4650, p-value = 0.0715
Null hypothesis retained
95% confidence interval for r.jk - r.jh: -0.0195 0.1349
Null hypothesis retained (Lower boundary <= 0)

hittner2003: Hittner, May, and Silver's (2003) modification of Dunn and
Clark's z (1969) using a backtransformed average Fisher's (1921) Z
procedure
z = 1.4650, p-value = 0.0715
Null hypothesis retained

zou2007: Zou's (2007) confidence interval
95% confidence interval for r.jk - r.jh: -0.0145 0.1006
Null hypothesis retained (Lower boundary <= 0)
```

Appendix D: Calculation of Pearson correlation coefficient of the imputed dataset

The Pearson correlation coefficient of the imputed datasets was calculated with Harel's Fisher's r to z transformation utilizing the normal distribution of z (539). This was realized with the user-written STATA command '*mibeta*' with the option '*fisherz*' (539). In normal cases, the Pearson correlation coefficient can be obtained with the routine STATA command, *'correlate',* however it does not work after imputation since it is not a supported post estimation command. The linear regression of the imputed data also does not provide the R-square statistic (the squared correlation between the observed and expected values of the dependent variable) so that the correlation coefficient cannot be obtained from this way either (the square root of R-square statistic). Due to the abnormal distribution of the R^2 for the 30 imputations, it may be inappropriate to simply averaging the R^2 values (539). Harel (574) suggested to use the Fisher's r to z transformation method to firstly transform the R^2 from each of the imputed datasets into z, which is supposed to have a normal distribution, then the mean of the z values among the imputed datasets is transformed back into an r, which becomes the Pearson correlation coefficient. This was realized with the user-written STATA command '*mibeta*' with the option '*fisherz*' (539).

STATA code:

```
Mibeta mi_diff_eq5d_score mi_diff_pdq39SI, fisherz
```

Appendix E: Checking normality of residuals

To choose whether it is appropriate to conduct Pearson correlation or Spearman correlation, the normality of the residual of regressions of EQ-5D/ICECAP-O predicted PDQ-39 were checked. Five imputed datasets (n=1, 3, 10, 20 and 30) were checked.

The results include:

1. regression of change of ICECAP-O, predicted by the change of PDQ-39
2. regression of change of EQ-5D-3L, predicted by the change of PDQ-39

MI: wide form, year 2 and years 4.

For each regression, three methods were undertaken:

- Shapiro-Wilk W test
- Kernal density estimate
- Scatter plot of the residual against the fitted value of the response variable (ICECAP-O/EQ-5D)

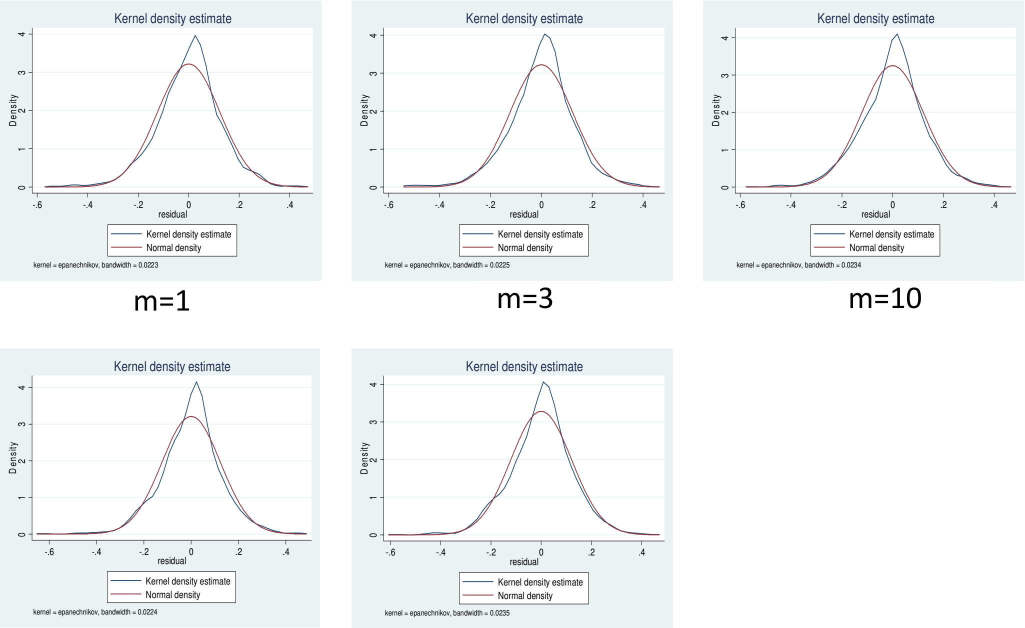
1. Change of ICECAP-O predicted by change of PDQ-39

- Shapiro-Wilk W test

Table Appendix E-1: Shapiro-Wilk W test result to determine the normality of the residuals of change of ICECAP-O predicted by change of PDQ-39

Imputed dataset # (m=)	Variable	Obs	W	V	z	Prob>z
1	residual	1238	0.986	10.576	5.891	0.00000
3	residual	1238	0.983	13.276	6.459	0.00000
10	residual	1238	0.989	8.37	5.307	0.00000
20	residual	1238	0.987	9.725	5.682	0.00000
30	residual	1238	0.990	7.876	5.155	0.00000

- Kernel density estimate

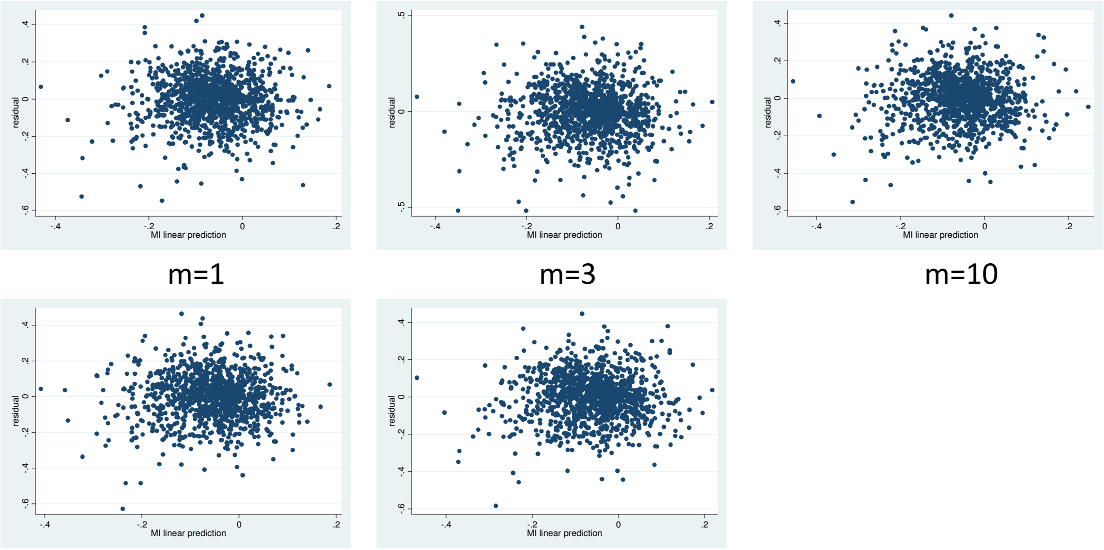


m=1 m=3 m=10

m=20 m=30

Figure Appendix E-1: Kernel density graph in each of the randomly selected imputed datasets (m=1,3,10,20,30) to determine the normality of the residuals of change of ICECAP-O predicted by change of PDQ-39

- Scatter plot of the residual against the fitted value of the ICECAP-O



m=1 m=3 m=10

m=20 m=30

Figure Appendix E-2: scatter plot of residual against the fitted value of the ICECAP-O in each of the randomly selected imputed datasets (m=1,3,10,20,30).

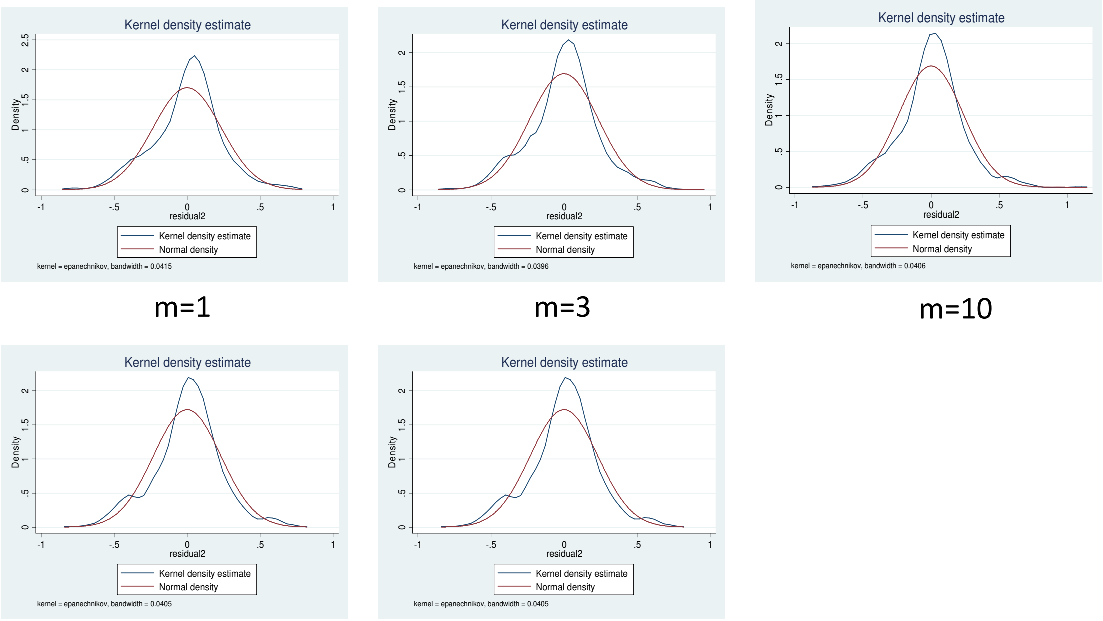
2. Change of EQ-5D-3L vs. PDQ-39

- Shapiro-Wilk W test

Table Appendix E-2: Shapiro-Wilk W test result to determine the normality of the residuals of change of EQ-5D-3L predicted by change of PDQ-39

Imputed dataset # (m=)	Variable	Obs	W	V	z	Prob>z
1	residual	1238	0.986	10.576	5.891	0.00000
3	residual	1238	0.983	13.276	6.459	0.00000
10	residual	1238	0.989	8.370	5.307	0.00000
20	residual	1238	0.987	9.725	5.682	0.00000
30	residual	1238	0.990	7.876	5.155	0.00000

- Kernel density estimate



m=1

m=3

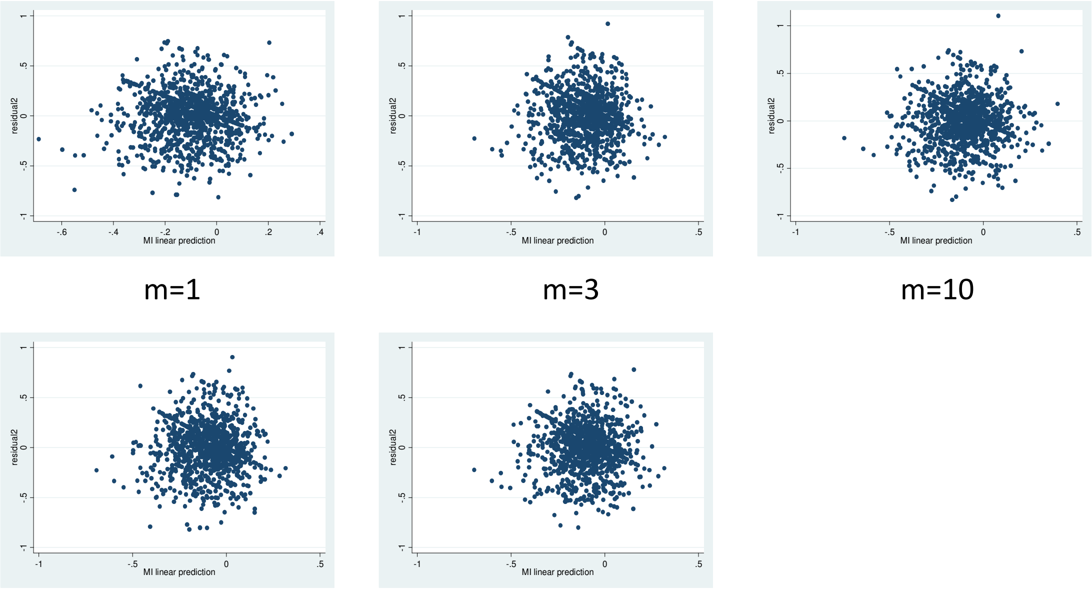
m=10

m=20

m=30

Figure Appendix E-3: Kernel density graph in each of the randomly selected imputed datasets (m=1,3,10,20,30) to determine the normality of the residuals of change of EQ-5D-3L predicted by change of PDQ-39

- Scatter plot of the residual against the fitted value of the ICECAP-O



m=1

m=3

m=10

m=20

m=30

Figure Appendix E-4: scatter plot of residual against the fitted value of the EQ-5D-3L in each of the randomly selected imputed datasets (m=1,3,10,20,30).

Appendix F: Regression analysis of responsiveness: complete case analysis

This appendix contains the regression analysis result with the complete case analysis for determining which PDQ-39 dimensions are strong predictors of the change of ICECAP-O (Table Appendix F-1) /EQ-5D-3L (Table Appendix F-2) over five years.

Table Appendix F – 1: Regression analysis to predict the change of ICECAP-O index score from the change of the PDQ-39 eight dimensions (complete case analysis)

Change of PDQ-39 dimensions to predict ICECAP-O	Coefficient	SE.	P value	95% CI
Mobility	-0.00154	0.00082	0.063	-0.00316 ,0.00008
ADL	-0.00116	0.00074	0.121	-0.00263 ,0.00031
Emotional wellbeing	**-0.00223**	0.00092	**0.016**	-0.00405 ,-0.00042
Stigma	0.00020	0.00094	0.835	-0.00167 ,0.00207
Social support	**-0.00245**	0.00087	**0.006**	-0.00418 ,-0.00072
Cognition	-0.00108	0.00076	0.16	-0.00259 ,0.00043
Communication	**0.00155**	0.00076	0.044	0.00004 ,0.00307
Bodily discomfort	-0.00083	0.00069	0.231	-0.00219 ,0.00053
Constant	-0.00948	0.01752	0.589	-0.04418 ,0.02522

Table Appendix F – 2: Regression analysis to predict the change of EQ-5D index score from the change of the PDQ-39 eight dimensions (complete case analysis)

Change of PDQ-39 dimensions to predict EQ-5D	Coefficient	SE.	P value	95% CI
Mobility	**-0.00247**	0.00094	**0.009**	-0.00431 ,-0.00062
ADL	**-0.00197**	0.00093	**0.036**	-0.0038 ,-0.00013
Emotional wellbeing	**-0.00364**	0.00112	**0.001**	-0.00583 ,-0.00144
Stigma	-0.00152	0.00102	0.137	-0.00352 ,0.00049
Social support	-0.00003	0.00107	0.98	-0.00213 ,0.00207
Cognition	-0.00132	0.00091	0.149	-0.0031 ,0.00047
Communication	**0.00201**	0.00086	0.02	0.00032 ,0.0037
Bodily discomfort	-0.00135	0.00075	0.071	-0.00282 ,0.00012
Constant	-0.02108	0.01959	0.283	-0.05965 ,0.01749

Appendix G: Permission for including published materials in this thesis

This appendix contains the permissions obtained for the following pieces of published content from the publisher that are included in this thesis. They are:

- Part of the content in Chapter 4 has been published by Springer in the journal 'Quality of life research':

Xin Y & McIntosh E. (2017) Assessment of the construct validity and responsiveness of preference-based quality of life measures in people with Parkinson's: a systematic review. Quality of Life Research, 26(1): 1-23.

- Two figures are cited from the following book chapter:

Zumbo BD, Chan EKH. Setting the Stage for Validity and Validation in Social, Behavioral, and Health Sciences: Trends in Validation Practices. In: Zumbo BD, Chan EKH, editors. Validity and Validation in Social, Behavioral, and Health Sciences. 1 ed: Springer International Publishing; 2014.

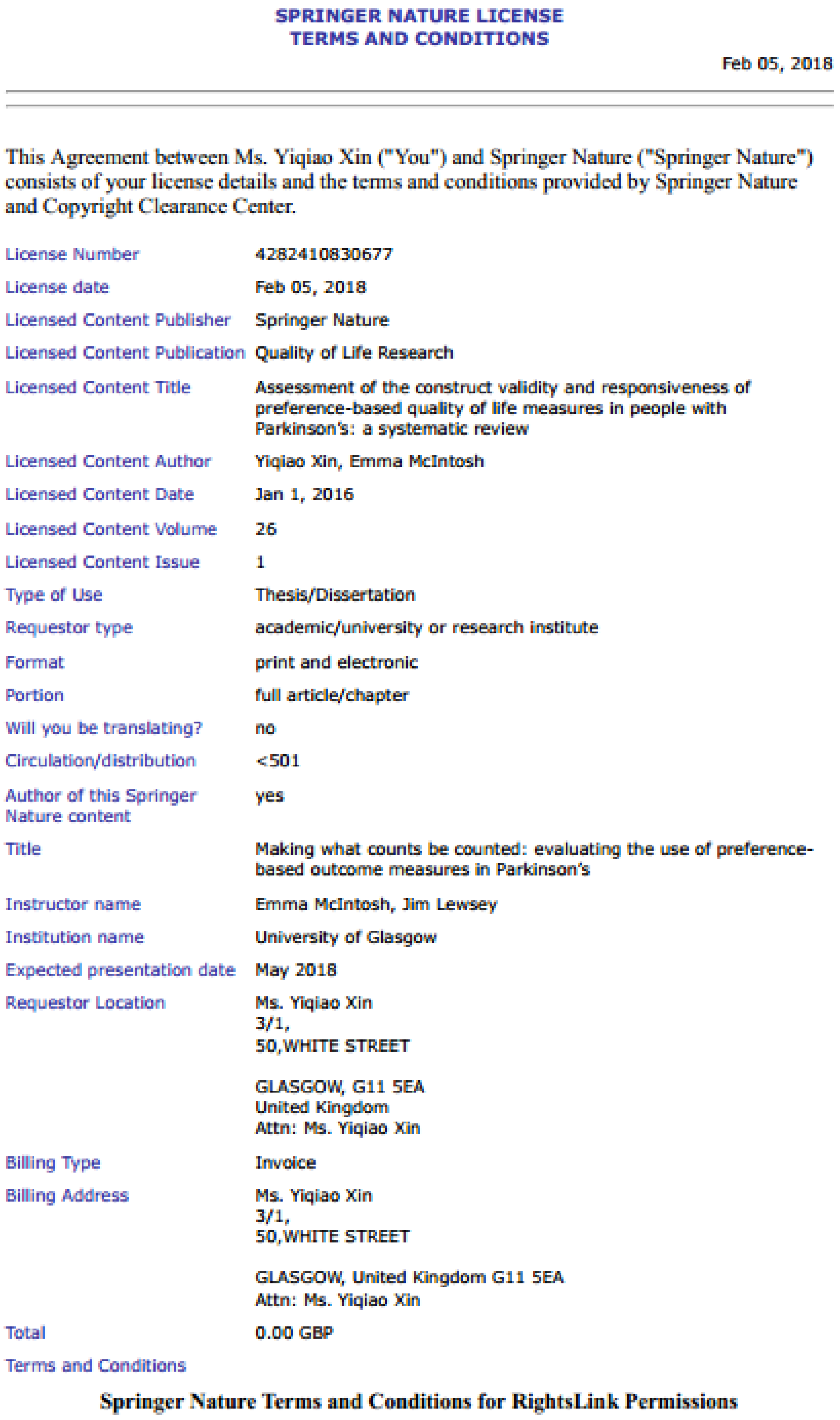
- One figure is cited from each of the following publications:

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. J Clin Epidemiol 2010;63:737-745. 2012.

Smith GT. On construct validity: issues of method and measurement. Psychological assessment. 2005;17(4):396-408.

- One figure is cited from the following archived material:

Gumber A, Ramaswamy B, Ibbotson R, Ismail M, Thongchundee O, Harrop D, et al. Economic, Social and Financial Cost of Parkinson's on Individuals, Carers and their Families in the UK. Project report. Centre for Health and Social Care Research, Sheffield Hallam University; 2017



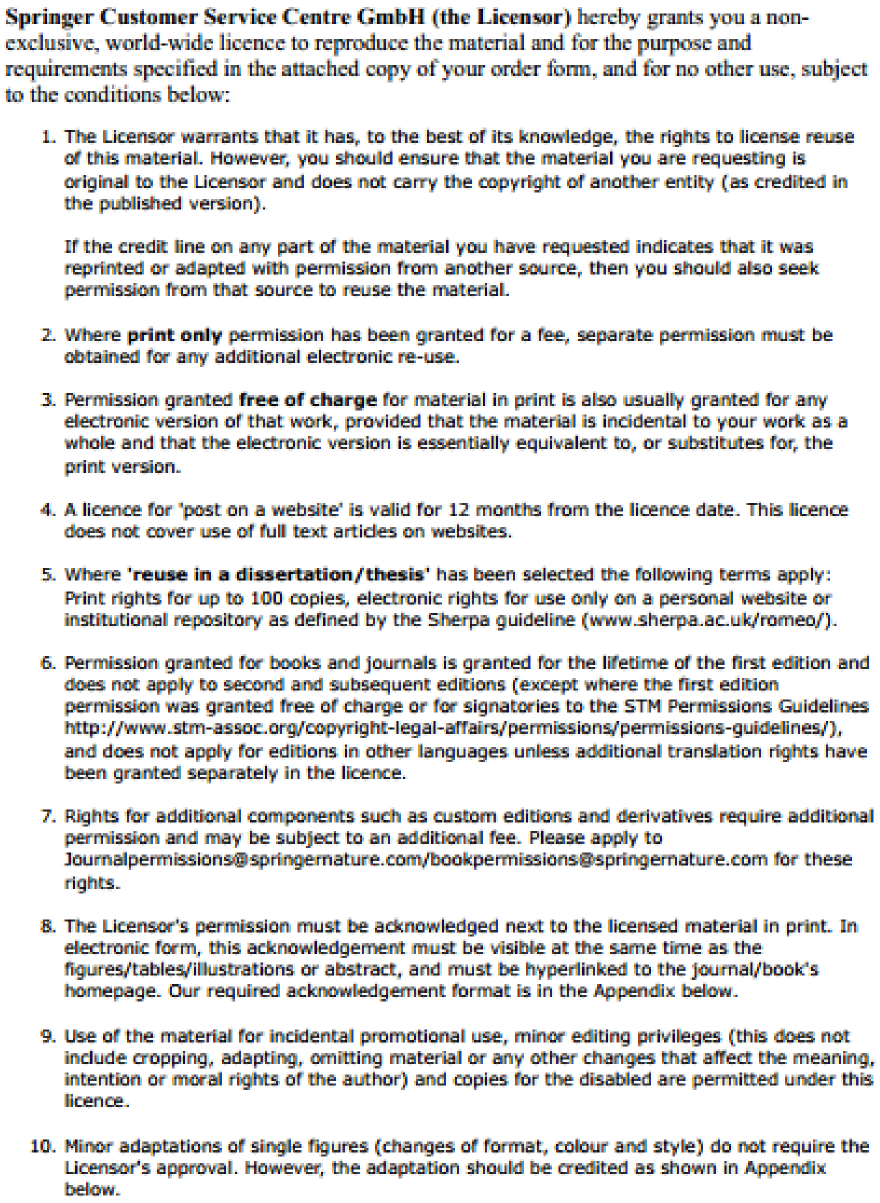
SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Feb 05, 2018

This Agreement between Ms. Yiqiao Xin ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4282410830677
License date	Feb 05, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Quality of Life Research
Licensed Content Title	Assessment of the construct validity and responsiveness of preference-based quality of life measures in people with Parkinson's: a systematic review
Licensed Content Author	Yiqiao Xin, Emma McIntosh
Licensed Content Date	Jan 1, 2016
Licensed Content Volume	26
Licensed Content Issue	1
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	full article/chapter
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	yes
Title	Making what counts be counted: evaluating the use of preference-based outcome measures in Parkinson's
Instructor name	Emma McIntosh, Jim Lewsey
Institution name	University of Glasgow
Expected presentation date	May 2018
Requestor Location	Ms. Yiqiao Xin 3/1, 50,WHITE STREET GLASGOW, G11 5EA United Kingdom Attn: Ms. Yiqiao Xin
Billing Type	Invoice
Billing Address	Ms. Yiqiao Xin 3/1, 50,WHITE STREET GLASGOW, United Kingdom G11 5EA Attn: Ms. Yiqiao Xin
Total	0.00 GBP
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions



Springer Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

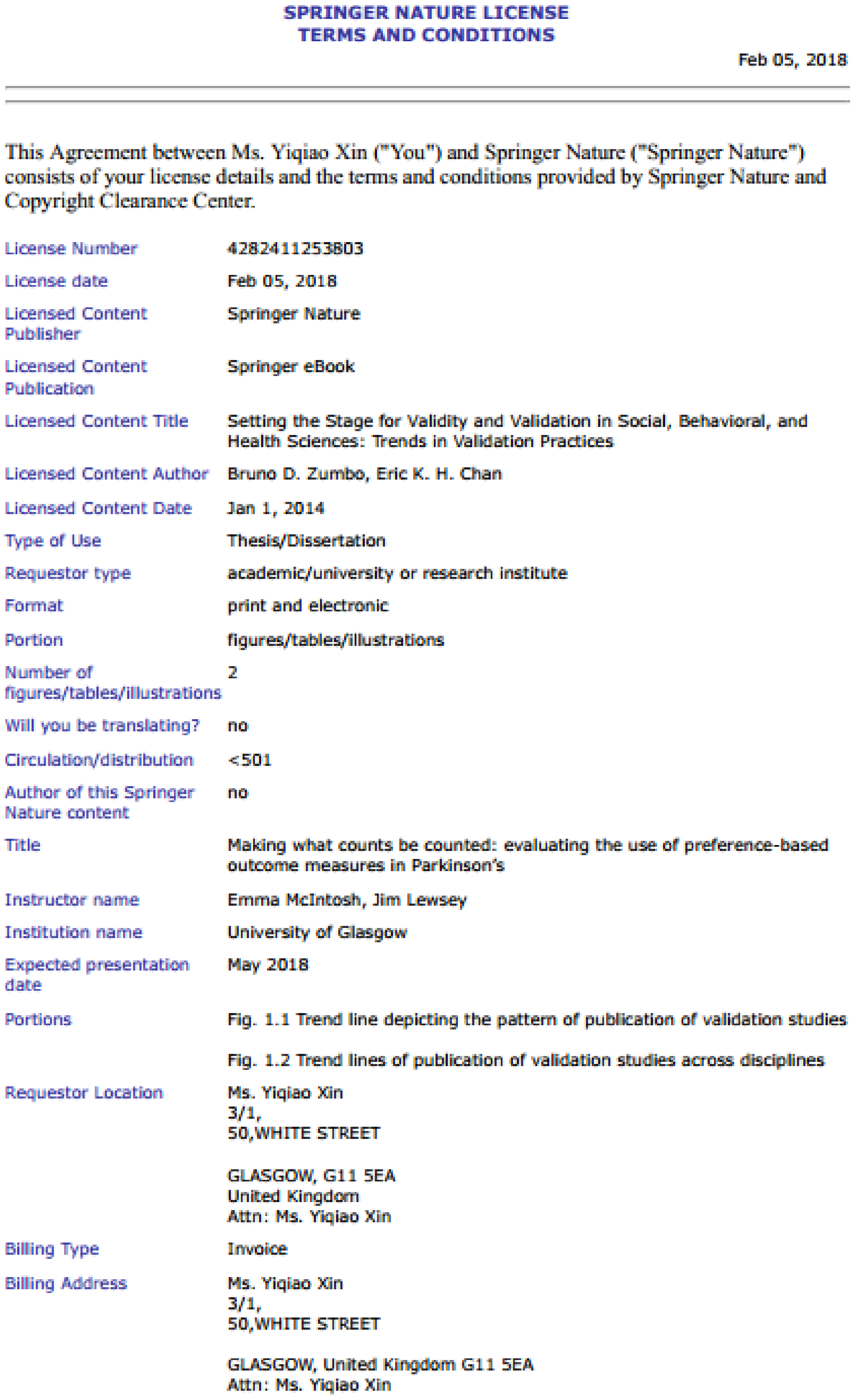
6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

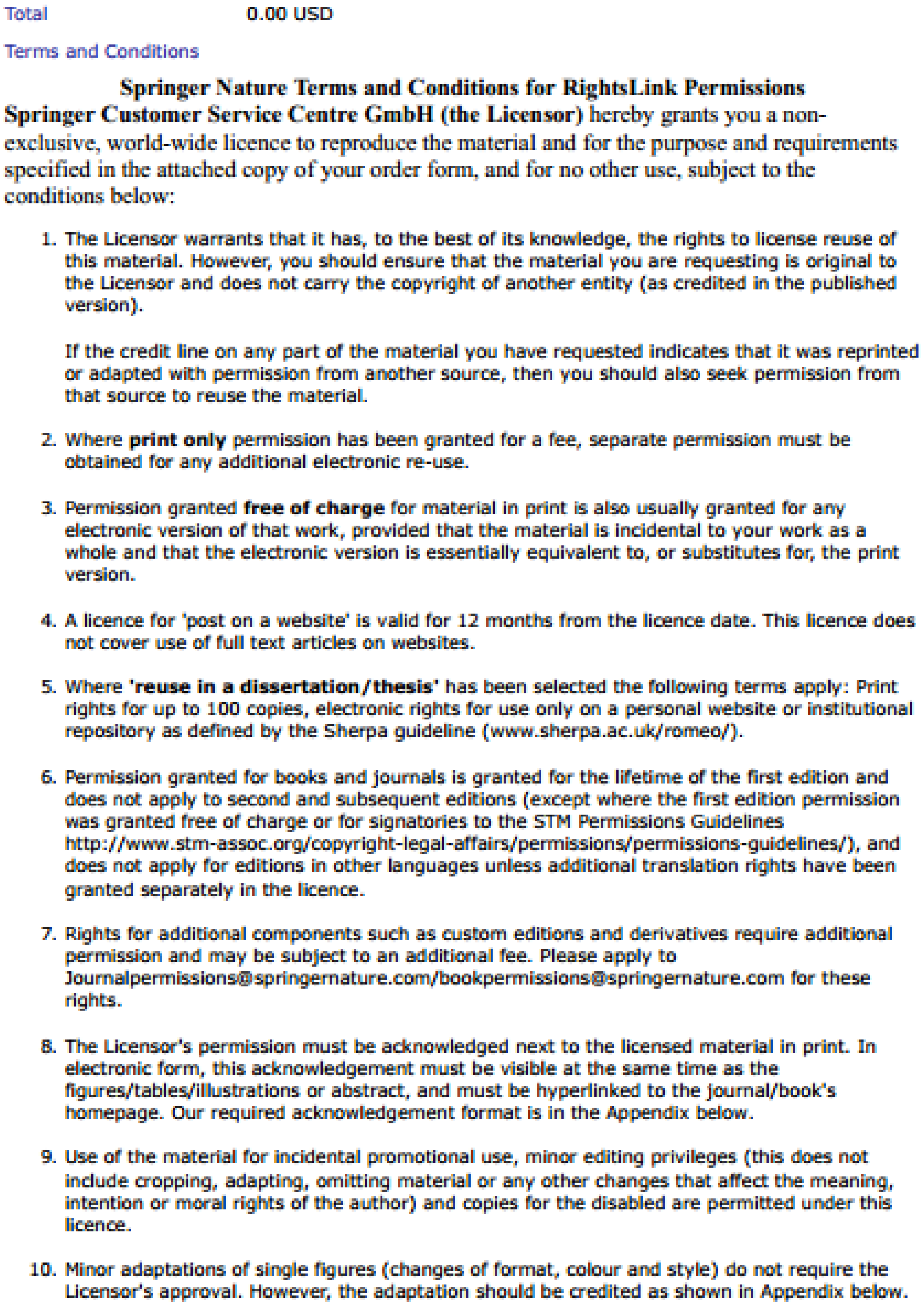


SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Feb 05, 2018

This Agreement between Ms. Yiqiao Xin ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4282411253803
License date	Feb 05, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Setting the Stage for Validity and Validation in Social, Behavioral, and Health Sciences: Trends in Validation Practices
Licensed Content Author	Bruno D. Zumbo, Eric K. H. Chan
Licensed Content Date	Jan 1, 2014
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Making what counts be counted: evaluating the use of preference-based outcome measures in Parkinson's
Instructor name	Emma McIntosh, Jim Lewsey
Institution name	University of Glasgow
Expected presentation date	May 2018
Portions	Fig. 1.1 Trend line depicting the pattern of publication of validation studies Fig. 1.2 Trend lines of publication of validation studies across disciplines
Requestor Location	Ms. Yiqiao Xin 3/1, 50,WHITE STREET GLASGOW, G11 5EA United Kingdom Attn: Ms. Yiqiao Xin
Billing Type	Invoice
Billing Address	Ms. Yiqiao Xin 3/1, 50,WHITE STREET GLASGOW, United Kingdom G11 5EA Attn: Ms. Yiqiao Xin



Total 0.00 USD

Terms and Conditions

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

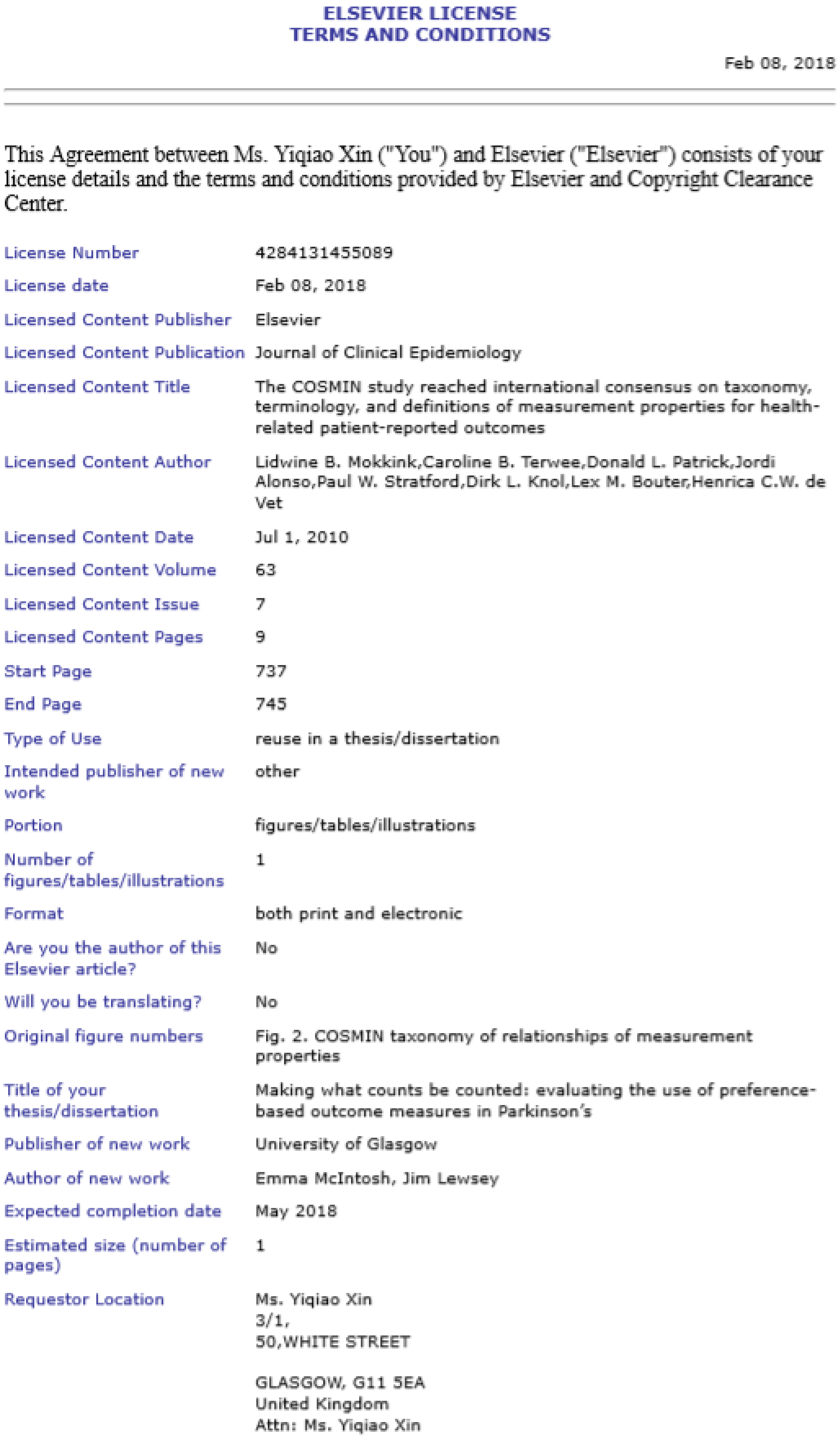
6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

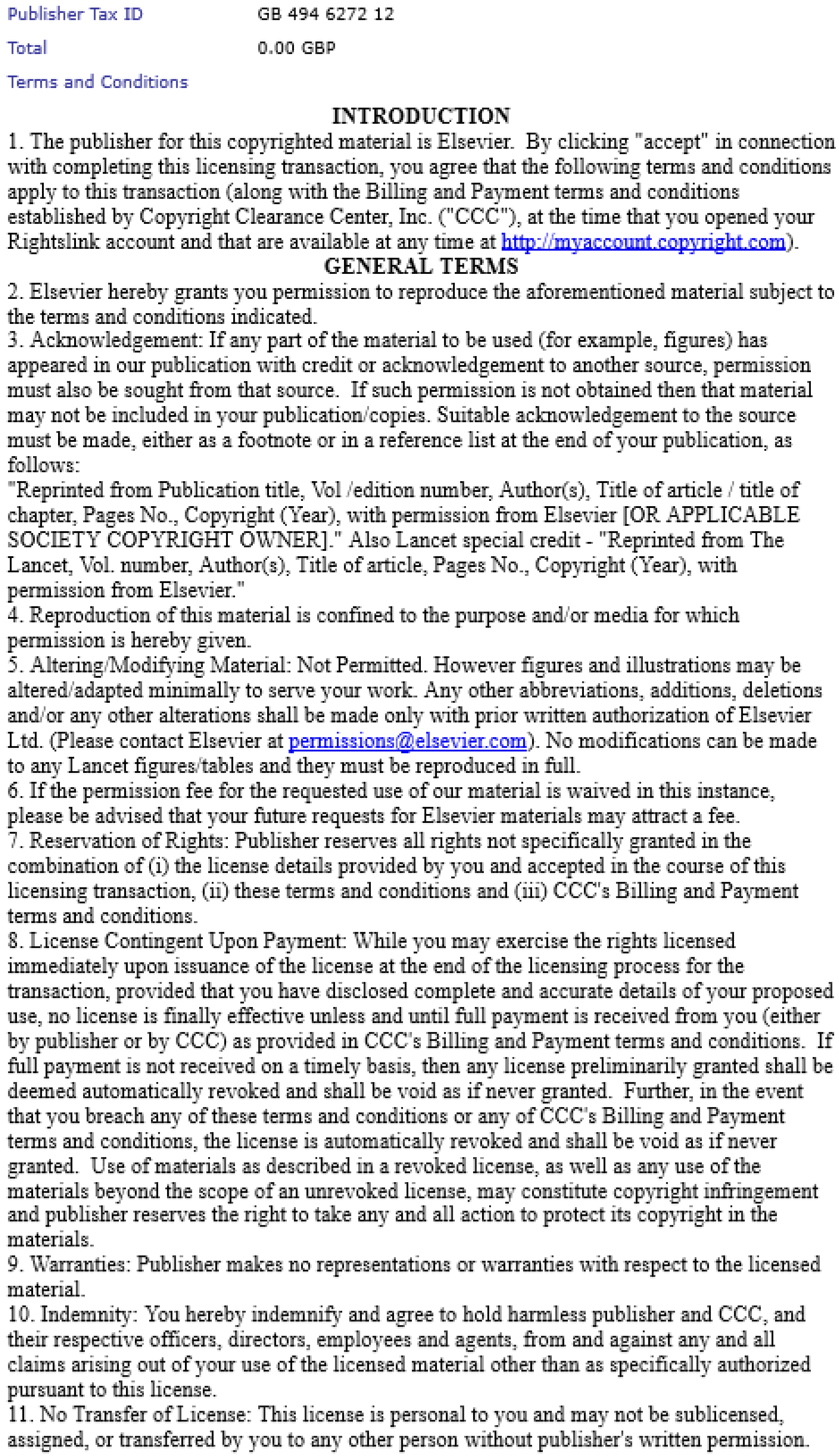


ELSEVIER LICENSE
TERMS AND CONDITIONS

Feb 08, 2018

This Agreement between Ms. Yiqiao Xin ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4284131455089
License date	Feb 08, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Clinical Epidemiology
Licensed Content Title	The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes
Licensed Content Author	Lidwine B. Mokkink,Caroline B. Terwee,Donald L. Patrick,Jordi Alonso,Paul W. Stratford,Dirk L. Knol,Lex M. Bouter,Henrica C.W. de Vet
Licensed Content Date	Jul 1, 2010
Licensed Content Volume	63
Licensed Content Issue	7
Licensed Content Pages	9
Start Page	737
End Page	745
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig. 2. COSMIN taxonomy of relationships of measurement properties
Title of your thesis/dissertation	Making what counts be counted: evaluating the use of preference-based outcome measures in Parkinson's
Publisher of new work	University of Glasgow
Author of new work	Emma McIntosh, Jim Lewsey
Expected completion date	May 2018
Estimated size (number of pages)	1
Requestor Location	Ms. Yiqiao Xin 3/1, 50,WHITE STREET GLASGOW, G11 5EA United Kingdom Attn: Ms. Yiqiao Xin



Publisher Tax ID	GB 494 6272 12
Total	0.00 GBP

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

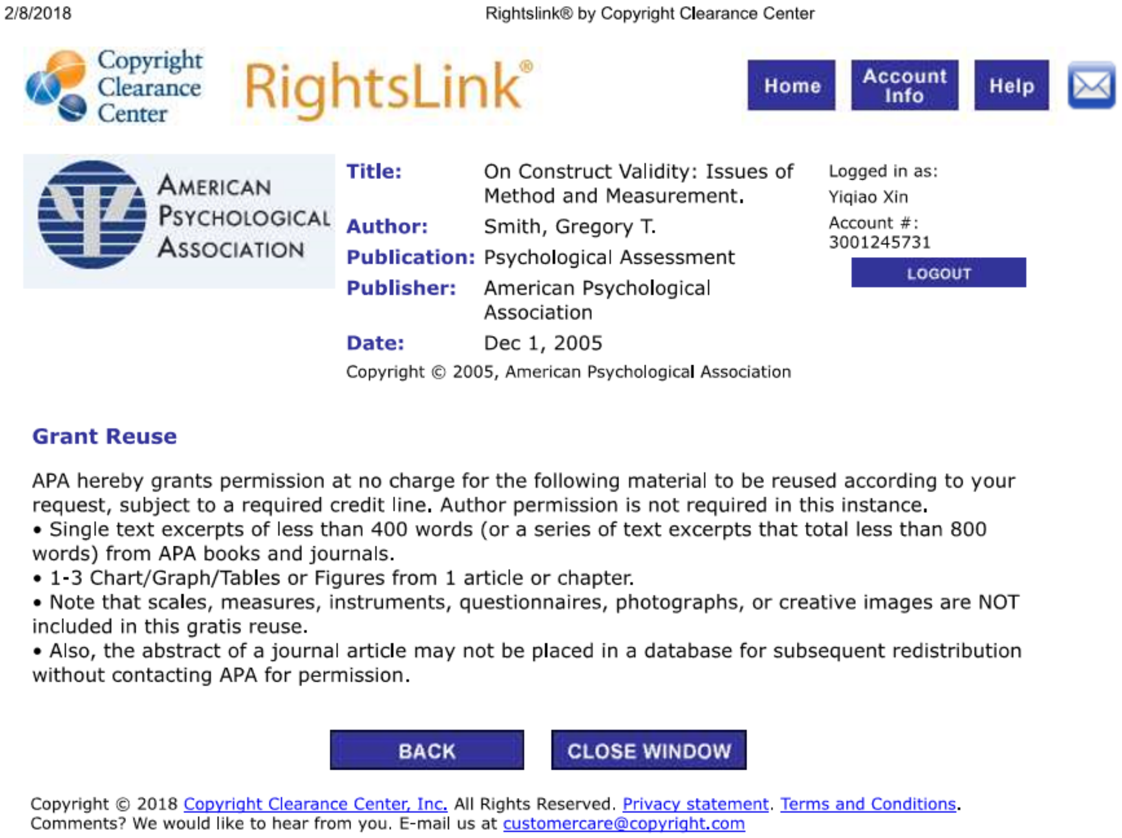
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.


Copyright Clearance Center
RightsLink®
Home
Account Info
Help
American Psychological Association
LOGOUT
BACK
CLOSE WINDOW

2/8/2018 Rightslink® by Copyright Clearance Center

Title:	On Construct Validity: Issues of Method and Measurement.
Author:	Smith, Gregory T.
Publication:	Psychological Assessment
Publisher:	American Psychological Association
Date:	Dec 1, 2005

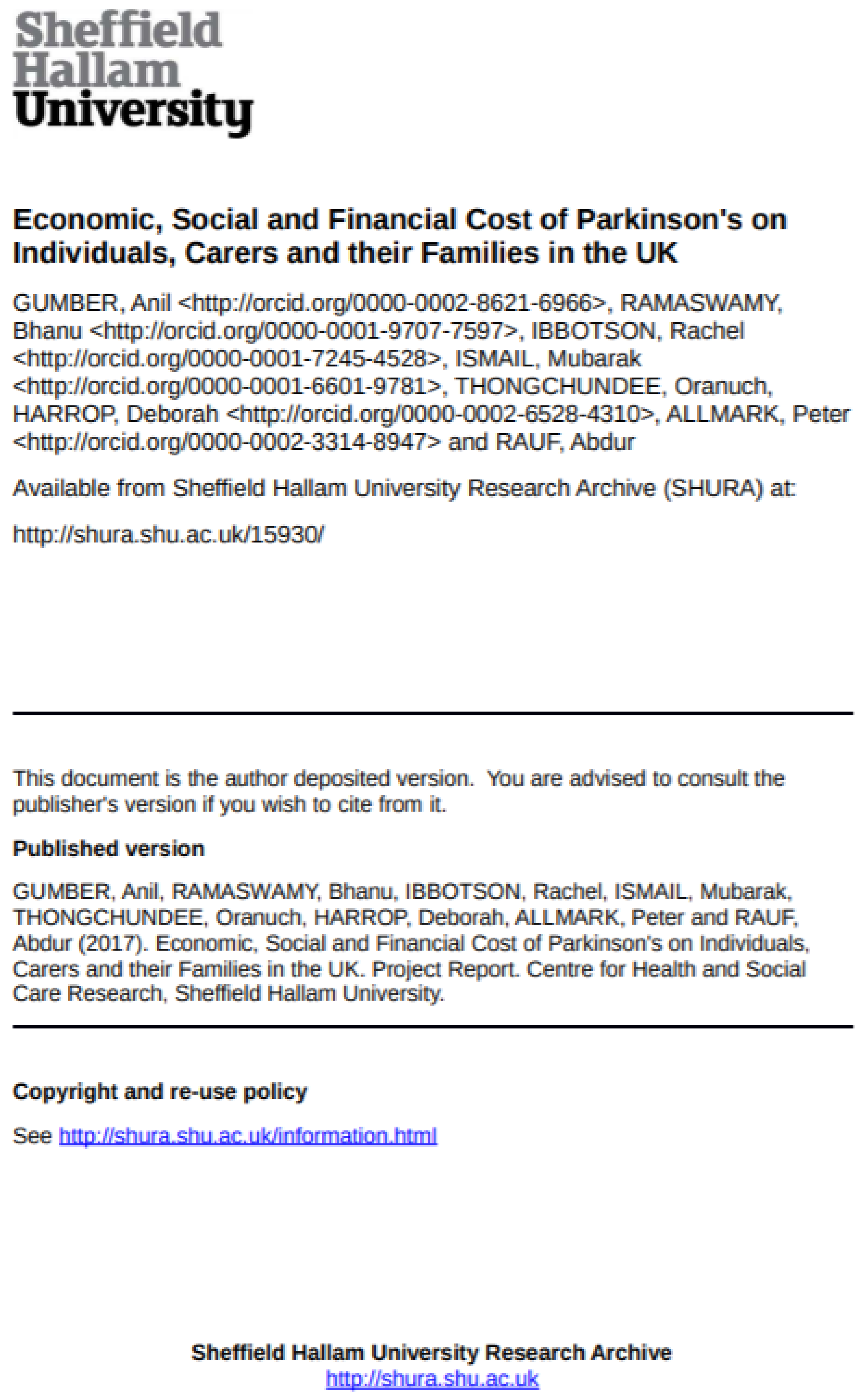
Copyright © 2005, American Psychological Association

Logged in as:
Yiqiao Xin
Account #:
3001245731

Grant Reuse

APA hereby grants permission at no charge for the following material to be reused according to your request, subject to a required credit line. Author permission is not required in this instance.

- Single text excerpts of less than 400 words (or a series of text excerpts that total less than 800 words) from APA books and journals.
- 1-3 Chart/Graph/Tables or Figures from 1 article or chapter.
- Note that scales, measures, instruments, questionnaires, photographs, or creative images are NOT included in this gratis reuse.
- Also, the abstract of a journal article may not be placed in a database for subsequent redistribution without contacting APA for permission.



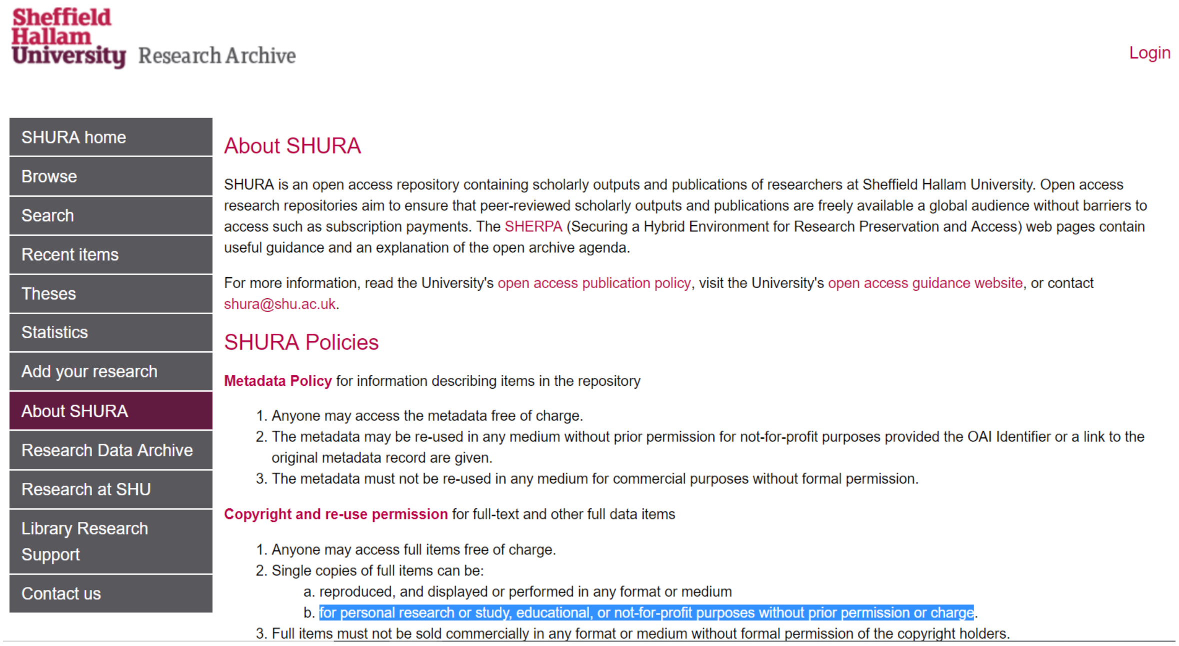
Economic, Social and Financial Cost of Parkinson's on Individuals, Carers and their Families in the UK

GUMBER, Anil <http://orcid.org/0000-0002-8621-6966>, RAMASWAMY, Bhanu <http://orcid.org/0000-0001-9707-7597>, IBBOTSON, Rachel <http://orcid.org/0000-0001-7245-4528>, ISMAIL, Mubarak <http://orcid.org/0000-0001-6601-9781>, THONGCHUNDEE, Oranuch, HARROP, Deborah <http://orcid.org/0000-0002-6528-4310>, ALLMARK, Peter <http://orcid.org/0000-0002-3314-8947> and RAUF, Abdur

Published version

GUMBER, Anil, RAMASWAMY, Bhanu, IBBOTSON, Rachel, ISMAIL, Mubarak, THONGCHUNDEE, Oranuch, HARROP, Deborah, ALLMARK, Peter and RAUF, Abdur (2017). Economic, Social and Financial Cost of Parkinson's on Individuals, Carers and their Families in the UK. Project Report. Centre for Health and Social Care Research, Sheffield Hallam University.

http://shura.shu.ac.uk/information.html



About SHURA

SHURA is an open access repository containing scholarly outputs and publications of researchers at Sheffield Hallam University. Open access research repositories aim to ensure that peer-reviewed scholarly outputs and publications are freely available a global audience without barriers to access such as subscription payments. The SHERPA (Securing a Hybrid Environment for Research Preservation and Access) web pages contain useful guidance and an explanation of the open archive agenda.

For more information, read the University's open access publication policy, visit the University's open access guidance website, or contact shura@shu.ac.uk.

SHURA Policies

Metadata Policy for information describing items in the repository

1. Anyone may access the metadata free of charge.
2. The metadata may be re-used in any medium without prior permission for not-for-profit purposes provided the OAI Identifier or a link to the original metadata record are given.
3. The metadata must not be re-used in any medium for commercial purposes without formal permission.

Copyright and re-use permission for full-text and other full data items

1. Anyone may access full items free of charge.
2. Single copies of full items can be:
   a. reproduced, and displayed or performed in any format or medium
   b. for personal research or study, educational, or not-for-profit purposes without prior permission or charge.
3. Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.