April 23, 2019

Analysis of patient-reported outcome following hip and knee replacement – interpretation, strengths, limitations and solutions



Anqi Gao

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2018

"...I am happy if I have made a small contribution." – Stephen Hawking

Acknowledgement

If it turned out that I had actually made a small contribution to society upon completion of my DPhil, the first two people I need to thank would be my academic supervisors, Professor Andrew Price and Professor David Beard. Their wealth of knowledge goes far beyond patient reported outcome measures, orthopaedics and clinical trials. This work would not have been successful without their guidance and support. Furthermore, my sincere appreciation goes to my thesis examiners, Professor Dominic Furniss and Professor Andrew Metcalf. Their insightful questions and suggestions lead to considerable improvement of this work. Moreover, I would like to express my deep gratitude to Professor Afsie Sabokbar for all the support, encouragement and useful advice in the last four years.

I would like to thank Professor Daniel Lunn from the Department of Statistics for introducing me to the fascinating R programming language and logistic regression, both of which proved extremely useful in my study later on.

Having spent the vast majority of my last four years at the Botnar Research Centre, I would like to thank my Botnar colleagues, particularly Lottie Davies, Ines Rombach, Jiyang Li, Sue King, Vandana Ayyar Gupta, Elena Benedetto and Laura Diment, for keeping me entertained and going home feeling positive every day.

I am also very lucky to have met some brilliant people outside the Botnar, special thanks goes to Simon Carr, Johana Musalkova, Lavinia Sangiorge, Judith Guo, Siddharth, and Nicola Dotti, for tolerating my quirkiness and lifting me up when I was at my nadir, through jokes, laughs and heartfelt conversations.

Lastly, a very special thanks goes to my parents for their unwavering support throughout my life.

Abstract

Total knee replacement (TKR) and total hip replacement (THR) are surgical procedures frequently performed to alleviate pain and restore function of the affected joint for people with end-stage osteoarthritis (OA). In order to make a fully informed decision, patients and providers need to acquire a thorough understanding of the short- and long-term success or failure rate of the intervention. In previous studies, various methods have been used to analyse patient-reported outcome (PRO) following TKR and THR, generating a wide range of success rates for the same population.

The thesis aimed to address the following three questions:

1. How to define a "successful outcome" following TKR and THR? What are the limitations associated with the current analysis methods?
2. How best to monitor success rate following arthroplasty in the long-term?
3. Is there an alternative way of defining a "successful outcome" that can overcome the limitations in the current methods? If so, what is the potential usage of this novel approach in a clinical context?

To assess the success rate of knee replacement and hip replacement operation, the conceptual framework of "sufficient response" was introduced. A "sufficient response" consists of three components — an instrument used to measure the outcome, a response criterion and its associated cut-off score. Using this framework, the first part of the thesis explored the short-term outcome following TKR and THR, as well as the stability and lability of long-term outcome following TKR. The limitations associated with the current methods for measuring the prevalence of success were also examined.

In the second part of this report, the Oxford patient-reported success index (OxPReSI), a novel, alternative method of defining sufficient response was introduced. OxPReSi was constructed upon two anchor questions, which are designed to address patients' self-perceived improvement and satisfaction. Application of OxPReSi divides patients into three categorical outcome groups — high benefit, low benefit and no benefit.

The last part of the thesis investigates how OxPReSi can be used pre-operatively during the decision-making period to help patients and clinicians decide on the optimal timing and type of surgery. The results demonstrate that: I) For TKR patients, a pre-operative OKS of above 17 is desirable; once pre-operative score drops below 17, the proportion of "high benefit" decreases significantly. The negative effect of low pre-operative OHS is much more minor for the outcome of THR. II) For patients with medial compartmental osteoarthritis, the outcome of partial knee replacement appeared to be superior compared with total knee replacement .

Declaration

This thesis is entirely my own work, and describes my own research.

Anqi Gao, Wolfson College

List of Abbreviations

PROM Patient reported outcome measures

HRQoL Health related quality of life

WHO World Health Organization

ICF International Classification for Functioning

KSS Knee Society Score

HHS Harris Hip Score

NHS National Health Service

TKR Total knee replacement

THR Total hip replacement

OKS Oxford Knee Score

OHS Oxford Hip Score

EQ-5D EuroQol 5-dimension

WOMAC Western Ontario and McMaster Universities Arthritis Index

OKS-APQ Oxford Knee Score - Activity and Participation Questionnaire

SF-12 Short form 12-item survey

KOS-ADLS Knee Outcome Survey - Activities of Daily Living Scale

OxPReSi Oxford Patient Reported Success Index

TOPKAT Total and Partial Knee Arthroplasty Trial

PRO Patient Reported Outcome

FDA Food and Drug Administration

SIP Sickness Impact Profile

NHP Nottingham Health Profile

CHIP-CE Child Health and Illness Profile - Child Edition

SEIQoL Schedule for the Evaluation of Individual Quality of Life

MHIQ McMaster Health Index Questionnaire

AIMS Arthritis Impact Measurement Scales

MCID Minimal Clinically Important Difference

MID Minimal Important Difference

CID Clinically Important Difference

CIR Clinically Important Responder

SF-36 Short form 36-item survey

EQ-VAS EuroQol Visual Analogue Scale

EQ-5D-3L EuroQol 5-dimension 3-level

EQ-5D-5L EuroQol 5-dimension 5-level

KOOS Knee Injury and Osteoarthritis Outcome Score

HOOS Hip Injury and Osteoarthritis Outcome Score

VAS Visual Analogue Scale

SEM Standard Error of Measurement

SD Standard Deviation

PASS Patient acceptable Symptom State

MIC Minimal Important Change

OMERACT Outcome Measures in Rheumatoid Arthritis Clinical Trials

OARSI Osteoarthritis Research Society International

HSCIC Health and Social Care Information Centre

PCS Pain Component Subscale

FCS Function Component Subscale

EFA Exploratory Factor Analysis

KAT Knee Arthroplasty Trial

ANOVA Analysis of Variance

CI Confidence Interval

SDT Signal Detection Theory

MPI Minimal Perceptible Improvement

MAS Minimal Acceptable Satisfaction

IQR Interquartile range

RCT Randomised Controlled Trial

PKR Partial knee replacement

VIF Variance Inflation Factor

OR Odds Ratio

MDC Minimal Detectable Change

ADL Activity of Daily Living

ML Machine Learning

EHR Electronic Health Record

HES Hospital Episode Statistics

EVT Extreme Value Theory

Contents

1 Introduction **1**

1.1 Overview of joint replacement 2

1.2 Origin of medical outcome study 2

1.3 Medical outcomes in the modern era 3

1.3.1 Health related quality of life (HRQoL) 4

1.3.2 The ICF model 5

1.4 Psychometric properties 7

1.4.1 Validity 8

1.4.2 Reliability 9

1.4.3 Responsiveness and sensitivity to change 10

1.5 Measuring outcome in total joint replacement 10

1.5.1 Assessing longevity and revision rate 11

1.5.2 Assessing function using clinician-administered outcome measures 12

1.5.3 Assessing function using PROMs 12

1.6 Important questions for PROM-related research 13

1.7 The scope of this thesis 15

1.8 Structure of thesis 16

1.9 Link to the next chapter 17

2 Literature Review **18**

2.1 History of PROMs 19

2.2 Basic concepts 21

2.2.1 A conceptual model of HRQoL 21

2.2.2 The conceptual framework of PROM 22

2.2.3 A new classification system for PROMs 23

2.3 Different types of PROMs 26

2.3.1 Generic instruments 26

2.3.2 Disease-specific instruments 27

2.3.3 Population-specific instruments 27

2.3.4 Individualised instruments 28

2.3.5 Summary items 28

2.3.6 Comparative evaluation of instruments 29

2.4 Implementation and benefits of PROMs 30

2.5 Interpretation of PROM data 31

2.6 PROMs in orthopaedics 34

2.6.1 Generic PROMs 35

2.6.2 Condition-specific PROMs 37

2.6.3 Other patient-reported measures related to outcome 41

2.7 Link to the next chapter 42

3 Defining a "successful" outcome after TKR/THR and improved estimation of the proportion of patients achieving this **43**

3.1 Introduction 44

3.1.1 Identifying benefit - measuring whether patients feel better. . 48

3.1.2 PASS – measure whether patients "feel satisfied" 49

3.1.3 Modified OMERACT-OARSI set of responder criteria – measure whether patients "responded" to treatment 50

3.2 Methods . . . 52
3.2.1 Data Source . . . 52
3.2.2 Outcome measures . . . 55
3.2.3 Criteria for "sufficient response" . . . 56
3.2.4 Statistical Analysis . . . 59
3.3 Results . . . 59
3.3.1 Study population characteristics . . . 59
3.3.2 Proportion of TKR patients with a good outcome . . . 61
3.3.3 Proportion of THR patients with a good outcome . . . 62
3.4 Discussion . . . 65
3.4.1 Main findings . . . 65
3.4.2 Advantages and disadvantages of different response criteria . . . 66
3.4.3 Response criteria based on patients' experience only . . . 69
3.4.4 Strengths and weaknesses . . . 69
3.4.5 Clinical implications . . . 71
3.5 Link to next chapter . . . 72

4 Temporal characteristics of self-reported outcome after TKR 74

4.1 Introduction . . . 75
4.2 Methods . . . 76
4.2.1 Data Source . . . 76
4.2.2 Criterion for a "sufficient response" . . . 78
4.2.3 Statistical analysis . . . 79
4.3 Results . . . 80
4.3.1 The group-level average OKS between one year and ten years after TKR . . . 80
4.3.2 Fluctuation in long-term post-TKR outcome . . . 82
4.3.3 When do patients achieve sufficient response following TKR? . . . 86

4.3.4 When do patients achieve optimal outcome following TKR? . 87
4.4 Discussion . . . 88
4.4.1 Main findings . . . 88
4.4.2 What this study can add . . . 91
4.4.3 Strength and limitations . . . 92
4.4.4 Conclusion . . . 93
4.5 Link to next chapter . . . 95

5 The relationship between the metrics of transition question, satisfaction question and self-reported outcome scores 96

5.1 Introduction . . . 97
5.2 Methods . . . 100
5.2.1 Data Source . . . 100
5.2.2 Statistical Analysis . . . 101
5.3 Results . . . 102
5.3.1 Distribution of transition and satisfaction in TKR and THR patients . . . 102
5.3.2 Further partitioning of patients based on responses to anchor questions . . . 104
5.3.3 Relationship between transition and satisfaction in TKR and THR patients . . . 105
5.3.4 Relationship between PROM scores and transition . . . 110
5.3.5 Relationship between PROM scores and satisfaction . . . 111
5.4 Discussion . . . 113
5.4.1 Main findings . . . 113
5.4.2 Conclusion . . . 117
5.5 Link to next chapter . . . 118

6 A novel method of assessing self-reported outcome following TKR/THR – the Oxford patient reported success index 119

6.1 Introduction . . . 120

6.2 Methods . . . 122

6.2.1 Data Source . . . 122

6.2.2 Measuring subjective perception of TKR/THR outcome . . . 123

6.3 Results . . . 126

6.3.1 Application of the new method to the national PROMs database . . . 126

6.3.2 Association between OxPReSi and the OKS/OHS . . . 127

6.4 Discussion . . . 131

6.4.1 Main findings . . . 131

6.4.2 Validity of the OxPReSi . . . 132

6.4.3 OxPReSi versus threshold-based methods . . . 134

6.4.4 Limitations . . . 135

6.5 Conclusion . . . 135

6.6 Link to next chapter . . . 136

7 Application of the modified OxPReSi to RCT data – comparing the outcome following partial and total knee replacement 137

7.1 Introduction . . . 138

7.2 Methods . . . 140

7.2.1 Data Source . . . 140

7.2.2 Outcome Measures . . . 141

7.2.3 Analysis . . . 144

7.3 Results . . . 145

7.3.1 Application of the modified Oxford patient reported success index (OxPReSi) to TOPKAT data . . . 145

7.3.2 Comparing the outcome of PKR and TKR analysed using the OxPReSi with results from the original trial 146
7.3.3 Comparing the outcome of PKR and TKR analysed using different criteria 147
7.3.4 Comparing the outcome of TOPKAT-TKR patients with results from the NHS PROM data 149
7.4 Discussion 150
7.4.1 Analysis of the TOPKAT data using the modified Oxford Patient Reported Success Index (OxPReSi) 150
7.4.2 Comparison of the results using OxPReSi with the original trial results 151
7.4.3 Analysis of the TOPKAT data using the MIC and PASS criteria. 152
7.4.4 Relationship between the modified OxPReSi and PASS score . 153
7.4.5 Comparing the outcome of RCT data with cohort data 154
7.5 Conclusion 155
7.6 Link to the next chapter 156

8 The effect of pre-operative PROM score on post-operative outcome 157

8.1 Introduction 158
8.2 Methods 160
8.2.1 Data source 160
8.2.2 Statistical method 161
8.3 Results 162
8.3.1 Description of study sample 162
8.3.2 Subgroups based on pre-operative scores 165
8.3.3 Enumerative study 165
8.3.4 Stratified study 167

8.3.5 Multiple logistic regression analysis . . . 169
8.4 Discussion . . . 172
8.4.1 Main findings . . . 172
8.4.2 Clinical relevance . . . 173
8.4.3 Strength and limitations . . . 174
8.5 Conclusion . . . 175

9 Conclusions and future directions **176**
9.1 Conclusions . . . 177
9.1.1 Defining and quantifying the "sufficient response" . . . 178
9.1.2 Analysis of longitudinal data . . . 180
9.1.3 The Oxford patient-reported success index (OxPReSi) . . . 181
9.1.4 Application of OxPReSi to the Total and Partial Knee Arthroplasty Trial data . . . 185
9.1.5 The optimal pre-operative condition for maximum post-operative benefit . . . 187
9.2 Future directions . . . 187
9.2.1 Defining and quantifying the "sufficient response" . . . 187
9.2.2 Analysis of long-term outcome following TKR . . . 189
9.2.3 Future of the Oxford Patient Reported Success Index . . . 190
9.2.4 Machine learning for research in patient-reported outcome following total joint arthroplasty . . . 191

Bibliography **194**

A Chapter Four supplementary material **222**
A.1 Ethics approval of the knee arthroplasty trial (KAT) study . . . 222
A.2 Analysis of data containing missing values . . . 223
A.2.1 Analytical methods used in this study . . . 223

A.2.2 Results from complete-case analysis 223
A.2.3 Conclusion 226

B Chapter 7 supplementary material **227**

C Instruments for collection of patient-reported outcome data in orthopaedics (hip and knee) **256**

D Chapter 8 supplementary material **268**

List of Figures

1.1 Interactions between the components of ICF (WHO 2001) 6

1.2 Using the ICF classification on a knee OA patient 7

2.1 PRO and HRQL publications 1976-2013 (Mapi, an ICON plc company). 20

2.2 Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995; 273: 59-65. 21

2.3 Diagram of the conceptual framework of a PRO instrument (Deshpande et al 2011) 22

3.1 Flow chart showing how TKR data were sorted 53

3.2 Flow chart showing how THR data were sorted 54

3.3 Proportions of TKR and THR patients with successful outcome – depending on the criteria, 72% to 90% of TKR patients and 77% to 95% of THR patients have obtained a "sufficient response" following their operation. 64

4.1 Initially 2,217 patients were recruited and received operation with allocated type of implant; the lightest shade represents the maximum sample cohort size. The medium shade represents those patients whose post-operative OKS were recorded in the KAT database – these were the number of patients who actually remained in the trial each year. The darkest shade represents those patients, with at least three years of follow-up data, whose post-operative OKS were included in the final analysis. 78

4.2 The mean pre-operative and post-operative OKS following TKR each year. Error bars represent 95% confidence interval. 80

4.3 Box-whisker plot of the distribution of all participants' post-operative OKS observed at 10 years conditional on their post-op OKS at one year. There is a strong, positive linear correlation between patients' post-op OKS at ten years and at one year (r=0.64, β_1=0.72 (0.64,0.80)). 82

4.4 Bar chart showing the long-term outcome of those 71% patients who achieved a successful outcome (i.e. Δ OKS$\geq$11) at one-year follow up. Data from patients with at least three years of OKS follow-up data and a sufficient response (ΔOKS $\geqslant$ 11) at year-one post-operation were included in the analysis. (n = 977) 84

4.5 Bar chart showing long-term outcome of the 29% patients who failed to achieve a successful outcome (i.e. ΔOKS$\leq$10) at one year post-TKR. Data from patients with at least three years of OKS follow-up data and failed to achieve a sufficient response at year-one post-operation were included in the analysis. (n = 382) 85

4.6 Bar chart showing the distribution of the earliest time patients achieve a sufficient response (ΔOKS $\geqslant$ 11). Data from patients with at least three years of OKS follow-up data and achieved a sufficient response at some point between one-year and ten-year post-operation were included in the analysis (n = 1205). 87

4.7 Bar chart illustrating the time it takes for patients to reach their optimal outcome (i.e. highest post-op OKS). Data from patients with at least three years of OKS follow-up data were included in the analysis (n = 1359). 88

5.1 Comparing the distribution of TKR vs THR patients' transition rating (i.e. self-perceived change in symptoms) at six-month post operation. The predominant rating choice of both TKR and THR patients is "much better". 102

5.2 Comparison of the distribution of TKR vs THR patients' satisfaction at six-month post operation. Over 40% THR patients rated their operation as "excellent", and approximately a quarter of TKR patients did so. The majority of both TKR and THR patients selected one of the three favourable response choices ("excellent", "very good" and "good"). 103

5.3 A 3-D graphical representation of data in Table 5.4 , showing the proportion (represented by height of the prism) of different satisfaction levels (represented by different shades) within each transition category for TKR patients. 107

5.4 A 3-D graphical representation of data in table 5.5 , showing the proportion (represented by height of the prism) of different satisfaction levels (represented by different shades, with the darkest being the most satisfied) within each transition category for THR patients. . . 109

5.5 Box-whisker plots illustrating how Δ OKS/OHS change according to patients' transition rating. Box width is proportional to patient number in its respective transition category. There is a low to moderate, positive correlation between patients' OKS/OHS and their transition rating, with the relationship being slightly stronger in TKR (TKR: $\rho = 0.57$, THR: $\rho = 0.45$). 111

5.6 Box-whisker plots illustrating how Δ OKS/OHS (y-axis) change according to patients' satisfaction level (x-axis). The width of each box is proportional to patient number in its respective category. Results from analysis of variance indicates that there is a moderate, positive, linear correlation between patients' change in PROM scores and their satisfaction (TKR: $\rho = 0.55$, THR: $\rho = 0.44$) 112

6.1 The Oxford Patient Reported Success Index – an anchor-based index for assessing the outcome of TKR/THR (MPI = minimal perceptible improvement, MAS = minimal acceptable satisfaction). 125

6.2 The proportion of TKR patients versus THR patients in each outcome group (2014-2015 data). There is a greater proportion of THR patients obtaining a "high benefit" compared to TKR patients. . . . 127

6.3 Comparison of the change in OKS of each outcome category for TKR patients with the change in OHS for THR patients in corresponding outcome group (2014-2015). The width of the box is proportional to the number of patients. (TKR–n=43,115 THR–n=40,917) 130

7.1 The two-step modified Oxford Patient Reported Success Index (MPI = minimal perceptible improvement, MAS = minimal acceptable satisfaction). 143

7.2 Outcome of PKR vs TKR patients from TOPKAT assessed by the Oxford Patient Reported Success Index. 146

8.1 The distribution of post-operative outcome (categorised as "high benefit", "low benefit" and "no benefit") in TKR (left) and THR (right) patients according to their pre-operative OKS/OHS. 166

8.2 Comparing the change in the prevalence of sufficient response (i.e. high benefit) between TKR and THR patients, as pre-operative OKS/OHS increases. 167

8.3 Comparing the change in the prevalence of "sufficient response" between TKR and THR patients stratified by age and gender, as pre-operative OKS/OHS increases. The two panels on the left are THR results, and TKR results are shown on the right. Results from female patients are shown at the top, and results from males shown at the bottom. Results from different age groups are shown in different colours as indicated by figure legends. 168

9.1 The levels measured by the OKS/OHS are marked by red circles, and the levels covered by the OxPReSI method are marked by blue squares. (Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995; 273: 59-65.) 183

A.1 Bar chart showing the long-term outcome of the patients who achieved a successful outcome at one-year follow up. (complete-case analysis, n=328) 224

A.2 Bar chart showing the long-term outcome of patients who had a "failed" outcome at one-year post-TKR. (complete-case analysis (n=121) 225

List of Tables

2.1 A classification system for patient reported outcome measures: axis and categories (Valderas 2008) 24

3.1 Common outcome measures, response criteria, and associated cut-off point for TKR. The last column shows the outcome in terms of percentage of those who have done well. The values were reported by different groups using different data sets, different instruments and different response criteria. 46

3.2 Common outcome measures, response criteria, and associated cut-off point for THR. The last column shows the outcome reported by different groups using different data sets, different instruments and different response criteria. 47

3.3 Threshold values for each criteria for TKR 57

3.4 Threshold values for each criteria for THR 58

3.5 Baseline Descriptive Statistics for TKR patients (NHS PROMs 2014-2015) 60

3.6 Baseline Descriptive Statistics for THR patients (NHS PROMs 2014-2015) 61

3.7 Proportion of TKR patients with a good outcome by criterion 62

3.8 Proportion of THR patients with a good outcome by criterion 63

3.9 Proportion of TKR and THR patients with a good outcome judging by satisfaction and transition anchor questions 63

5.1 Common response criteria, anchor questions and anchor-based methods that investigators have used to estimate the associated cut-off values. 98

5.2 According to their responses to both anchor questions, all TKR patients were separated into 25 divisions. Patients within each division rated satisfaction and change in symptom level identically. Almost three quarters of patients selected one of the three divisions in the top left corner – much better-excellent, much better-very good, and much better-good. 17% of TKR patients experienced either lack of improvement or lack of satisfaction or neither. 104

5.3 According to their responses to both anchor questions, all THR patients were separated into 25 divisions. Patients within each division rated satisfaction and change in symptom level identically. 85% of total THR patients selected one of the three divisions in the top left corner – much better-excellent, much better-very good ,and much better-good. 9% patients experienced either lack of improvement or lack of satisfaction or neither. 105

5.4 Comparing the distribution of satisfaction level for TKR patients within each transition category. For TKR patients who experienced an improvement ("much better" or " a little better") in symptoms after operation, the predominant response choices for satisfaction were "very good" and "good". Level of satisfaction drops as patients' experience regarding change in symptoms becomes less favourable – between 70% and 90% of those who experienced no improvement or worsening in symptoms (i.e. "same", "a little worse" or "much worse") felt unsatisfied (i.e. "fair" or "poor"). Overall relationship between transition and satisfaction in TKR was $\tau = 0.61$ (n=46094). 106

5.5 Comparing the distribution of satisfaction level for THR patients within each transition category. Patients who experienced an improvement in symptoms – feeling "much better" or " a little better" post- operatively, predominantly rated their satisfaction as "excellent" and "good", respectively. Level of satisfaction drops as patients' experience regarding post-operative symptom level worsens – between 60% and 90% of those whose symptoms did not improve or declined (i.e. "same", "a little worse" or "much worse") felt unsatisfied (i.e. "fair" or "poor") after THR. Overall correlation between transition and satisfaction in THR was $\tau = 0.49$ (n=44015). 108

5.6 The change in OKS and OHS from pre- to post-operation for TKR and THR patients with different transition ratings. The greatest difference in Δ OKS/OHS is seen between patients who felt "much better" and those feeling " a little better". Also, patients who felt " the same" or "a little worse" also exhibited a positive score change. . 110

5.7 The change in OKS and OHS for TKR and THR patients with different levels of satisfaction. 112

6.1 Comparing OxPReSi with existing threshold-based methods of measuring post-operative outcome, in the context of the "sufficient response" framework. 124
6.2 Proportion, mean Δ OKS, interquartile range (IQR) of Δ OKS and effect size for TKR patients in each outcome category (classified using OxPreSi) from 2012 to 2015. 128
6.3 Proportion, mean Δ OHS, interquartile range (IQR) of Δ OHS and effect size for THR patients in each outcome category (separated using OxPreSi) from 2012 to 2015. 129

7.1 Comparison of the original Oxford Patient Reported Success Index with the modified version 143
7.2 The response criteria used in this study – OxPReSi, MIC and PASS score. They each provide a different benchmark for a "sufficient response" to the intervention. 144
7.3 Comparison of the results from the original trial with results obtained using the OxPReSi method. 147
7.4 Proportion of PKR and TKR patients reaching a desired outcome level assessed using different criteria. 147
7.5 Proportion of patients succeeded and failed to meet the PASS score cut-off amongst those classified as having "high benefit" (top row) and those classified as having "low/no benefit" (bottom row). 148
7.6 Comparison of the pre-operative score of patients having conflicting outcomes (according to the PASS score and OxPReSi criterion) with the average pre-operative score of the trial. 149
7.7 Comparison of the outcome of TOPKAT-TKR and NHS PROMs data in terms of the proportion of patients achieving a "sufficient response" according to various criteria. 150

8.1 Descriptive statistics of patients whose data were included (completers) and excluded (non-completers) for logistic regression modelling. 164

8.2 Subgroups based on pre-operative scores. 165

8.3 Multiple logistic regression analysis for investigating the effect of pre-operative OKS/OHS on post-operative outcome. 170

8.4 Table 8.4 above combined with Table 8.3 in the previous page show the results from multiple logistic regression analysis, the purpose of which is to determine the effect of pre-operative scores on post-operative TKR/THR outcome while controlling for other confounding factors (e.g. age, co-morbidities). The post-operative outcome is expressed as the chance of obtaining a "sufficient response" assessed using the new, anchor-based OxPReSi method. 171

A.1 Proportion achieving a sufficient response for the first time each year (amongst those who did meet the PASS criterion some time between year-one and year-ten). Results from complete-case analysis (bottom row, n=402) and from incomplete case analysis (n=1205) are highly similar. 226

D.1 Multinomial logistic regression analysis 269

D.2 Table D.2 above and combined with D.1 before show results from ordinal regression analysis, the purpose of which was to determine the effects of different factors on the odds of achieving a good or bad outcome comparing to the intermediate outcome, using the OxCom method. 270

Chapter 1

Introduction

1.1 Overview of joint replacement

The first total joint replacement was performed on the 20th of May, 1890 by Professor Themistocles Gluck on a 17-year-old lady, whose tuberculous knee joint was replaced by a hinged ivory prosthesis [1]. A year later, Gluck manufactured and implanted the first artificial hip. Since then, knee and hip arthroplasty have advanced markedly in terms of both implant design and surgical techniques [2]. In the last twenty years, there has been a surge in the number of patients undergoing knee or hip arthroplasty worldwide. Last year, approximately 200,000 total hip and knee replacement (TKR and THR) operations were performed across England and Wales [3], and well over 400,000 procedures were carried out in the United States [4]. As more and more patients suffering from joint problems opt for TKR and THR, effective assessment of surgical outcome becomes crucial.

Before delving into measuring the outcome of total hip and knee replacement (which will be discussed at length in Chapter Two), the remainder of this chapter provides an overview on measuring health and medical outcomes in general.

1.2 Origin of medical outcome study

The founder of modern medical outcome studies is Dr Ernest Codman, a pioneering surgeon who was highly accomplished in six areas of medicine – anaesthesiology, radiology, duodenal ulcer surgery, orthopaedic oncology, shoulder surgery, and the study of medical outcomes [5]. In 1905, Codman was the first physician to promulgate the need for medical outcomes studies. At the time, he described the problem of measuring outcomes in his "end result idea", which states:

"The common sense notion that every hospital should follow every patient it treats,

long enough to determine whether or not the treatment has been successful, and then to inquire, 'If not (successful), why not?' with a view to preventing similar failures in the future [6]."

Codman followed up on his new found passion by keeping track of his patients using "end results cards" – index cards containing basic demographic data, diagnosis, treatment received and outcome, on every patient he treated [7]. These "end results cards", which were deemed unnecessary by most medics at the time, marked the beginning of modern outcome studies and evidence-based medicine. Codman also wrote:

"We believe it is the duty of every hospital to establish a follow-up system, so that as far as possible the result of every case will be available at all times for investigation by members of the staff, the trustees, or administration, or by other authorized investigators or statisticians [7]."

1.3 Medical outcomes in the modern era

Since the 1970s , outcome measures have been playing an increasingly important role in medical practice, providing essential information for both clinical audit and academic research. The British Paediatric Association outcome measure working group defined "health status measurement", "outcome measurement" and "implied outcome measurement" as follows:

- " A **health status measurement** is a direct measure of some aspect of health in which improvement is sought, whether or not its relationship to any intervention, social or environmental circumstance is understood [8]."
- " An **outcome measurement** is a sub-type of health status measurement where changes in the measure are known (or at least believe) to be largely

attributable to a health service intervention [8]."

- " An **implied outcome measurement** is an indirect indicator of health that can be used as a valid proxy for an outcome measurement. In our view, the health status measurement may also be described as *'function status measurement'* and *'quality of life measure'* [8]."

Outcome measures are designed to assess quality of life, therefore the evolution of modern medical outcomes study takes place within the historical context of quality of life research [9].

1.3.1 Health related quality of life (HRQoL)

The first documented reference to quality of life is in A.C. Pigou's *The economics of Welfare* (1920) where he discusses the impact of government support on the well being of the poor and on the overall national finances. The book, despite possessing a fresh and unorthodox flavor, failed to prompt much response from the wider public [10]. Almost 30 years later, in 1948, the World Health Organization (WHO) redefined health as " A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity ". This new, broader definition of health sparked considerable interest in searching for new ways to measure HRQoL [10].

HRQoL was first mentioned in the medical literature by J.R Elkinton in the article " Medicine and the Quality of Life" published in the Annals of Internal Medicine in 1966, and the number of publications in HRQoL measurement has grown significantly since then [9]. In 1977, the US National Library of Medicine officially incorporated "quality of life" as a key word in their MEDLINE computer search system. Following a MEDLINE search for the period between 1966 and 1985, just

over three thousands references related to quality of life can be found. Following a search between 1986 and 1994 over ten thousand references related to quality of life are available [10].

When a patient's health is affected by diseases, preserving function and minimising disability becomes vital for optimizing their quality of life. The most widely recognised framework that provides a conceptual base for measuring health and disability is the International Classification of Functioning, Disability and Health (ICF), which was approved by the World Health Assembly for use in various sectors in 2001.

1.3.2 The ICF model

The ICF is a multi-purpose classification system designed to be used in a variety of disciplines and sectors, and across different cultures. The aims of the ICF include I) to provide a scientific basis for understanding health, health outcomes and determinants of health; II) establish a common language for describing health in order to improve communication between users working in different disciplines (e.g. clinicians, social workers, policy makers); III) permit comparison of data across countries and cultures [11]. In the ICF, function and disability are viewed as multi-dimensional concepts comprised of the following components:

- **Body functions and structures** of people, and impairments thereof – functioning/disability at the level of the body.
- **Activities** of people, and the activity limitations they experience – functioning/disability at the level of the individual.
- **Participation** of people in different areas of life, and participation restrictions they experience – functioning/disability at the level of societal involvement.
- Environmental factors, which are classified as either facilitators or barriers.

In this updated ICF model, functioning and disability are seen as a complex, bi-directional interactions between the specific condition (affecting the health of the individual) and other contextual factors, both environmental and personal. The new classification treats these dimensions as interactive and dynamic rather than linear or static. The basis of this new framework is best illustrated in Figure 1.1 proposed by the World Health Organization (WHO)[11].

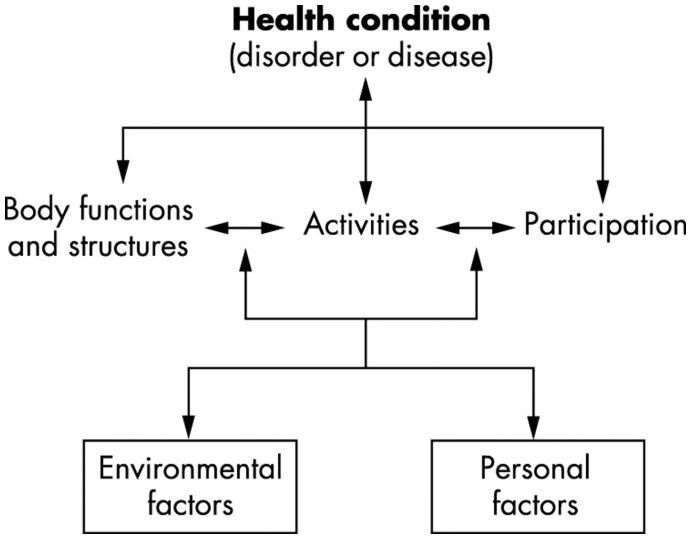


Figure 1.1: Interactions between the components of ICF (WHO 2001)

When applying the ICF classification individuals will not get a straight-forward 'yes' or 'no' answer about whether they are disabled. Instead, the ICF provides a "descriptive profile" of an individual's pattern of functioning [11]. Figure 1.2 illustrates how the ICF can be applied to a patient suffering from osteoarthritis (OA) at the knee.

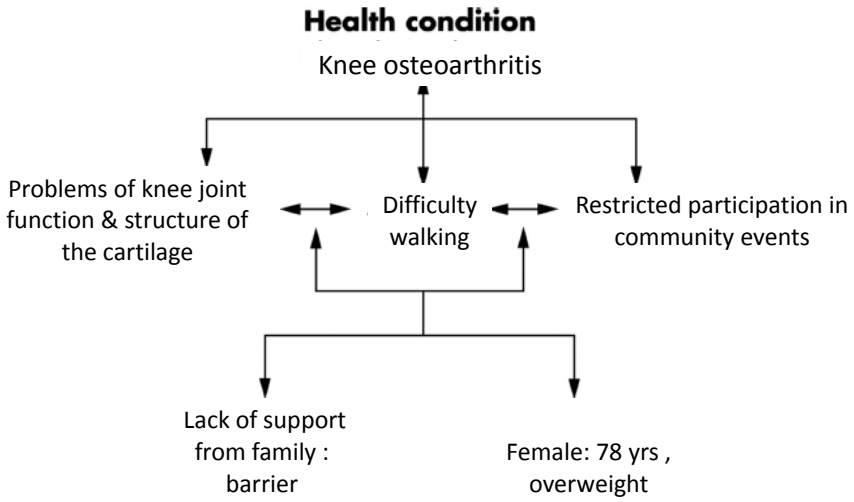


Figure 1.2: Using the ICF classification on a knee OA patient

Rather than only focusing on one's disease, the current ICF model describes how a disease affects a person's function at different component levels – the body, the individual and involvement in society. Different measurement tools are needed to assess the level of disability at each level. Clinical tests are usually performed in order to determine the severity of damage to the body structure, whereas the impact on 'activities' and 'participation' are usually best addressed by patient reported outcome measures (PROMs) [9].

1.4 Psychometric properties

Patient-reported outcome measure (PROM) is a relatively new phenomenon in outcome measure research. In the field of health measurement, traditionally, there is a preference for objective laboratory test results over subjective personal reports in determining an accurate diagnosis or assessing the effectiveness of an intervention. This bias in scientific medicine started to shift in the early 1990s, when physicians started to realise that personal reports on medical and social histories from the patients are pertinent resources in deciphering the epidemiology of a condition and

determining the preferred type of treatment by the patients [9]. This section presents an overview of the major psychometric properties that one needs to consider before selecting a specific PROM for their intended purpose.

1.4.1 Validity

Essentially, validity emphasizes that an instrument actually measures what it purports to. The validity of a specific measure can usually be assessed in three aspects – construct, content and criterion.

Construct validity refers to the extent to which the score of a particular measurement can truly represent the construct to be measured. A *construct* can be construed as a mini-theory that can potentially explain the relationships among various behaviours or attitudes [12]. Examples of *construct* include anxiety and irritable bowel syndrome (IBS), for which there are no "unequivocal" medical tests available, and therefore diagnoses are made based on the presence and absence of various observable symptoms.

Most PROMs are designed to measure constructs, entities that are not directly observable. For example, if Index A is a new questionnaire designed to assess "*outcome following TKR operation*", one prediction would be that patients with higher scores on Index A would demonstrate much greater range of motion at the knee, lower level of pain, less limitation on performing daily activities and more extensive participation in social events than patients with lower scores on Index A. If it turns out that there is no distinguishable difference in terms of pain and functional ability between patients with very high and very low scores on Index A, Index A would be considered to have poor construct validity.

Content validity refers to "the degree to which the instrument content adequately

reflects the construct that is being measured [13, 14]." In other words, when using a measurement with higher content validity, we are able to draw broader valid inferences about a person under various conditions than we would using a measurement with lower content validity. For example, grip strength has poor content validity when used as a measure to assess the symptom level of a rheumatoid patients, because it does not allow the clinician to make inferences about other aspects of the patients' condition, such as morning stiffness and pain [12].

Criterion validity is the correlation of a new scale with the "gold standard" measurement of the trait or disorder under study. The next question that naturally arises is if a gold standard already exists, why bother developing a new scale? Potential reasons that justify the labour of creating a new measure include the gold standard being too expensive, too dangerous or too time consuming. For example, to check for tuberculosis, chest X-ray is regarded as the gold standard test producing a definitive diagnosis. However, chest X-ray involves using expensive machines and exposing the patient to radiation. Therefore, the need for a cheaper and less-invasive measurement tool still exists [12].

1.4.2 Reliability

Reliability refers to the degree to which an instrument can produce consistent results on different occasions (involving differences in time, location, observer, etc) when there is no evidence of real change [12, 15]. In practice, the reliability of an instrument is assessed by its *reliability coefficient*, which is calculated as follows:

$$Reliability = \frac{\text{Subject Variability}}{\text{Subject Variability} + \text{Measurement Error}}$$

The reliability coefficient is a value, which indicates the ratio between the variability

between patients and the total variability, which is the sum of between-patient variability and measurement error. The value of reliability coefficient falls between 0 and 1; instruments with higher coefficients are considered to be more reliable [12].

1.4.3 Responsiveness and sensitivity to change

Responsiveness and *sensitivity to change* share commonalities, but should not be used interchangeably. Sensitivity to change refers to the ability of an instrument to detect change, regardless of whether the change is significant from a clinical point of view [16]. Responsiveness, on the other hand, is defined as the ability of an instrument to measure any clinically meaningful change in status [12, 17]. The question that follows is what is a 'clinically meaningful change'. Different definitions of minimal clinically important change/difference will be discussed extensively in later chapters of this thesis. Some authors have considered responsiveness and sensitivity to change as an aspect of validity, closely linked to criterion validity - viz. how does the change detected by this new instrument correlate with change measured by other existing instrument [12]?

This section has presented an overview of the major psychometric properties that one needs to consider when developing or selecting a measurement for a specific purpose. The next section focuses on the historical development in the way outcome is measured following total knee and hip replacement – from implant survival (i.e. revision) to clinician-reported outcome measures to PROMs.

1.5 Measuring outcome in total joint replacement

Since the 1970s, accompanying the advancement in the surgical aspect of total joint replacement procedures, there has been an evolution in how TKR and THR out-

comes are assessed.

1.5.1 Assessing longevity and revision rate

In the early 1970s, Professor Goran Bauer, the head of the Department of Orthopaedics in Lund University, foresaw the need to initiate a national registry to monitor the outcome of knee arthroplasty. The Swedish Knee Arthroplasty Project, the first nation-wide orthopaedic registry, was officially established in 1975, the goal was to collect data from patients across the country, evaluate the results and identify types and rates of failure [18, 19].

There was a lack of consensus on how "success" or "failure" should be *systematically* measured regarding TKR at the time - loosening? instability? infection? After some discussion, the orthopaedic community agreed upon using revision as the measurement for failure, as it is an event that can be recorded with precision and no ambiguity. Patients would need to come back to the same hospital where they received primary arthroplasty if a revision was needed [20]. A combination of registry data and other available patient information would allow researchers to identify inferiorities in implant design and surgical techniques.

However, there are some problems with using revision as the marker or equivalent of failure. Even though having an early revision unambiguously indicates a failure, the absence of revision does not always mean a successful operation. Revision surgery is a high-risk, high-cost procedure [21] that could be delayed (significantly or sometimes indefinitely) by multiple extrinsic factors, including patients being unfit for surgery, a lack of hospital resources, etc [22]. Notably, some patients may have a "failed" arthroplasty in terms of continuing pain and dysfunction, but not undergo revision.

1.5.2 Assessing function using clinician-administered outcome measures

In order to overcome the inherent limitations associated with using revision as a measure of failure, researchers started seeking alternative ways of assessing TKR, and clinician-administered outcome measures came into prominence, including the Knee Society Score (KSS), Harris Hip Score (HHS) and Knee Society roentgenographic evaluation System [20]. The main issue with these outcome measures is that the final judgement call is made by the clinicians, whose opinions are sometimes in discordance with those of the patients. Furthermore, results from studies conducted in various countries consistently indicate that patients are often "insightful judges" of their own health, morbidity and mortality [9].

1.5.3 Assessing function using PROMs

In 2008, the Darzi report was published – Lord Darzi proposed that "quality is to be at the heart of everything we (the National Health Service) do", and an important dimension of quality is effectiveness of care. He emphasised the importance of letting the patients themselves assess the effectiveness of their treatment using PROMs [23]. Following the release of the Darzi report, PROMs have been used as the primary outcome measure across England and Wales for all National Health Service (NHS) patients undergoing hip or knee replacement surgeries. The official instruments used in the NHS PROM program across England and Wales are the Oxford Knee Score or Oxford Hip Score (OKS or OHS) and EQ-5D (a standardized instrument developed by the EuroQol Group). The OKS and OHS are classified as "condition-specific" outcome measures [24], whereas the EQ-5D as a "generic" measure [25]. Other widely recognised PROMs include the WOMAC, Oxford Knee Score - Activity and Participation Questionnaire (OKS-APQ), 12-item short form survey (SF-12) and

Knee Outcome Survey – Activities of Daily Living Scale (KOS-ADLS) [26]. A detailed explanation of the advantages and disadvantages of PROMs in general, as well as, different types of PROMs used in orthopaedics will be presented in Chapter Two of this thesis.

1.6 Important questions for PROM-related research

First, there has been some debate regarding the definition and format of PROMs. For the purpose of this thesis, PROM refers to "an umbrella term that incorporates any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" [27]. Also, in this thesis, the term PROM are limited to fully standard questionnaires with fixed items and response options.

Over the last decade, the use of PROMs in orthopaedics grew significantly worldwide, but the extent to which PROMs have been implemented has been somewhat varied. In England, the NHS PROM program, which invites all patients undergoing TKR and THR to complete a PROM questionnaire before and after surgery, has been in operation since 2009 [28]. Given the scale of the NHS programme, it is the most extensive implementation of PROMs to date in the world, providing a substantial body of routinely collected national data. Therefore, the identification of any problems associated with analysis of these data, and lessons that can be learned, is particularly valuable [29]. Some important problems within PROM research that need addressing are:

- **Interpretation of short-term PROM scores** – the key to interpreting numerical PROM scores is the ability to place the magnitude of these scores in a meaningful context [30]. For example, what does it mean to a patient

who had TKR six months ago that their OKS is now 10 points higher than pre-operation? Various criteria in the form of categorical cut-off points have been developed in order to dichotomise patients into those who have achieved a 'successful' outcome and those who have not [31]. Some cut-offs are higher than others; the "gold standard" criterion still remains unknown. The general consensus is that TKR/THR provides extensive benefits to the majority of patients, but since the choice of criterion makes a substantial difference to this postulate (presented in terms of the proportion of 'successful' outcome), the actual prevalence of 'success' following TKR/THR is still undetermined.

- **Interpretation of long-term PROM scores** – Joint replacement surgeries are designed to provide benefits to patients in both the short and long term. So far, the vast majority of TKR outcome studies have focused on the short-term results, publication data on long-term post-operative pain and functional outcome following TKR are relatively scarce with flawed methods and inconsistent results.

- **Analysing information provided by anchor questions** – A global score is one derived from a response to a single global question, an example of which is "In general, how would you rate your overall health?". The choice of answers usually range between "excellent", "very good", "good" " fair" and "poor". By designating a score to each answer choice, a global score is provided [9]. Earlier studies in the quality of life in cancer patients have shown that results from these single item measures of health state are consistent with those from more complex multi-item PROMs [32, 33]. While global questions may refer to general health, an alternative usage is to "anchor" the question to a specific health concern, such as symptoms at the knee for knee OA patients [9]. In the NHS PROM program, beside the OKS/OHS and EQ-5D, two anchor questions addressing the transition of symptoms at the knee and satisfaction with the

operation are also included. These anchor questions are, without a doubt, the shortest, simplest and least expensive assessment. Despite evidence from previous publications supporting the validity of such single, global questions, so far anchor question data have mainly been used only for the purpose of deriving categorical cut-off scores in orthopaedics. They have not been used to address outcome after intervention. One major area is to explore whether there is a better way to make use of the large data provided by anchor questions.

1.7 The scope of this thesis

This thesis explores the limitations in the way PROM data containing the OKS/OHS and anchor questions (transition and satisfaction) have been analysed, and identify solutions that can overcome those deficiencies to evaluate post-TKR/THR outcome more accurately and comprehensively. The thesis consists of three parts:

1. Exploratory studies to address: a) variations in reported outcome, both short- and long-term, following TKR and THR due to inconsistencies in the way how a "successful outcome" is calculated. b) the problem of insufficient data and a lack of understanding of the relationship between PROM scores and anchor questions frequently used in the derivation of clinically important cut-off scores.

2. Introduction of a solution to the issue identified in (1) — the development of the Oxford Patient Reported Success Index (OxPReSi), a new outcome assessment method utilising and incorporating patients' responses to the transition and satisfaction anchor questions to supplement existing measures such as the OKS/OHS.

3. Application of the new method, OxPReSi to: a) investigate the effect of pre-

operative score on post-operative outcome (where outcome is measured using OxPReSi). b) review and confirm interpretation of one year outcome results from a large RCT of arthroplasty using the modified version of OxPReSi.

1.8 Structure of thesis

The structure of this thesis is as follows:

Chapter 2 presents a literature review of different outcome measures for assessing the outcome following total hip/knee replacement and the severity of symptoms at the knee/hip.

Chapter 3 first sets up the new framework of evaluating post-operative outcome following TKR/THR – in terms of the proportion of patients achieving a "sufficient response". This chapter then reports the results from a cross-sectional study where a single outcome measure – the OKS/OHS, and a set of four different response criteria and their associated cut-off scores were used to assess the sufficient response rate of a nationally collected population data at 6-months post-operation.

Chapter 4 presents results from a longitudinal study investigating the long-term outcome trajectory as measured by change over time in the proportion of patients having obtained a sufficient response and the proportion of those who failed to do so following TKR.

Chapter 5 examines the relationship between anchor questions and PROM scores, in addition to, the relationship between different anchor questions which are designed to address different aspects of outcome following TKR and THR.

Chapter 6 introduces the Oxford Patient Reported Success Index (OxPReSi) — a new outcome assessment method for identifying patients who have achieved a high

benefit, low benefit and no benefit.

Chapter 7 explores whether the OxPReSi, the new metric developed using population data, would be suitable for measuring outcome in randomised controlled trials (Total and Partial Knee Arthroplasty Trial). Specifically, does the new OxPReSi reveal the same or similar patterns as were identified by the original outcome measures for the TOPKAT study design.

Chapter 8 presents how the new method, OxPReSi, can help answer an old (but still unresolved) question – the effect of pre-operative PROM scores on post-operative outcome in TKR and THR patients.

Chapter 9 summarises the clinical relevance of the findings presented in this thesis, describes on-going work and provides recommendations for future work.

1.9 Link to the next chapter

This chapter presented the overall scope, specific objectives and the structure of this thesis. Also, a brief review of the historical development of medical outcomes study and the evolution of outcome measurement in the field of total hip and knee replacement were provided. The next chapter will discuss PROMs and a variety of outcome measures in orthopaedics in detail.

Chapter 2

Literature Review

2.1 History of PROMs

In the traditional health care delivery system, important decisions regarding what was good for the nation's health were made by government officials and medical experts. The consumers, on the other hand, did not have nearly as much influence as other more important stakeholders [9]. This traditional system started undergoing substantial reform towards the end of the 1980s after experts observed that the rise in the cost of care had not lead to reduced morbidity, mortality or disability [9].

In 1990, in an effort to investigate successful practice in medicine and how to make clinical decisions that are truly reflective of the needs of patients, Wennberg and associates conducted a pilot study on outcome research, where recruited patients were acknowledged as "critical contributors to their own diagnosis and treatment [9, 34]." Before treatment, patients were asked to provide information on their symptoms, functional status, health perception, etc. Throughout the trial, patients' subjective health assessment was essential in formulating diagnosis and treatment plans. The results demonstrated that patients often have a good idea of their own condition and what is the best course of treatment for them to follow [34]. Research performed by Cleary and colleagues indicated that patients are also "informative and useful evaluators" of the intervention they have received [35]. Moreover, over time there was an accumulation of evidence coming from different countries and cultures suggesting that people with serious health conditions were often better at predicting their future functional status than clinicians [9, 36].

This growing interest and understanding about the importance of measuring patients' own perceptions about their illness and quality of life is reflected in academic research. The first article incorporating "patient reported outcome (PRO)" in the title was published in 1976 [37], the first article with "HRQoL" in the title was published in 1982 [38]. Since then, the field of patient reported outcomes research has

expanded rapidly and the amount of publications has grown exponentially (shown in Figure 2.1 [39].

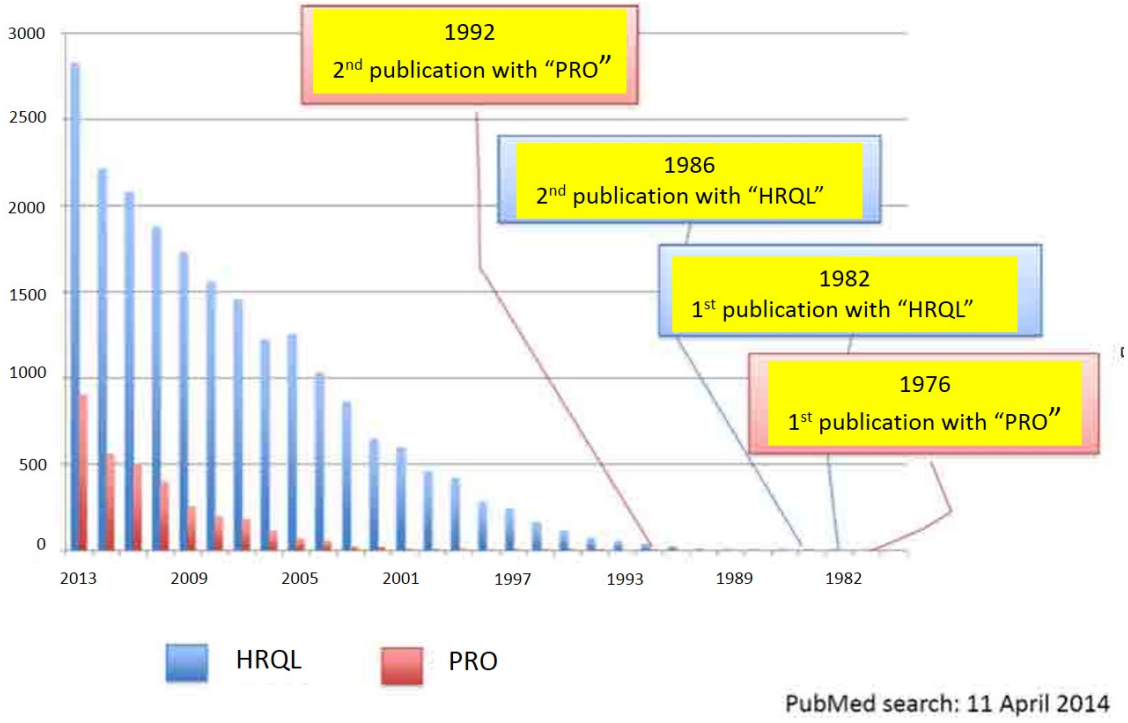


Figure 2.1: PRO and HRQL publications 1976-2013 (Mapi, an ICON plc company).

Nowadays, PROMs play an important role in research, audit and clinical governance. To ensure PROMs are applied and interpreted properly, it is essential for both researchers and clinicians to gain a comprehensive knowledge of this area. The next section discusses the Wilson & Cleary model of HRQoL, the PRO conceptual framework recommended by the FDA, and the new classification system for PROMs, all of which form the basis of modern PRO study.

2.2 Basic concepts

2.2.1 A conceptual model of HRQoL

In 1995 Wilson and Cleary developed the conceptual model of HRQoL, which categorised PROMs according to the underlying health concepts they represented, and proposed specific causal relationships between these concepts. The model conceptualises patient outcomes in five levels in order of increased complexity: biological and psychological variables, symptom status, functional status, general health perception and overall quality of life [40]. Details of the model is shown in Figure 2.2.

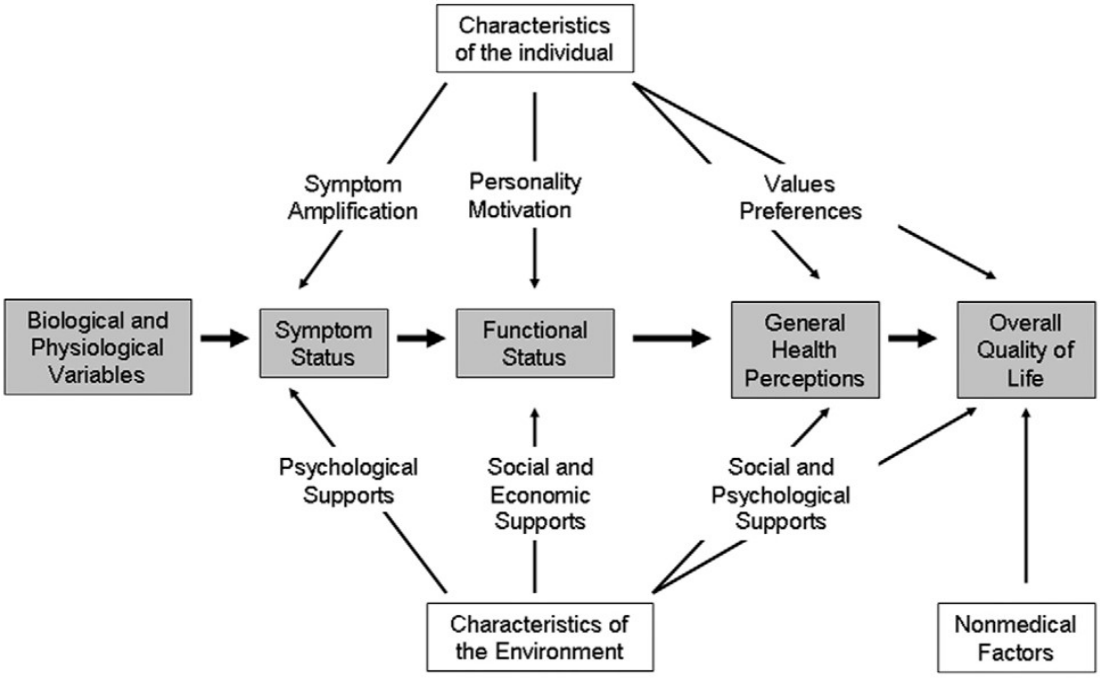


Figure 2.2: Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995; 273: 59-65.

The arrows in the figure do not imply the relationship is not reciprocal, neither does the absence of arrows between non adjacent levels imply lack of relationships. Wilson and Cleary emphasised that the schematic illustration of their conceptual

model only highlights the most dominant causal associations between the outcomes [40]. There is strong empirical evidence supporting the model's validity in various contexts, including the general population aged over 65, patients living with HIV, patients suffering from coronary heart disease etc [41].

2.2.2 The conceptual framework of PROM

In the final version of the "Guidance for Industry, PROM Use in Medical Product Development to Support Labeling Claims" released by the Food and Drug Administration (FDA), the conceptual framework for developing a PRO instrument was explained [42, 27]. Figure 2.3 illustrates a generic example of what the conceptual framework of a PROM instrument looks like: I) Domains 1 and 2 are found to be related to the general concept of interest. II) Four items have been selected to measure domain 1 and two items to measure domain 2.

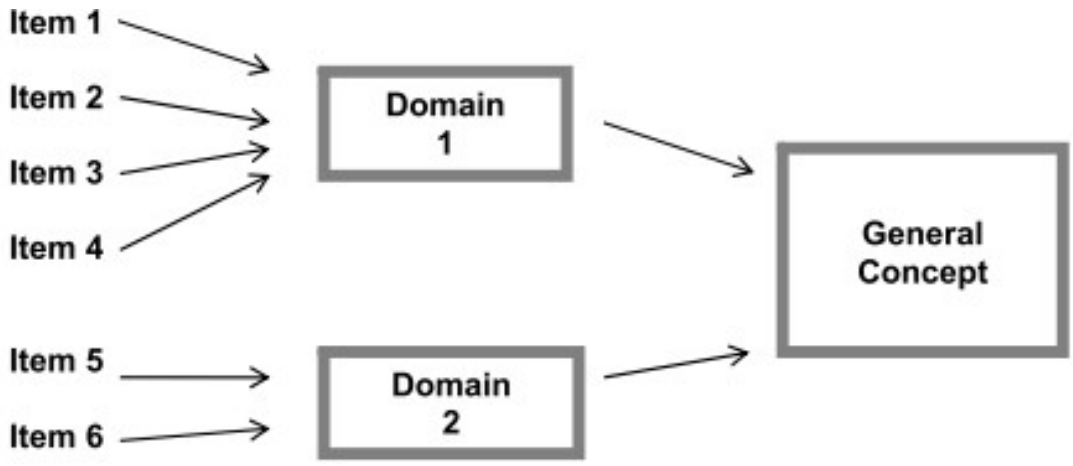


Figure 2.3: Diagram of the conceptual framework of a PRO instrument (Deshpande et al 2011)

The conceptual framework can be straightforward or complex, depending on the nature of the concept of interest. For more concrete and specific concepts (e.g. pain

intensity), a single domain and a single item may be a reliable and valid measure. For more general concepts (e.g. physical function), a single-item PRO instrument does not have sufficient face validity, in that case all the appropriate domains need to be identified. During the instrument development process, the relationship at the item level (between individual items), at the domain level (between domain 1 and 2) and across different levels (e.g. between item 2 and domain 1, between domain 1 and the general concepts) need to be carefully examined; the conceptual framework will continuously evolve and be confirmed over the course of its development.

2.2.3 A new classification system for PROMs

Building on the Wilson-Cleary model and the ICF model (introduced in Chapter One of this thesis), Valderas and colleagues developed a new PROM classification system consisting of axes and categories shown in Table 2.1 [41].

Axis	Categories
A. Construct	A.1. Symptoms (see Table 2b)
	A.2. Functional status (see Table 2c)
	A.3. Health perceptions
	A.4. Health related quality of life
	A.5. Other health related constructs
	Satisfaction with care
	Disadvantage
	Resilience
	Environmental
B. Population	B.1. Age
	(a) All ages
	(b) Children
	(c) Adolescents
	(d) Adults
	(e) Seniors
	B.2. Gender
	(a) All genders
	(b) Female
	(c) Male
	B.3. Disease (see Table 2b)
	B.4. Culture (country and language dyads)
C. Measurement	C.1. Metric
	(a) Psychometric
	(b) Econometric
	(c) Clinimetric
	(d) Other metrics
	C.2. Dimensionality
	(a) Index
	(b) Profile
	(c) Index and profile
	C.3. Adaptability
	(a) Completely standardized
	(b) Partially individualized
	(c) Completely individualized

Table 2.1: A classification system for patient reported outcome measures: axis and categories (Valderas 2008)

This classification system provides a comprehensive view of measures and can potentially facilitate the selection of the appropriate PROM to be used in research or

clinical practice [41]. The three guiding concepts (axes) in this model are: construct, population and measurement.

- **Construct** refers to the range of objects measured by the instruments, including symptom status, functional status, general health perceptions, HRQoL and any other health-related constructs.
- **Population** refers to the different groups of people the instrument is suited for. Population is divided into four major categories in terms of age, gender, diseases and culture.
- The concept **measurement** consists of three components – metric, dimensionality and adaptability.
  - *Metric* refers to the specific method used to assign numerical scores to the responses given by individuals and how final score is constructed.
  - *Dimensionality* is defined as the number of scores produced for each individual. When all the responses can be summarised in a single value, this type of instrument is an index. When multiple scores are needed to summarise the information on the instrument, it is referred to as a profile.
  - *Adaptability* refers to the degree to which the instrument can cater for the specific circumstances and preferences of each individual user.

The Sickness Impact Profile (SIP) is one of the first PROMs to be developed; when applying this classification system to the SIP, the results are as follows [41]:

- The constructs that SIP measures include symptom status, functional status and QoL.
- The SIP-targeted population include people of all ages, all genders and all

diseases.

- In terms of measurement, the SIP is psychometric, an index, and completely standardised (not adaptable).

2.3 Different types of PROMs

Currently there are three types of standardised PROMs — generic, disease-specific and population-specific. Non-standardised PROMs, including individualised measurements and summary items, have also been developed. In practice, clinicians often use a combination of different types of PROMs.

2.3.1 Generic instruments

Generic instruments are designed to measure patient's general health status with respect to their ability to function in everyday life and to participate in activities that they enjoy [43]. This type of instruments are designed to measure very broad aspects of health and therefore can potentially be applied across populations of different age, gender, disease or culture [44, 45]. Examples of generic PROMs include the SIP, Nottingham Health Profile (NHP) and the EuroQoL 5-dimension (EQ-5D).

The main advantage of generic instruments is that they are potentially capable of capturing the influence of co-morbidities on health, and detecting unexpected positive or negative effects of an intervention [46]. However, there is a trade-off between the breadth of scope and level of detail. Generic instruments often lack details, which limits their responsiveness, and potentially making them unable to detect subtle but clinically important changes when used as a primary outcome measure for a specific condition [47].

2.3.2 Disease-specific instruments

Disease-specific instruments measure the symptoms and limitations patients experience as a result of having a specific disease/condition [47, 48]. This type of instrument usually only contain items that are relevant to the particular disease they are designed for. Comparing to generic instruments, disease-specific instruments are more responsive, meaning that they are better at detecting clinically important changes in patients' health as a result of the treatment [49, 50]. One of the limitations of disease-specific measures is that their narrow focus (on the disease/condition of interest) makes them unable to detect side effects of a treatment. Also, a common approach for assessing the impact of a particular disease is to compare the scores of the patients (with this disease) with the scores of the general (healthy) population. In the case with disease-specific measures, because they are only valid for patients with a particular condition, the health status scores cannot be compared with the general public [51, 48].

2.3.3 Population-specific instruments

Population-specific instruments are designed to measure the quality of life of a specific demographic group, such as children or the elderly. An example of population-specific measurements is the Child Health and Illness Profile - Child Edition (CHIP-CE), which contains 45 items covering child health in five aspects - satisfaction, discomfort, resilience, risk avoidance and achievement [52]. Population-specific instruments may be more relevant and accessible to the targeted group. For example, the CHIP-CE features cartoon illustrations to help convey instructions to children. The disadvantage is similar to that observed with disease-specific measures – they are not applicable to the general public, and therefore cannot be used to compare the effectiveness of a particular treatment across different population groups [53].

2.3.4 Individualised instruments

The shortcoming shared amongst the three types of PROMs mentioned above is that patients' views and experiences are evaluated by standardised items, which were pre-determined by researchers. To overcome this issue, individualised instruments have been created to let patients identify their own concerns. The Schedule for the Evaluation of Individual Quality of Life (SEIQoL) was one of the earliest individualised instruments [54]. Administration of SEIQoL involves conducting a lengthy interview to persuade patients to identify their top five concerns regarding their quality of life. The SEIQoL index score showed a weak correlation (r=0.21) with the McMaster Health Index Questionnaire (MHIQ) [55], which measures health in three domains – physical, social and emotional. A similar level of association was observed when SEIQoL was compared with the Arthritis Impact Measurement Scales (AIMS) [56], a disease-specific measure of QoL for arthritis patients.

2.3.5 Summary items

Over the last few decades, a number of HRQoL questionnaires have been developed and tested for validity, reliability and responsiveness. However, the search for the "ideal measure" is still on-going. A major problem with most questionnaires is the large number of items – it is time-consuming for the patients to complete and for the clinicians/researchers to score. The shortest and simplest assessment is a single question that asks respondents to summarise diverse aspects of their health status [57, 9]. Transition items are a subtype of summary item: they ask the respondents to compare their current health state with a specific point in the past (e.g. post-surgery versus pre-surgery). An example of a transition item is "Thinking of any overall effects your arthritis may have on you, how would you describe yourself compared to the last time I interviewed you (three months ago)?" Response items

include “much better”, “slightly better”, “the same”, “slightly worse” and “much worse” [58]. While these summary items may refer to an individual’s general health, alternatively they can be anchored to a particular dimension (e.g. social well-being, psychological well-being) or a specific health concern.

The main advantage of summary items is that they are short, simple, quick to administer and to score. Comparing to multi-item measures, the brevity of summary items makes them cheaper to use and can potentially improve response rate [9]. Also, there is some evidence supporting the reliability and validity of summary items [59]. One of the limitations of summary items is a loss of sensitivity; there has been concerns that the number of response choices are too few to detect small but clinically important changes in health. Also, single-item scales provide less statistical information compared with more elaborate measures, which compromises their reliability [60].

2.3.6 Comparative evaluation of instruments

Since there is a large compendium of HRQoL measures, comparative evaluations become essential. Several studies have found standardised generic and disease-specific PROMs to be more sensitive to change than clinician-reported instrument [61]. Furthermore, in a large-scale comparative study conducted by Dunbar et al, four generic (NHP, SF-12, SF-36 and SIP) and three disease-specific instruments (Lequesne Index, OKS, and WOMAC) were given to 3600 patients randomly selected from the Swedish Knee Arthroplasty Registry, the response rate, completion rate, time required to complete, reliability and validity of these instruments were assessed. The SF-12 and OKS emerged as the best generic and disease-specific measures, respectively [62].

2.4 Implementation and benefits of PROMs

Whilst the usage of PROMs is on the rise amongst individual clinicians and hospitals, widespread use by the country's health care system is still largely restricted to the UK, Sweden and some parts of the US. In England, PROMs have primarily been used to measure outcome related to elective surgeries [63]. Since April 2009, PROM data have been collected nationally from patients undergoing the following surgeries: hip replacement, knee replacement, groin hernia repair, and varicose vein. Plans to extend the PROMs program over a broader range of conditions, including mental health problems, cancer care, asthma, COPD, diabetes, epilepsy, heart failure and stroke, are in place [64].

Since its initial implementation in 2009, PROMs data have proven to be useful for three different groups — *clinicians* seeking to better their practice; *patients* looking for the best treatment option possible; *researchers* looking to derive patterns and draw useful conclusions via data analysis [64].

PROM data collected using measurements with evidently satisfactory psychometric properties become a new addition in the clinicians' patient-management tool box. For initial assessment, PROMs provide baseline information on quality of life and disease severity from the patients' perspective. Once the treatment has started, whether it be a conservative or a surgical protocol, PROMs can be used to monitor disease progression and the effectiveness of treatment [65, 66]. Furthermore, because these questionnaires are filled out completely by patients, they can raise the subset of concerns that might be overlooked or considered unimportant by the clinicians had the PROMs program not been implemented [64].

PROMs data offer a rich set of information regarding treatment outcomes from different hospitals, enabling patients to make more informed decisions when selecting

a treatment facility [67]. Throughout the treatment process, the routine act of filling out PROMs questionnaires aids patients in remembering and being able to describe the improvement or deterioration of their illness [66]. Furthermore, PROMs data could help patients form reasonable expectations about how much they will improve after receiving a specific treatment [64].

Nationally collected PROM data contain a large number of patients with a relatively comprehensive profile for each individual (e.g. age, gender, BMI, pre-operative score, post-operative score, etc). Such datasets convey information representative of typical everyday clinical practice, which could be used for analysing treatment effectiveness and evaluating relevant health care policies [68]. Moreover, PROMs data are useful for translational research. For example, by examining the correlation between patients' characteristics (e.g. age, gender, ethnicity, etc) and improvement (or deterioration) in PROMs score after treatment, scientists would be able to identify specific subset of patient groups that would be more (or less) likely to benefit from this particular treatment procedure [64].

2.5 Interpretation of PROM data

In the 1980s, when PROM research was still in its infancy, change in PROM scores was usually expressed as a unit on an abstract scale, which does not convey much information unless it is correlated with something more concrete and interpretable. Also, statistical significance does not always guarantee clinical importance, so there was a demand for developing a systematic approach to document the level of change in PROM score that shows a clinically significant improvement.

In 1989, Jaeschke and colleagues were the first to introduce the term minimal clinically important difference (MCID), which was defined as the "smallest difference

in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management [69]." In the original study, the concept of MCID was derived in the context of interpreting the meaning of group-level change score , but the authors claimed that the MCID obtained can be used " in interpreting questionnaire scores, both in individuals and in groups of patients participating in controlled trials". Five years later, Juniper and colleagues suggested the name minimal important difference (MID) as an alternative to MCID, after recognising that the process (of deriving this threshold score) does not always involve any bona-fide clinical assessments or judgement [30].

Since its introduction, for over a decade, the terms MCID or MID were used as an umbrella term for the interpretation of both group-level and individual-level change [70]. In 2006, the draft FDA PRO Guidance was issued, which for the first time clearly delineated the distinction between group- and individual-level change – *MID* was to be used for interpreting group-level mean differences, and the *responder definition* was to be used for individual-level changes as a result of the treatment. Furthermore, the draft FDA guidance also suggested that in terms of magnitude, the responder definition (i.e. the smallest meaningful individual-level effect) would be a higher threshold than the MID (i.e. the smallest meaningful group-level effect) [70, 71].

In the final FDA PRO Guidance released in 2009, it was stated that regardless of whether the endpoint of the clinical trial is based on individual-level or group-level response, it is usually useful to show individual response. The *a priori* **responder definition** refers to the individual change in score (measured on a particular PROM) over a pre-determined time period that should be interpreted as clinically important. Such responder definition is determined empirically and may vary by target population or other design characteristics of the clinical trial. Therefore, the

responder definition of each instrument should be evaluated in the context of each clinical trial. The 2009 FDA PRO Guidance also included instructions regarding what methods are appropriate for calculating responder definition:

- The primary method for deriving responder definition should be anchor-based. Anchor-based approaches explore the association between the targeted concept of the PROM and the concept measured by the anchors. For practical reasons, the anchors chosen should be easier to interpret than the PROM itself. A common anchor is patients' transition ratings (i.e. ratings of change administered at different periods of time). Response choices to anchor questions range from worse to the same and better. For example, the average change in PROM score for persons who rate their condition as "a little better" can be used as a responder definition.

- Another set of approaches to deriving responder definition are distribution-based methods, such as using effect size, half of standard deviation or standard error of measurement to define a meaningful change on an instrument. Distribution-based methods can be used to classify PROM score changes as small, medium or large, but should be considered as supportive (i.e. supplementary to anchor-based approach) and not be used independently in determining responder definition.

- Lastly, The FDA Guidance (2009) also recommends using the cumulative distribution curve to display the entire range of responses to the treatment. In a cumulative distribution graph, the change in PROM score from baseline is shown on the X-axis, and the percent of patients experiencing that score change is shown on the Y-axis. Various responder definitions can potentially be identified along the cumulative distribution curve.

Even though the draft PRO Guidance made clear distinction between group- and

individual-level clinically important change (i.e. responder definition versus MID), inconsistent usage of the term MID has continued – in the field of PRO, MID is still erroneously being used for these two concepts interchangeably sometimes. In 2015, two new terminologies attempting to provide a concise way of describing these two interpretations of change scores has been suggested – "clinically important difference" and "clinically important responder" [70]. The new terms eliminated any reference to "minimal". "Clinically important difference" (CID) is used to describe the differences in scores between two treatment groups that can be considered as clinically significant; the term "clinically important responder" (CIR) refers to the amount of change a patient would have to report to indicate that they have experienced a significant benefit from the treatment. Since their conception, the words "CID" and "CIR" have been well received; these terms make intuitive sense and emphasise the clinical interpretations of different thresholds instead of the magnitude of the thresholds. The use of "CID" and "CIR" is still limited, MID and MCID are still mainstream, but this situation will hopefully change in the next decade.

2.6 PROMs in orthopaedics

As mentioned in Chapter one, over the years, the focus on outcome measurement in orthopaedics has shifted from revision rate to patient reported outcomes (PRO). The number of PROMs utilised in orthopaedic registries and trials has expanded greatly in the last two decades, some of which are generic measurements, whilst others are specifically designed for orthopaedic patients. The following sections provide a detailed description of some frequently used instruments in orthopaedic research.

2.6.1 Generic PROMs

Generic measurements usually cover multiple HRQoL dimensions and are applicable to patients with various conditions. Generic instruments frequently used in Orthopaedics include the Short Form health surveys (SF-36), the EuroQol 5-dimension (EQ-5D) and the Nottingham Health Profile (NHP). The main advantage of generic PROMs is that, when used in combination with generic health measures, they are capable of detecting the impact of co-morbidities [46].

2.6.1.1 Short Form 36 health surveys (SF-36)

The SF-36 covers 8 dimensions of health (physical functioning (PF); role limitations because of physical health problems (RP); bodily pain (BP); general health perceptions (GH); vitality (VT); social functioning (SF); role limitations because of emotional problems (RE); and mental health (MH)), with 2 to 10 items for each dimension [72, 73]. The practical and psychometric properties of SF-36 in clinical trials involving OA patients has been thoroughly examined, and it has been proven to be psychometrically sound for assessing the outcome following orthopaedic surgeries [74, 72]. Jones and colleagues have found that patients' SF-36 total scores improve substantially after receiving TKR or THR. However, the majority of the change came from pain and physical function dimensions; whilst the improvement in psychosocial dimensions was much smaller [75]. Also, a greater amount of improvement in terms of SF-36 scores and a higher satisfaction rate were observed in THR patients compared to TKR patients, a pattern seen consistently over time in different studies [75].

Following the popularity of SF-36, a shortened version consisting of only twelve items was developed in 2011. This instrument was designed to reduce burden for the respondents whilst maintaining adequate validity, responsiveness and reliability.

The SF-12 still covers 8 dimensions, with only 1 to 2 items assessing each dimension [76, 77]. Today the SF-12 is a preferred measurement instrument for routine follow-ups in joint registries [78, 79].

2.6.1.2 EuroQol 5-dimension (EQ-5D)

The EuroQol 5-dimension 3-level (EQ-5D-3L) measure was developed by the EuroQol group in 1990 as a measure of health utilities [80], and it is the only generic instrument used in the NHS PROM program [64]. The EQ-5D-3L is comprised of two components, the EQ-5D descriptive system and the visual analogue scale (EQ-VAS). The EQ-5D-3L descriptive system covers 5 dimensions, assessing the level of difficulty patients have with mobility, self-care, usual activities, pain/discomfort and anxiety/depression [81]. Response options are "no problem", "some problem" and "unable to do/extreme problem", which are represented by digits 1,2, and 3, respectively. The patients are asked to choose the response that suits their current health state best. When collecting data, digits for all 5 dimensions are first combined into a 5-digit number (i.e. the patient's "health profile"), which is then converted to a 1-digit summary index number by applying a formula that adds weights to each of the response options in each dimension [80].

The EQ-VAS addresses the patient's current self-perceived health state via a vertical visual analogue scale, where 0 represents the worst imaginable health state and 100 represents the best imaginable health state [80].

In 2005, an updated version of the original EQ-5D-3L was developed, where the response options for each of the 5 dimensions were extended to five levels – no problems, slight problems, moderate problems, severe problems, and extreme problems. The new version, known as EQ-5D-5L, has shown better sensitivity to change and lower ceiling effects than EQ-5D-3L. Also, some of the wording in the original ver-

sion has been changed slightly to enhance consistency and comprehensibility. With these improved features, implementation of EQ-5D-5L in clinical trials and registries is anticipated to increase [80].

2.6.1.3 Nottingham Health Profile (NHP)

The Nottingham Health Profile (NHP) was one of the oldest generic instruments, intended to assess patients' own perception regarding their emotional, social and physical health problems. Part I of the NHP is consisted of 38 items assessing health status in 6 dimensions: energy level, pain, emotional reaction, sleep, social isolation and physical mobility [82]. Part II addresses whether patients' present health state is causing problems in seven areas of life (including work, house chores, social life, family life, sex life, hobbies and vacation) [83]. Response categories to each item are simply "yes" or "no". The NHP was frequently used in orthopaedic research [84], but has been gradually replaced by the shorter EQ-5D and SF-12.

2.6.2 Condition-specific PROMs

Many hip and knee replacement surgery patients are older and have co-morbidities, which has been defined as "any distinct additional entity that has existed or may occur during the clinical course of a patient who has the index disease under study" [85]. In knee and hip arthroplasty research, the index disease is end stage hip/knee osteoarthritis, and comorbidities refer to a range of other health problems that can impact patients' pain, function, and other dimensions of HRQoL.

2.6.2.1 The Oxford Knee Score (OKS) and Oxford Hip Score (OHS

The OHS was developed in 1997, followed by the development of the OKS in 1998. Both instruments are condition-specific PROMs and were developed using patient interview approach with the intended purpose of assessing outcomes of THR/TKR in terms of patients' own perception on pain and function [86, 87]. Both OHS and OKS consist of 12 items with five Likert-type responses. Each item is scored from 0 (worst pain or highest functional limitation) to 4 (no pain or no functional limitation), the total score (ranging between 0 and 48) is obtained by summing the scores for each item [86, 87].

The OKS was originally created to be used as a one summary score; but it was later confirmed (using factor analysis) that it could be divided into pain and functional subscales [88]. There are five items in the function subscale (consisted of questions 2, 3, 7, 11, 12), and the total subscale score range between 0 (worst) and 20 (best). The pain subscale is consisted of the remaining seven items (questions 1, 4, 5, 6, 8, 9, 10), and the total subscale score can range between 0 (worst) to 28 (best) [88].

The OKS and OHS were the first outcome measurement instruments specifically developed for use in patients undergoing hip or knee replacement operations. Both instruments have demonstrated high level of reliability, validity and sensitivity to change, and have been used extensively in clinical trials, as well as, national registry PROMs programs [86, 87, 89, 90].

One of the main issues with the OKS and OHS is their potential ceiling effect, which refers to the phenomenon where 15% or more individuals in a sample obtaining the maximum score possible. The results from different studies have been inconclusive. Using data of 100 patients collected from a single academic centre, Jean-Yves et al found the ceiling effect of the OKS to be 33% [91] in 2013. The ceiling effect was defined as the rate of cases having a score within one standard deviation away

from the max (48). In 2015, using a national population data (n=71,607), Harris et al reported that 2.7% of patients achieved the maximum possible OKS [92]. Elsewhere, a slightly higher value of 8% has been reported by Hamilton et al regarding the proportion achieving the highest OKS possible in a prospective cohort study [93]. Combining all previous results from the literature, it is evident that despite a considerable amount of TKR patients are able to achieve very high post-operative scores (i.e. above 40), far less than 15% achieve the highest score possible (48), and therefore does not exhibit a ceiling effect.

However, it is important to acknowledge that it has been over twenty years since the OKS has first been developed, and the demographics of TKR patients have changed. There has been an increase in the proportion of younger patients, for whom the OKS may not be the most appropriate measure. PROMs specifically designed for patients with higher functional abilities may be more suitable, as they are, in theory, more discriminating for high functioning patients [92].

2.6.2.2 The Western Ontario and McMaster Universities Arthritis Index (WOMAC)

Unlike the OKS and OHS, the Western Ontario and McMaster Universities Arthritis Index (WOMAC) was developed to assess pain, disability, and joint stiffness in patients with osteoarthritis [94]. The WOMAC consists of 24 items, 5 of which are dedicated to pain, 2 to stiffness and 17 to physical function. Each question is scored on a scale of 0-4, corresponding to: none (0), mild (1), moderate (2), severe (3), and extreme (4). The final score for each subscale is calculated by adding up scores for individual items; possible score ranges between 0 and 20 for pain, 0 and 8 for stiffness, 0 and 68 for physical function. Lower scores indicate lower pain, less stiffness, better function, and vice versa [94].

The WOMAC has been proven to be a valid, reliable and responsive measurement in assessing outcome following arthroplasty [94, 95]. Since it was first developed in 1982, the WOMAC has been translated into over 90 languages worldwide and is frequently used in clinical trials, as well as, some large local registries in the US and Canada [96]. In 2003, a shortened version of WOMAC, consisting of 12 items (including all 5 original items on pain and 7 of the original 17 items on physical function) was developed and validated in a TKR/THR patient population [97].

2.6.2.3 The Knee injury and Osteoarthritis Outcome Score and Hip disability and Osteoarthritis Outcome Score

The Knee Injury and Osteoarthritis Outcome Score (KOOS) was developed as an "extension and comprehensive alternative" to the WOMAC, specifically assessing HRQoL in young and middle-aged patients with ACL or meniscus injury, or post-traumatic osteoarthritis in the following five dimensions: I) pain, II) symptoms including swelling, grinding or clicking upon movement and restricted range of motion, III) stiffness and activities of daily living, IV) function, sports and recreational activities and V) quality of life [98]. All items in KOOS are scored from 0 (best) to 4 (worst), and the five health dimensions are scored independently. Next, each subscale score is transformed to a 0-100 scale with 0 indicating the most extreme problems at the knee and 100 indicating no problem [98].

The HOOS is an adaptation of the KOOS, also an extension to the WOMAC with the intended purpose of evaluating symptoms and functional limitations associated with the hip [99]. HOOS measures the same five health dimensions, and has the same scoring system as the KOOS. Both measures have proven to be a valid and reliable instrument for assessing outcome following TKR/THR [99, 100]. More over, comparing to the WOMAC, the KOOS and HOOS have shown higher responsiveness

and sensitivity to change in young and middle-aged patients undergoing TKR/THR [99, 100]. Both HOOS and KOOS are being used in some registries in Europe and the United States, but they may not be appropriate for routine follow-ups as the large number of items (40 items in HOOS, 42 in KOOS) render these two questionnaires cumbersome to complete.

In 2008, the KOOS and HOOS physical function short forms were developed (KOOS-PS containing 7 items, HOOS-PS containing 5 items), which only included items from the physical function dimension

2.6.3 Other patient-reported measures related to outcome

In addition to true, multi-item PROMs that directly measure various dimensions of HRQoL, there are also single-item measures designed to assess a specific aspect of the outcome, such as overall satisfaction with treatment, willingness to undergo the operation again, or fulfillment of expectation. The idea of using a single question to assess the outcome of total knee or hip replacement is conceptually appealing as they are easy to administer, cause minimal disruption to the flow of clinical care and most of the questions and responses are meaningful to patients [9]. However, the response to a single item question may be affected by age, sex and the patients' emotional state on the day of assessment [101]. The two questions incorporated into the NHS PROM program address transition of symptoms and satisfaction.

- The **transition item** is only used post-operatively in TKR/THR, asking patients to make retrospective comparisons – " how is the symptom in your affected knee/hip now comparing to pre-operation". Response choices range from "much worse" to "much better". Currently, the transition item is often used in conjunction with a standard PROM in order to assign meaning to certain magnitudes of change on that instrument [70]. For example, the dif-

ference in mean change score (from pre- to post-operation) between patients who reported feeling " a little better" and those reporting feeling "the same" has been used as the clinically important difference on a group level [102].

- The **satisfaction item**, like the transition item, is also used during post-operation only, asking patients to rate their satisfaction' responses range from "very dissatisfied" to "very satisfied" or from "poor" to "excellent" [102, 96, 78]. There is an increased focus on post-operative satisfaction nowadays. Although alleviating pain and improving function are the main purpose of total joint replacement operations, the importance of these goals may vary from one patient to another depending on their individual expectation from the surgery [96]. A high level of satisfaction from the patient is ultimately what every clinician strives for. However, it has been found that the level of satisfaction may be affected by a multitude of factors, such as pre-operative expectation [103, 104, 93, 105], depression [106], hospital experience [93] and postoperative complications [105].

2.7 Link to the next chapter

This chapter presented the important concepts in developing PROMs, different types of PROMs, implementation and interpretation of PROMs, as well as, established PROMs in orthopaedics. The next chapter will first introduce a new framework of evaluating post-operative outcome following TKR/THR, and then present an exploratory study aiming to determine the true range of success rate following TKR and THR.

Chapter 3

Defining a "successful" outcome after TKR/THR and improved estimation of the proportion of patients achieving this

3.1 Introduction

Currently, over 169,000 patients undergo either TKR or THR across the UK every year [107], and the number is estimated to rise substantially in the next two decades. With such soaring demand within the population, and a substantial cost to the health care system, the clinical effectiveness of TKR and THR have always been under scrutiny. Whereas the technical aspect of TKR and THR (judged by implant failure or revision rate) is exceptional, outcome in terms of patient-reported pain, function, and satisfaction appears less so [108]. A wide range of success rate has been reported in previous studies. Beswick and colleagues found that the proportion of unfavourable long-term pain outcome ranged between 7% and 23% for THR, between 10% and 34% for TKR [109]. An even lower success rate has been reported by Hawker et al, claiming that only 53% patients achieve a good outcome following TKR or THR [110]. On the other end of the spectrum, a success rate as high as 86% following TKR and 94% following THR have also been found [111].

All of the aforementioned studies described "outcome" of TKR and THR in terms of the percentage of patients who have or have not achieved a "sufficient response". However, the exact definition of a "sufficient response" varies from one study to another — ranging from a significant reduction in pain level measured on the WOMAC-pain subscale to "feeling satisfied" measured by a visual analog scale (VAS). To describe a "sufficient response", three components are essential – the outcome measure, the response criterion and the cut-off score.

The outcome measure is the instrument used to measure treatment outcome in the specific study. There are a number of joint-specific outcome measures. The OKS, WOMAC, the Knee injury and Osteoarthritis Outcome Score (KOOS) and the Knee Society score (KSS) are amongst the most commonly used instruments

for measuring the outcome of TKR; the OHS, WOMAC, the hip disability and osteoarthritis outcome score (HOOS) are frequently used instruments in THR.

The response criterion is a benchmark for a "sufficient response" to treatment or intervention. If patients' response to TKR or THR was presented on a horizontal axis, the choice of a response criterion essentially amounts to picking a location along the response axis, which represents the distinction between sufficient and insufficient response. "Sufficient response" is assumed when patients' response are greater than this criterion, and an "insufficient response" when patients' response are less than the criterion.

When the response is measured by patients' post-operative scores or a difference between pre- and post-operative scores on a particular instrument, the response axis is representative of scores. The score corresponding to the location of the response criterion becomes the cut-off score. Each cut-off score is specific to a single response criterion.

To derive the criterion-specific cut-off score, there are two main methodological approaches – distribution-based and anchor-based. Distribution-based approaches rely on relating the difference between pre-operative and post-operative scores to some measure of variability, such as the standard error of measurement (SEM) or standard deviation (SD). Anchor-based approaches involve tying the change in patients' scores on a particular instrument to a relevant and more clearly understood measure, also known as the " anchor question" (e.g. global transition rating) [112] [102]. Response options to an anchor question are usually presented in the form of a 0-100 VAS, or qualitative expressions such as "much better", "about the same" or "much worse"[113]. Tables 3.1 and 3.2 presents some common outcome measures, response criteria, and cut-off scores used in the evaluation of TKR and THR.

Outcome measure	Response criteria	Criterion-specific cut-off	Outcome
OKS	Minimal clinically important difference (MCID)	1/2 SD of baseline summary score [114]	
	MCID – pain	1/2 SD of baseline pain score [114]	
	MCID – function	1/2 SD of baseline function score [114]	
	Minimal important change (MIC)	Δ OKS $\geqslant$ 9.22 [102]	
	Patient acceptable symptom state score (PASS) – Δ OKS	Δ OKS $\geqslant$ 11 [115]	67.1% (6 months) [115]
	PASS – post-op OKS	post-op OKS $\geqslant$30 [115]	71.7% (6 months)[115]
Non-specific	OMERACT-OARSI responder criteria	50% relative and 20 points absolute improvement in pain or function [116] OR	86.3% (12 months) [111]
		20% relative and 10 points absolute improvement in pain and function [116] OR	
		20% change in global assessment [116]	
WOMAC	MCID-pain	normalised ΔWOMAC-pain $\geqslant$ 22.87 [117]	66.2% (at 6 months)[117]
	MCID-function	normalised ΔWOMAC-function $\geqslant$ 19.01 [117]	68.7% (at 6 months) [117]
	PASS-function	normalised post-op WOMAC-function $\leqslant$ 68 [111]	67.7% (at 12 months) [111]
	PASS-pain	normalised post-op WOMAC-pain $\leqslant$ 75 [111]	75.0% (at 12 months) [111]

Table 3.1: Common outcome measures, response criteria, and associated cut-off point for TKR. The last column shows the outcome in terms of percentage of those who have done well. The values were reported by different groups using different data sets, different instruments and different response criteria.

Outcome measure	Response criteria	Criterion-specific cut-off	Outcome
OHS	Minimal clinically important difference (MCID)	1/2 SD of baseline summary score [114]	
	MCID – pain	1/2 SD of baseline pain score [114]	
	MCID – function	1/2 SD of baseline function score [114]	
	Minimal important change (MIC)	Δ OHS $\geqslant$10.63 [102]	
	Patient acceptable symptom state score (PASS) – ΔOHS	ΔOHS $\geqslant$ 14 [115]	70.4% (6 months) [115]
	PASS – post-op OHS	post-op OHS $\geqslant$35	73.7% (6 months)[115]
Non-specific	OMERACT-OARSI responder criteria	50% relative and 20 points absolute improvement in pain or function [116] OR	94.6% (12 months) [111]
		20% relative and 10 points absolute improvement in pain and function [116] OR	
		20% change in global assessment [116]	
WOMAC	MCID-pain	normalised ΔWOMAC-pain $\geqslant$ 29.26 [118]	75.0% (at 6 months)[118]
	MCID-function	normalised ΔWOMAC-function $\geqslant$ 26.54 [118]	72.3% (at 6 months)[118]
	PASS-function	normalised post-op WOMAC-function $\leqslant$ 69 [111]	69.1% (at 12 months)[111]
	PASS-pain	normalised post-op WOMAC-pain $\leqslant$ 80 [111]	80.0% (at 12 months)[111]

Table 3.2: Common outcome measures, response criteria, and associated cut-off point for THR. The last column shows the outcome reported by different groups using different data sets, different instruments and different response criteria.

As shown in tables 3.1 and 3.2, multiple response criteria can be developed for a single outcome measure. Common response criteria include the MCID, MIC, PASS and the modified OMERACT-OARSI responder criteria, each of which views "sufficient response" from a different angle. The following sections explain the conceptional foundation of each response criterion and some of the methodologies used to derive the associated cut-off score.

3.1.1 Identifying benefit - measuring whether patients feel better.

3.1.1.1 Minimal clinically important difference

The concept of measuring whether patients "feel better" using minimal clinically important difference (MCID), was first introduced by Jaeschke, defining it as "the smallest difference in score in the domain of interest which patients perceive as beneficial . . . "[69]. Having conducted a systematic review of health-related quality of life instruments for chronic diseases, Norman et al identified that (in most cases) half of the standard deviation (SD) of the average baseline score would be a good estimate of the cut-off score for MCID [114] [119]. This is a distribution-based threshold value; the cut-off score for MCID can also be derived using anchor-based approaches [117] [118].

3.1.1.2 Minimal important change

On the conceptual level, minimal important change (MIC) refers to the " clinically relevant change (improvement or deterioration) in health status in a single group or a single individual over time" [102]. On the methodological level, the cut-off points associated with MIC have been identified by first stratifying patients according to

their response to the transition anchor question " Overall, how are your hip/knee problems now, compared to before your operation?", and then calculating the mean change in PROM score for the group that selected the "(feeling) a little better" option [102].

3.1.2 PASS – measure whether patients "feel satisfied"

In addition to the MIC, which measures whether patients feel sufficiently better, a further option is to measure whether patients feel satisfied; the concept of patient acceptable symptom state (PASS) score was based on this idea [120]. The PASS values have been identified in the context of evaluating the outcome of total joint replacement operations based on patients' change in their OHS and OKS, and is defined as the value of the OHS and OKS beyond which patients consider themselves in a satisfactory state [115].

In the original study, Judge et al used data from the South West London Elective Orthopaedic Centre (EOC), which is a purpose-build NHS treatment centre specialising in orthopaedic operations. The dataset included information on 3082 THR patients and 3608 TKR patients, who underwent surgery between 2004 and 2009. Regarding the methodology of deriving the associated cut-off point, the first step is to divide patients according to their response to the satisfaction anchor question – with 0 being not satisfied and 100 being very satisfied; 50 was used as the dividing mark between those who feel satisfied and those who do not. In step two, ROC curve analysis (i.e. best cut point) was used to determine PASS thresholds. In this case, the continuous variable, the absolute change in OKS/OHS or post-operative OKS/OHS was used to discriminate between two mutually exclusive states of experience – "satisfactory" or " not satisfactory". Once the threshold was identified, patients who obtained a final score or change in score equal to or greater than that

cut-off would be considered to have achieved a sufficient response, and vice versa [115].

3.1.3 Modified OMERACT-OARSI set of responder criteria – measure whether patients "responded" to treatment

OMERACT stands for the Outcome Measures in Rheumatoid Arthritis Clinical Trials, it is an "international, informally organised network" initially established in 1992 with the goal of "improving outcome measures in rheumatology". Since 1992, this multinational network has been consistently organising its own biannual conference across the world. During the conference, experts work in groups to discuss and update new data-driven recommendations on measurements for rheumatological conditions and treatments [121].

OARSI is the acronym for Osteoarthritis Research Society International, the primary international organization for scientists and health care professionals interested in the prevention and treatment of osteoarthritis through promoting research, education and communication of cutting-edge knowledge [122].

In 2004, Pham et al published their study, where they examined and modified the former OMERACT-OARSI approved responder criteria using a data-driven/expert opinion combination approach. The conclusion was a proposal of a simplified definition of symptomatic improvement in OA patients described as follows [116]:

If the patients achieved a relative improvement in pain or function of at least 50%, and an absolute change of at least 20 points, they are categorised as a responder. One thing of note is that all scores are normalised to a 0-100 scale with 100 being the favourable condition.

In the instance where the patients failed to obtain either the relative improvement

or absolute change, they would be considered a responder if they satisfied two of the three criteria:

- Improvement in pain $\geqslant 20\%$, and absolute change in pain $\geqslant 10$ points.
- Improvement in function $\geqslant 20\%$, and absolute change in function $\geqslant 10$ points.
- Improvement in global assessment $\geqslant 20\%$, and absolute change $\geqslant 10$ points.

Once the criterion for a " sufficient response" has been set and the associated cut-off point derived, post-operative outcome can be reported in terms of the proportion of patients having achieved or failed to achieve a "sufficient response". More and more studies have adopted this new outcome reporting approach as it is more meaningful than reporting the group-level average change score or final score only. The current problem is that the data sources, outcome measures and response criteria used in such studies differ from one another, and there is a wide variation regarding the outcome following TKR and THR published by different studies. This variation could reflect the difference in how a good outcome is calculated rather than a true difference in study populations.

The aim of the study in this chapter is to comprehensively evaluate the outcome following TKR and THR using a new approach, viz. employment of four well-established criteria to calculate the proportion of patients achieving a "sufficient response" at six-month post operation. In addition, the study aimed to gain more insight into the advantages and limitations of each metric by comparing the results obtained from various criteria, and ultimately provide a more accurate estimation of the proportion of patients achieving successful outcome.

The null hypothesis to be tested is:

The wide variation in the reported "success" rate following TKR and THR reflects a true difference in study populations, and therefore is not a result of the difference

in how a good outcome is calculated.

3.2 Methods

3.2.1 Data Source

This study used the Patient Reported Outcome Measures (PROMs) Record Level Knee Replacement and Hip Replacement Data, which are prospectively collected on a national scale through the NHS. The data was accessed via the Health and Social Care Information Centre (HSCIC) official website. Full PROMs data reports and detailed guidelines on methodology is also available online [107]. Between April 2014 to March 2015, the data of 47,743 TKR patients and 45,554 THR patients were recorded. Patients were invited to complete a pre-operative (baseline) OKS or OHS Questionnaire, and then a follow-up questionnaire at 6-month post-operation. Beside the OKS/OHS, patients were also asked to answer two additional anchor questions addressing transition and satisfaction. For this study, only patients with a complete set of information (answers to pre- and post-operative OKS questionnaires and anchor questions) were included in the analysis, this reduced the sample size to 46,094 and 44,015 for TKR and THR, respectively. Figures 3.1 and 3.2 illustrate how TKR and THR data were sorted.

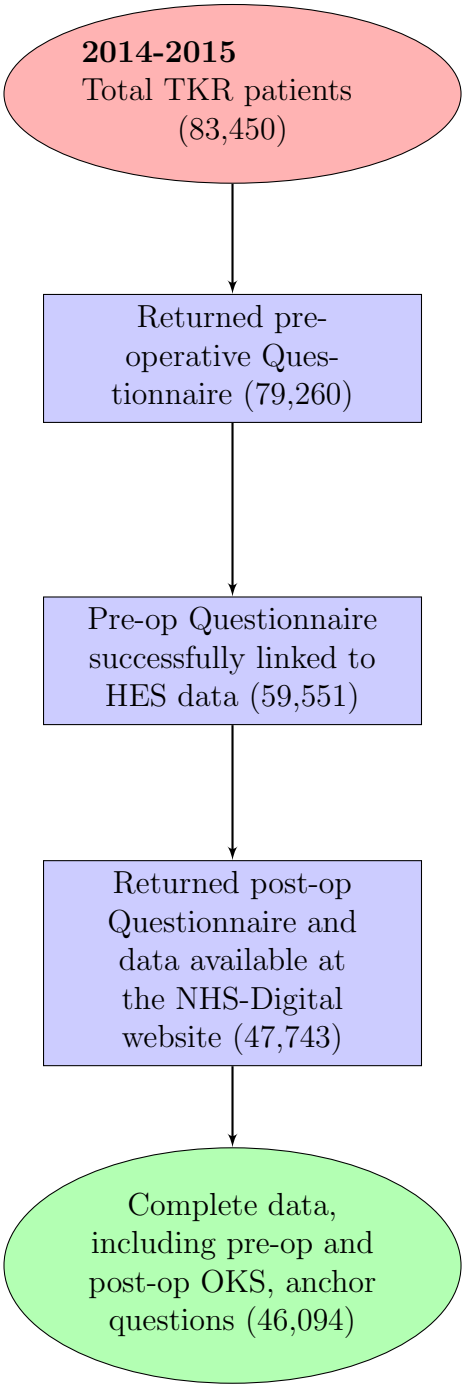


Figure 3.1: Flow chart showing how TKR data were sorted

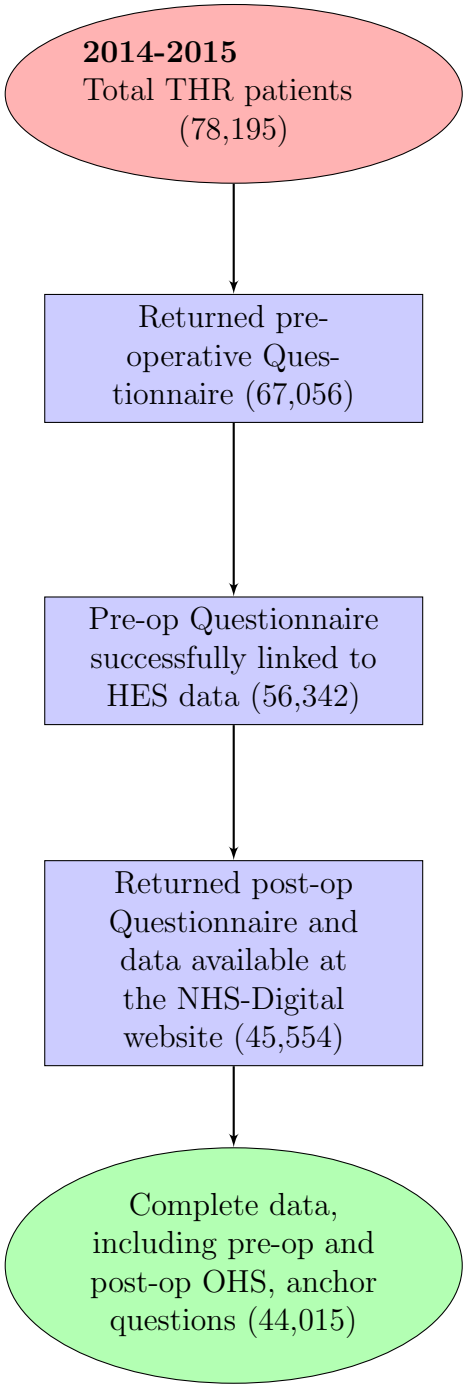


Figure 3.2: Flow chart showing how THR data were sorted

3.2.2 Outcome measures

3.2.2.1 The OKS and OHS summary score

Details of the OKS and OHS questionnaires can be found in section 2.6.2.1 of the Literature Review (Chapter Two).

3.2.2.2 The OKS pain and function subscales

To assess pain and function separately, the OKS-pain component subscale (OKS-PCS) and OKS-function component subscale (OKS-FCS) were used [88]. The OKS-PCS is consisted of items 1,4,5,6,8,9 and 10; the OKS-FCS is consisted of items 2,3,7,11 and 12. The OKS-PCS and OKS-FCS were created based on results from rigorous exploratory and confirmatory factor analysis; both subscales have already been validated [88].

3.2.2.3 The OHS pain and function subscales

In order to examine outcome in the domains of pain and function independently, the OHS-pain component subscale (OHS-PCS) and OHS-function component subscale (OHS-FCS) were used [88]. The OHS-PCS is consisted of items 1,8,9,10,11,12; the OHS-FCS includes items 2,3,4,5,6,7. Similar to the OKS subscales, the OHS-PCS and OHS-FCS were created based on results from thorough factor analysis, and the validity of both scales have been tested [88].

3.2.2.4 The anchor questions

In addition to the OKS and OHS, the following two anchor questions are used in the current NHS TKR/THR Post-operative Survey Questionnaire, one addresses

patients' level of satisfaction and the other enquires about perceived "transition".

1. "How are the problems now in the operated knee/hip compared to pre-operation?" Please select one of the following : 'much better', 'a little better', 'about the same', 'a little worse' and 'much worse'.

2. "Rate your overall level of satisfaction with the operation by selecting one of the following: 'excellent', 'very good', 'good', 'fair' and 'poor' [123].

3.2.3 Criteria for "sufficient response"

Currently, there are no universally-approved criteria/standards defining a "sufficient response" following TKR and THR. In this study, we used the four established criteria outlined earlier in the chapter:

- MCID – ΔOKS or ΔOHS must reach 1/2 SD of the mean baseline OKS or OHS for the total patient population [114].
- Δ OKS/ ΔOHS must reach the minimal important change (MIC) of OKS/OHS – the average score change associated with patients who answered "a little better" to the transition anchor question [102].
- Δ OKS or 6-month post-operative OKS must reach PASS score, which separates patients who "feel satisfied" from those who do not [115].
- The modified OMERACT-OARSI criteria, which categorises patients as "responders" and "non-responders" based on patients' relative and absolute change in pain and function [116].

Tables 3.3 and 3.4 summarise the corresponding cut-off scores for each criterion.

Table 3.3: Threshold values for each criteria for TKR

Criterion	Threshold value
MCID – 1/2 SD (baseline)	
OKS-Summary	$*\Delta$OKS $\geqslant$3.86
OKS-PCS	$*\Delta$OKS $\geqslant$2.35
OKS-FCS	$*\Delta$OKS $\geqslant$1.83
MIC	
OKS-Summary	ΔOKS $\geqslant$9.22
PASS score	
Change in OKS-Summary	ΔOKS $\geqslant$ 11
Post-op OKS-Summary	Post-op OKS $\geqslant$ 30
Modified OARSI	
Relative OKS-PCS change	20% increase in OKS-PCS
Relative OKS-FCS change	20% increase in OKS-FCS
Absolute OKS-PCS change	ΔOKS-PCS (standardised) $\geqslant$ 10
Absolute OKS-FCS change	ΔOKS-FCS (standardised) $\geqslant$ 10
OR	
Relative OKS-PCS change	$\geqslant$ 50% increase in OKS-PCS
Absolute OKS-PCS change	ΔOKS-PCS (standardised) $\geqslant$ 20
OR	
Relative OKS-FCS change	$\geqslant$ 50% increase in OKS-FCS
Absolute OKS-FCS change	ΔOKS-FCS (standardised) $\geqslant$ 20

$*$These cut-off scores were determined using PROMs TKR 2014-2015 data by the author; all other cut-offs were identified in previous studies using different data sources.

Table 3.4: Threshold values for each criteria for THR

Criterion	Threshold value
MCID – 1/2 SD (baseline)	
OHS-Summary	$*\Delta$OHS $\geqslant$4.16
OHS-PCS	$*\Delta$OHS $\geqslant$2.13
OHS-FCS	$*\Delta$OHS $\geqslant$2.37
MIC	
OHS-Summary	ΔOHS $\geqslant$10.63
PASS score	
Change in OHS-Summary	ΔOHS $\geqslant$14
Post-op OHS-Summary	Post-op OHS $\geqslant$35
Modified OARSI	
Relative OHS-PCS change	20% increase in OHS-PCS
Relative OHS-FCS change	20% increase in OHS-FCS
Absolute OHS-PCS change	ΔOHS-PCS (standardised) $\geqslant$ 10
Absolute OHS-FCS change	ΔOHS-FCS (standardised) $\geqslant$ 10
OR	
Relative OHS-PCS change	$\geqslant$ 50% increase in OHS-PCS
Absolute OHS-PCS change	ΔOHS-PCS (standardised) $\geqslant$ 20
OR	
Relative OHS-FCS change	$\geqslant$ 50% increase in OHS-FCS
Absolute OHS-FCS change	ΔOHS-FCS (standardised) $\geqslant$ 20

$*$These cut-off scores were determined using PROMs THR 2014-2015 data by the author; all other cut-offs were identified in previous studies using different data sources.

3.2.4 Statistical Analysis

Data were analysed using R 3.3.0 GUI programming language. Pre-operative characteristics were calculated using proportions and means as appropriate. The mean change in OKS/OHS-Summary, OKS/OHS-FCS and OKS/OHS-FCS scores were calculated for TKR and THR patients separately. The proportion of patients who have achieved our criteria for "good outcome" were calculated with 95% confidence interval for TKR and THR patients separately.

3.3 Results

3.3.1 Study population characteristics

3.3.1.1 TKR patients

Between April 2014 and March 2015, a total of 83,450 patients across England and Wales underwent TKR [123], of which 47,743 patients' record level data were available at the NHS-Digital website [107]. Among these 47,743 patients, 46,094 completed both pre- and post-operative OKS Questionnaires, as well as, the two anchor questions addressing transition and satisfaction. These 46,094 patients' (repliers) data were used for data analysis in this study. Data from the remaining 1,649 patients (non-repliers) whose information was incomplete (missing either OKS or anchor question response) were not used. The flow chart (Figure 3.1) illustrates how TKR data were sorted. Baseline characteristics of repliers and non-repliers were compared using the chi-squared test (for proportional data such as age and gender) and t-test (for numerical data such as the pre-operative OKS/OHS). Quantile-quantile plot was constructed and confirmed normal distribution regarding pre-operative scores. Statistically significant differences were observed between

included and excluded data: I) Repliers had better pre-operative OKS-Summary scores, 19.17 vs 18.17. Such difference however, does not constitute clinical significance. II) Repliers were younger with 12% aged 80 and above; the proportion of octogenarian/nanogenarian was over 20% among non-repliers. III) There was a greater proportion of females among non-repliers. Detailed results are shown in Table 3.5.

		TKR Patients Used	TKR Patients Not Used	P-value
Age n(%)	below 60	4301(10.0%)	97(6.4%)	<0.05
	60 – 69	15835(36.7%)	414(27.2%)	<0.05
	70 – 79	17763(41.2%)	701(46.0%)	<0.05
	80+	5256(12.2%)	312 (20.5%)	<0.05
	missing	2939	125	
Gender n(%)	Male	18123(42.0%)	563 (36.9%)	<0.05
	Female	25032(58.0%)	961(63.1%)	
	missing	2939	125	
Pre-op score	mean (95% CI)	19.17(19.10,19.24)	18.17(17.68,18.66)	< 0.05
	SD	(7.74)	(7.91)	
	missing (n)	0	630	

Table 3.5: Baseline Descriptive Statistics for TKR patients (NHS PROMs 2014-2015)

3.3.1.2 THR patients

A total of 78,195 patients underwent THR between April 2014 and March 2015 [123], of which 45,554 patients' record level data were available at the NHS-Digital website [107]. Only patients who completed pre- and post-operative OHS Questionnaires, along with the two anchor questions were classified as repliers, and their data were used for data analysis in this study. 44,015 of the 45,554 patients were repliers and the rest were non-repliers. The flow chart (Figure 3.2) illustrates how THR data were sorted. Baseline characteristics of repliers and non-repliers were compared using the chi-squared test and t-test as approriate. Similar to TKR patients, Statistically

significant differences were observed between included and excluded data: I) Repliers had better pre-operative OHS-Summary scores, 18.3 vs 16.8. II) Repliers were significantly younger with 14% below 60 and 13% over 80. Only 8% non-repliers were younger than 60 and over 20% were over 80. III) In terms of gender distribution, non-repliers had a slightly higher proportion of female patients. Detailed results are shown in Table 3.6.

		THR Patients Used	THR Patients Not Used	P-value
Age n(%)	below 60	5906(14.4%)	113(7.8%)	<0.05
	60 – 69	13843(33.8%)	378(26.2%)	<0.05
	70 – 79	15849(38.7%)	654(45.3%)	<0.05
	80+	5319(13.0%)	300 (20.8%)	<0.05
	Missing	2939	125	
Gender n(%)	Male	16207(39.6%)	460 (31.8%)	<0.05
	Female	24710(60.4%)	985(68.2%)	
	Missing (n)	2939	125	
Pre-op score	mean (95% CI)	18.34(18.42,18.26)	16.84(16.29,17.40)	<0.05
	SD	(8.33)	(9.14)	
	Missing	0	531	

Table 3.6: Baseline Descriptive Statistics for THR patients (NHS PROMs 2014-2015)

3.3.2 Proportion of TKR patients with a good outcome

For the MCID criterion, 89% patients achieved the threshold in OKS-summary score, almost 91% achieved the cut-off in OKS-PCS, and 79% met the threshold in OKS-FCS. For the MIC criterion, 75% of patients met the threshold. When evaluated using the PASS-change score, which was previously identified to be 11 (OKS-Summary), 72% met that threshold. As for the PASS-final score, 75% achieved the cut-off of 30 points [115]. 83% patients met the modified OARSI criteria. Moreover, the proportion of satisfied patients (those who answered "excellent","very good" or "good" to the satisfaction anchor question) were found to be 85%, which is within

the range set by these threshold-based criteria mentioned before. The proportion of patients who felt better overall after operation (those who answered " much better" or " a little better" to the transition anchor question) were found to be 89%. Detailed results are presented in Table 3.7, Table 3.9 and Figure 3.3

Criterion	**TKR** $n = 46094$ %(95% CI)
MCID (Norman et al 2003)	
OKS Summary	89.2(88.9,89.5)
OKS-PCS	90.5(90.2,90.7)
OKS-FCS	78.7(78.3,79.0)
MIC (Beard et al 2015)	
OKS Summary	75.2(74.8,75.6)
PASS score (Judge et al 2012)	
Change in OKS Summary	72.1(71.7,72.5)
Final OKS Summary	74.5 (74.1,74.9)
Modified OARSI	83.0 (82.7,83.4)

Table 3.7: Proportion of TKR patients with a good outcome by criterion

3.3.3 Proportion of THR patients with a good outcome

For the 1/2 SD criterion, 93% patients achieved the threshold in OHS-summary score, 95% achieved the cut-off for OHS-PCS, and 87% met the threshold in OHS-FCS. 85% of total patients achieved the MIC for THR. When assessed using the PASS-change score, which was identified to be 14 points, 78% met that threshold. As for the PASS final score (at 6-month post-operation), 77% achieved the cut-off of 35 points [115]. 91% THR patients met the modified OARSI criteria. Lastly,

the proportion of satisfied patients (those who answered "excellent","very good" or "good" to the satisfaction anchor question) were found to be 93%, which is close to the results obtained using the Modified OARSI or 1/2 SD criteria. The proportion of patients who felt better overall after operation (those who answered "much better" or " a little better" to the transition anchor question) were found to be 95%. Detailed results are presented in Table 3.8, Table 3.9 and Figure 3.3.

Criterion	**THR** $n = 44015$
	% (95% CI)
MCID (Norman et al 2003)	
OHS Summary	93.5(93.3,93.8)
OHS-PCS	94.6(94.4,94.8)
OHS-FCS	86.7(86.4,87.1)
MIC (Beard et al 2015)	
OHS Summary	84.9(84.6,85.3)
PASS score (Judge et al 2012)	
Change in OHS Summary	78.1(77.1,78.5)
Final OHS Summary	76.6(76.2,77.0)
Modified OARSI	91.3(91.0,91.6)

Table 3.8: Proportion of THR patients with a good outcome by criterion

Criterion	**TKR** ($n = 46094$)	**THR**($n = 44015$)
	% (95% CI)	%(95% CI)
Satisfaction	85.4 (85.1,85.7)	92.5(92.3,92.8)
Transition	89.3 (89.0,89.6)	95.2 (95.0,95.4)

Table 3.9: Proportion of TKR and THR patients with a good outcome judging by satisfaction and transition anchor questions

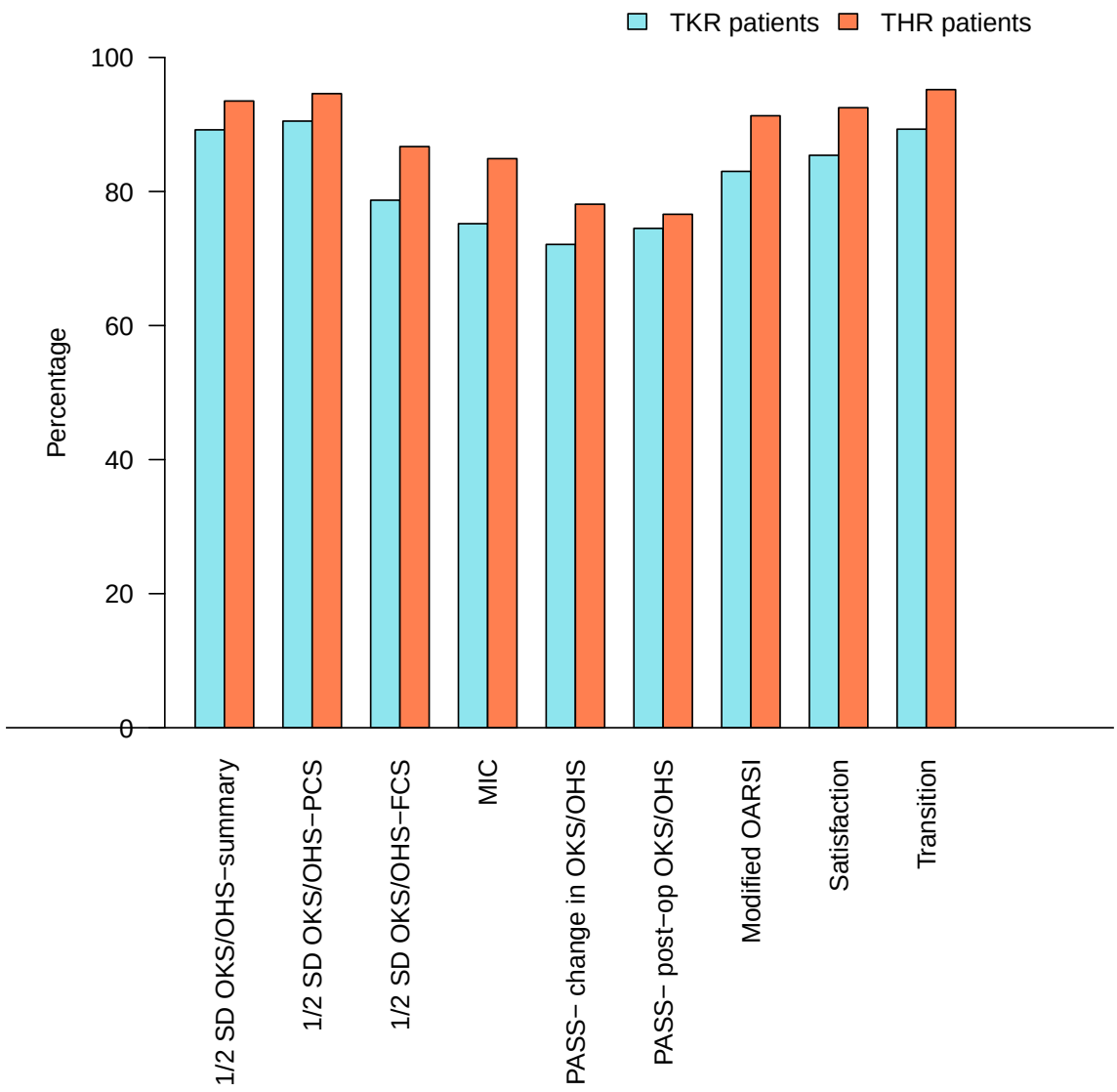


Figure 3.3: Proportions of TKR and THR patients with successful outcome – depending on the criteria, 72% to 90% of TKR patients and 77% to 95% of THR patients have obtained a "sufficient response" following their operation.

3.4 Discussion

3.4.1 Main findings

In previous studies, a wide range of values have been reported regarding the proportion of TKR and THR patients having achieved "sufficient response" from their operation. The definition of "sufficient response" vary according to the choice of outcome measure, response criteria and the methodology used to derive the cut-off scores associated with the criteria. Our study used a large population data, a well-recognised joint-specific outcome measure, and a set of four different response criteria to assess the outcome following TKR and THR. This is the most comprehensive and considered evaluation of the true success of TKR and THR yet. We also investigated the effect of different criteria choices on intervention success rate by comparing results obtained using different criteria.

Our results indicate that the levels of improvement in terms of pain, function, overall perceived benefits and satisfaction is high (over 70%) following TKR and THR. When the outcome is defined as the proportion of patients having achieved "sufficient response", the outcome is closely tied to what response criterion is being used. Different response criteria define "sufficient response" differently, which leads to different criterion-specific cut-off scores. A higher cut-off score would be met by a lower percentage of patients, and therefore is a more conservative estimate of success rate. Following the same logic, a lower cut-off score is likely to amplify success rate. For TKR, PASS – Δ OKS is associated with the highest cut-off scores, which were met by 72% of total TKR patients. The MCID for OKS-function subscale is associated with the lowest cut-off, which was achieved by as many as 90% of all TKR patients. Regarding THR, the highest cut-off score was derived from PASS – post-op OHS criterion, which was met by 77% of total THR patients. Similar to

TKR, the MCID for OHS-function subscale is associated with the lowest cut-off, which was achieved by 95% of total THR patients.

Consistent with results from other studies [124, 105, 125, 126], our data showed that baseline scores (measured on the OKS and OHS) were similar in TKR and THR patients, but the proportion of good outcome was significantly higher in THR patients. Within each procedure, pain improves more significantly than function [127], this holds true for every criterion applied.

3.4.2 Advantages and disadvantages of different response criteria

3.4.2.1 MCID – 1/2 SD

A higher proportion of patients achieved the MCID (1/2 SD) than those achieving the MIC and PASS score criteria (detailed results are shown in Table 3.7 and Table 3.8). The fundamental difference between MCID – 1/2 SD and the other two criteria is that a cut-off score of 1/2 SD is purely distribution-based, whereas MIC and PASS score contain an anchor-question component. The OKS and OHS are condition-specific questionnaires with a high degree of responsiveness. Sometimes patients' scores improve but they do not perceive a change in their overall health state. This "systematic subjective pessimism" phenomenon has been reported previously by Angst et al [128]. In their study, patients' change in WOMAC score were linked to their response to the transition question, which asks participants to assess how they are feeling today comparing to three months ago (as far as their arthritis is concerned). On average, patients who replied "equal" to the transition question had a small improvement in their WOMAC score [128]; the author attributed this discrepancy to patients being inherently "pessimistic". However, in this context,

saying that the questionnaire scores are overly optimistic would be equally valid. This small but systematic difference between patients' change score and subjective perception of change in health status demonstrates the need to incorporate external anchor questions when assessing treatment outcome, and therefore it would not be appropriate to use MCID alone as the responder criterion. Moreover, in 2009 the FDA recommended that for clinical studies, the *a priori* "responder definition" (i.e. A PROM cut-off score or change score indicating clinically significant benefit from the treatment) should be derived using an anchor-based approach; cut-off scores obtained using distribution-based methods should only be used for "supportive" purpose [27].

3.4.2.2 OMERACT-OARSI responder criteria

Regarding the modified-OARSI criteria, using the OKS and OHS subscales [88], we have identified 91% THR patients and 83% TKR patients as "responders"– patients who have felt a clinically significant improvement regarding pain and function in response to the surgery. Escobar et al conducted a similar study where they investigated the percent of patients meeting OARSI criteria using a different dataset. Escobar and colleagues found that the prevalence of good outcome was 95% for hip OA patients and 86% for knee OA patients, which is slightly higher than, but still comparable to, the proportion we have identified [111]. Escobar and colleagues used WOMAC pain and function subscales as their outcome measures, and change score was defined as the difference between pre- and one-year post-treatment WOMAC-pain or WOMAC-function scores. Despite the difference in outcome measures, data source, and time of assessment, the proportions of "responders" obtained in these two studies were in the same range. This suggests that the modified OARSI criteria, which defines "sufficient response" in terms of both relative and absolute change in pain and function domains, can be adapted to the scoring system of various out-

come measures and produces a more or less consistent success rate. However, not all outcome measures can be clearly divided into pain and function components. Also, the methodology associated with this criterion appears to be the most cumbersome to implement.

3.4.2.3 PASS score

As mentioned in introduction, the values of PASS score (Δ OKS $\geqslant$ 11, Δ OHS $\geqslant$ 14, post-op OKS $\geqslant$ 30, post-op OKS $\geqslant$ 35) were derived using the ROC curve analysis. In this case, the area under the curve (AUC) was found to be high, ranging between 0.80 and 0.86, indicating that patients' OKS/OHS is able to correctly identify those with a satisfactory outcome from those without. Also, our results suggest that PASS score is the most conservative marker of success (amongst the four criteria applied in this study) which, to some extent, offers a level of reassurance – a patient who has achieved success by the PASS score criterion would most likely be deemed to have had successful treatment when judged by other criteria.

However, there are a few issues associated with the derivation of the PASS score. Regarding the methodology, in the original study, patients were first asked to rate their post-operative satisfaction on a scale of 0 to 100, and then a cut-off of 50 was used to classify patients into "satisfied" and "non-satisfied" categories. This raises the question whether it was appropriate to convert the 100-level NRS satisfaction data into a binary variable. There is also a lack of evidence supporting the use of a score of 50 as the threshold. Furthermore, on a conceptual level, the PASS score defines a "sufficient response" as being satisfactory with the operation. Even though patient satisfaction has seemingly become the most important outcome parameter in recent years, it is well known that satisfaction is affected by a multitude of factors, including patients' mental state and socioeconomic status, cultural back-

ground, experience with other services in the hospital, etc [129]. Therefore, using patient satisfaction as the basis for assessing post-operative outcome may not be the most appropriate.

3.4.3 Response criteria based on patients' experience only

The application of any conventional response criteria (e.g. MCID, MIC, PASS score and OMERACT-OARSI) in assessing post-operative outcome becomes possible only after the associated cut-off scores have been identified. In addition to the four traditional, threshold-based criteria, we also incorporated two purely patient's experience-centred variables – satisfaction and transition (which is also referred to as "self-perceived improvement" or " global rating scale" sometimes). The proportion of *improved* and *satisfied* patients ranged between 85% and 95%, which is comparable to the results obtained here using threshold-based criteria.

3.4.4 Strengths and weaknesses

The strength of this study lies in the size and comprehensiveness of the data, which confers a high validity to the results. This study used large, prospectively collected orthopaedic datasets containing comprehensive pre- and post-operative demographical and clinical information on all patients across England and Wales who underwent TKR and THR between April 2014 and March 2015 and volunteered to participate in the NHS program. Also, the condition-specific outcome measures chosen for this study, the OKS and OHS, possess high validity and responsiveness. The OKS and OHS subscales were based on Harris et al's work, where they used a scientifically sound methodology, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), to separate the 12-item questionnaires into two domains - pain and function [88]. Previous subscales were created based on wording of the question (i.e.

whether it "sounds" more closely associated with pain or function) without formal statistical analysis [130, 131].

Missing data constitutes the main weakness of this study – amongst all patients whose data were recorded in the NHS, 1,649 TKR and 1,539 THR patients' data were excluded in this analysis due to missing values. As shown in Tables 3.6 and 3.5, for both TKR and THR, the non-repliers as a group are older and more female-dominant, and have a slightly lower pre-operative score.

Missing data is a common problem in clinical and epidemiological research; the following are the three main types of missing data [132].

- **Missing Completely at random (MCAR)** refers to the situation when the probability that the data are missing is not related to either the specific value which is supposed to be obtained or the set of observed responses.
- Data are defined as **missing at random (MAR)** when the probability that the data are missing is related to the set of observed responses, but is not related to the specific missing values which is supposed to be obtained.
- **Missing not at random (MNAR)** occurs when the probability that the data is missing is related to the specific missing values which is supposed to be obtained. For example, people with high salaries generally do not want to disclose their incomes in surveys, resulting in a large proportion of missing data.

In the case with MCAR, removing missing values and analysing the observed data will produce unbiased results. Unbiased results can sometimes be obtained for MAR data. For MNAR data, excluding missing data in analysis always produce biased results. It is difficult to determine the underlying mechanism of missing data by using observed data only. In the case with NHS-PROM data, it is reasonable to

assume the nature of missing data to be MAR [132].

Multiple imputation by chained equations (MICE) is a practical approach to the problem of missing data. It aims to allow for the uncertainty about the missing data by first creating several different plausible imputed data sets, and then combining results obtained from each of them. The true values of the missing data can never be ascertained, only estimated [133, 132].

The multiple imputation procedure requires the user to model the distribution of each variable with missing values, in terms of the observed data. The validity of multiple imputation results is heavily dependent on such modelling being done carefully and appropriately. Flawed imputation procedure can seriously undermine the validity of the research and potentially lead to to false conclusions [133]. When there is no flaws in the technical aspect of multiple imputation, larger data with a greater number of variables and a smaller percentage of missing value tend to lead to better imputated datasets (i.e. closer to the real values of the missing data).

Multiple imputation was not performed in this study. Considering the size of the original datasets, and the low percentage of missing data (3%), the validity of the results were unlikely to have been affected significantly. In the future, it is still worthwhile to perform the same analysis on a properly imputated dataset.

3.4.5 Clinical implications

In order to make a fully informed decision when deciding whether to undergo joint replacement surgery or not, patients need to acquire a thorough understanding of the potential success or failure rate of the intervention [134]. This study has exposed an important issue regarding the analysis and reporting of outcome following TKR and THR. For a single dataset where a single outcome measure is used, depending on

the response criteria on which evaluation of the intervention is based, post-operative outcome can appear rather different. Also, in the field of total joint replacement, several response criteria have been developed, but each one has its own merits and drawbacks, and a gold-standard is still lacking.

To obtain a more comprehensive evaluation of post-operative outcome, it may be reasonable to incorporate a variety of response criteria. Moreover, when reporting outcome following TKR or THR, the outcome measure, response criterion and its associated cut-off score, as well as methodology used for deriving the cut-off, must all be specified. Under circumstances where the use of multiple outcome measures and criteria is not feasible, the decision of which response criterion to use should be made on an individual basis. It is advisable to identify the dimension of the outcome that is most important to the patient (e.g. pain, function or overall satisfaction), and then use the response criterion that matches their priority.

Lastly, it has become apparent that the level of success is dependent on the criterion and method chosen, and the variability in reported success rate of the interventions (on the same population) suggest a new approach may be required. A standardised approach would be useful and some groups have attempted to address this need [135]. There may be a case for directly using experience-oriented anchoring questions such as transition or satisfaction to define a "sufficient response". There is little doubt, in the end the best solution is to ask the patient whether the treatment has been successful or not from their perspective.

3.5 Link to next chapter

This chapter first set up the new framework of evaluating post-operative outcome following TKR/THR – in terms of the proportion of patients achieving a "sufficient

response". The definition of "sufficient response" vary according to I) the choice of outcome measure, II) the response criterion and III) the criterion-associated cut-off score.

In this study, the OKS and OHS questionnaire, and four different response criteria, namely the MCID, MIC, PASS, modified-OARSI, were used to assess the short-term outcome of TKR/THR for all patients recorded in the NHS. The results of this chapter indicate that depending on the selected response criterion, sufficient response rate at 6-month post-operation ranges between 72% and 90% for TKR patients, between 77% and 95% for THR patients. More stringent criteria lead to lower response rate and vice versa. PASS-ΔOKS is shown to be the most conservative criterion for TKR.

The next chapter reports the exploration of the long-term outcome trajectory following TKR. Outcome has been assessed using the evaluation framework set up in this chapter, viz. OKS being used as the outcome measure, PASS-ΔOKS being used as the response criterion, and $\Delta\text{OKS} \geq 11$ being used as the cut-off score.

Chapter 4

Temporal characteristics of self-reported outcome after TKR

4.1 Introduction

TKR is an intervention designed to provide benefits to knee OA patients in both the short and long term. To date, the vast majority of TKR outcome studies have focused on the short-term results, usually between six months to two years post-operation [111] [110]. Previous publication data on long-term post-operative pain and functional outcome following TKR are relatively scarce and the results are inconsistent.

The inconsistency is reflected in the literature – some studies claim that there is no observable change in patients' long-term outcome compared to their short-term outcome, but some report otherwise. Using the Swedish knee arthroplasty registry data from patients who underwent TKR between 1981 and 1995, Robertsson and colleagues found that only 8% of these patients were dissatisfied regarding their operation. The satisfaction rate was constant for up to 17 years post-op [136]. In contrast, results from Ritter et al [137] showed significant decline between 3rd and 12th year in patients' Knee Society Score (KSS)- functional component. A study conducted by Gandhi et al showed that patients' post-operative average Western Ontario and McMaster Universities Index (WOMAC) scores stayed relatively constant until the fourth year and then declined gradually [138]. Using the Oxford Knee Score (OKS) as the primary outcome measure, Williams et al examined the change in scores for each individual questions on the OKS questionnaire. Optimal score occurred during the first or second year in most cases, with the exception of night pain. After year two, a steady, but statistically, insignificant yearly decrease in average individual-question score was observed for all items except kneeling [139].

Furthermore, the methods used in previous studies have potential flaws – each year, the group-level average PROM score was calculated and then compared with one another [137, 138, 139, 140]. The problem with using raw numerical scores from

quality of life (QoL) instrument is that the magnitude of these scores (or change in scores) is not placed in a context that is meaningful for clinicians or patients [30]. Chapter Three of this thesis introduced a more meaningful way of describing post-operative outcome for the population (not individuals) – in terms of the proportion of patients having achieved a "sufficient response"; the higher the proportion, the better the outcome. The three essential components for defining a "sufficient response" include I) an appropriate outcome measure, II) a pre-determined response criterion and III) a criterion-associated cut-off score.

So far, in joint replacement research, the method of dichotomising patients into broad categories according to their PROM scores has mainly been used to report short-term outcome. The aim of the study described in this chapter is to use such (dichotomising) method to examine the long-term outcome following TKR – focusing on change over time in the proportion of TKR patients having obtained a "sufficient response" and the proportion of those who failed to do so each year. In addition, the time it takes for patients to achieve their optimal outcome was also explored.

The null hypothesis to be tested is:

The proportion of TKR patients with a "successful" outcome and those with a "failed outcome" according to the PASS score criterion remains more or less constant from one year to the next during the first ten years following knee replacement operation.

4.2 Methods

4.2.1 Data Source

The study used data from the Knee Arthroplasty Trial (KAT)– a partial factorial, multi-centre randomised controlled trial with the original objective of comparing

clinical effectiveness and cost-effectiveness of three different design features of the TKR implant [141, 142]. Between July 1999 and Jan 2003, participants were recruited and randomly allocated to different treatment groups. The KAT data set contains post-operative OKS from year-one to year-ten, as well as substantial information regarding patients' clinical and demographical characteristics prior to operation (e.g. age, sex, pre-operative mobility issues, ASA physical status grade etc). Initially, 2217 patients were recruited and received their respective operation; the number decreased in subsequent years mainly due to death and non-response. Figure 4.1 illustrates the number of patients remaining in the trial each year and the number of patients included in our analysis. Patients with at least three years of follow-up data were used for final analysis.

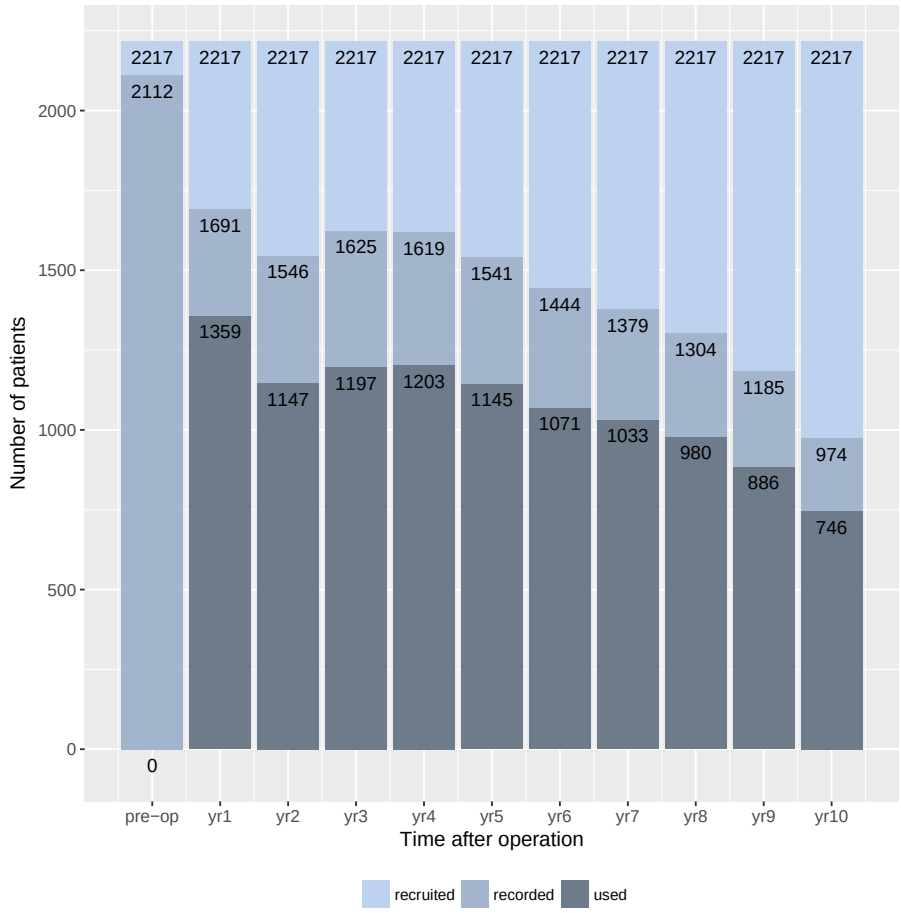


Figure 4.1: Initially 2,217 patients were recruited and received operation with allocated type of implant; the lightest shade represents the maximum sample cohort size. The medium shade represents those patients whose post-operative OKS were recorded in the KAT database – these were the number of patients who actually remained in the trial each year. The darkest shade represents those patients, with at least three years of follow-up data, whose post-operative OKS were included in the final analysis.

4.2.2 Criterion for a "sufficient response"

Currently, there are no universally-approved criterion defining what a "sufficient response" is following TKR. The response criterion used in this study is the PASS score, with an associated cut-off score of $\Delta\text{OKS} \geqslant 11$. The PASS score was first identified by Judge et al, and aimed to separate patients who "feel satisfied" from those

who do not [115]. The reason for choosing PASS score as the response criterion in this study is that the PASS score criterion has been shown to give more conservative estimates than other criteria (refer to Chapter Three for details); in other words, if someone has achieved the PASS score cut-off, their post-operative outcome would most likely be deemed successful when judged by other criteria. In this chapter, patients who succeeded in satisfying the PASS score criterion are labelled as having a "successful outcome", those who failed to meet the criterion labelled as having a "failed outcome".

4.2.3 Statistical analysis

The programming language R (version 3.4.1) was used for all data analyses and figure plotting. One way analysis of variance (ANOVA) was used to examine how average post-operative OKS varies with time (i.e. from year one to year ten post-operation). The correlation between year-one and year-ten post-op OKS was assessed using a linear regression model fitted by the least squares method. To examine the long-term outcome, patients were first divided into two groups according to their year-one ΔOKS (using the PASS score criterion mentioned before) – "success" and "failure". After that, the outcome trajectory of the "success" and "failure" groups in subsequent years were assessed separately in terms of the proportion of patients with "successful", "failed" and "fluctuating" outcomes. "Fluctuating" outcome is defined as "having met the PASS score criterion in the previous year but failed to do so in current year, or vice versa".

4.3 Results

4.3.1 The group-level average OKS between one year and ten years after TKR

The change in average post-operative OKS over time is depicted in Figure 4.2. Results from an ANOVA analysis indicates that there is a small but statistically significant ($p<0.001$) decrease in group-level mean OKS over the course of ten years following TKR – after post-op year one, each subsequent year passed is associated with a 0.17 point drop in OKS (regression coefficient = -0.17(-0.24,0.10)). However, this small, but steady decrease over time is not clinically significant.

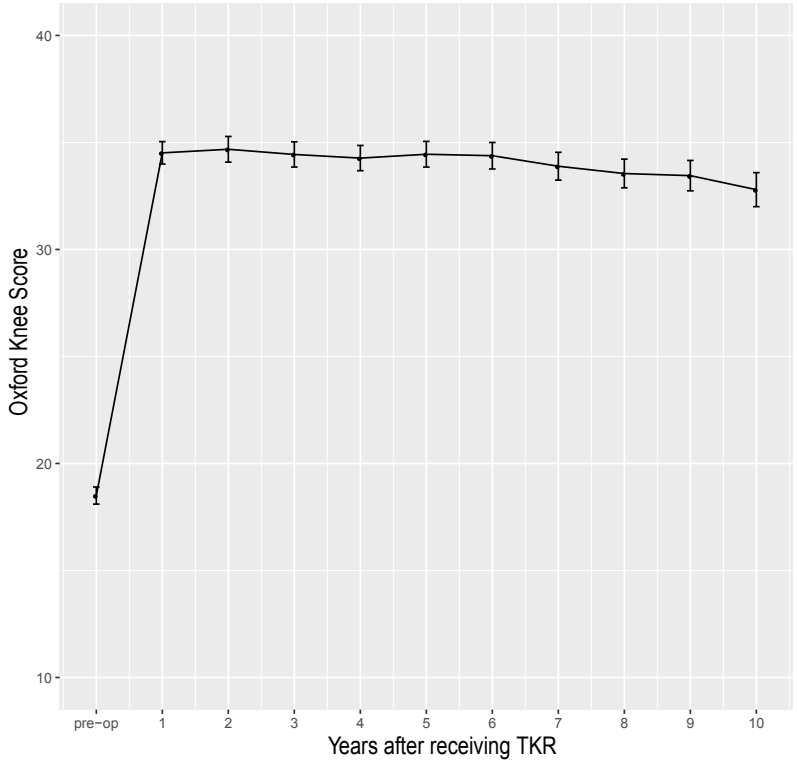


Figure 4.2: The mean pre-operative and post-operative OKS following TKR each year. Error bars represent 95% confidence interval.

.

To further explore the nature of the relationship between patients' short-term and long-term outcome following TKR, we constructed a box-whisker plot (Figure 4.3), showing the distribution of patients' OKS at ten years post-operation conditional on their year-one post-operative OKS. Data from 746 patients, for whom both post-operative year-one and year-ten OKS follow-up data were available, were used in this analysis. The width of the box is proportional to the number of patients in that score range. One year after TKR, the majority of patients achieved a post-op OKS of higher than 34 points, only a few scored under 25 points. The length of the box and "whisker" reflects how widely spread patients' OKS was at ten years given their one-year score. Not only is a higher first year post-operative OKS correlated with higher ten-year score, it is also associated with a narrower range of variation at year-ten follow up. A correlation coefficient of 0.64 was identified by the linear regression model, suggesting a moderate, positive linear correlation between post-op OKS at one year and ten years. Moreover, the r-squared value of 0.41 shows that approximately 40% of the variation in OKS at year-ten post-op can be explained by the variation in OKS at year-one.

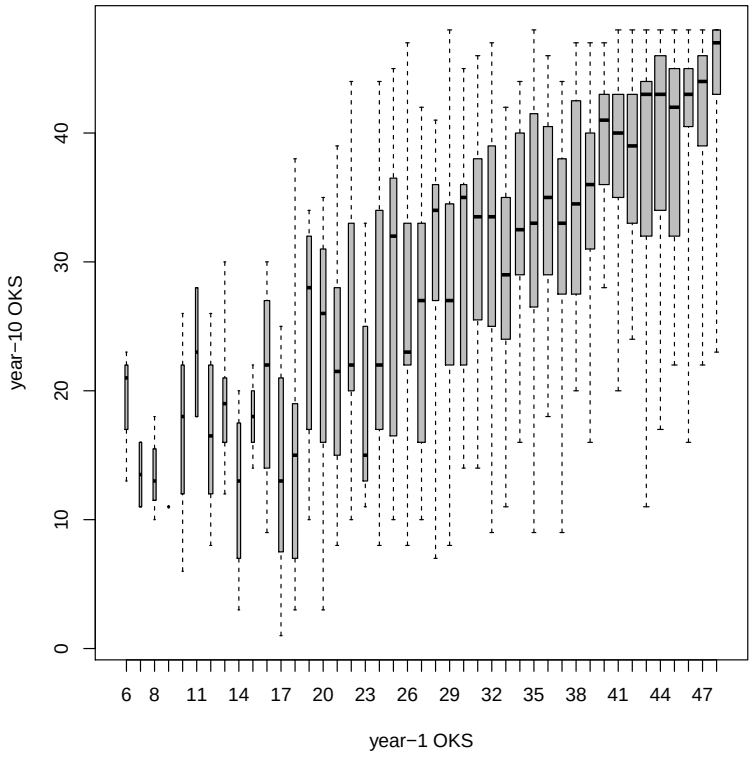


Figure 4.3: Box-whisker plot of the distribution of all participants' post-operative OKS observed at 10 years conditional on their post-op OKS at one year. There is a strong, positive linear correlation between patients' post-op OKS at ten years and at one year (r=0.64, β_1=0.72 (0.64,0.80)).

4.3.2 Fluctuation in long-term post-TKR outcome

In this study, fluctuation refers to the situation where patients achieved a "successful outcome" in one year (i.e. obtaining a post-operative OKS that is at least 11 points greater than pre-operative OKS), but failed to do so in the following year, or the other way around. Applying the cut-off score associated with the PASS criterion (ΔOKS $\geq$11), 71% of total patients achieved a "successful outcome" at one year post TKR; Figure 4.4 illustrates how these patients' outcome have progressed for the next ten years. Each colour in Figure 4.4 represents a different outcome category:

1. Bright pink represents patients with a successful outcome in both the previous and current year.

2. Bright cyan represents patients with a failed outcome in both the previous and current year.

3. Shaded pink represents the first type of fluctuators – patients had a failed outcome in previous year but achieved a successful outcome in the current year.

4. Shaded cyan represents the second type of fluctuators – patients had a successful outcome in previous year but failed to maintain it in the current year.

In the first year, all patients had a successful outcome, represented by the leftmost bar coloured in pink. By year two, 9.7% of these patients' ΔOKS dropped below 11 points – they are the early "fluctuators" (second bar from left, shaded cyan colour). Subsequently, both types of fluctuations exist each year. This fluctuating phenomenon continued until the tenth year, and the fluctuator population was kept in the range of 11% to 16%. The proportion of failed outcome (ΔOKS $\leq$11) gradually increased from 10% to a quarter from year two to year ten.

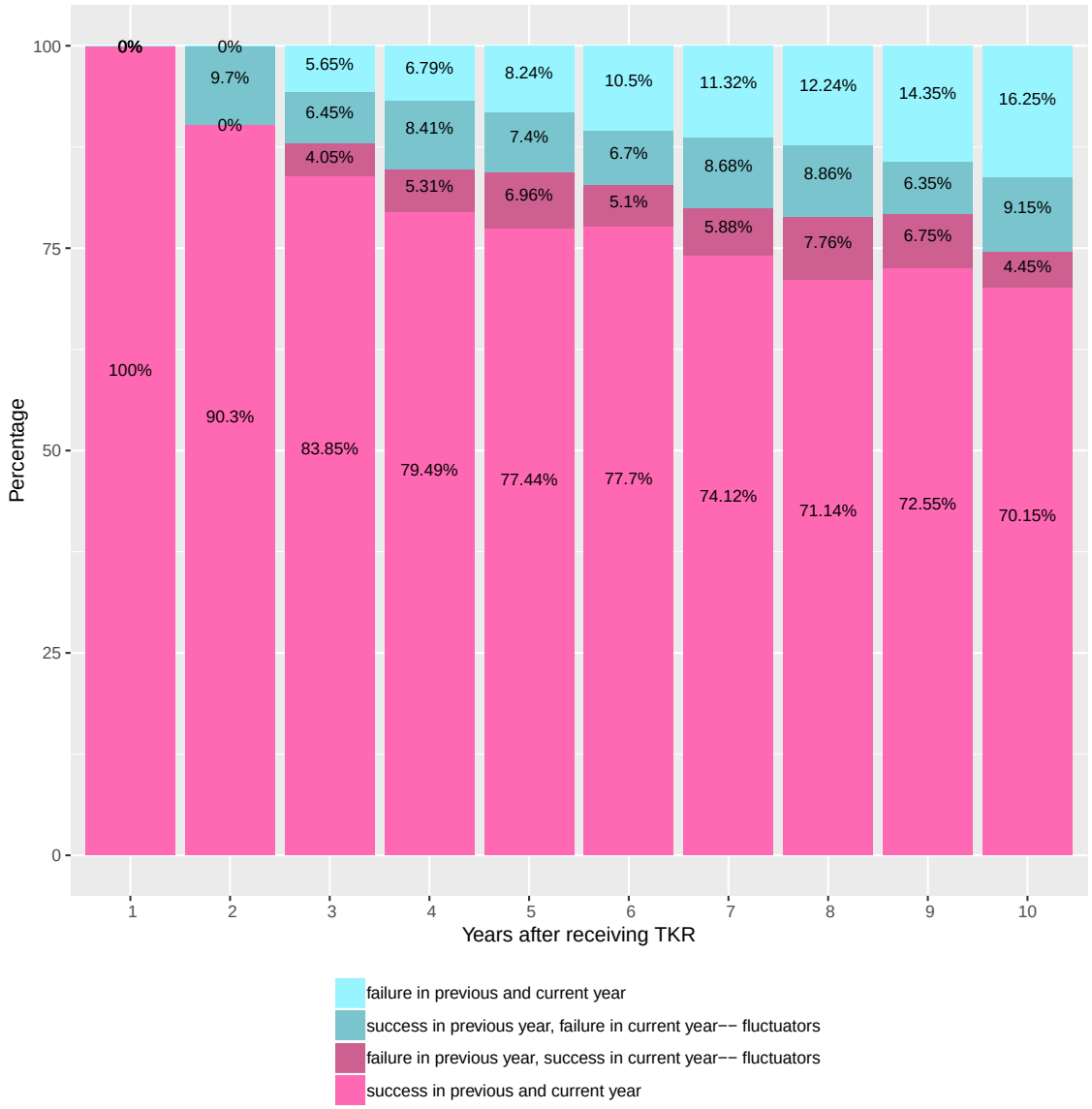


Figure 4.4: Bar chart showing the long-term outcome of those 71% patients who achieved a successful outcome (i.e. Δ OKS$\geq$11) at one-year follow up. Data from patients with at least three years of OKS follow-up data and a sufficient response (ΔOKS $\geqslant$ 11) at year-one post-operation were included in the analysis. (n = 977)

While 71% of total patients achieved a "successful" outcome at one-year post-operation, 29% of patients failed to meet the PASS criterion (ΔOKS $\geq$11); Figure 4.5 depicts how these patients' outcome have progressed for the next ten years. The same colour code from Figure 4.4 was used in this Figure 4.5.

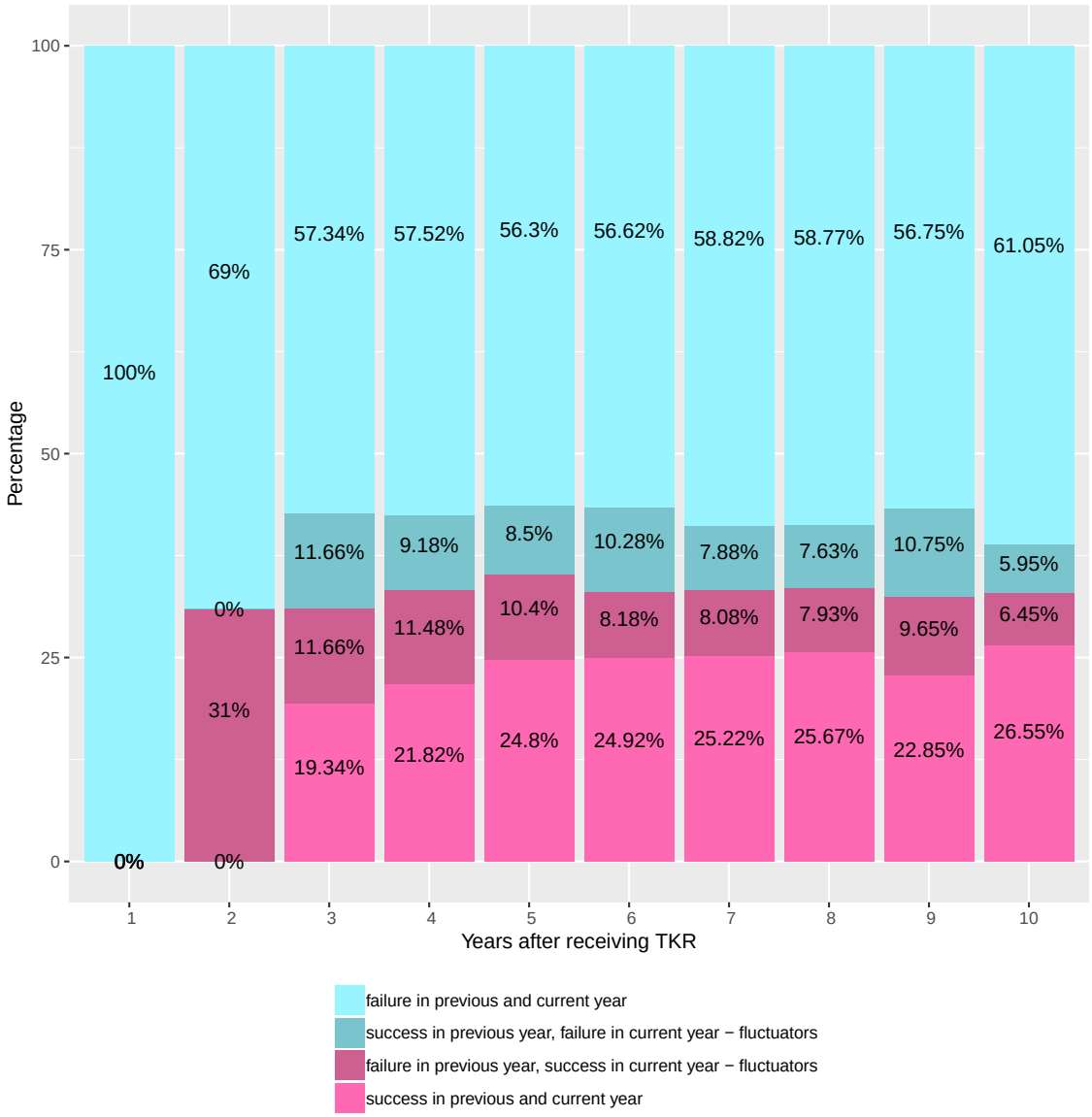


Figure 4.5: Bar chart showing long-term outcome of the 29% patients who failed to achieve a successful outcome (i.e. ΔOKS$\leq$10) at one year post-TKR. Data from patients with at least three years of OKS follow-up data and failed to achieve a sufficient response at year-one post-operation were included in the analysis. (n = 382)

In the first year, all patients (designated as "failed" at one-year follow up), are represented by the leftmost bar coloured in bright cyan. By year two, 31% of these patients managed to improve and meet the PASS score cut-off, they are the early

fluctuators (second bar from left, shaded pink colour). This fluctuating phenomenon persisted until year ten, but the fluctuating population gradually diminished from 30% to 12% in the course of ten years. The overall ratio of successful versus failed outcome was maintained at 1:2 from the second until the 10th year.

4.3.3 When do patients achieve sufficient response following TKR?

Results from the previous section (4.3.2) have indicated that following TKR, between post-operative year-one and year-ten, a sizeable proportion of patients fluctuate between "successful" and "failed" outcomes. In particular, Figure 4.5, shows that approximately one third of those patients with a failed outcome in year one can improve later on and achieve sufficient response at some point between two and ten years post-operation. This finding leads us to examine the time at which patients first achieve a sufficient response, our results are shown in Figure 4.6.

Amongst all patients who achieved the PASS criteria at some point between year-one and year-ten post-operation, 80% met the cut-off score of $\Delta\text{OKS} \geqslant 11$ in first year, 9% did so in second year, and by year-five over 97% patients would have achieved the cut-off score. A very small fraction (less than 3%) of patients achieve the PASS criteria between year-six and year-ten.

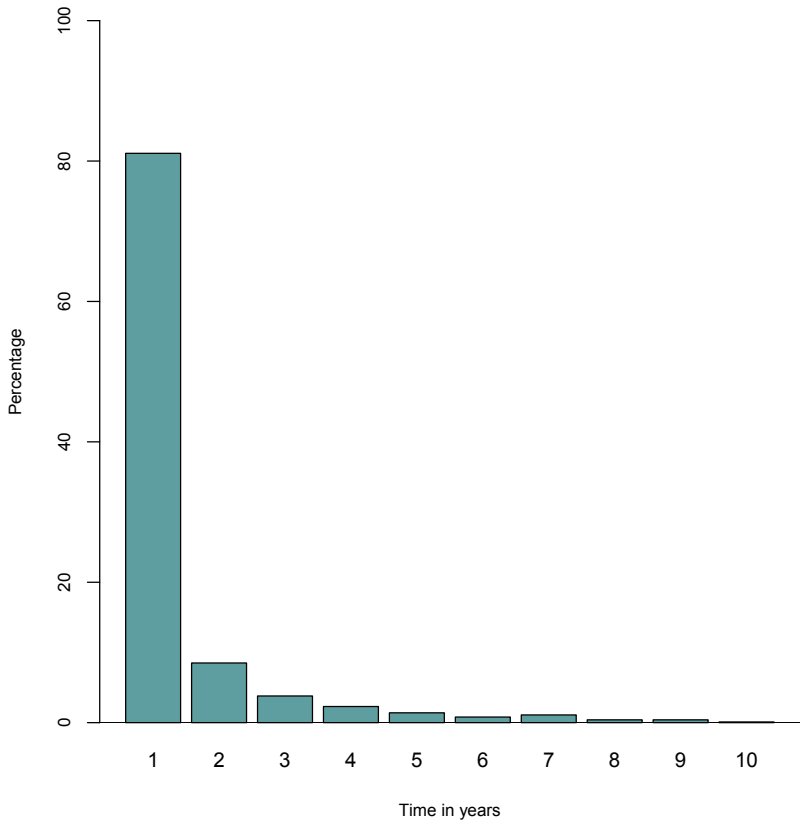


Figure 4.6: Bar chart showing the distribution of the earliest time patients achieve a sufficient response (ΔOKS $\geqslant$ 11). Data from patients with at least three years of OKS follow-up data and achieved a sufficient response at some point between one-year and ten-year post-operation were included in the analysis (n = 1205).

4.3.4 When do patients achieve optimal outcome following TKR?

According to the results from the previous section (4.3.3), amongst those patients having obtained a sufficient response at some point between year-one and year-ten post-TKR, 90% of them have done so within the first two years. This section explores the amount of time needed to reach best outcome post-TKR, and a rather different result was revealed. Figure 4.7 is a bar chart illustrating the time it takes for patients to reach their highest post-operative OKS, which was used as a proxy for their pain and functional outcome. A quarter of the patients reached best outcome

within the first year; 15% took two years to achieve that, 73% managed to achieve their best outcome within the first five years, and over a quarter took more than five years to reach their optimal outcome.

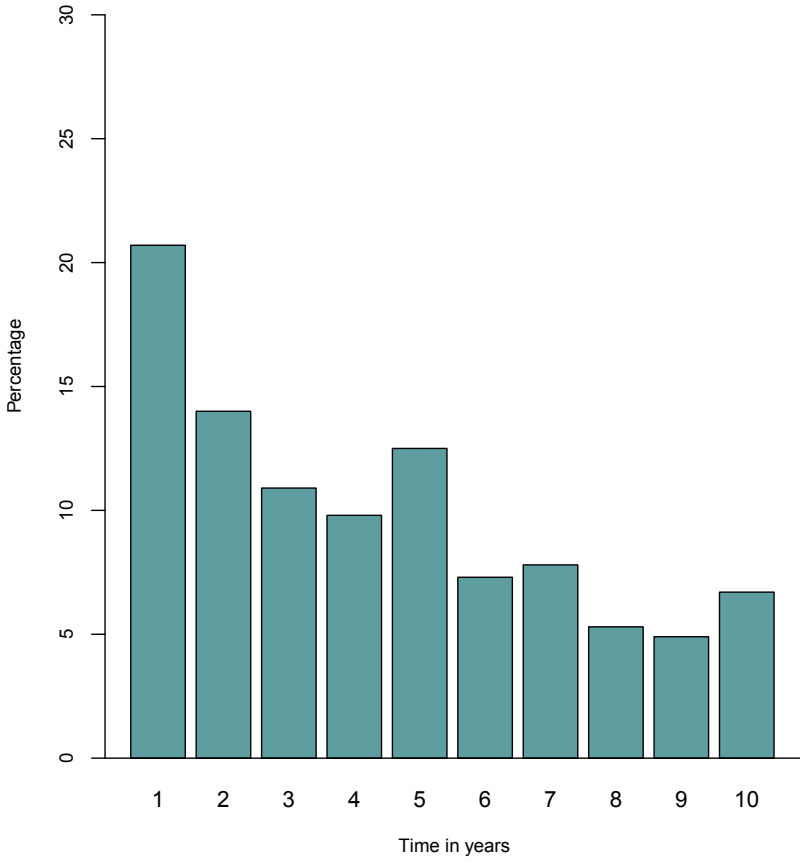


Figure 4.7: Bar chart illustrating the time it takes for patients to reach their optimal outcome (i.e. highest post-op OKS). Data from patients with at least three years of OKS follow-up data were included in the analysis (n = 1359).

4.4 Discussion

4.4.1 Main findings

There is a perception, with some available evidence [143, 136, 144, 138], that the majority of patients (aside from catastrophic failure) undergoing knee arthroplasty

attain maximum benefit (highest level of outcome) by one year following surgery, and there is little deterioration or improvement after that. The first part of this study, using post-operative OKS data from a large, pragmatic, multicentred- RCT, confirmed this established pattern. Our results indicate that most improvement in terms of change in OKS after knee replacement is achieved in the first year after surgery, followed by a period of relative stability for up to ten years post arthroplasty. There is a gradual and statistically significant drop in OKS over the longer term. However, from a clinical perspective, such change (1.9 points over a ten-year period) does not reach clinical significance [102]. Further evidence of this temporal stability in outcome is provided by the relationship between first year scores and tenth year scores. A moderate, positive correlation was found between patients' first and tenth-year post-op OKS, despite a wider distribution of scores in the tenth year (shown in Figure 4.3).

Unfortunately, the pattern of relative stability is derived from the use of simple raw OKS values which have the limitations previously outlined in the Introduction section. The second part of this study used an alternative approach to examine TKR patients' long-term outcome, one that involves appropriate classification of patients based on whether they have achieved a sufficient response or not, labelled as "successful" or "failed" outcome. The response criterion used in this study is the patient acceptable symptom state (PASS) score, which incorporates a patient-centred dimension of what is a successful outcome, including the notion of satisfaction, and is referenced off the individual patients themselves.

Using this new approach, the relative stability in outcome over time, seen using other metrics and analyses, is not replicated. According to the PASS score criterion, 71% of patients achieved a successful outcome at one year post-operation. By the end of the second year approximately 10% of these patients failed to maintain their success. This trend continues with the proportion of worsening patients gradually

increased from 10% to 25% from year two to year ten. Of significant interest is the "change in status" of many patients. Patients can be seen to move between classifications from one year to the next and over time. It is the first time these fluctuations in outcome over time have been reported. Data show that the size of this "fluctuating" population is fairly consistent at between 11% and 16% of the sample. Looking at the patterns of outcome in more detail it can be seen that patients can be dichotomised into two groups if they achieve a successful outcome in year one following surgery. The patients who remain in the "success" category throughout could be labelled "early achievers" and those that vacillated between success and failure in subsequent years, labelled "fluctuators".

Substantial early change in status was also observed in patients who were classified as having a failed outcome in the first year from operation (just over a quarter of the population). Amongst these early "failures", by the second year 31% of them had improved sufficiently to be reclassified as having a successful outcome. As shown, fluctuations between "success" and "failure" continued over the full ten years but the overall ratio of successful to failed outcome was maintained at 1: 2 until the tenth year. The number of patients fluctuating between good and bad from one year to the next (the fluctuating cohort) diminished gradually as time went on. The new classification of patients according to long-term outcome trajectory allows specific sub-groups to be identified for TKR patients who do not do well in the first year after operation: I) patients who do not improve from initial poor outcome – "failures" II) patients who improve and maintain their successful outcome status – "late achievers". III) patients who vacillate between "success" and "faiure" over time – "fluctuators".

A further analysis explored the time taken for patients to meet the response criterion of $\Delta\text{OKS} \geqslant 11$. Amongst the subset of patients who achieved the PASS criteria at some point between year-one and year-ten (around 89% of total population) post-

operation, 80% attained the cut-off score in first year, 9% did so in the second year, and by year-five over 97% patients have achieved a sufficient response. This finding supplements the fluctuating phenomenon described earlier – the vast majority of patients achieve the response criterion in the first five years, after which some maintain their good outcome status, but some vacillate between good and bad outcome until the tenth year.

Lastly, we examined the time taken for patients to achieve their best outcome (i.e. highest post-operative OKS), and a rather different picture was brought to light. Between 20% and 25% of patients reached their highest outcome level at year one following operation, about 15% accomplished it between year one and two, just under 30% acquired their best OKS between year three and five. Over a quarter of patients took over five years to attain their highest level of outcome (see Figure 4.7). These results suggest that when not taking PASS score cut-off into consideration, patients continue making small improvement well beyond the first couple of years post-operation.

4.4.2 What this study can add

The most important message from this work is that the outcome achieved at one year does not necessarily represent later outcome after knee replacement. Moreover, a substantial number of patients (approximately 15 to 20% patients) will vacillate between categories of "success" and "failure" over time and can be further distinguished by their pattern of change. The four types – early achievers, late achievers, failures and fluctuators can be used to further our understanding of longer term outcome after knee replacement and help decision making both in the clinic and for professional policy setting.

For patients who are seen in clinic with non-specific difficulties, a level of reassur-

ance can be given that changes in outcome over time are both commonplace and expected. Patients struggling in the first year following operation can be given the reassurance that, particularly between years one and two, they have a good chance of experiencing an improvement in their condition. Conversely, the study shows that although obtaining a successful outcome early on provides a good indicator for success in the long term, it is not entirely guaranteed. This could be communicated to patients prior to surgery and in subsequent clinic appointments.

4.4.3 Strength and limitations

There are several strengths of this study including the relatively large data size with 1359 patients included in the final analysis. The original condition-specific outcome measures chosen for this study, the OKS, possesses high validity and responsiveness [62, 145, 146] and the PASS score for change is a strong valid metric to use for dichotomising patients into "successful" and "failed" categories. The PASS score was also derived from a large, recently-collected cohort data (the South West London Elective Orthopaedic Centre database, n=3608). Furthermore, intra-rater reliability (i.e. reproducibility) of the OKS was shown to be high [86]. In theory, an outcome measure with low intra-rater reliability would result in a relatively high proportion of "false fluctuators" in the sort of analysis conducted in this chapter. In the case with the OKS, the test-retest sample correlation was 0.92. The difference between the means of the first and second test taken 24 hours apart was not significantly different from 0. Therefore, the "fluctuator" group identified in this study are most likely to represent patients whose outcome truly vacillate from one year to the next.

The main limitation of this study is the high proportion of missing data from the original trial (shown in Figure 4.1), which could lead to a reduction in the power and precision of the estimate [147]. 2217 patients received operation, we decided

to include patients with at least three years of follow-up data in the final analysis, which reduced the total sample size in first year to 1359. Furthermore, as there were more short-term data available than long-term data, the sample size decreased each year. To verify our results, all longitudinal analyses were repeated using a subset of patients (n=449), for whom a complete set (all ten years) of data were available. Detailed results from complete-case analysis are shown in Appendix A.2.3. Incomplete-case analysis and complete-case analysis produced highly similar results in all instances. This indicates that patients with incomplete follow-up data do not behave significantly differently from those with complete follow-up data, but behaviour of those recruited patients whose information are completely missing from the trial database remains undetermined.

Another limitation of this study is that patients who have received a revision operation at some point during the 10-year follow-up were not treated differently from those not needing a revision. This may not have been the most appropriate decision, as revision surgery can lead to a sizeable change in patients' OKS afterwards. However, the revision rate in this trial was low, and therefore may not have had an impact on the final results/conclusion.

4.4.4 Conclusion

This is the first study that has investigated and demonstrated the fluctuating and dynamic nature of long-term outcome following TKR. Contrary to previous evidence, our results suggest that there is a high degree of variation in the trajectory of post-TKR outcome. A substantial proportion of patients fluctuate between successful and failed outcomes over time. Knowledge of enhanced detail and identification of temporal trends in outcome of knee replacement can usefully be employed in the clinic to inform and reassure patients. Long-term monitoring using self-reported

methods, including patient referenced metrics of success, are important.

4.5 Link to next chapter

Following the framework introduced in Chapter Two, which evaluates TKR/THR outcome in terms of the proportion of patients achieving a "sufficient response", this chapter explored the long-term (up to ten years) change in the proportion of patients achieving a sufficient response and the proportion of those failing to do so.

As explained in Chapter Two, a sufficient response is defined by three essential components – a validated PROM, the response criterion and the criterion-associated cut-off point (in terms of PROM score). Anchor questions are often used in the process of deriving cut-off scores. After having used the concept of sufficient response to assess short- and long-term outcome, the next chapter will take a step back and examine the relationship between anchor questions and PROM scores in detail, which will offer some insight into the credibility of cut-off scores commonly used in measuring outcome following total joint replacement.

Chapter 5

The relationship between the metrics of transition question, satisfaction question and self-reported outcome scores

5.1 Introduction

Over the last decade, the use of PROMs in orthopaedics has grown significantly worldwide. However, despite PROM's surging popularity, there have been some challenges in its utilisation. One of the difficulties associated with using PROMs in a clinical setting is the interpretation of the scores [148]. The key to interpreting numerical PROM scores is placing the magnitude of these scores in a context that is meaningful to both the clinicians and the patients [30].

Chapter Three of this thesis elucidated a supposedly more meaningful way of reporting the outcome following TKR/THR – in terms of the proportion of patients having achieved a "sufficient response". Three components are essential in defining a sufficient response – the outcome measure, a response criterion and a cut-off score associated with the criterion. There are two primary approaches for estimating the cut-off scores: distribution-based and anchor-based methods. The anchor-based methods are generally agreed as more valid, as they effectively capture patients' own assessment of treatment by incorporating relevant, more easilly understood external indicators (a.k.a. the anchor questions) [148]. Anchor questions usually address a specific aspect of treatment, including level of satisfaction, change in symptom level (i.e. transition), fulfillment of expectation, etc.

Using different response criteria and a variety of anchor-based approaches, a number of associated cut-off points have been developed, some of which have been applied in clinical trials and contributed to the development of clinical guidelines [149]. Table 5.1 summarises the common response criteria and anchor-based approaches that investigators have used to estimate the cut-off scores.

Outcome measure (PROM)	**Response criteria**	**Anchor question**	**Method used to determine cut-off point**	**Relationship between PROM score and anchor**
WOMAC, SF-36	MCID [118]	Transitional question about joint improvement on a five-point likert scale [118].	Mean difference between pre-op and 6-month post-op PROM scores for patients who reported feeling "somewhat better" about their affected joint [118]	Not assessed
OKS, SF-12 pain component score	MCID [150]	Two questions addressing satisfaction with level of improvement in I) pain II) ability to perform regular activities; response options are in a five-point likert-scale format [150].	Linear regression analysis of how mean post-operative PROM score changes according to change in satisfaction level; slope of the regression line used as the cut-off point [150].	Spearman's correlation coefficient between PROM score and pain relief (r = 0.51-0.56), PROM score and functional improvement (r = 0.56-0.60) [150].
WOMAC, EQ-5D, KSS	"Treatment success", a combined criteria [151].	A list of four questions addressing post-op satisfaction, functional improvement, pain relief and willingness to undergo the procedure again [151].	ROC curve analysis to identify the PROM score that best predicts "treatment success" [151].	Not assessed specifically. The AUC was 0.62-0.75 and 0.69-0.83 at 2-month and 12-month post-op respectively [151].
OKS, OHS	PASS score	Rate overall satisfaction with the outcome of surgery between 0 (extremely dissatisfied) and 10 (extrememly satisfied) [152].	Dichotomise satisfaction score into dissatisfied (<5) and satisfied ($\geqslant 5$); then use ROC curve analysis to identify the OKS/OHS that best predicts satisfaction [152].	Pearson's correlation coefficient between satisfaction score and OKS/OHS , r = 0.64 (OKS) and r=0.52 (OHS) [152].
WOMAC function subscale	Minimal clinically important improvement (MCII) [153]	Transition question asking patients to evaluate their response to treatment on a five-point likert scale [153].	Cumulative percentage analysis to identify the 75th percentile of the Δ WOMAC from patients rating their treatment response as "good", this value was then used as cut-off [153].	Not assessed

Table 5.1: Common response criteria, anchor questions and anchor-based methods that investigators have used to estimate the associated cut-off values.

As shown in Table 5.1, most common anchor questions in the field of TKR and THR address transition (a.k.a. global rating of change in symptoms) and satisfaction, and response choices are usually presented in the format of five-point likert scale or numerical rating scale (NRS). The main methods for cut-off point estimation include I) optimal cut-point from ROC curve analysis, and II) using the mean change in PROM score of the group that reported a small improvement – feeling "a little better" or "somewhat better". Despite the far-reaching applications of some of the previously defined "sufficient responses" (refer to Table 3.1 and Table 3.2 in Chapter Three for more details), including MCID or PASS-scores, the credibility and reliability of the methods used to derive the associated cut-off scores have not been evaluated sufficiently. Moreover, currently there is no instrument or standards available for such credibility assessment [148].

Some recommendations were made regarding the anchor selection process. Revicki et al suggested that anchor questions should be chosen based on evidence that the anchors are relevant to the disease of interest, and have a "non-trivial" association with the specific PROM used in the study [154]. Also, method triangulation has been proposed – the use of multiple anchors to derive cut-off points, and assess whether there is any convergence of the derived results [154, 155]. In the case where method triangulation is implemented and different cut-off points were generated, more importance should be assigned to the thresholds derived from anchors with a closer conceptual link to the outcome measure [154].

Unfortunately, as indicated in the last column of Table 5.1, the relationship between outcome measure and anchor questions of interest are regularly insufficiently/ inappropriately assessed (or completely neglected) [151, 153] [118]. Therefore, it is worth thoroughly reviewing and examining the link between anchor questions and change in PROM scores. The aim of the work in this chapter is to examine the relationship between anchor questions and PROM scores, as well as, the relation-

ship between different anchor questions addressing different dimensions of outcome. Such exploration is the first step to gain some insight into the credibility of cut-off scores commonly utilised in the field of total joint replacement.

The null hypothesis to be tested is:

There is a moderate to strong, positive correlation between the following metrics of post-operative outcome:

- PROM scores and self-reported satisfaction
- PROM scores and self-perceived transition of symptoms
- Self-reported level of satisfaction and transition of symptoms

TKR/THR patients' PROM scores and level of satisfaction

5.2 Methods

5.2.1 Data Source

This study used the same dataset as in Chapter Three – Record Level Knee Replacement and Hip Replacement Data, which were prospectively collected on a national scale through the NHS PROMs program between April 2014 and March 2015. Only patients with a complete set of information (answers to pre- and post-operative OKS questionnaires and anchor questions) were included in the analysis; the final sample size was 46094 and 44015 for TKR and THR, respectively.

5.2.1.1 The anchor questions

Beside the OKS and OHS, the following two anchor questions also feature in the NHS TKR/THR post-operative survey questionnaire.

1. "How are the problems now in the operated knee/hip compared to pre-operation?" Please select one of the following : 'much better', 'a little better', 'about the same', 'a little worse' and 'much worse'.

2. "How would you describe the results of your operation?" Please select one of the following: 'excellent', 'very good', 'good', 'fair' and 'poor' [123].

The first question addresses patients' self-perceived symptomatic change (in the affected joint), which is also known as "transition of symptoms". The second question measures patients' level of satisfaction.

5.2.2 Statistical Analysis

All data were analysed and graphs constructed using R 3.3.0 GUI programming language, except for the three-dimensional bar graphs, which were constructed using MatLab version R2015b (8.6.0). Correlations between level of satisfaction and level of change in symptoms at affected joint was assessed using Kendall's tau. Correlations between satisfaction level and change in PROM scores (difference between baseline and 6-month post-operative OKS/OHS) was calculated using Spearman's rank correlation coefficient. Correlations between change in OKS/OHS and patients' global rating of change (a.k.a. transition) were calculated using the same method. The mean change in score with 95% confidence interval (CI) for patients in each satisfaction/transition group were computed and compared. All analyses were performed for TKR and THR separately.

5.3 Results

5.3.1 Distribution of transition and satisfaction in TKR and THR patients

The distribution of "transition of symptoms" in TKR and THR patients at six-month post-operation is shown in Figure 5.1. The vast majority, 86% of THR and 73% of TKR patients, described the symptom level at their affected joint as "much better" compared to pre-operation.

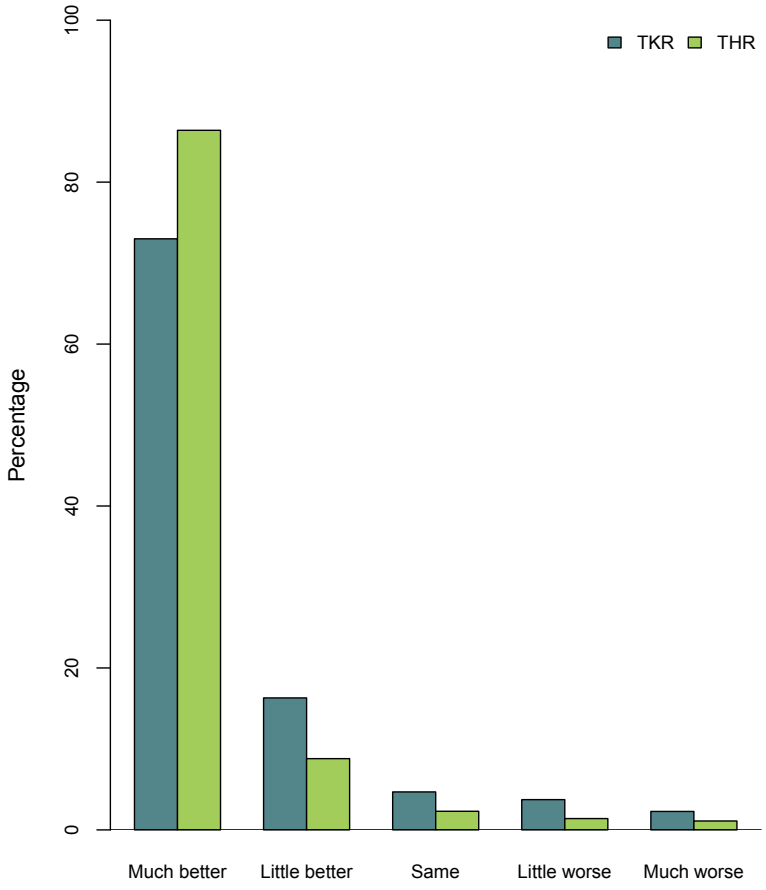


Figure 5.1: Comparing the distribution of TKR vs THR patients' transition rating (i.e. self-perceived change in symptoms) at six-month post operation. The predominant rating choice of both TKR and THR patients is "much better".

The distribution of satisfaction in TKR and THR patients at six-month post-

operation is shown in Figure 5.2. Comparing to transition, satisfaction ratings are more evenly spread across different categories. For THR patients, 42% rated their surgery as "Excellent"; the proportion is significantly lower, only 26%, amongst TKR patients. Over 90% of THR patients selected a favourable category ("excellent", "very good" or "good"), and approximately 85% of TKR patients rated their satisfaction favourably.

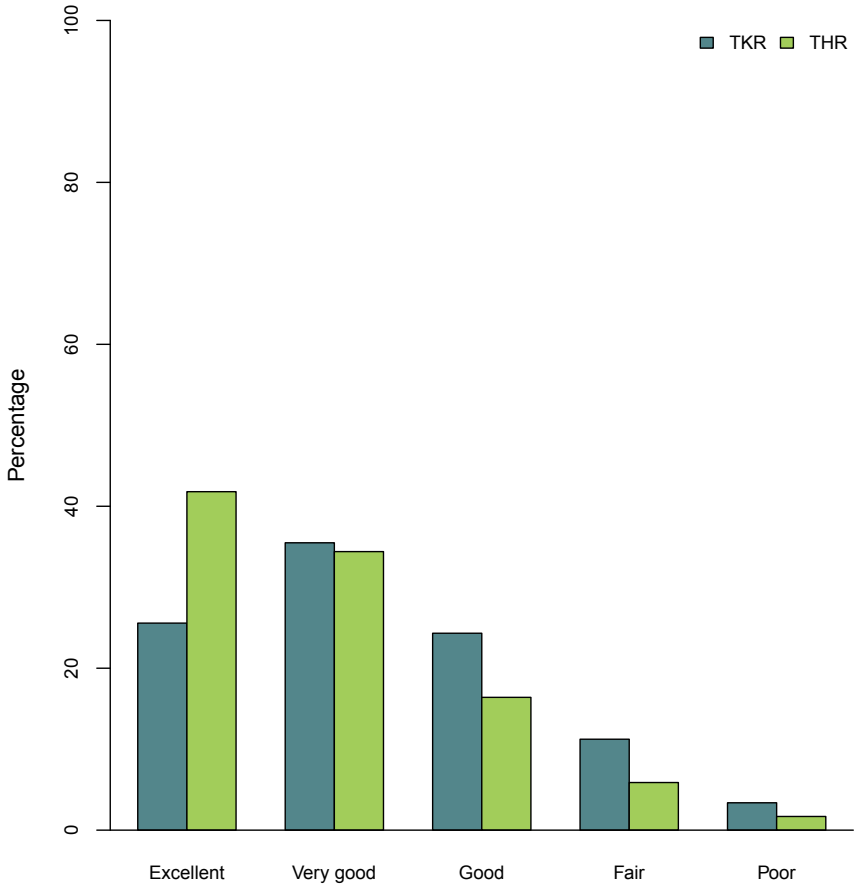


Figure 5.2: Comparison of the distribution of TKR vs THR patients' satisfaction at six-month post operation. Over 40% THR patients rated their operation as "excellent", and approximately a quarter of TKR patients did so. The majority of both TKR and THR patients selected one of the three favourable response choices ("excellent", "very good" and "good").

5.3.2 Further partitioning of patients based on responses to anchor questions

Having examined the distribution of transition and satisfaction independently, patients were then separated into 25 transition-satisfaction combined divisions. Within each division, patients answered both anchor questions identically. The divisions representing desirable outcomes are located in the top-left corner (circled in red). The majority of divisions (19/25) represent a non-ideal outcome – dissatisfied, unimproved, or both. The number of TKR and THR patients in each division are shown in Tables 5.2 and 5.3, respectively. In terms of patient distribution, over 80% are concentrated in the six desirable outcome divisions, with the rest disbursing over other non-ideal outcomes.

		Transition					
		Much better	A little better	Same	A little worse	Much worse	Total
Satisfaction	Excellent	11610	131	13	20	14	11788
	V. good	15277	939	89	39	17	16361
	Good	6335	4051	563	229	30	11208
	Fair	396	2290	1280	1013	198	5177
	Poor	19	110	216	424	791	1560
	Total	33637	7521	2161	1725	1050	46094

Table 5.2: According to their responses to both anchor questions, all TKR patients were separated into 25 divisions. Patients within each division rated satisfaction and change in symptom level identically. Almost three quarters of patients selected one of the three divisions in the top left corner – much better-excellent, much better-very good, and much better-good. 17% of TKR patients experienced either lack of improvement or lack of satisfaction or neither.

		Transition					
		Much better	A little better	Same	A little worse	Much worse	Total
Satisfaction	Excellent	18248	94	29	7	3	18381
	V. good	14487	548	75	14	11	15135
	Good	4892	1928	290	86	21	7217
	Fair	366	1201	519	360	94	2540
	Poor	46	85	96	151	364	742
	Total	38039	3856	1009	618	493	44015

Table 5.3: According to their responses to both anchor questions, all THR patients were separated into 25 divisions. Patients within each division rated satisfaction and change in symptom level identically. 85% of total THR patients selected one of the three divisions in the top left corner – much better-excellent, much better-very good ,and much better-good. 9% patients experienced either lack of improvement or lack of satisfaction or neither.

5.3.3 Relationship between transition and satisfaction in TKR and THR patients

To examine the relationship between post-operative satisfaction and transition, patients were first stratified based on their responses to the transition anchor question. Next, distribution of satisfaction within each strata were computed and shown in Tables 5.4 and 5.5, as well as, Figures 5.3 and 5.4. In addition, overall correlation between satisfaction and perceived transition were calculated using Kendall's tau, $\tau = 0.61$ for TKR patients, and $\tau = 0.49$ for THR patients.

		Transition				
		Much better	A little better	Same	A little worse	Much worse
Satisfaction(%)	Excellent	34.5	1.7	0.6	1.2	1.3
	V. good	**45.4**	12.5	4.1	2.3	1.6
	Good	18.8	**53.9**	26.1	13.3	2.9
	Fair	1.2	30.4	**59.2**	**58.7**	18.9
	Poor	0.0	1.5	10.0	24.6	**75.3**
	Total	100.0	100.0	100.0	100.0	100.0

Table 5.4: Comparing the distribution of satisfaction level for TKR patients within each transition category. For TKR patients who experienced an improvement ("much better" or " a little better") in symptoms after operation, the predominant response choices for satisfaction were "very good" and "good". Level of satisfaction drops as patients' experience regarding change in symptoms becomes less favourable – between 70% and 90% of those who experienced no improvement or worsening in symptoms (i.e. "same", "a little worse" or "much worse") felt unsatisfied (i.e. "fair" or "poor"). Overall relationship between transition and satisfaction in TKR was $\tau = 0.61$ (n=46094).

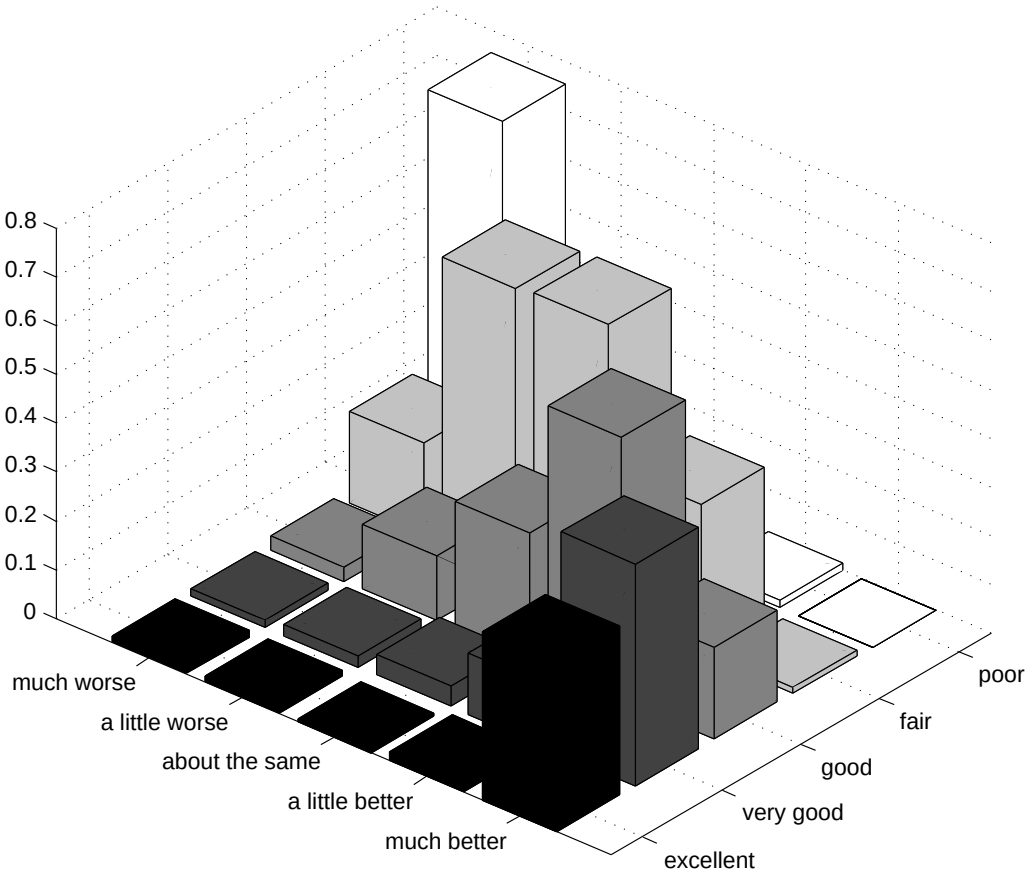


Figure 5.3: A 3-D graphical representation of data in Table 5.4 , showing the proportion (represented by height of the prism) of different satisfaction levels (represented by different shades) within each transition category for TKR patients.

In Figure 5.3, the height of the prism reflects the proportion of patients in each satisfaction level **within each transition category**. For example, the tall, white prism in the back corner shows that amongst those patients who felt "much worse" after TKR (which only accounts for 2% of total patient population), over 75% rated their satisfaction as "poor". The darkest prism in the front corner indicates that amongst those feeling "much better" post-TKR, a third of them also rated their surgery as "excellent". The satisfaction profiles for patients who rated their symptoms as " about the same" or "a little worse" exhibit significant similarities with each other.

		Transition				
		Much better	A little better	Same	A little worse	Much worse
Satisfaction(%)	Excellent	**48.0**	2.4	2.9	1.1	0.6
	V. good	38.1	14.2	7.4	2.3	2.2
	Good	12.9	**50.0**	2.9	13.9	4.3
	Fair	1.0	31.1	**51.4**	**58.3**	19.1
	Poor	0.0	2.2	9.5	24.4	**73.8**
	Total	100.0	100.0	100.0	100.0	100.0

Table 5.5: Comparing the distribution of satisfaction level for THR patients within each transition category. Patients who experienced an improvement in symptoms – feeling "much better" or " a little better" post- operatively, predominantly rated their satisfaction as "excellent" and "good", respectively. Level of satisfaction drops as patients' experience regarding post-operative symptom level worsens – between 60% and 90% of those whose symptoms did not improve or declined (i.e. "same", "a little worse" or "much worse") felt unsatisfied (i.e. "fair" or "poor") after THR. Overall correlation between transition and satisfaction in THR was $\tau = 0.49$ (n=44015).

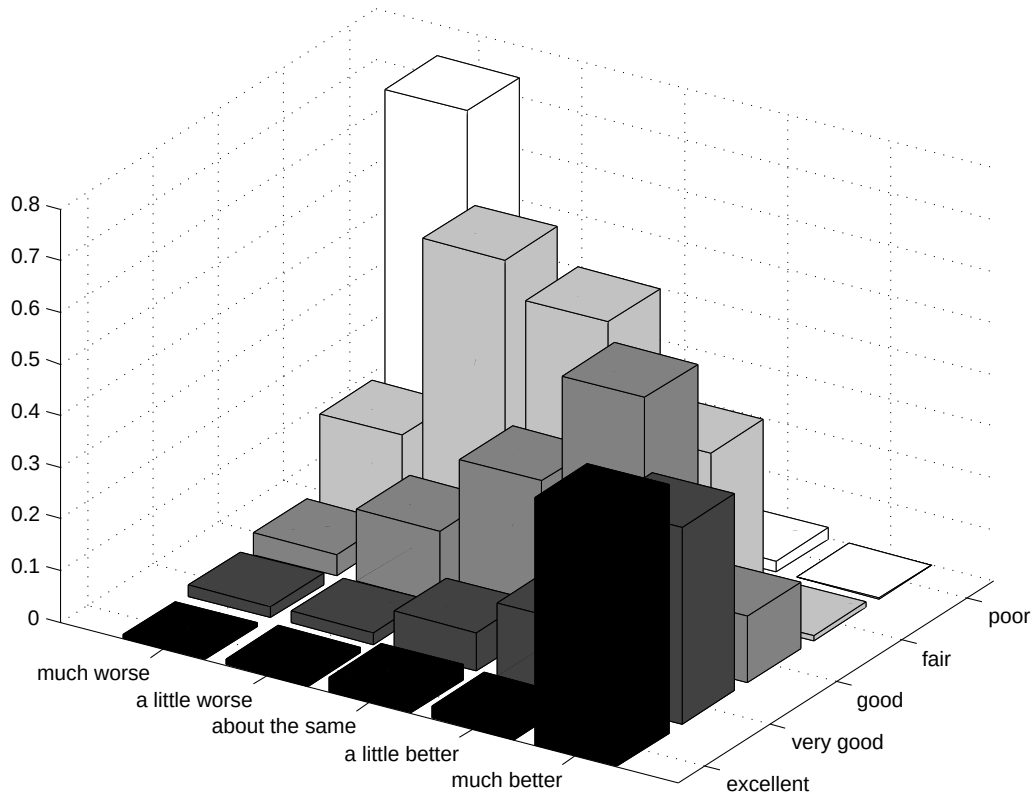


Figure 5.4: A 3-D graphical representation of data in table 5.5 , showing the proportion (represented by height of the prism) of different satisfaction levels (represented by different shades, with the darkest being the most satisfied) within each transition category for THR patients.

Similar to Figure 5.3, in Figure 5.4, the height of the prism reflects the proportion of THR patients in each satisfaction level **within each transition category**. For example, the tall, white prism in the back corner shows that amongst those patients who felt "much worse" following THR (which only accounts for 1% of total THR patients), over 70% rated their satisfaction as "poor". The darkest prism in the front corner indicates that amongst those feeling "much better", just under 50% of them also rated their surgery as "excellent". This result suggests that even the most positive response to the transition question does not guarantee a high level of satisfaction. Furthermore, the satisfaction profiles for THR patients who rated their symptoms as "about the same" or "a little worse" exhibit significant similarities with each other. On the other hand, the satisfaction profile of patients who felt "much

better" was considerably better than that of patients feeling "a little better".

5.3.4 Relationship between PROM scores and transition

There was a significant difference in the change in OKS/OHS (Δ OKS/Δ OHS) according to patients' responses to the transition anchor question. A greater increase in OKS/OHS was obtained by those with better ratings of transition (Table 5.6). The correlation between change in OKS/OHS and transition rating is illustrated in Figure 5.5, and the correlation coefficient was calculated using Spearman's rank correlation, $\rho = 0.57$ in TKR and $\rho = 0.45$ in THR.

		Transition				
		Much better	A little better	Same	A little worse	Much worse
Δ OKS	Mean	19.35	10.02	5.15	2.07	-2.69
	95% CI	(19.27, 19.44)	(9.86, 10.19)	(4.85, 5.44)	(1.74, 2.40)	(-3.13, -2.26)
Δ OHS	Mean	22.97	11.42	5.95	2.41	-2.39
	95% CI	(22.88, 23.06)	(11.17, 11.67)	(5.42, 6.49)	(1.81, 3.01)	(-3.13, -1.66)

Table 5.6: The change in OKS and OHS from pre- to post-operation for TKR and THR patients with different transition ratings. The greatest difference in Δ OKS/OHS is seen between patients who felt "much better" and those feeling " a little better". Also, patients who felt " the same" or "a little worse" also exhibited a positive score change.

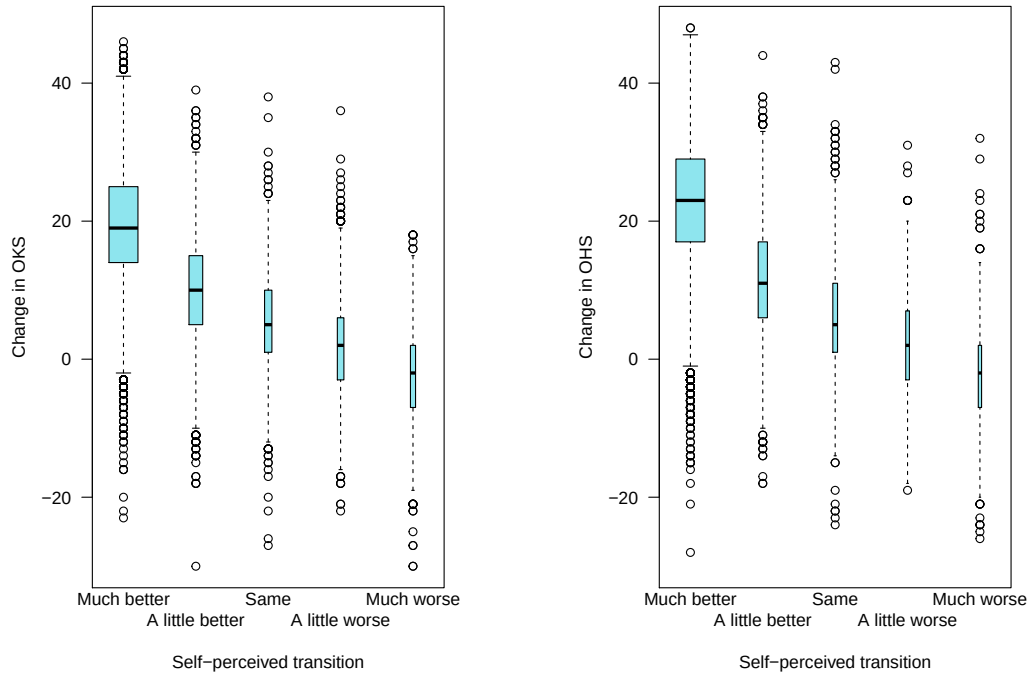


Figure 5.5: Box-whisker plots illustrating how Δ OKS/OHS change according to patients' transition rating. Box width is proportional to patient number in its respective transition category. There is a low to moderate, positive correlation between patients' OKS/OHS and their transition rating, with the relationship being slightly stronger in TKR (TKR: $\rho = 0.57$, THR: $\rho = 0.45$).

5.3.5 Relationship between PROM scores and satisfaction

There was a significant difference in the change in OKS/OHS according to patients' responses to the satisfaction anchor question. A greater increase in OKS/OHS was obtained by those feeling more satisfied with the operation (Table 5.7). The correlation between change in OKS/OHS and level of satisfaction is illustrated in Figure 5.6, and the correlation coefficient was calculated using Spearman's rank correlation, $\rho = 0.55$ in TKR and $\rho = 0.44$ in THR.

		Satisfaction				
		Excellent	Very good	Good	Fair	Poor
Δ OKS	Mean	22.05	18.26	13.23	6.18	-0.54
	95% CI	(21.91, 22.20)	(18.14, 18.38)	(13.08, 13.38)	(5.97, 6.39)	(-0.92, -0.16)
Δ OHS	Mean	25.13	21.42	15.96	8.68	0.89
	95% CI	(25.00, 25.25)	(21.28, 21.56)	(15.75, 16.17)	(8.33, 9.02)	(0.20, 1.58)

Table 5.7: The change in OKS and OHS for TKR and THR patients with different levels of satisfaction.

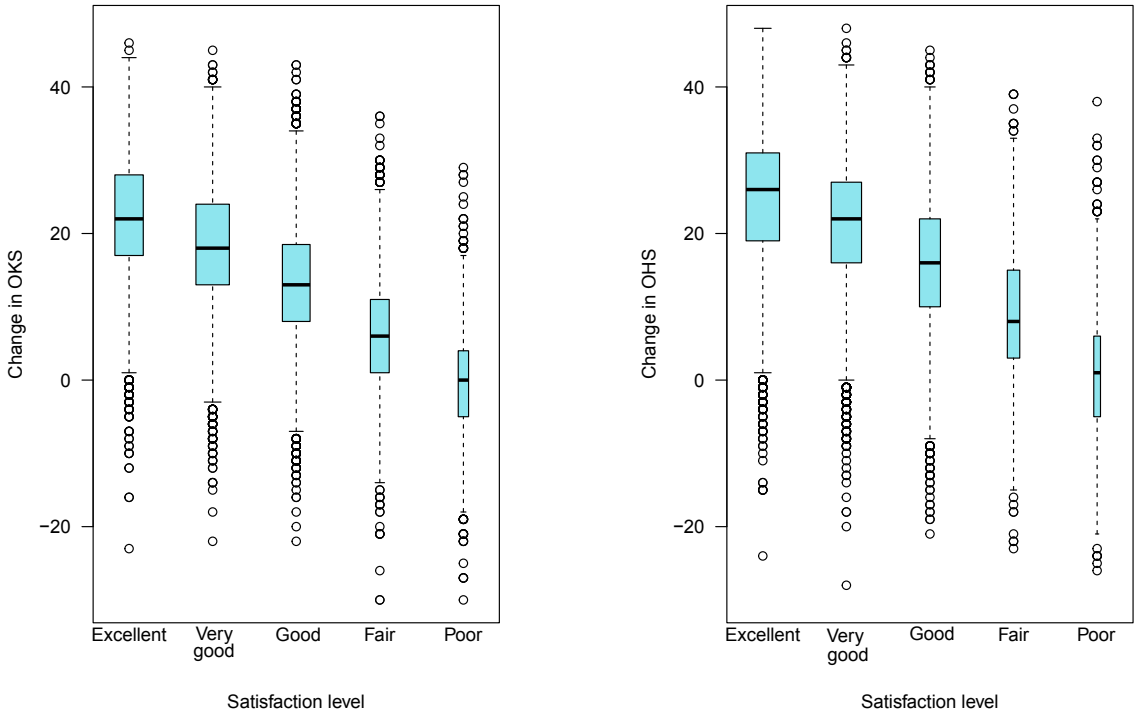


Figure 5.6: Box-whisker plots illustrating how Δ OKS/OHS (y-axis) change according to patients' satisfaction level (x-axis). The width of each box is proportional to patient number in its respective category. Results from analysis of variance indicates that there is a moderate, positive, linear correlation between patients' change in PROM scores and their satisfaction (TKR: $\rho = 0.55$, THR: $\rho = 0.44$)

5.4 Discussion

5.4.1 Main findings

This is the first study to use a large national registry data set to thoroughly examine the relationship between TKR and THR patients' PROM scores and their responses to anchor questions, as well as, the level of concordance between responses to anchor questions addressing different aspects of the outcome (e.g. transition and satisfaction).

5.4.1.1 Distribution of transition and satisfaction

Approximately 85% of THR and 75% TKR patients reported a considerable improvement in symptoms (feeling "much better") post-operation. Regarding satisfaction, 90% THR and 85% TKR patients rated their level of satisfaction positively ("excellent", "very good" or "good"). Our results are partially in agreement with previous findings from Clement et al, who assessed patients' satisfaction with their pain relief and functional improvement separately. Clement and colleagues used a much smaller (n=505) dataset from a single study center and found that 87% of total patients are satisfied with pain and 75% satisfied with functional gain [150]. The large size (n>45,000) and multi-centred nature of our data confer more credibility to these results, but this 10% difference in satisfaction between pain and function underscores the importance of assessing post-operative satisfaction with specific domains in addition to evaluating the overall satisfaction.

5.4.1.2 Relationship between transition and satisfaction

Further analysis of post-operative outcome involved partitioning of patients based on their responses to both transition and satisfaction questions – 80% TKR and 90% THR patients reported feeling both improved and satisfied after operation. Upon closer examination, there is considerable overlap between the satisfaction profile of patients who rated the change in their symptom as "the same" and those feeling " a little better" – for both groups, over 50% described their satisfaction level as "fair" (Table 5.4, Table 5.5). However, there was noticeable difference between the satisfaction profile of patients who rated transition of symptoms as " much better" and those who felt " a little better". For both TKR and THR, 80% of those feeling much better at 6-month post-operation rated satisfaction as "very good" or "excellent", under 20% of patients feeling a little better gave such high ratings regarding satisfaction (Table 5.4, Table 5.5). Such discrepancy in satisfaction level casts doubt on some of the current TKR/THR outcome reporting approaches, which combine the "much better" and " a little better" patients into one single category, and regard them together as having obtained a sufficient response [102] [118] [153]. An alternative grading system that takes both transition and satisfaction into consideration simultaneously may well be more appropriate.

Kendall's rank correlation (Kendall's tau) was used to assess the overall level of association between patients' satisfaction and transition. Kendall's tau is to some extent similar to Spearman's rank correlation; it is a non-parametric measure of correlation between two ranked variables. What is unique about Kendall's tau is that it represents a proportion – the proportion of concordant pairs minus the proportion of discordant pairs. The computed correlation coefficient is a direct reflection of the degree of concordance between the two variables – the greater the number of concordant pairs, the stronger the correlation coefficient will be, and vice versa

[156]. Our results show that the correlation between satisfaction and transition for TKR and THR are $\tau = 0.61$ and $\tau = 0.49$, respectively. According to the equation $\tau = (C - D)/(C + D)$, where C represents the number of concordant pairs and D the number of discordant pairs [156], the proportion of discordant pairs in our data can be calculated – 20% for TKR, and 25% for THR. These results suggest that the transition and satisfaction anchor questions, despite being related, are conceptually distinctive, and should be used in a combination in clinical practice.

5.4.1.3 Relationship between PROM scores and transition/satisfaction

The correlation between patients' PROM scores and satisfaction/transition was assessed using Spearman's rank correlation (a.k.a. Spearman's rho), which is the same as the more well-known Pearson's r, except calculated on the ranks of the observations. Spearman's correlation is suitable for data containing variables with a non-parametric distribution (e.g. skewed or ordinal), and has the advantage of not specifically assessing linear correlation, but any type of association in geneneral [157]. The correlation between change in PROM score and satisfaction was found to be 0.55 and 0.44 for TKR and THR patients, respectively; correlation between change in PROM score and transition was 0.57 for TKR and 0.45 for THR patients. According to the current "rules of thumb" for interpreting the size of a correlation coefficient, there is a low, positive correlation (ρ between 0.3 and 0.5) between the change in PROM score and anchor questions for THR patients, and a moderate, positive correlation (ρ between 0.5 and 0.7) for TKR patients [158]. Lastly, as ρ^2 is the percentage of variability of the data that is explained by the association between the two variables [157], in this case approximately 20% of the variation in Δ OHS and 30% of the variation in Δ OKS can be explained by the change in patients' satisfaction level or transition rating. Our results indicate that although it is intuitive to link greater improvement in PROM score with greater change in symptoms or

higher level of satisfaction, the strength of the relationship between these metrics is actually weaker than expected.

In addition, it is worth noting that both TKR and THR patients who reported feeling "the same" after operation exhibited a positive change in their OKS/OHS. A plausible explanation is that the OKS/OHS are highly sensitive, condition-specific measures specifically designed for patients undergoing knee/hip replacement, whereas the transition question, on the other hand, addresses symptoms and health perception on a global level. Minute functional gain and amelioration in pain would most likely be detected by the OKS/OHS, and shown as a positive score change. However, such minor improvement may not be adequate to change patients' perception on their overall symptom level/ health status.

Previously, as part of a study to define the patient acceptable symptom state (PASS) score, Keurentjes et al also examined the correlation between OKS/OHS and satisfaction . Satisfaction was measured on a numerical rating scale (NRS) with 0 being extremely dissatisfied and 10 being extremely satisfied, and the correlation coefficient was found to be 0.52 with OHS and 0.64 with OKS [152]. Comparing to our results, the correlation obtained by Keurentjes is marginally stronger, such discrepancy is likely due to the difference in how satisfaction was measured — one used a numerical scale with eleven possible levels, and one used a five-point likert scale. After assessing the correlation, Keurentjes and colleagues dichotomised each patients' satisfaction score ($\geqslant 5$ for "acceptable level" and < 5 for "unacceptable level" of satisfaction), and then used ROC curve analysis to identify the PASS cut-off score. This raises the question whether it was appropriate to present satisfaction data in the eleven-level NRS format when calculating correlation between satisfaction and PROM score. The 0 to 10 NRS was merely a way of collecting satisfaction data, which were then converted into a binary variable – satisfied ($\geqslant 5$) and unsatisfied (< 5). The strength of the association between anchor question and PROM score

directly impacts on the credibility of the final cut-off score. When assessing this relationship, it may be more appropriate to use anchoring data in the final format adapted for cut-off derivation as opposed to the format in which raw data were collected.

5.4.2 Conclusion

As explained in Chapter Three, a meaningful way of describing the outcome following TKR or THR is to report the proportion of patients who have achieved a "sufficient response". Three components are essential in defining "sufficient outcome" – the outcome measure, the response criterion, and the criterion-associated cut-off score. Among the two general approaches for estimating cut-off score (distribution-based vs anchor-based), the anchor-based methods are, in general, regarded as more valid [148]. Currently, a variety of anchor questions and derivation methods have been used, and a range of cut-off scores have been identified. However, there is a lack of standards/guidelines regarding what type of anchor questions are appropriate for which patient-reported outcome measures.

Using multiple approaches, this is the first in-depth investigation into the relationship between two widely-used anchor questions and the OKS/OHS. Our results indicate that there is a significant, but only low to moderate, non-linear correlation between these different metrics of assessment for TKR/THR outcome. Whilst anchors are important and have substantial merits for other aspects of analysis, findings from this study raises issues about identifying cut-off points solely based on anchor questions. What is the minimum correlation between anchor and PROM score required for the purpose of cut-off score derivation? Lastly, in comparison to PROM scores, anchor questions are often more easily understood, and precisely address aspects of the outcome that are fundamentally important to patients. There

may be a case for using a combination of anchor questions directly to define a "sufficient response", instead of using them for the sole purpose of finding cut-off scores.

5.5 Link to next chapter

This chapter is a thorough investigation into the relationship between patients' PROM scores following TKR/THR and their responses to the two most widely used anchor questions in the NHS PROM program – transition of symptoms and satisfaction. The results showed a weak to moderate positive correlation between these different metrics. Our findings suggest that the current practice of using anchor questions for the purpose of deriving criterion-associated cut-off points, may not be the most appropriate.

Patients' responses to anchor questions usually directly reflect their experiences during the operation. The next chapter will explore an alternative way of defining "sufficient response" following TKR/THR – basing response criteria and cut-off point on anchor question responses.

Chapter 6

A novel method of assessing self-reported outcome following TKR/THR – the Oxford patient reported success index

6.1 Introduction

As explained in earlier chapters, numerous condition-specific and generic PROMs have been developed over the years to measure patient-reported outcomes following total joint replacement operation [145]. Although the majority of clinicians would agree that a better PROM score usually reflects better outcome (i.e. higher satisfaction, less pain and better function), a score remains meaningless until given an appropriate context. Chapter Three of this thesis describes in detail a conceptual framework that facilitates the analysis and reporting of post-operative outcome to be done in a more meaningful way – in terms of the proportion of patients achieving a "sufficient response" (i.e. a successful outcome). Currently, the framework of "sufficient response" is built upon three components – the multi-item outcome measure (e.g. OKS, WOMAC), the response criterion (e.g. the MCID, PASS, MIC) and a cut-off point. Previous results have shown that when the same outcome measure is used, the proportion of sufficient response can vary widely depending on the associated cut-off score for a specific response criterion.

Despite the differences in stringency and the conceptual foundation they are developed upon, these common response criteria can all be brought together by extended application of the signal detection theory (SDT) from the field of psychology. SDT breaks down all decision-making processes into two parts – I) acquiring information and II) selecting a criterion [159]. In the case of evaluating post-operative outcome, the decision required is to ascertain whether a patient has obtained a sufficient response or not. The first component – information acquired, is the patient's response to surgery presented in the format of OKS/OHS – a higher post-operative score or a greater improvement in score is usually interpreted as a better outcome, but not always the case [115, 149]. The second component, the criterion, such as the MCID, MIC, PASS (with their associated cut-off scores) are essentially different

criteria located at different points along the OKS/OHS response axis. It is considered a "sufficient response" when patients' acquired information (i.e. OKS/OHS) is greater than the pre-selected criterion location, and an "insufficient response" when patients' response falls short of the criterion mark.

In the above SDT scenario, the information acquired is always patients' PROM scores – OKS/OHS. The only difference is the selected location of criterion – the cut-off scores associated with MCID, MIC, and PASS differ from one another. Even though all of these categorical cut-off points have previously been regarded as valid response criteria [115, 102], findings from Chapter Three exposed the limitations of assessing outcome using thresholding methods, and the work in Chapter Five demonstrated that there is only a weak to moderate correlation between patients' OKS/OHS scores and their self-perceived change in symptoms and satisfaction. Moreover, the time and resources it takes to complete and score multi-item PROM questionnaires, and then categorise the scores according to pre-selected criterion is substantial. The ideal measure has been described as the "the clinical equivalent of the Swiss army knife", something short that can fit various circumstances [57]. The shortest outcome measures are global questions that enquire about patients' overall health status. To date, global questions have only played an auxiliary role in outcome assessment – as an anchor to help derive the cut-off scores, even though findings from previous studies have shown that global questions provide information consistent with that obtained using more elaborate questionnaires.

The purpose of this chapter is to provide a solution to overcome the problems with using categorical cut-off scores to measure outcome by introducing the Oxford Patient Reported Success Index (OxPReSi), a new outcome assessment method for identifying patients into three categories – (those who have achieved)" substantial benefit", "low benefit "and "no benefit". The development of OxPReSi follows the SDT model – addressing the task of deciding whether patients have achieved a

successful outcome into two parts, information acquisition and criterion selection. Unlike other existing methods, in OxPReSi, the information acquired is patients' responses to two anchor questions, which address transition of symptoms and satisfaction with surgery.

6.2 Methods

6.2.1 Data Source

The study used the Patient Reported Outcome Measures (PROMs) Record Level Knee Replacement and Hip Replacement Data, which are prospectively collected on a national scale through the NHS. These data were accessed via the Health and Social Care Information Centre (HSCIC) official website (Now NHS Digital). Full PROMs data reports and detailed guidelines on methodology is also available online [107]. Between April 2012 to March 2015, the data of over 13,500 TKR patients and 12,900 THR patients were recorded. Patients were invited to complete a pre-operative (baseline) OKS or OHS Questionnaire, and then a follow-up questionnaire at 6-month post-operation. Beside the OKS/OHS, patients were also asked to answer two additional anchor questions addressing transition and satisfaction. For this study, only patients with a complete set of information (answers to pre- and post-operative OKS questionnaires and both anchor questions) were included in the analysis. Sample sizes included in final data analysis for TKR were 39,720 (2012-13), 44,114 (2013-14), 43,115 (2014-15). As for THR, a total of 37,208 (2012-13), 42,871 (2013-14), 40,917 (2014-15) patient data were analysed each year.

6.2.2 Measuring subjective perception of TKR/THR outcome

6.2.2.1 The anchor questions

Beside the OKS and OHS, the following two anchor questions were incorporated into the National PROMs collection; One addresses patients' level of satisfaction with treatment outcome and the other enquires about perceived "transition" in symptoms from pre-intervention to post-intervention. The questions and instructions are given below:

1. "How are the problems now in the operated knee/hip compared to pre-operation?" Please select one of the following : 'much better', 'a little better', 'about the same', 'a little worse' and 'much worse'.
2. "Rate your overall level of satisfaction with the operation by selecting one of the following: 'excellent', 'very good', 'good', 'fair' and 'poor' [123].

6.2.2.2 Analysis

Following the "sufficient response" framework first introduced in Chapter Three, in Oxford patient-reported success index (OxPReSi), the outcome measures include the transition and satisfaction anchor questions. The transition anchor question was used to assess whether patients' symptoms have sufficiently improved, and the response choice " a little better" was set as the cut-off for the minimal perceptible improvement (MPI) criterion. As a result, those who answered "much better" or " a little better" to the transition anchor question were considered to be **improvers**, and those who selected "about the same", "a little worse" or "much worse" were considered to be **non-improvers**. With regard to level of satisfaction, the response choice "good" was selected as the cut-off for the minimal acceptable satisfaction (MAS)

criterion. Consequently, patients who rated their surgery outcome as "excellent", "very good" or "good" were categorised as **satisfied**, and those who chose "fair" or "poor" as **unsatisfied**. Our choices of criteria were justified following methods used to calculate MCID and minimal clinically important improvement (MCII) in the literature [102, 128, 153, 160]. Table 6.1 below illustrates how the development of the OxPReSi follows the same conceptual framework as other threshold-based methods, but differs in each component.

Outcome measure	Response criteria	Criterion-specific cut-off
Transition anchor question	Minimal perceptible improvement (MPI)	(Symptoms are) "a little better"
Satisfaction anchor question	Minimal acceptable satisfaction (MAS)	(Rate satisfaction as) "Good"
OKS/OHS	Minimal important change (MIC)	ΔOKS $\geqslant$ 9.22, ΔOHS $\geqslant$ 10.63 [102]
	Patient acceptable symptom state score (PASS) – ΔOKS/OHS	ΔOKS $\geqslant$ 11, ΔOHS $\geqslant$ 14 [115]

Table 6.1: Comparing OxPReSi with existing threshold-based methods of measuring post-operative outcome, in the context of the "sufficient response" framework.

With the Oxford Patient-reported Success Index (OxPReSi), the assessment of patient-reported outcome following knee or hip replacement requires two steps. As shown in Figure 6.1, in step one, patients are dichotomised into two classes, "improved" and "unimproved", based on their answers to the transition question. In step two, each class is further divided according to their self-reported satisfaction level. To achieve a "high benefit", patients must provide a positive answer to both anchor questions (i.e. feeling improved AND satisfied). If a patient reports feeling better but unsatisfied, the operation is considered to have delivered a "low benefit".

If patients do not feel better, regardless of their level of satisfaction, these patients are considered to have gained "no benefit". Priority was given to the transition of symptom anchor question over satisfaction because as previously reported, patients' satisfaction can be affected by a multitude of factors, such as socio-economic status [161], cultural background [162], mental status [163] and hospital experience [129], none of which is directly related to the TKR or THR operation itself. The transition of symptoms anchor question specifically addresses change in the health status of the affected joint as a result of TKR/THR.

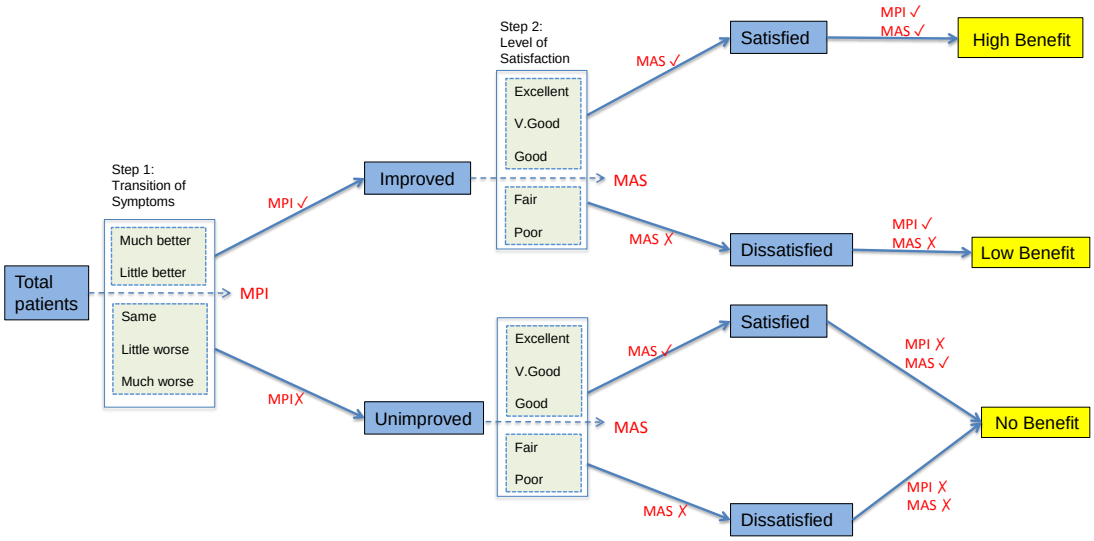


Figure 6.1: The Oxford Patient Reported Success Index – an anchor-based index for assessing the outcome of TKR/THR (MPI = minimal perceptible improvement, MAS = minimal acceptable satisfaction).

Application of this success index results in three categorical outcome groups – **high benefit, low benefit** and **no benefit**. It is then possible to characterize each group and validate the classification using the average change in score (OKS or

OHS) and average effect size of each outcome category. Effect size was calculated as the difference between mean pre- and post-operative score, divided by the pooled standard deviation [164]. All data were analysed using R 3.3.0 GUI programming language. Pre-operative characteristics were calculated using proportions and means as appropriate.

6.3 Results

6.3.1 Application of the new method to the national PROMs database

After applying the Oxford patient reported success index (OxPReSi) to the 2012 to 2015 NHS PROMs data on TKR patients, results are as follows: 82 - 83% have met both the minimal perceptible improvement (MPI) and the minimal acceptable satisfaction (MAS) cut-off level, and therefore were classified as having obtained a "high benefit". 6-7% reached the MPI but were not satisfied with their surgery, and therefore were classified as obtained a "low benefit". Approximately 11% patients failed to acquire the MPI, as a result, they were categorised into the "no benefit" group. In THR patients, 91% were classified as having a high benefit having met both the MPI and MAS cut-off criteria; 4% met only the MPI criterion but not the MAS, and therefore fell into the "low benefit" category; 5% failed to reach the MPI criterion cut-off, and therefore were classified as having "no benefit". The annual proportional statistics for TKR and THR databases are shown in Table 6.2 and Table 6.3, respectively. The results demonstrated a high level of temporal stability – the distribution of TKR and THR patients across three outcome categories from 2012 to 2015 was highly consistent on a yearly basis. Comparing the outcome of TKR with that of THR, there is a higher proportion of THR patients achieving a

high benefit (i.e. sufficient response) than TKR patients; such comparison is shown in Figure 6.2.

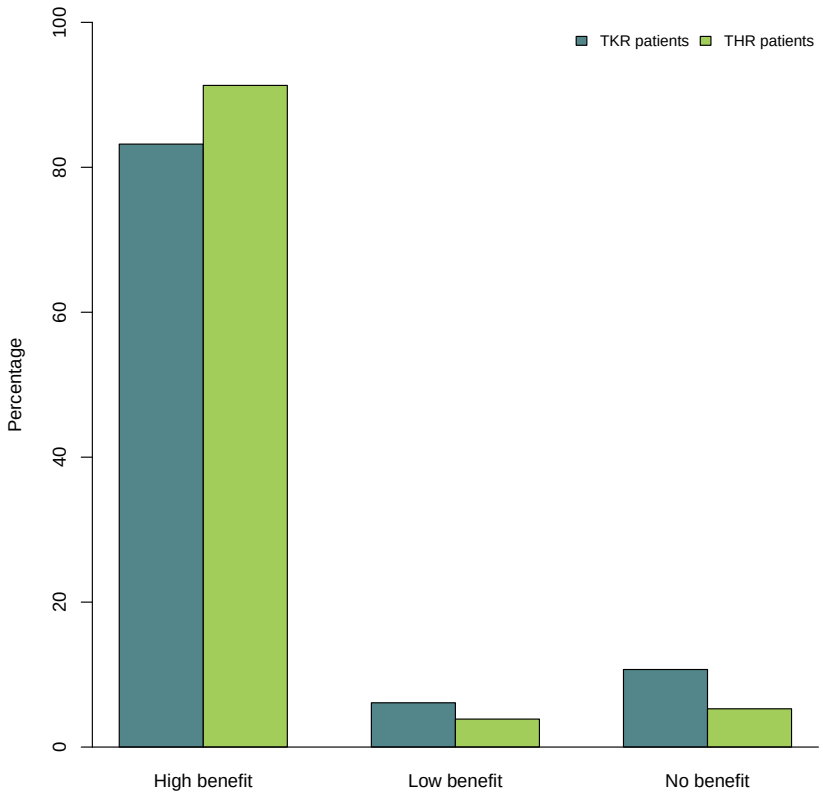


Figure 6.2: The proportion of TKR patients versus THR patients in each outcome group (2014-2015 data). There is a greater proportion of THR patients obtaining a "high benefit" compared to TKR patients.

6.3.2 Association between OxPReSi and the OKS/OHS

Having identified the proportion of TKR/THR patients with high benefit, low benefit and no benefit, this section presents the effect size and change in score on the OKS/OHS for patients in each outcome category. Individuals in the "high benefit" outcome group had a very large improvement (18.5 points for TKR and 22.3 for THR), with a large effect size (2.4 and 2.8 for TKR and THR respectively). The

Δ OKS/OHS and effect size for patients in the "low benefit" category are almost exactly half of the "high benefit" group. Patients in the "no benefit" group, who reported feeling unimproved and unsatisfied post-operation, showed a small but positive change in OKS/OHS and an effect size of approximately 0.3. Furthermore, the size of the interquartile range of Δ OKS/OHS for each outcome group is between 9 and 13 points, with very little overlap between groups. Lastly, similar to the proportion data, our results showed a high level of temporal stability – the mean group-level Δ OKS/OHS from 2012-2015 for each outcome category was highly consistent on a yearly basis. Detailed results are shown in Table 6.2 and Table 6.3.

Outcome	Year	Proportion in %	Δ OKS	IQR	Effect Size
High benefit	2012-13	81.5(81.1,81.9)	18.6(18.5,18.7)	13 to 24	2.37(2.35,2.39)
	2013-14	82.7(82.3,83.0)	18.5(18.4,18.6)	13 to 24	2.41(2.39,2.41)
	2014-15	83.2(82.8,83.4)	18.3(18.2,18.4)	13 to 24	2.40(2.28,2.42)
Low benefit	2012-13	6.9 (6.6,7.1)	9.1 (8.8,9.4)	4 to 14	1.21 (1.15,1.27)
	2013-14	6.2(6.0,6.5)	9.5(9.2,9.8)	5 to 14	1.27(1.21,1.33)
	2014-15	6.1(5.9,6.3)	9.0(8.7,9.3)	4 to 14	1.19 (1.13,1.25)
No benefit	2012-13	11.6(11.3,11.9)	2.2(2.0,2.4)	-3 to 7	0.26(0.22,0.30)
	2013-14	11.1(10.8,11.4)	2.4(2.2,2.6)	-2 to 7	0.30(0.26,0.34)
	2014-15	10.7(10.4,11.0)	2.4(2.2,2.6)	-2 to 7	0.29(0.25,0.33)

Table 6.2: Proportion, mean Δ OKS, interquartile range (IQR) of Δ OKS and effect size for TKR patients in each outcome category (classified using OxPreSi) from 2012 to 2015.

Outcome	Year	Proportion in %	Δ OHS	IQR	Effect Size
High benefit	2012-13	90.4(90.1, 90.7)	22.2(22.1,22.3)	16 to 29	2.78(2.76,2.81)
	2013-14	90.5(90.3,90.8)	22.3 (22.2,22.4)	16 to 29	2.78(2.76,2.81)
	2014-15	91.3 (91.0,91.6)	22.3(22.2,22.4)	16 to 29	2.86 (2.84,2.88)
Low benefit	2012-13	4.0 (3.8, 4.2)	11.2 (10.8, 11.6)	6 to 17	1.34 (1.26, 1.42)
	2013-14	4.1 (3.9, 4.3)	11.7 (11.3, 12.0)	6 to 17	1.42 (1.35, 1.49)
	2014-15	3.9 (3.7, 4.0)	11.6 (11.2, 12.0)	6 to 17	1.42 (1.48, 1.33)
No benefit	2012-13	5.6 (5.3, 5.8)	2.9 (2.5, 3.3)	-2 to 9	0.29 (0.24 to 0.36)
	2013-14	5.3 (5.1, 5.6)	3.1 (2.7,3.5)	-2 to 8	0.32 (0.26, 0.38)
	2014-15	4.8 (4.6, 5.0)	3.0 (2.6, 3.4)	-3 to 8	0.31 (0.25, 0.37)

Table 6.3: Proportion, mean Δ OHS, interquartile range (IQR) of Δ OHS and effect size for THR patients in each outcome category (separated using OxPreSi) from 2012 to 2015.

The data also allows comparison between operation type. The mean change in OKS for TKR patients (2014-2015) in each outcome category were compared with the mean change in OHS for THR patients (2014-2015) in the corresponding group, results are presented in Figure 6.3. Not only is there a higher proportion of THR patients than TKR in the "high benefit" category, the average change in score within each category is also higher amongst THR patients. For the "high benefit" group, the difference between the mean ΔOHS and ΔOKS exceeds 4 points, and is greater than the minimal important difference (MID). This gap between THR and TKR patients is diminished in the "low benefit" and "no benefit" groups. Figure 6.3 illustrates this phenomenon.

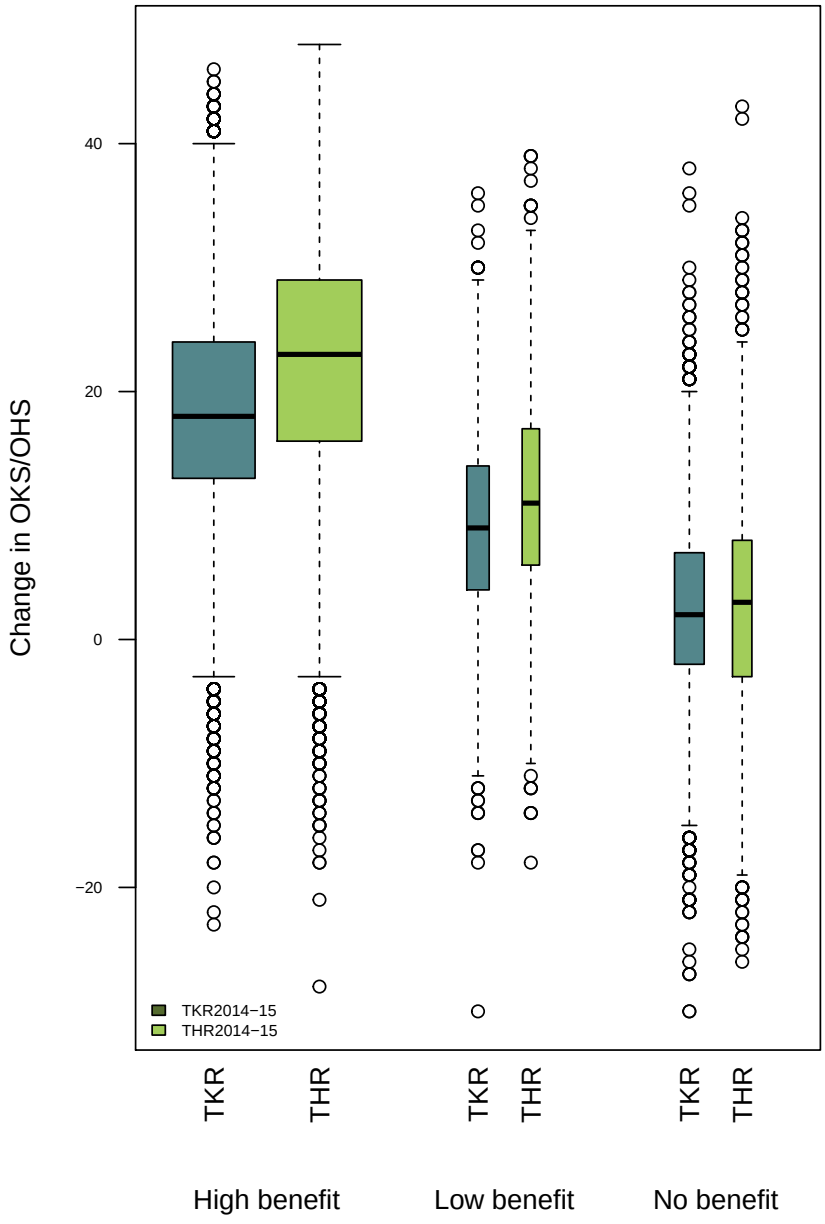


Figure 6.3: Comparison of the change in OKS of each outcome category for TKR patients with the change in OHS for THR patients in corresponding outcome group (2014-2015). The width of the box is proportional to the number of patients. (TKR–n=43,115 THR–n=40,917)

6.4 Discussion

6.4.1 Main findings

This chapter introduces a new method of assessing outcome following TKR/THR – the Oxford patient reported success index (OxPReSI). Unlike other traditional methods, which use patients' scores on a particular multi-item questionnaire as the main source of information for outcome assessment, the OxPReSi is based on patients' responses to the transition and satisfaction anchor questions. Application of this index divides patients into three *categorical* outcome groups:

- High benefit – having met both minimal perceptible improvement (MPI) and minimal acceptable satisfaction (MAS) criteria, i.e. feeling improved and satisfied.
- Low benefit – having met the MPI only, i.e. feeling improved but unsatisfied.
- No benefit – failed to achieve the MPI, i.e. feeling no improvement in symptom level [1].

Regarding the distribution of patients across the three outcome groups, for TKR patients, 83% achieved high benefit, 6% achieved low benefit and 11% experienced no benefit. As for THR patients, as high as 91% achieved high benefit, 4% had low benefit and 5% experienced no benefit. Our results indicate that the proportion of patients experiencing successful outcome is significantly higher amongst THR patients comparing to TKR. This trend is consistent with findings from previous studies that have used traditional methods [109]. Lastly, results from Chapter Three, where four widely-accepted, threshold-based criteria were used to determine the

[1] The proportion of patients who responded that they felt unimproved but still satisfied with their operation was under 0.5%. This contradictory response from a very small proportion of patients was most likely due to clerical error or misunderstanding. It was decided to merge these patients with the "no benefit" group.

proportion of THR and TKR patients experiencing a "sufficient response", demonstrated that depending on the criteria used, the prevalence of good outcome ranged between 72% and 90% for TKR patients, between 77% and 95% for THR patients. The percentage values obtained in this study fall within those ranges.

Following the traditional definition of a "sufficient response", which is built upon the multi-item outcome measure (e.g. OKS, WOMAC), the response criterion and its associated cut-off point, patients are separated into two categories only – those who have achieved a sufficient response and those who failed. With the OxPReSi, a third category, the "low benefit" group, has been created. The addition of the "low benefit" group is valuable, as it is likely to represent those patients with a borderline successful/failure outcome – the operation is beneficial enough for them to recognise that their affected joints have improved compared to pre-operation, but not to the extent that it would make them feel happy/satisfied.

6.4.2 Validity of the OxPReSi

The traditional way of assessing the criterion validity of a new scale of interest is to compute its correlation with other established measures (a.k.a the criterion measure) of the trait, ideally a "gold standard" which has already been validated and accepted. Criterion validity is further divided into *concurrent* and *predictive* validation. With concurrent validation, both the new measurement of interest and the criterion measure are administered at the same time. With regard to predictive validation, the criterion measure will not be given until some time later in the future [12].

The widely accepted OKS/OHS present patients with a numerical outcome between 0 and 48, but the Oxford Patient Reported Success Index produces a three-level categorical outcome. This difference in the type of outcome data produced by these

two metrics makes it difficult to compute the "correlation" in the conventional sense. Therefore, an alternative method was used, namely the exploration of the group-level Δ OKS and effect size for patients in each OxCom outcome category

For both TKR and THR patients, individuals in the "high benefit" outcome group obtained a substantial increase in OKS/OHS accompanied with a large effect size. Patients in the "low benefit" category had a mean increase in OKS and OHS of 9 and 11 points respectively, which would be a borderline case when judged using the minimal important change criterion (ΔOKS $\geqslant$9.22, ΔOHS $\geqslant$10.63) identified previously [102]. Patients in the "no benefit" group (i.e. feeling neither improved nor satisfied), showed a small but positive change in OKS/OHS and a small effect size of approximately 0.3. Such phenomenon, where patients expressing negative perceptions of their surgery but at the same time showing a positive change in PROM scores, have been reported previously [102, 128]. The OKS/OHS are highly sensitive, condition-specific measures, whereas the transition and satisfaction anchor questions focus on patients' overall perception of the surgery and health state. Patients in "no benefit" category might have had minuscule improvement in function and pain, but the amount of change falls short of the MPI and MAS cut-offs. Lastly, our results showed that Δ OKS/OHS from 2012 to 2015 for each outcome category was highly consistent on a yearly basis.

The OxPReSi prioritises patient experience over PROM scores, assessing post-operative outcome purely by patients' satisfaction and self-perceived global change in symptoms. Despite not taking OKS/OHS into consideration during the assessment/classification phase, our results indicate that OxPReSi is closely associated with the OKS/OHS. However, the wide range of ΔOKS/ΔOHS within each outcome group suggests that there are still some fundamental differences between anchor-based metric and score-based metric.

6.4.3 OxPReSi versus threshold-based methods

As discussed early in Chapter Three, several cut-off scores associated with various response criteria have been calculated using different methods, but each one has its own merits and drawbacks, and a gold-standard is still lacking. Like other threshold score-based methods for measuring post-operative outcome, the OxPReSi describes a "sufficient response" in terms of three components – the outcome measure, the response criteria and a criterion-related cut-off (see Table 6.1 for details). The outcome measure refers to the *instrument* used to measure treatment outcome in the specific study; the fundamental difference between OxPReSi and other threshold-based methods is the *format* of the selected instrument.

All existing threshold-based methods rely on lengthy multi-item questionnaires, which can be time- and resource-consuming to administer and score. OxPReSi, on the other hand, relies on two global questions addressing different aspects of the outcome – change in symptoms and satisfaction with operation. Patients' responses to the transition and satisfaction questions have conveniently been incorporated into the National PROMs data collection. Results from multiple studies consistently demonstrated that single, global questions can provide information that is consistent with much longer questionnaires [9].

Another problem with threshold-based methods is that categorical cut-off scores are most likely to be baseline-dependent. It has been shown that patients with more severe pre-op pain and functional impairment need to improve their pre-op score by a greater amount in order to reach a satisfactory state in comparison to those with less severe symptoms pre-operatively [115]. Using a single "universal" threshold score for everyone can be misleading; adjusting threshold-values according to patients' baseline pain and function level may be a more sensible approach. Currently, most criterion-associated cut-off scores developed to-date are *not* baseline-specific.

6.4.4 Limitations

In hindsight, more patient and public involvement (PPI) sessions should have been conducted while making decisions on what the final OxPReSi framework should look like. A hallmark of patient-reported outcome measure is that patient opinions are sought throughout the development process.

Furthermore, the most debatable element of the OxPReSi is the categorization of those patients who perceived themselves to be "unimproved" but "satisfied"; it would not be unreasonable to consider them as having achieved "low benefit". However, in this study, "transition of symptoms" have been given priority over "satisfaction", and anyone who considered themselves "unimproved" were assumed to have acquired "no benefit" from their knee/hip replacement operation. To test such assumption, a sensitivity analysis combining the patients who are "unimproved and satisfied" into the "low benefit" group would be useful. One way to assess the effect of such change is to calculate the mean OKS of the new "low benefit" group (with the addition of the "unimproved and satisfied" cohort) with that of the original group. Smaller change in the mean OKS of the "low benefit" group suggests lower sensitivity of the OxPReSi framework to this assumption.

6.5 Conclusion

To summarise, in the Oxford patient reported success index, the transition and satisfaction anchor questions are combined using a two-step algorithm, giving priority to transition of symptoms. This new approach – first classifying patients into "high benefit", "low benefit" and "no benefit" groups based on their responses to anchor questions, and then reporting the outcome in terms of patients' distribution across the three categorical groups, is potentially more patient-centred, more accurate and

save significant time and resources than traditional threshold-based methods. The proportion of patients achieving a good outcome using this new method is 83% for TKR and 91% for THR, which is in agreement with results from previous studies evaluating the prevalence of good outcome following total joint replacement.

6.6 Link to next chapter

The next chapter explores whether a modified version of the OxPReSi is suited for measuring outcomes in a randomised controlled trial (RCT) – the Total and Partial Knee Arthroplasty Trial (TOPKAT) designed to compare the clinical effectiveness of partial knee replacement with that of total knee replacement.

Chapter 7

Application of the modified OxPReSi to RCT data – comparing the outcome following partial and total knee replacement

7.1 Introduction

Currently, there is a mixed consensus among surgeons regarding the most appropriate surgical option for patients with late-stage medial-compartmental OA of the knee. Some surgeons firmly believe total knee replacement (TKR) to be the more superior option [165, 68], while some believe that partial knee replacement (PKR) leads to better outcome [166, 167, 168]. The history of partial or unicompartmental knee replacement (PKR) commenced in the early 1950s when McKeever and Elliot first introduced the "segmental approach" in knee arthroplasty [169]. Modern PKR implants started in 1972 after the introduction of Marmor's hemiarthroplasty, and has evolved rapidly in the past three decades.

Proponents of PKR believe that by only replacing the damaged component of the knee, PKR is a less aggressive procedure, and the operated knee functions closer to a natural knee, once recovered. Some studies have also reported PKR to be associated with lower post-operative morbidity and faster recovery [170, 171]. With the prevalence of knee osteoarthritis on the rise, knee replacement is becoming increasingly common. Being able to choose the right intervention for the right patient is the first and also most crucial step in ensuring a satisfactory post-operative outcome. Existing literature on this issue consists of retrospective studies and incomplete comparisons focusing on one specific aspect (e.g. revision rate, complication or alignment) [172, 173, 174] [175]. Studies comparing patients' self-reported health-related quality of life (HRQoL) following PKR versus HRQoL following TKR are especially scarce.

In order to make a fair comparison of these two surgical options regarding patient-reported outcome, two components are essential:

- Valid and reliable data, preferentially from a randomised controlled trial (RCT)

design.

- A comprehensive method for evaluating post-operative outcome.

The Total or Partial Knee Arthroplasty Trial (TOPKAT) is a multicentre, randomised controlled trial specifically designed to compare the clinical and cost effectiveness of PKR vs TKR. In the registered original trial, the primary outcome measure is the Oxford Knee Score (OKS), which is an outcome measure specifically developed to assess pain and functional ability in patients following knee replacement surgery. Despite its proven validity and responsiveness, direct usage of the OKS in TOPKAT may not be ideal. One of the potential limitations is that, assuming a statistically significant difference was found in post-operative score between PKR and TKR patients, any statistical significance cannot be interpreted as clinical difference, and the question regarding clinical effectiveness of PKR and TKR would remain insufficiently answered. The Oxford Patient Reported Success Index (OxPReSi) (introduced in Chapter Six) is a novel measurement tool to assess the outcome of joint replacement that may potentially be more meaningful to patients, and provide greater insight for comparisons of outcome. The OxPReSi was developed using the NHS PROMs data, and prior to this work had not been used to compare treatment effects such as in an RCT setting.

The aim of the work in this chapter was to analyse whether a modified version of the Oxford Patient Reported Success Index (OxPReSi), a metric developed using a population data, can be used to measure and compare outcomes in an RCT, such as the Total and Partial Knee Arthroplasty Trial (TOPKAT). Results (regarding the superiority between PKR and TKR) obtained using OxPReSi were compared with results from the original trial (where OKS was the primary outcome), as well as results obtained using existing threshold-based criteria, in order to gain more insight into the advantages and disadvantages of each metric. Lastly, the OxPReSi

was used to compare the outcome of TOPKAT data (an RCT) with the outcome of the NHS PROM data (a population data).

The null hypothesis to be tested is: Applying the OxPReSi as the outcome measure to TOPKAT data, there is no significant difference between the outcome following TKR compared to the outcome following PKR.

7.2 Methods

7.2.1 Data Source

In this study, one-year follow up data from the Total or Partial Knee Arthroplasty Trial (TOPKAT) was used, which is the first large, multi-centre randomised controlled trial specifically designed to compare the clinical effectiveness of TKR with that of PKR for patients with medial compartment osteoarthritis of the knee. A total of 528 patients (264 in each arm) from 27 sites across the UK were recruited between Jan 2010 and Sept 2013. Sixty three surgeons were involved, and outside randomisation of type, they were free to choose which brand of implant to use. The trial employed a combined equipoise/expertise-based approach. An "equipoise surgeon" is someone with sufficient experience in performing both PKR and TKR. A "TKR expert" or "PKR expert" is a surgeon with sufficient skill level in performing only TKR or PKR, respectively. In TOPKAT, TKR experts and PKR experts worked as pairs to form an "equipoise unit". Patients allocated to the PKR group would be operated either by a surgeon in equipoise, who is capable of delivering both operations or, by a PKR expert out of individual equipoise but forming part of an equipoise unit. Similarly, patients allocated to the TKR group would be operated either by a surgeon in individual equipoise or by a TKR expert from an equipoise

unit. More details regarding the protocol of the trial, including exclusion inclusion criteria, the concept of equipoise and equipoise unit have been published previously [176]. For this study, only patients with a complete set of information at one year (answers to pre- and post-operative OKS questionnaires and anchor questions) were included in the analysis. This reduced the sample size to 478, with 236 TKR and 242 PKR patients.

7.2.2 Outcome Measures

For the original one-year follow-up trial analysis, the OKS was the primary outcome. In addition, a variety of secondary outcome measures were also used to assess patients' function, including activity level, pain, satisfaction, transition of symptoms, general health and complication/re-operation rate. In this study, we extracted only the Oxford Knee Score (OKS) questionnaire and two external anchor questions addressing "transition" and "satisfaction". Details of the OKS can be found in Chapter Two (literature review) of this thesis. The anchor questions in TOPKAT were the same as those used in the NHS PROMs program, but the answer choices are slightly different. Details of the questions and answer choices used in TOPKAT are as follows:

- Transition question: "How are the problems related to your knee now compared with before your knee surgery?"
  - Answers: No problems at all; Much better; Slightly better; No change; Slightly worse; Much worse
- Satisfaction question: "How satisfied are you with your knee?"
  - Answers: Very satisfied; Satisfied; Dissatisfied; Uncertain

The difference regarding the response choices to the transition question is that the

answer " No problem at all" is provided in TOPKAT, but absent in the NHS PROMs data collection. As for the satisfaction question in TOPKAT, the response choices "very satisfied", "satisfied", "dissatisfied" and "uncertain" replaced the original choices (including "excellent", "very good", "good", "fair" and "poor") used in the NHS PROMs data.

Therefore, to enable the Oxford Patient Reported Success Index (OxPReSi) to be used for analysis of TOPKAT data, some minor modifications to the original cut-off points were made. In the new version, to reach the minimal perceptible improvement (MPI) criterion, patients must have replied to the transition anchor questions with "slightly better" or more positive; to reach the minimal acceptable satisfaction (MAS) criterion, patients must have replied to the satisfaction question with the answer "satisfied" or "very satisfied".

The modified version of the two-step OxPReSi is largely the same as the original (Chapter Six). In step one, patients are split into two classes, "improved" and "unimproved", based on their answers to the transition question. In step two, each class is further divided according to their satisfaction level. To achieve a "high benefit" outcome, patients must provide a positive answer to both anchor questions (i.e. feeling improved and satisfied).

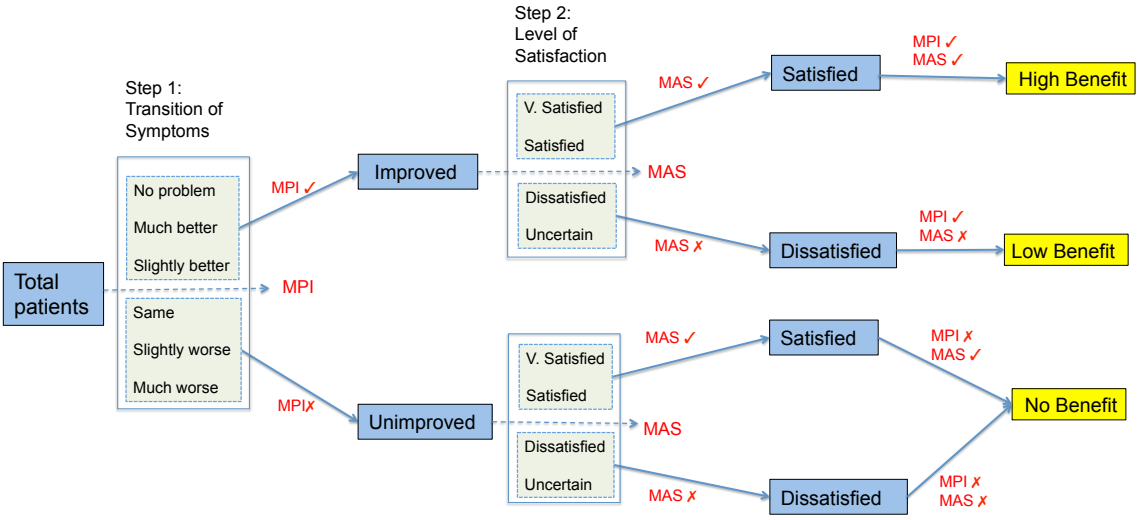


Figure 7.1: The two-step modified Oxford Patient Reported Success Index (MPI = minimal perceptible improvement, MAS = minimal acceptable satisfaction).

The only difference between the modified and the original versions lies in the exact definition of "improved", "unimproved", "satisfied" and "dissatisfied". Table 7.1 provides more details of the distillation of the categories.

	Transition		Satisfaction	
Version	Improved	Unimproved	Satisfied	Dissatisfied
Original	Much better	About the same	Excellent	Fair
	A little better	A little worse	Very good	Poor
		Much worse	Good	
Modified	No problem	No change	Very satisfied	Dissatisfied
	Much better	A little worse	Satisfied	Uncertain
	A little better	Much worse		

Table 7.1: Comparison of the original Oxford Patient Reported Success Index with the modified version

To measure the percentage of PKR and TKR patients achieving a "sufficient response"(in addition to OxPReSi), two other established criteria, the minimal im-

portant change (MIC) and patient acceptable symptom state score (PASS) (shown in Table 7.2) were used in this study. More details on how to define a sufficient response can be found in Chapter Three.

Outcome measure	Response criterion	Criterion-associated cut-off
Transition anchor question	Minimal perceptible improvement (MPI)	(Symptoms are) "slightly better"
Satisfaction anchor question	Minimal acceptable satisfaction (MAS)	Feeling "satisfied" with surgery
OKS	MIC	$\Delta OKS \geqslant 9.22$
	PASS-Δ OKS	$\Delta OKS \geqslant 11$

Table 7.2: The response criteria used in this study – OxPReSi, MIC and PASS score. They each provide a different benchmark for a "sufficient response" to the intervention.

7.2.3 Analysis

R version 3.4.1 programming language was used for all data analyses. The outcome of PKR and TKR patients in terms of the proportion obtaining a "high benefit", "low benefit" and "no benefit" according to the OxPReSi method was calculated. The overall trend in terms of whether PKR or TKR is more superior was compared with one-year results from the original trial, where OKS was the primary outcome measure. The outcome was also assessed using two OKS threshold-based criteria, the MIC and PASS score. The chi-squared test was used to compare the proportion of TKR patients obtaining a "high benefit" assessed using the OxPReSi, or "sufficient response" assessed using MIC/PASS–ΔOKS score, versus the proportion amongst PKR patients in the TOPKAT trial. The level of overlap between patients classified as achieving "high benefit" according to the OxPReSi method and those classified as sufficient responders according to the PASS-ΔOKS score were also investigated

in more detail. Patients with "discordant results", viz. achieving a "high benefit" but failing to meet the PASS-ΔOKS cut-off, or reaching the PASS-ΔOKS cut-off but failing to obtain a "high benefit", were identified, and their pre-operative scores were compared with the trial average using the t-test. Quantile-quantile plot of pre-operative scores was constructed and normal distribution was confirmed. Lastly, the outcome of TKR patients in TOPKAT was compared with the outcome of TKR patients in the NHS PROMs program (2014-2015).

7.3 Results

7.3.1 Application of the modified Oxford patient reported success index (OxPReSi) to TOPKAT data

The distribution of both PKR and TKR patients across the three outcome groups (categorised using the modified OxPReSi method) is shown in Figure 7.2. The percentage values of each group is shown later in Table 7.4. Patients in the "high benefit" group reported feeling improved and satisfied post-operatively. Patients in the "low benefit" group reported feeling improved but dissatisfied. Patients in the "no benefit" group reported feeling unimproved following their PKR/TKR operation. These data allow comparison of outcome between PKR and TKR at one-year follow up.

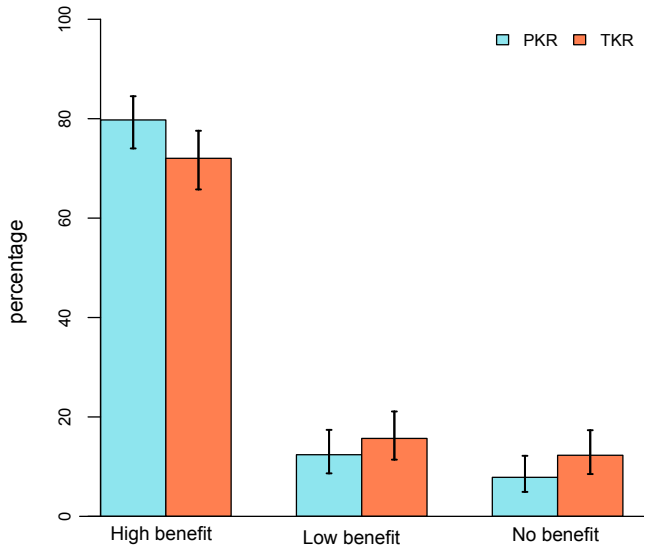


Figure 7.2: Outcome of PKR vs TKR patients from TOPKAT assessed by the Oxford Patient Reported Success Index.

7.3.2 Comparing the outcome of PKR and TKR analysed using the OxPReSi with results from the original trial

In the Total and Partial Knee Arthropasty Trial, the primary measurement of clinical outcome was Oxford Knee Score (OKS). Results obtained using the OxPReSi were compared with the the original one-year trial results, details are shown in Table 7.3. Chi-squared test was used to analyse whether there was a statistically significant difference between the "high benefit", "low benefit" and "no benefit" rates in TKR and PKR patients according to the OxPReSi.

Evaluating Criterion	PKR n=242	TKR n=238	PKR vs TKR Difference	p-value
OKS at 1-yr	36.9 (SD=9.9)	35.1 (SD=10.3)	1.9 (95% CI: 0.2, 3.6)	0.03
OxPReSi method				
–High benefit, % (95% CI)	79.8 (74.0,84.5)	72.0 (65.8,77.6)	7.8 (-0.3, 15.8)	0.06
–Low benefit, % (95% CI)	12.4 (8.7,17.4)	15.7 (11.4,21.1)	-3.3 (-9.9, 3.4)	0.37
–No benefit, % (95% CI)	7.8 (4.9,12.2)	12.3 (8.5,17.3)	-4.5 (-10.2, 1.4)	0.14

Table 7.3: Comparison of the results from the original trial with results obtained using the OxPReSi method.

7.3.3 Comparing the outcome of PKR and TKR analysed using different criteria

In addition to the OxPReSi, the outcome of the PKR and TKR patients in TOP-KAT were also analysed using other threshold-based methods, namely the minimal important change (MIC) and patient acceptable symptom state (PASS) score. The proportion of PKR and TKR patients reaching the cut-off of each criterion, including the OxPReSi, is shown in Table 7.4. Chi-squared test was used to analyse whether there was a statistically significant difference between the "sufficient response" rates in TKR and PKR patients according to different criteria.

Evaluating Criterion	PKR n=242 %(95% CI)	TKR n=238 %(95% CI)	PKS vs TKR Difference	p-value
OxPReSi method				
–High benefit	79.8 (74.0,84.5)	72.0 (65.8,77.6)	7.8 (-0.3, 15.8)	0.06
ΔOKS threshold-based				
-MIC	79.3 (73.6,84.2)	73.7 (67.5,79.1)	5.6 (-2.4, 13.6)	0.18
-PASS-ΔOKS	78.5 (72.7,83.4)	69.0 (62.7,74.8)	9.5 (1.2, 17.7)	0.03

Table 7.4: Proportion of PKR and TKR patients reaching a desired outcome level assessed using different criteria.

Using the OxPReSi, the prevalence of PKR patients achieving "high benefit" is

almost 8% higher than TKR patients, but this difference is not statistically significant. A non-statistically significant 5.6% difference was observed between PKR and TKR patients when the MIC criterion was applied. The difference between PKR and TKR was widened to 9.5% and reached statistical significance when the PASS-ΔOKS criterion was used.

The relationship between patients classified as "high benefit" using the OxPReSi and those meeting the PASS score cut-off of ΔOKS ⩾11 is shown in Table 7.5. Of the 478 patients, 363 (75.9%) achieved "high benefit", 115 did not. Amongst the "high benefit" achievers, 88.2% also met the PASS score criterion by improving OKS by at least 11 points, but 11.8% failed to do so. Amongst the 115 patients classified as having "low/no benefit", the majority also failed to reach the PASS score cut-off, but 28.7% did obtain a change in OKS of 11 points or above. Although there is considerable overlap, patients achieving "high benefit" according to the OxPReSi are not necessarily the same as those reaching the ΔOKS ⩾11 PASS cut-off score.

		PASS (ΔOKS ⩾11)		
		Yes (n,%)	No (n,%)	Total (n,%)
OxPReSi	High benefit	320 (88.2%)	43 (11.8%)	363 (100.0%)
	Low/no benefit	33 (28.7%)	82 (71.3%)	115 (100%)

Table 7.5: Proportion of patients succeeded and failed to meet the PASS score cut-off amongst those classified as having "high benefit" (top row) and those classified as having "low/no benefit" (bottom row).

There is evidence suggesting that TKR patients with lower pre-operative score require a greater absolute change in score, and patients with high pre-operative score require a lower change in score, in order to achieve post-operative satisfaction [115]. To confirm this finding, the mean pre-operative scores for patients with two types of discordant outcomes – I) "high benefit" (i.e. feeling improved and satisfied) but ΔOKS <11 and, II) "low/no benefit" (i.e. feeling unimproved or dissatisfied) but

ΔOKS $\geqslant$11, were calculated and compared to the average pre-operative score of the trial using the t-test. Results are shown in Table 7.6 – the pre-operative score of patients with type I discordance is significantly higher than the trial average, whereas the pre-operative score of those with type II discordance is significantly lower than the trial average. Therefore, data in Table 7.6 strengthens the evidence regarding the baseline-dependency nature of cut-off scores.

Type of discordance	OxPReSi	PASS (ΔOKS $\geqslant$11)	Baseline OKS	t-test p-value
Type I (n=43)	High benefit	No	24.8 (21.9, 27.8)	0.0004
Type II (n=33)	Low/no benefit	Yes	15.5 (13.4, 17.6)	0.002
All patients (n=478)	-	-	19.1 (18.5, 19.7)	-

Table 7.6: Comparison of the pre-operative score of patients having conflicting outcomes (according to the PASS score and OxPReSi criterion) with the average pre-operative score of the trial.

7.3.4 Comparing the outcome of TOPKAT-TKR patients with results from the NHS PROM data

The outcome (assessed using the MIC, PASS and OxPReSi methods) of the patients from TOPKAT receiving TKR was compared with the outcome of the 2014-15 NHS PROM data, results are shown in Table 7.7. There is no significant difference between the proportion of patients achieving "sufficient response" judged by the MIC and PASS criteria, but the prevalence of "high benefit" is significantly higher amongst the NHS-PROM patients than in TOPKAT-TKR patients.

		Data Source		χ^2-test
		TOPKAT-TKR (n=236)	NHS PROM (n=46094)	p-value
Criterion	OxPReSi	72.0%	83.2%	< 0.001
	MIC	73.7%	75.2%	0.60
	PASS	69.0%	72.1%	0.31

Table 7.7: Comparison of the outcome of TOPKAT-TKR and NHS PROMs data in terms of the proportion of patients achieving a "sufficient response" according to various criteria.

7.4 Discussion

7.4.1 Analysis of the TOPKAT data using the modified Oxford Patient Reported Success Index (OxPReSi)

This study investigated whether the Oxford patient reported success index (OxPReSi), a metric developed using a population data (collected nationally through the NHS PROM program), can be adapted to measure outcomes in randomised cotrolled trials (RCT) such as the Total and Partial Knee Arthroplasty Trial (TOPKAT). As explained in Chapter Six, the OxPReSi classifies patients into broad categorical groups by using patients' responses to the transition and satisfaction anchor questions as the main source of information for outcome assessment. The results showed that with minor modification, the OxPReSi can be applied to RCT data. Comparing to the original version, the modified OxPReSi differs in terms of the response choices that constitute "improvement" and "satisfaction" (refer to Table 7.1 for details). Applying the modified OxPReSi to the patients from TOPKAT allocated to TKR, 72% achieved a high benefit, 16% had low benefit and 12% had no benefit. As for those patients from the trial allocated to PKR, 80% obtained high benefit, whilst 12% and 8% had low and no benefit, respectively. The 8% difference

between the proportion of patients experiencing a "sufficient response" (i.e. 'high benefit" in this case) in PKR and TKR was notable, but not found to be statistically significant according to the chi-squared test (p =0.06).

7.4.2 Comparison of the results using OxPReSi with the original trial results

The results obtained using the OxPReSi were consistent with the results from the original trial (where mean OKS at one-year follow-up was the primary outcome measure) in terms of the overall trend. The outcome of PKR is superior to TKR – higher post-operative OKS and greater proportion of "high benefit" was found in PKR patients compared to TKR. However, the results appeared disparate considering the *magnitude* of difference between PKR and TKR – a statistically significant p-value (p=0.03) was observed in the original trial, but not in the current study using the OxPReSi as the main outcome measure (p=0.06).

The p-value refers to the probability of getting a difference equal to or larger than the one observed, if the null hypothesis is true. A lower p-value suggests stronger evidence against the null hypothesis. The insignificant p-value obtained in this study is most likely a result of the trial being under-powered. Power analysis is the name given to the process for determining the sample size for a research study. Sample size is calculated in a different way for continuous or categorical variables. The outcome variable in the current study, the OxPReSi, is a categorical variable. Post-operative OKS, the outcome variable of the original RCT, is a continuous variable, and power analysis was performed accordingly. Since continuous variable is statistically more powerful than categorical variable, a sample size that has been calculated to be large enough to detect a "true" effect when a continuous outcome variable is used, may become inadequate when outcome measure is substituted with a categorical

variable.

In addition, currently there may be an over-emphasis on the importance of p-value. In the case of comparing patient-reported outcome of PKR and TKR, the p-value may be less important than the actual value of difference and the overall trend. In clinical practice, comparing to the original trial, results from this study (using the OxPReSi as the primary outcome measure) offers a higher level of granularity and is possibly more informative for patients. The majority of patients will not be able to comprehend the difference between a post-operative OKS of 34 and 36 despite the statistical difference. In comparison, an 8% difference between patients receiving "high benefit" is much easier to interpret.

7.4.3 Analysis of the TOPKAT data using the MIC and PASS criteria.

To confirm the results obtained using the OxPReSi, existing threshold-based methods, including the minimal important change (MIC) and patient acceptable symptom state (PASS) scores were also utilised in the analysis and comparison of PKR and TKR outcome. When the PASS score (ΔOKS $\geq$ 11) criterion was selected, a nearly 10%, statistically significant (p =0.03) difference was observed between the "sufficient response" rate in PKR and TKR patients. However, when the MIC criterion (ΔOKS $\geq$ 9.22) was used, the proportional difference between PKR and TKR narrowed to under 6% and was not statistically significant (p =0.18).

Analysis results obtained using the modified version of OxPReSi showed the same pattern as results obtained using existing threshold score-based methods – there is a consistently higher proportion of PKR patients achieving the desired outcome than TKR patients. Depending on the method of assessment, the difference in "sufficient response" rate between PKR and TKR ranged between 5 and 10%. The superiority

of PKR appeared most prominent when the PASS-ΔOKS criterion was selected. Regarding the importance of the p-value, in March 2016, the American Statistical Association (ASA) issued a warning advising researchers to "avoid drawing scientific conclusions or making policy decisions based on P values alone" [177]. As shown in this study, with the same dataset, different levels of statistical significance can be obtained using different cut-offs. Each individual value, the overall pattern and trend are just as important, if not more so than the p-value, in drawing conclusions.

7.4.4 Relationship between the modified OxPReSi and PASS score

The patient acceptable symptom state (PASS) score, refers to "the value of the OKS beyond which patients consider themselves in a satisfactory state". It shares common ground with the Oxford patient reported success index (OxPReSi) in the sense that it is anchored to the patients' personal experience. However, placing these two methods in the conceptual framework that describes a "sufficient response", the PASS and OxPReSi clearly differ from each other in all three components – the outcome measure, the response criterion and the cut-offs (shown in Table 7.2). Despite the fundamental differences, the results have shown that, the proportion of "sufficient responders" in the TOPKAT data, calculated using modified OxPReSi and PASS, differ by less than 3%.

The relationship between patients classified as having a "high benefit" using the OxPReSi and those meeting the PASS score cut-off of ΔOKS $\geqslant$11 was examined, and the results indicate that, despite considerable overlap, patients achieving "high benefit" are not necessarily the same as those reaching the ΔOKS $\geqslant$11 PASS cut-off score. Approximately 16% of patients had *discordant* results, of which 9% had type I discordance – "high benefit" with a change in OKS of less than 11 points. These

patients had a significantly higher than average pre-operative score. 7% had type II discordance –"low/no benefit" with a change in OKS of 11 points or over. The pre-operative score of these patients were significantly lower than average.

Our results confirmed previous findings regarding the baseline-dependency of threshold values. Patients with more severe symptoms (i.e. lower pre-operative score) need a larger absolute increase in PROM score in order to reach a sufficiently satisfactory state than patients with less severe symptoms. Therefore, it is important to take patients' baseline clinical characteristics into consideration when deriving cut-off scores for the purpose of separating patients that have achieved a "sufficient response" from those who have not. Currently, it is a common practice to identify *one* cut-off score for each response criterion, and then apply this single threshold value universally to all patients' data of interest. Such disregard for patients' baseline condition can potentially produce misleading results. The OxPReSi, on the other hand, assesses outcome based on patients' post-operative, self-reported level of satisfaction and perception of change in symptoms only, and does not rely on PROM scores. The threshold levels of OxPReSi, including the MPI and MAS, are therefore applicable across all patients with a range of baseline scores.

7.4.5 Comparing the outcome of RCT data with cohort data

Comparing the outcome of TOPKAT-TKR with the outcome of the 2014-2015 NHS PROM-TKR, there is no significant difference between the proportion of patients achieving "sufficient response", assessed by the MIC and PASS score metrics. However, the prevalence of "high benefit" amongst the NHS-PROM patients is over 11% greater than the TOPKAT-TKR patients (shown in Table 7.7). Such statistically significant discrepancy in post-operative outcome between TOPKAT and PROM data is most likely attributed to the different wording of response choices used in

the original and modified version of the OxPReSi:

- "How satisfied are you with your knee"
  - Original OxPReSi (used in NHS PROMs data) : Excellent, very good, good, fair, poor
  - Modified OxPReSi (used in TOPKAT): Very satisfied, satisfied, uncertain, dissatisfied

The main problem with the response choices given in the NHS PROMs questionnaire is that it is unbalanced – there are three favourable and two unfavourable choices. Rating scales in general should be balanced – presenting equal number of favourable and unfavourable items. The TOPKAT rating scale is balanced [178]. Therefore, from a purely statistical point of view, for NHS PROMs-TKR patients, they have 60% chance of being classified as satisfied, 40% chance of being unsatisfied. In contrast, the probability of being satisfied/unsatisfied for TOPKAT patients stands at 50-50. A severely unbalanced rating scale with response items "perfect", "excellent", "extremely good", "very good", "good", "fair" and "poor" would most likely result in a positive outcome.

7.5 Conclusion

Using first-year data from the Total or Partial Knee Arthroplasty Trial (TOPKAT) as a medium, results from this study demonstrate that with some minor modification, the Oxford patient reported success index (OxPReSi) is a useful potential measure of outcomes in a randomised controlled trial, particularly for patient understanding.

OxPReSi can effectively separate participants in TOPKAT into three clinically

meaningful groups – "high benefit", "low benefit" and "no benefit". The results obtained in this study confirmed the original trial results – the outcome following PKR is superior over TKR. Results obtained using the MIC and PASS response criteria also showed a consistent pattern – PKR outcome is better than TKR in terms of "sufficient response" rate.

Furthermore, this study showed that the benchmark score for post-operative satisfaction is more dependent on baseline level than previously thought. Also, comparing to the average group-level post-operative OKS and traditional threshold-based methods, the OxPReSi has some advantages. It provides more granularity, is easier to interpret, and less sensitive to variations in pre-operative scores.

One caveat about using the OxPReSi method to compare the outcome of two different treatments is that the results of OxPReSi may be sensitive to change in either the wording of the transition/satisfaction questions or the wording of the response choices. Therefore, it is best to ensure that identical wordings are used when making such comparisons.

7.6 Link to the next chapter

Chapter Six and Seven thoroughly explained the development of the Oxford patient reported success index, as well as its application in population and RCT data. The next chapter explores the effect of pre-operative PROM scores on TKR and THR patients' probability of achieving a "high benefit" post-operatively.

Chapter 8

The effect of pre-operative PROM score on post-operative outcome

8.1 Introduction

Chapters Six and Seven explained the development and usage of the Oxford Patient reported success index (OxPreSi), a new way of defining "sufficient response" following TKR/THR – by using an anchor-based criterion that combines minimal perceptible improvement (MPI) and minimal acceptable satisfaction (MAS) in a sequential fashion. This new method can potentially assess post-operative outcome in a way that is comprehensive and more meaningful to patients. Using this new anchor-based assessment tool, 17% of TKR and 9% of THR patients were identified as having failed to achieve a sufficient response post-operatively. In Chapter Three, using four different response criteria, the prevalence of unsuccessful outcome for TKR was found to be in the range of 10% to 28%, and between 5% and 22% for THR. Considering the increasingly high demand for TKR and THR, and the cost of these operations to the health care system, it is important to identify the factors associated with operation failure in order to minimise the prevalence of non-desirable outcome.

A number of factors have been found to impact post-operative outcome, and one of the strongest predictors is pre-operative pain and functional ability. Previous studies indicate that higher pre-operative pain and function scores are associated with better post-operative outcome, but worse pre-operative scores lead to greater magnitude of improvement after surgery comparing to their baseline level [179, 180]. The limitation with these studies is that most of them used PROM scores as a proxy for post-op outcome. These PROM scores may be a fair assessment of pain and function. However, patient-reported outcome is a multi-layered concept as shown in the Wilson-Cleary model (Chapter Two), and therefore assessment by standard PROMs may not be comprehensive enough.

Furthermore, in terms of the timing of surgery, if TKR/THR is performed too early in the course of a patient's history of OA, there is the risk of prosthesis failure and long-term complications related to surgery. This might reduce patient's quality of life to a level that is below their pre-surgery baseline. On the other hand, if surgery is performed too late, the function of a patient's affected joint may have declined so much that the amount of benefit surgery can provide would be compromised [180]. The optimal timing of surgery is still unclear and remains the subject of several research programmes.

With the backdrop of patients' outcome following TKR/THR being influenced by a multitude of demographic and clinical characteristics (e.g. co-morbidities, age, gender, etc), the aim of the work reported in this chapter was to further explore one important factor, namely the effect of pre-operative PROM scores on post-operative outcome in TKR and THR patients. With the rapidly increasing demand for orthopaedic surgeries across the world and limited resources, it is important to ensure that operations are performed on the right people at the right time. The intention was to find the threshold level of pre-operative pain and function that maximises the proportion of patients experiencing successful outcome, viz. the optimal pre-operative condition needed in terms of function and pain for maximum post-operative benefit. The new method of defining "sufficient response" described in Chapter Six, the Oxford Patient Reported Success Index, is used to assess proportion of successful outcome.

The null hypothesis to be tested is:

The proportion of patients with a successful outcome (i.e. "high benefit") following knee or hip replacement operation is not affected by their pre-operative Oxford Knee/Hip Scores.

8.2 Methods

8.2.1 Data source

This study used the Patient Reported Outcome Measures (PROMs) Record Level Knee Replacement and Hip Replacement Data, which are prospectively collected on a national scale through the NHS. A more detailed description of the methodology regarding data collection and data linkage is available online [107]. Before their operation, patients were invited to complete the Knee/Hip Surgery Questionnaire I (Q-I), where patients' pre-operative OKS/OHS and other characteristics including age at operation (40 to 49, 50 to 59, 60 to 69, 70 to 79 and 80+), gender, previous replacement surgery on the same knee, other co-morbidities and, presence of any disability were collected. At 6-month post-operation, patients were invited to complete the Knee/Hip Surgery Questionnaire II (Q-II). In addition to the OKS/OHS, patients were also asked to answer two additional anchor questions addressing transition and satisfaction in the Q-II.

Between April 2014 to March 2015, the data of 47,743 TKR patients and 45,554 THR patients were recorded. For the descriptive study, patients who completed both pre- and post-operative OKS/OHS questionnaires and anchor questions were included in the analysis, which reduced sample size to 46,094 and 44,015 for TKR and THR, respectively. Baseline descriptive statistics of patients whose data were included in the enumerative study were compared with those whose data were excluded, details of this comparison are shown in Tables 3.5 and 3.6 (Chapter Three). For the multiple logistic regression study, sample sizes were further reduced to 40,661 and 38,150 for TKR and THR, respectively. This reduction took place because only data from patients who completed the co-morbidities and other demographic characteristic questions (completers) were used for constructing the multiple logistic regression

model. A comparison of baseline descriptive statistics between the completers and non-completers is shown in Table 8.1.

8.2.2 Statistical method

For the descriptive study, the outcome is an ordinal variable consisting of three possible values – "high benefit", "low benefit" and "no benefit". Patients with "high benefit" have achieved both the minimal perceptible improvement (MPI) and minimal acceptible satisfaction (MAS) – they are satisfied improvers. Patients achieving "low benefit" have only met the MPI criterion – they are unsatisfied improvers. Patients experiencing"no benefit" failed to obtain the MPI, and therefore are considered as non-improvers. Details on how each outcome category is defined can be found in Chapter Six.

R version 3.4.1 programming language was used for all data analyses. Multiple logistic regression was conducted to examine the effect of pre-operative PROM scores (namely the OKS and OHS) on post-operative outcome. To control for confounding variables, most factors that have been reported to have an influence on the outcome of TKR/THR in the literature (including pre-operative scores, age, gender, pre-operative disability, previous surgery status and presence of co-morbidities such as diabetes and heart disease) were included as exposure variables in this study. As for the outcome variable, "low benefit" and "no benefit" were collapsed into one category – the "insufficient response". By doing so, the outcome was converted into a binary variable (which is essential for building a logistic regression model), consisting of "sufficient" (i.e. high benefit) and "insufficient" responses. Collinearity amongst exposure variables was checked by variance inflation factor (VIF), and model fit was assessed by pseudo R value. In addition, a multinomial logistic regression model, using the "low benefit" group as baseline, was also fitted, detailed results are shown

in Appendix D.

Based on their pre-operative scores, patients were divided into twelve subgroups. The score interval and the percentage of patients in each group is shown in Table 8.2. Pre-operative score, which was originally a continuous variable, was transformed into an ordinal variable (with twelve categories). Such categorisation allowed analysis of the effect of each individual score category on post-operative outcome.

8.3 Results

Chapter Three described a more meaningful way of reporting outcome following TKR/THR – in terms of the proportion of patients having achieved a "sufficient response". Chapter Six introduced a new method of defining "sufficient response", which is based on patients' responses to anchor questions. Utilising these new approaches of measuring and reporting outcome, the results in this chapter show how post-operative outcome changes according to patients' baseline OKS/OHS following TKR/THR. The first part of this section presents results from a descriptive study, whilst the second part of this chapter explains the multiple logistic regression study in detail.

8.3.1 Description of study sample

The 2014-2015 knee and hip replacement data from the HSCIC website contained information on 47,743 TKR and 45,554 THR patients. For the multiple logistic regression study, only patients who completed both pre- and post-operative questionnaires were included, reducing sample size to 40,661 and 38,150 for TKR and THR, respectively. Baseline demographic details are described in Table 8.1. There were small differences between patients that did and did not complete the pre- or

post-operative questionnaires. In both TKR and THR patients, the completers were younger with a higher proportion of male patients than the non-completers. Also, the prevalence of diabetes, heart disease and lung disease is approximately 3-4% greater amongst completers than non-completers. However, there were no clinically significant differences in pain and function at baseline as assessed by the OHS/OKS [1], nor for comorbidities including depression, previous surgery and circulation.

[1]The 0.5-point difference in pre-operative score is still statistically significant when assessed using the students' t-test; this statistical significance most likely results from the large sample size.

Variable	TKR		THR	
	Included (n=40661)	Excluded (n=7082)	Included (n= 38150)	Excluded (n= 7404)
Gender, n (%)				
-Male	17183 (42.3)	1503 (37.2)	15176 (39.8)	1491 (35.4)
-Female	23478 (57.7)	2515 (62.8)	22974 (60.2)	2721 (64.6)
-Missing, n	-	3064	-	3192
Age, n (%)				
-Under 60yo	4063 (9.9)	335 (8.3)	5489 (14.4)	530 (12.6)
-60-69yo	14999 (36.9)	1250 (31.1)	12983 (34.0)	1238 (12.6)
-70-79yo	16705 (41.1)	1759 (43.8)	14743 (38.6)	1760 (41.8)
-Over 80yo	4894 (12.0)	674 (16.8)	4935 (13.0)	684 (16.2)
-Missing, n	-	3064	-	3192
Baseline OKS/OHS				
-mean (s.d.)	19.2 (7.7)	18.7 (7.8)	18.0 (8.5)	18.2 (8.5)
-missing, n	-	630	-	534
Depression, n (%)				
-Yes	3647 (9.0)	568(8.0)	3159 (8.3)	548 (7.4)
-No	37014 (91.0)	6514 (92.0)	34991 (91.7)	6856 (92.6)
Previous Sx, n (%)				
-Yes	2886 (7.1)	510 (7.8)	3321 (8.7)	559 (8.1)
-No	37775 (92.9)	6057 (92.2)	34829 (91.3)	6311 (91.9)
Diabetes, n (%)				
-Yes	5497 (13.5)	711 (10.0)	3623 (9.5)	481 (6.5)
-No	35164 (86.5)	6371 (90.0)	34527 (90.5)	6923 (93.5)
Heart disease, n (%)				
-Yes	4271 (10.5)	497 (7.0)	3617 (9.5)	512 (6.9)
-No	36390 (89.5)	6585 (93.0)	34533 (90.5)	6892 (93.1)
Circulation, n (%)				
-Yes	2696 (6.6)	430(6.1)	2068 (5.4)	341 (4.6)
-No	37965 (93.4)	6652 (93.9)	36082 (94.6)	7063 (95.4)
Lung disease, n (%)				
-Yes	3961 (9.7)	460 (6.5)	3270 (8.6)	405 (5.5)
-No	36700 (90.3)	6622 (93.5)	34880 (91.4)	6999 (94.5)
Disability, n (%)				
-Yes	22605 (55.6)	3889 (54.9)	21622 (56.7)	4010 (54.2)
-No	18056 (44.4)	3193 (45.1)	16528 (43.3)	3394 (45.8)

Table 8.1: Descriptive statistics of patients whose data were included (completers) and excluded (non-completers) for logistic regression modelling.

8.3.2 Subgroups based on pre-operative scores

TKR and THR patients were divided into subgroups based on their pre-operative OKS/OHS, details are presented in Table 8.2. Patients in the same subgroups have pre-operative scores within the same 4-point (MCID: ΔOKS/OHS ≈ 4) interval.

Subgroup	Pre-operative score interval	Proportion	
		TKR (%)	THR (%)
1	0-4	1.57	2.71
2	5-8	6.54	9.46
3	9-12	13.12	14.57
4	13-16	17.51	17.63
5	17-20	18.82	17.44
6	21-24	17.69	14.96
7	25-28	12.43	10.93
8	29-32	7.27	6.98
9	33-36	3.54	3.21
10	37-40	1.15	1.39
11	41-44	0.31	0.48
12	45-48	0.04	0.22

Table 8.2: Subgroups based on pre-operative scores.

8.3.3 Enumerative study

The proportion of patients experiencing "high benefit", "low benefit" and "no benefit" within each pre-operative score subgroup were computed, and results are shown in Figures 8.1. For TKR patients, as baseline scores improve from the 0-4 range to the 17-20 range, an increase in the proportion of "high benefit" from 72% to 82%, and a decrease in "no benefit" from 20% to 12% were observed. Outcome distri-

bution becomes relatively stable once pre-operative OKS reaches above the 17-20 range. For THR patients, there is a slight increase in "high benefit" from 83% to 87% as baseline OHS improves from the 0-4 range to the 17-20 range. No other obvious difference in outcome was observed across groups with different baseline levels.

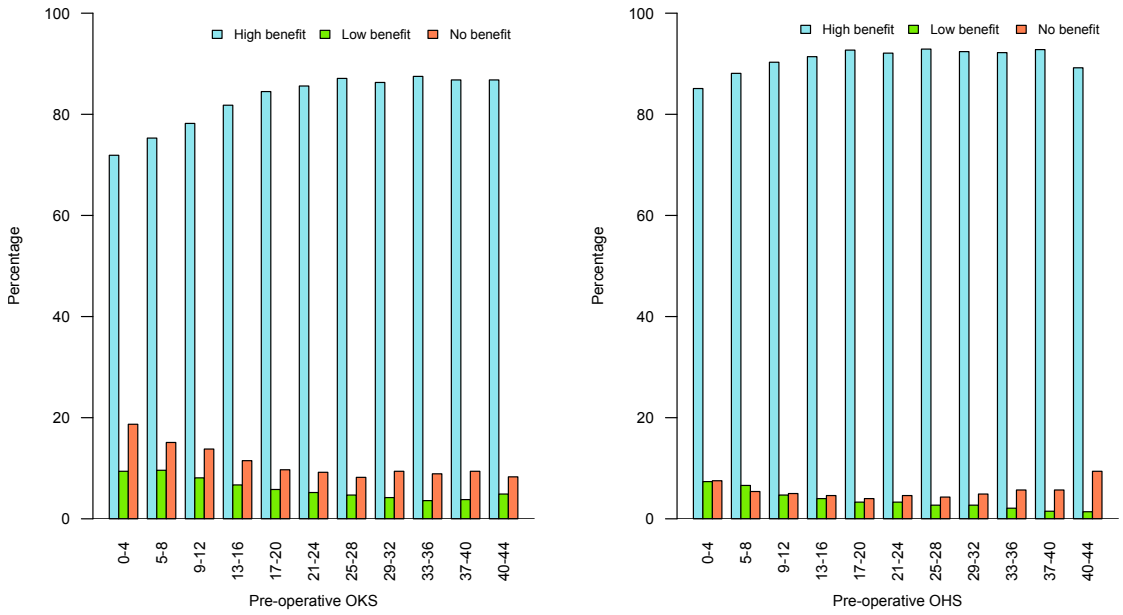


Figure 8.1: The distribution of post-operative outcome (categorised as "high benefit", "low benefit" and "no benefit") in TKR (left) and THR (right) patients according to their pre-operative OKS/OHS.

The difference between TKR and THR patients becomes more apparent upon closer examination of the proportion of "sufficient response" (i.e. high benefit) in patients with different ranges of pre-operative scores. Figure 8.2 shows that a decrease in pre-operative OKS from 25 to 4 is accompanied with a 15% drop in sufficient response rate. In comparison to TKR, the outcome of THR is much less sensitive to change in pre-operative score.

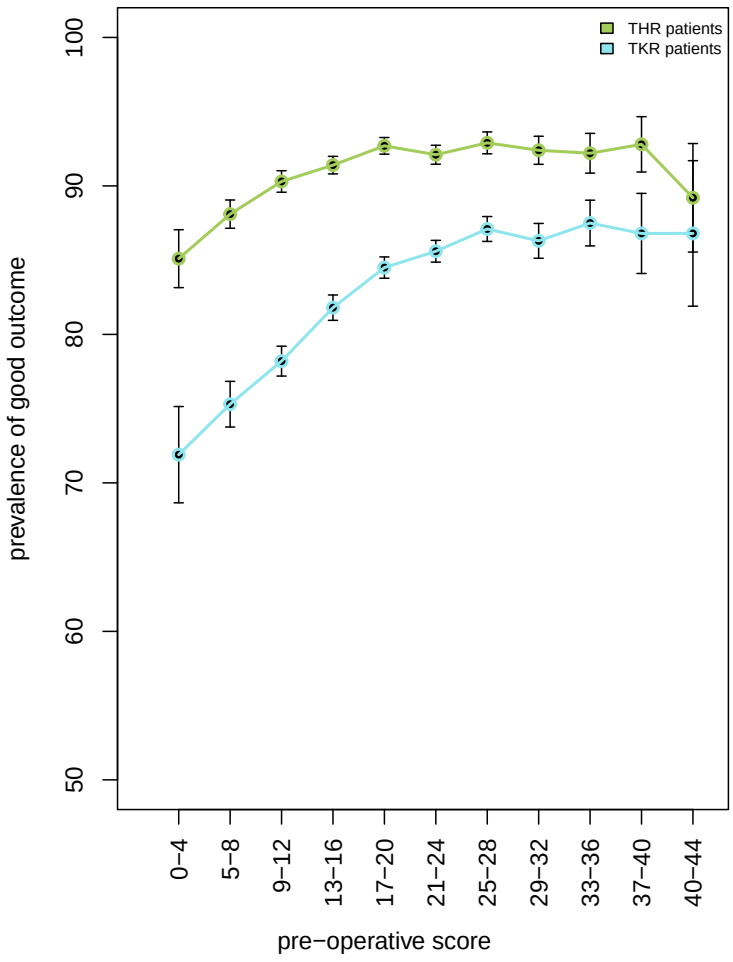


Figure 8.2: Comparing the change in the prevalence of sufficient response (i.e. high benefit) between TKR and THR patients, as pre-operative OKS/OHS increases.

8.3.4 Stratified study

To further explore how outcomes change according to pre-operative OKS/OHS, patients were stratified by age and gender, two factors that have been shown to affect post-operative outcome [181]. Results are shown in Figure 8.3, which again highlights the difference between TKR and THR – regardless of age and gender, there is a noticeable decrease in the proportion of "sufficient response" in TKR

patients with low pre-operative scores (below 17), whereas such a drop in sufficient response rate is not seen in THR patients with low baseline scores.

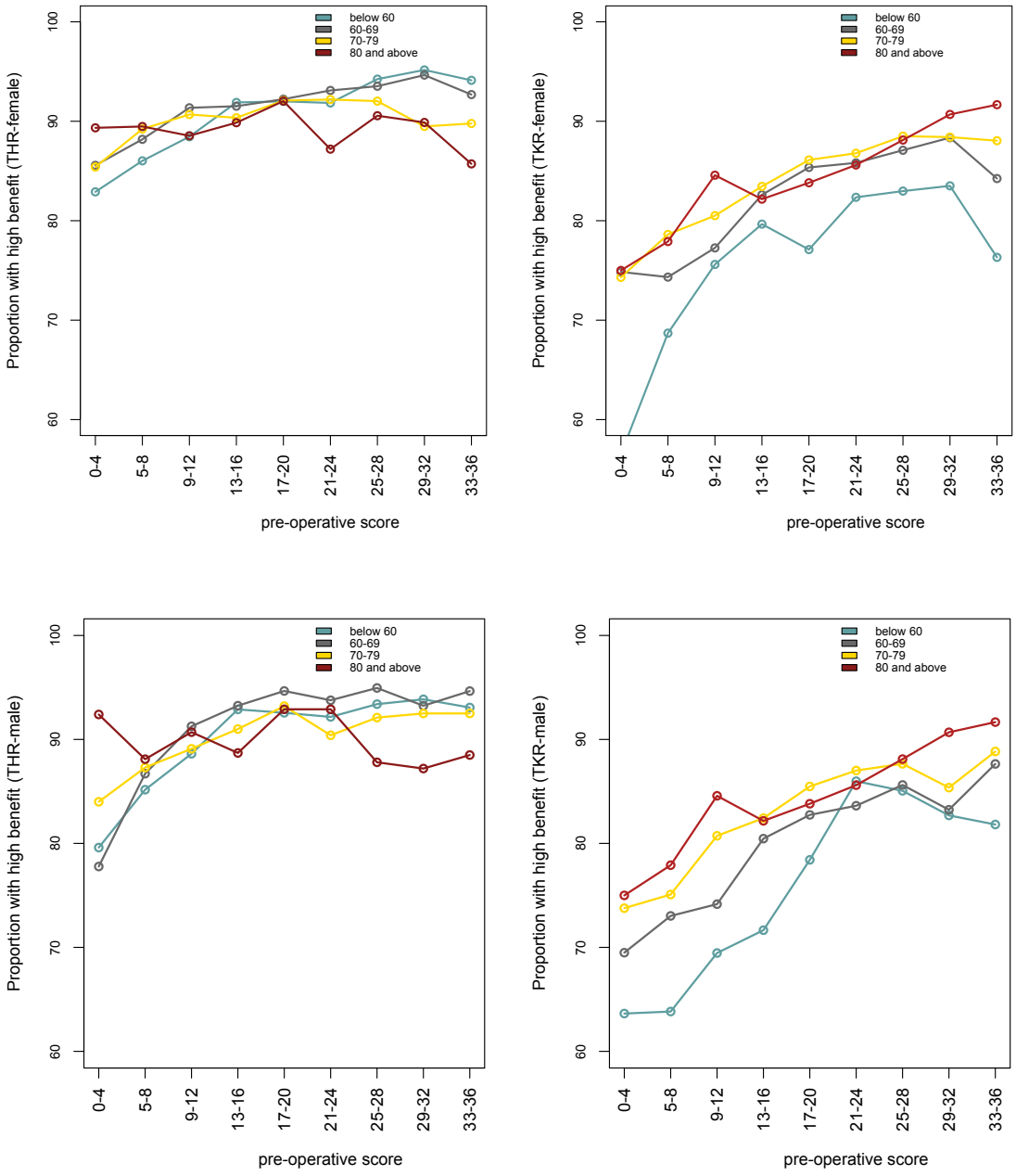


Figure 8.3: Comparing the change in the prevalence of "sufficient response" between TKR and THR patients stratified by age and gender, as pre-operative OKS/OHS increases. The two panels on the left are THR results, and TKR results are shown on the right. Results from female patients are shown at the top, and results from males shown at the bottom. Results from different age groups are shown in different colours as indicated by figure legends.

8.3.5 Multiple logistic regression analysis

Results from previous sections have consistently demonstrated that low baseline OKS is associated with a drop in the percentage of "sufficient response" amongst TKR patients post-operatively. On the other hand, the outcome of THR is not significantly impacted by low pre-operative PROM scores. However, before attributing any difference in outcome to pre-operative score, it is important to control for other factors that could also affect the outcome. The stratification strategy partially solves the problem, as it restricts comparisons to individuals with the same age band and gender. However, besides age and gender, previous studies have indicated that other factors such as co-morbidities can also affect post-operative outcome [2][181, 182].

To examine the effect of pre-operative score on post-operative outcome, while controlling for other potential confounding factors, two multiple logistic regression models were fitted – one for TKR patients and one for THR. In order to make post-operative outcome into a binary variable, the "low benefit" and "no benefit" were collapsed into one category – "insufficient response". Age, gender, pre-operative pain level and a host of co-morbidities were included as exposure variables. Results are shown in Table 8.3 with the exposure variables (i.e. factors that can potentially affect the outcome of operation) listed in the first column, regression coefficients and odds ratios for TKR in the second and third columns and, regression coefficients and odds ratios for THR in the fourth and fifth columns.

[2]Outcome was measured and reported using the traditional approach in these studies – in terms of the magnitude of the change in the score on multi-item questionnaires.

Variable	TKR (n=40,661)		THR (n=38,150)	
	Coefficient (95% CI)	OR (95% CI)	Coefficient (95% CI)	OR (95% CI)
Gender				
-Female	0	1.00	0	1.00
-Male	0.07* (-0.13,-0.02)	0.93(0.88, 0.98)	0.11** (0.03,0.19)	1.12 (1.04,1.21)
Age				
-Under 60yo	0	1.00	0	1.00
-60-69yo	0.29*** (0.21,0.38)	1.34 (1.23,1.46)	0.20*** (0.08, 0.31)	1.22 (1.09, 1.37)
-70-79yo	0.49 *** (0.41,0.58)	1.64 (1.50,1.79)	0.11 (-0.00, 0.22)	1.12 (1.00, 1.25)
-Over 80yo	0.62*** (0.51,0.73)	1.86(1.66,2.08)	0.12(-0.01, 0.26)	1.13 (0.99, 1.30)
Baseline OKS/OHS				
-0-4	-0.53*** (-0.73,-0.34)	0.59(0.48,0.72)	-0.43*** (-0.63, -0.22)	0.65(0.53, 0.80)
-5-8	-0.41*** (-0.53,-0.30)	0.66(0.59,0.74)	- 0.33***(-0.47, -0.19)	0.72 (0.62, 0.83)
-9-12	-0.35*** (-0.44,-0.25)	0.71(0.64,0.78)	-0.20**(-0.33, -0.07)	0.82 (0.72-0.94)
-13-16	-0.14** (-0.23,-0.05)	0.87(0.79,0.95)	-0.15*(-0.28, -0.03)	0.86 (0.76, 0.98)
-17-20	0.00	1.00	0.00	1.00
-21-24	0.08 (-0.01,0.17)	1.08(0.98,1.19)	-0.17*(-0.31, -0.04)	0.84(0.74,0.97)
-25-28	0.21*** (0.10,0.32)	1.23(1.11,1.38)	-0.08(-0.23, 0.07)	0.92(0.79,1.07)
-29-32	0.11 (-0.02,0.23)	1.11(0.98,1.26)	-0.17(-0.34, 0.01)	0.85 (0.71, 1.00)
-33-36	0.22 (0.04,0.40)	1.25 (1.04,1.50)	-0.16(-0.39, 0.07)	0.85(0.68, 1.08)
-37-40	0.11 (-0.16,0.41)	1.12 (0.85,1.51)	-0.04(-0.37, 0.32)	0.96(0.69, 1.38)
-41-44	0.14 (-0.40,0.75)	1.15 (0.67,2.11)	-0.22(-0.71, 0.32)	0.80(0.49, 1.38)
Pre-op disabilities				
-No	0.00	1.00	0.00	1.00
-Yes	-0.26*** (-0.36, -0.20)	0.77(0.73,0.82)	-0.37***(-0.45,-0.28)	0.69(0.64, 0.75)
Previous Sx				
-No	0.00	1.00	0	1.00
-Yes	-0.51*** (-0.60, -0.42)	0.60(0.55,0.66)	-1.05***(-1.15, -0.95)	0.35(0.32,0.38)

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8.3: Multiple logistic regression analysis for investigating the effect of pre-operative OKS/OHS on post-operative outcome.

Variables	TKR		THR	
	Coefficient (95% CI)	OR (95% CI)	Coefficient (95% CI)	OR (95% CI)
Depression				
-No	0.00	1.00	0.00	1.00
-Yes	-0.26*** (-0.34, -0.17)	0.77(0.71, 0.84)	-0.54*** (-0.64, -0.43)	0.59(0.53, 0.65)
Heart disease				
-No	0.00	1.00	0.00	1.00
-Yes	-0.21*** (-0.29, -0.13)	0.81 (0.75, 0.88)	-0.30*** (-0.41, -0.19)	0.74(0.66, 0.83)
Lung disease				
-No	0.00	1.00	0.00	1.00
-Yes	-0.11** (-0.20, -0.03)	0.89 (0.82, 0.97)	-0.19** (-0.31, -0.07)	0.83 (0.74, 0.93)
Circulation problem				
-No	0.00	1.00	0.00	1.00
-Yes	-0.53*** (-0.62, -0.44)	0.59 (0.54, 0.65)	-0.50*** (-0.62, -0.37)	0.61(0.54, 0.69)
Diabetes				
-No	0.00	1.00	0.00	1.00
-Yes	-0.27*** (-0.34, -0.19)	0.77(0.72, 0.82)	-0.18** (-0.30, -0.07)	0.83 (0.74, 0.93)
Pain[3]				
-None	0.00	1.00	-	-
-Very mild	-0.12*** (-0.19,-0.05)	0.89 (0.83,0.95)	-	-
-Mild	-0.37*** (-0.51,-0.22)	0.69 (0.60,0.81)	-	-
-Moderate	0.12 (-0.21, 0.47)	1.12 (0.81,1.60)	-	-
-Severe	0.07 (-0.56,0.80)	1.07 (0.57,2.21)	-	-

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8.4: Table 8.4 above combined with Table 8.3 in the previous page show the results from multiple logistic regression analysis, the purpose of which is to determine the effect of pre-operative scores on post-operative TKR/THR outcome while controlling for other confounding factors (e.g. age, co-morbidities). The post-operative outcome is expressed as the chance of obtaining a "sufficient response" assessed using the new, anchor-based OxPReSi method.

Consistent with the results from the descriptive study, results from multiple logistic regression analysis indicated that for both TKR and THR patients, after having

[3]Pre-operative pain was not included as an exposure variable in the THR model, as all associated pain levels showed insignificant P-values ($P \geqslant 0.05$)

adjusted for other factors (including age, gender, other comorbidities, etc), a sub-17 pre-operative score (assessed by the OKS/OHS questionnaire) was associated with being significantly less likely to achieve a sufficient response, i.e. feeling improved and satisfied; the negative impact of having a low pre-operative score is more prominent in TKR than in THR patients. In comparison to TKR patients in the lowest pre-op score range (between 0 and 4), the odds of obtaining a sufficient response is 1.7 times greater for patients with average pre-operative scores (in the range of 17 and 20), having controlled for other exposure factors. Equally important, once the pre-operative score has reached the 17-20 region, the odds of having a successful outcome stays relatively stable, and there is no statistically significant difference between patients' with very high baseline scores (e.g. 41-44), and those with baseline scores close to the mean.

In addition to pre-operative OKS/OHS, a number of factors were identified to significantly affect patients' chance of obtaining a sufficient response. Younger age, depression, disability, previous surgery, and other co-morbidities including lung disease, diabetes, circulation problems and heart disease all significantly decrease patients' odds of obtaining a sufficient response.

8.4 Discussion

8.4.1 Main findings

Conforming with findings from several previous studies [181, 183], the results confirmed that pre-operative pain and function as measured by the Oxford Knee Score, is an important factor of post-operative outcome. Using a large national joint registry data and the new anchor-based criterion for sufficient response consisted of minimal acceptable satisfaction (MAS) and minimal perceptible improvement (MPI)

(Chapter Six), results from both an enumerative study and a statistical modelling analysis indicate that for TKR patients with a pre-operative OKS of 17 or above, the proportion experiencing a successful outcome, i.e. achieving a sufficient response – feeling better and satisfied, remains above 85%. Once patients' pre-operative OKS drops below 17 points, sufficient response rate starts decreasing at a rate of 2.5% per every 4-point interval (shown in Figure 8.2).

As for the outcome of THR, in contrast to previous findings showing how worse pre-operative pain and function is associated with worse post-operative outcome [180, 179], results from this enumerative study indicate that the negative effect of a low pre-operative OHS on post-operative outcome appears to be less substantial compared to TKR, but is still statistically significant according to our multiple logistic regression model (Table 8.3). The proportion of THR patients with a successful outcome remains above 85% regardless of their baseline score. The relationship between baseline scores and post-operative outcome for TKR and THR are compiled in Figure 8.2 – the gap in "sufficient response" percentage between the two surgeries widens as pre-operative score decreases below the 17-20 threshold interval.

Following TKR/THR, in addition to pre-operative PROM scores, this study has identified several other determinants of outcome, including age, co-morbidities, and previous surgery, all of which are in agreement with previous studies [184, 185, 186]. Older, mentally and physically fit patients with minimal co-morbidities and no previous surgery have a higher chance of obtaining a sufficient response following surgery.

8.4.2 Clinical relevance

Findings from this study can be readily applied to clinical practice – for TKR patients who are seen with medium-level function and pain, effort needs to be made

to provide them the intervention in a timely manner, before their symptoms deteriorate. For TKR patients with pre-operative scores well below average, management of expectations needs to be appropriately carried out prior to surgery. Equally, THR patients with low pre-operative scores can be given the reassurance that, there is a high chance (above 85%) of obtaining a high benefit from surgery regardless of their pre-operative condition.

8.4.3 Strength and limitations

This is the first study investigating the effect of pre-operative PROM score, while using the new anchor-based criterion incorporating MAS and MPI as benchmarks for a successful outcome. A common shortcoming of previous studies investigating the effect of pre-operative functional ability on post-operative outcome is that outcome was often solely evaluated by PROM scores [180, 179, 181], including SF-36 physcial/social function, WOMAC-pain/function/total and OKS. PROM scores can assess patients' pain and function to a certain extent, but they can only reflect part of the total experience following surgery. In the last decade, patient satisfaction has become an important parameter for measuring outcome following elective surgery [187, 188, 189].

Another limitation of using a PROM score to evaluate post-operative outcome is the potential ceiling effect [181]. On health-related quality of life outcome measures with fixed-end scale, there is sufficient room for improvement for patients with poor pre-operative scores, but not enough of a ceiling for patients with high baseline scores. For example, someone with a pre-operative OKS of 4 points can gain a maximum of 44 points; but someone with a pre-operative score of 40 can improve their score by a maximum of 8 points only. In this case, the 8 points may be an underestimate of how much functional gain the patient has actually made from their operation.

Such ceiling effect is not present in our anchor-based criterion, which directly relies on patients' own evaluation of their satisfaction and perceived benefit from surgery.

With regard to limitations, due to the relatively small (but still clinically significant) difference between sufficient response rate of the low baseline score group and that of the high baseline score group, our multiple logistic model has a low R squared value od 0.04, indicating low predictive power. However, this only shows that our model is not suitable for outcome-predicting purposes; it does not weaken the validity of our finding regarding the impact of baseline score on post-operative outcome.

The other limitation of this study is that the impact of pre-operative score on long-term outcome trajectory following TKR/THR could not be studied here, as only six-month post-operative data were available. Chapter Four of this thesis explained how outcome achieved at one year does not necessarily represent later outcome. Since TKR and THR are operations designed to provide benefit both in the short term and in the long run, it is imperative to know how pre-operative score impacts on sufficient response rate in the long-term.

8.5 Conclusion

The most important message from this work is that a pre-operative score (measured on the OKS) of 17 points or lower can significantly decrease the patient's chance of obtaining a "high benefit" following TKR. In comparison to THR, the effect of low pre-operative score (measured on the OHS) on the outcome of THR is much less substantial. For knee OA patients, a pre-operative score of 17 or above, which is close to the current national average (19.2), may be an optimal threshold level (for surgery) that can maximise patients' chance of obtaining a successful outcome at six-months post-operation without compromising on the magnitude of improvement.

Chapter 9

Conclusions and future directions

9.1 Conclusions

With the rapidly rising demand for total knee and total hip replacement surgeries (TKR/THR) and a considerable cost to the health care system, the clinical effectiveness of TKR and THR are constantly under scrutiny. In previous studies, various methods have been used to analyse patient-reported outcome following TKR and THR, resulting in a wide range of success rates for the same population. To enable patients and providers make fully informed decisions at each stage of the intervention, proper analysis and interpretation of patient-reported outcome data is vital.

At the beginning of this thesis, the conceptual framework, "sufficient response", was introduced. A "sufficient response" consists of three components – an instrument used to measure the outcome, a response criterion and its associated cut-off score. Using this framework, the first part of the thesis explored the short-term outcome following TKR and THR, as well as the stability and lability of long-term outcome following TKR. The limitations associated with the current methods for measuring the prevalence of success were also examined.

The second part of this thesis presented a novel method of defining sufficient response, the Oxford Patient-Reported Success Index (OxPReSI), which was constructed upon two anchor questions designed to address patients' self-perceived improvement and satisfaction. Application of OxPReSi divides patients into three categorical outcome groups — high benefit, low benefit and no benefit. The OxPReSi was then applied to a randomised controlled trial data, viz. the Total and Partial Knee Arthroplasty Trial (TOPKAT), to compare the outcome of total and partial arthroplasty for patients with medial compartmental osteoarthritis. In comparison to traditional outcome measures and response criteria (e.g. the OKS, MIC, PASS), the OxPReSi is more informative, more interpretable and insensitive to change in baseline conditions.

The last part of the thesis investigated the effect of pre-operative factors, specifically pre-operative PROM scores measured on the Oxford Knee/Hip Score Questionnaires (OKS/OHS), on post-operative outcome measured by the OxPReSi. The results showed that for patients considering TKR, timing is a key component for optimising post-operative outcome – the probability of achieving a "high benefit" can be reduced by 15% if surgery was performed too late and patients' symptoms had significantly declined. The negative effect of low pre-operative OHS is much more minor for the outcome of THR.

9.1.1 Defining and quantifying the "sufficient response"

It has been recognised that, simply reporting the mean post-operative score or change in score at the group level, is not meaningful to patients or even some clinicians who do not possess in-depth knowledge of the instrument. A preferred approach would be to first establish a threshold for "successful" or "satisfactory" or "acceptable" outcome, and then report the results in terms of the proportion of patients that have achieved such threshold [115]. In 1989, Jaeschke and colleagues developed the first threshold for the purpose of patient reported outcome classification – the ' minimal clinically important difference" (MCID). In the past two decades, there has been a significant increase in the number of terminologies and definitions related to the minimal important difference, namely the minimal important change (MIC), the clinically significant change, the clinically important difference (CID), the minimal detectable change(MDC), the smallest real difference (SRD) [190] and the patient-acceptable symptom state score (PASS) just to name a few [191]. The original data set and mathematical method used to derive these thresholds differ from one another. As a result, the values of these thresholds are highly variable, and some are easier to achieve than others.

In general, in order to provide high quality care to patients, it is helpful to know how much change in HRQoL is sufficient to justify continuing the current treatment, modifying the treatment, or stopping the treatment and starting a new one. In orthopaedic research, when measuring patient-reported outcome following TKR/THR operations, it is of utmost importance to determine whether the change in their HRQoL due to surgery has been sufficient. Various attempts have been made to deduce this deceptively straightforward threshold value – MID, MIC, MDC, PASS etc [191, 192, 115].

In Chapter Three of this thesis, the confusing collective of threshold-related terminologies are consolidated into a conceptual framework centred on "sufficient response". In this framework, a "sufficient response" is defined by three components – the outcome measure, the response criterion and its associated cut-off score (see Tables 3.1 and 3.2 for details). Using a large, recent, nationally collected dataset (2014-15 NHS PROM data), the only PROM specifically developed for patients undergoing TKR/THR (OKS/OHS), and a set of four different response criteria (1/2 SD, MIC, PASS score and OMERACT-OARSI), the study found that the prevalence of sufficient response was between 72% and 90% amongst TKR patients, and between 77% and 95% amongst THR patients. Within each procedure, pain improves more prominently than function. This is the most comprehensive, accurate and considered evaluation of the true success rate of TKR and THR to date. Our results also confirmed the suspicion that the outcome, expressed as the proportion of patients achieving a sufficient response, vary rather significantly according to the choice of outcome measure, response criterion and its associated cut-off.

In the past two decades, a dozen different versions of "sufficient response" regarding TKR/THR outcome have appeared in the literature; it is likely that more will be identified in the coming years [191]. While it may be appealing to adopt a single response criterion and its associated cut-off score as the "gold-standard" or "rule

of thumb", results from our study have demonstrated that each version of response criterion has its own merits and drawbacks. For the same patient population whose TKR/THR are assessed by a single outcome measure, the outcome can appear rather different depending on which response criterion has been used to evaluate the intervention.

9.1.2 Analysis of longitudinal data

Chapter Four presents a longitudinal study that investigates the temporal characteristics of TKR patients' self-reported outcome from one-year to ten-year post-operation. Unlike previous studies that used group-level average OKS values as an assessment of outcome, long-term outcome following TKR was examined using an alternative approach; one that involves proper classification of patients based on whether they have achieved a sufficient response or not, labelled as "successful" or "failed" outcome. Using the conceptual framework introduced in Chapter Three, a "sufficient response" is defined by the OKS (as the outcome measure), the patient acceptable symptom state (PASS) score (as the response criterion), and OKS $\geqslant 11$ (as the cut-off). The PASS score incorporates a patient-centred dimension of what is a satisfactory state.

Our results demonstrate that the outcome achieved at one year does not necessarily represent later outcome after TKR. In the first year, 71% of patients achieved a successful outcome and 29% failed to do so. In subsequent years, a substantial number of patients (approximately 15-20%) can be seen to move between classifications (in both directions) from one year to the next and over time. According to patients' long-term outcome trajectory, specific sub-groups were identified:

- *Early achievers* – patients who achieved a successful outcome in year one and remained in the "success"category throughout.

- *Late achievers* – patients who failed to achieve a successful outcome in year one, but improved later on and maintained their "success" status.
- *Failures* – patients who never managed to achieve a successful outcome.
- *Fluctuators* – patients who vacillate between "success" and "failure" over time.

Findings from this chapter enhance our knowledge regarding the temporal trends in outcome following knee replacement, which can be useful in clinical practice. Patients with sub-optimal outcome in the first year following operation can be reassured that there is a good chance that their condition will improve, especially between the first and second year. On the other hand, patients with successful outcome in the first year should also be informed that although achieving success early on is highly desirable, long-term success is not guaranteed. Long-term monitoring using self-reported methods, especially patient-referenced metrics of success, is indispensable.

9.1.3 The Oxford patient-reported success index (OxPReSi)

The first part of this thesis introduced the concept of "sufficient response", and how this concept can facilitate interpretation of PROM data. However, findings from part one of the thesis also raised the question whether a standard, multi-item questionnaire is the best tool for measuring HRQoL in TKR/THR. Was there a better alternative? These questions were addressed in part two of this thesis.

In the last two decades there has been a rapid increase in the number of sophisticated instruments for measuring patient-reported health status. However, the implementation of these instruments in routine clinical practice has not being growing as fast. One of the main reasons is concern over the amount of time and resources needed to administer and score such questionnaires. Even though a great many measures of

health status have been developed and validated, the search for the ideal measure has continued [191].

The length of questionnaires makes them time-consuming to complete and score, and is most likely the major factor that limits their practical use [9]. The shortest assessment is a single item (a.k.a. a global question) that requires an overall description of health status. By assigning a score to each response category (to the global question), a global score is provided. Initially, there were concerns over whether a global score would have sufficient validity. Results from multiple studies consistently demonstrated that single, global questions can provide information that is consistent with more elaborate questionnaires [9, 193]. Lastly, if the scores obtained from completing a complex health status measure will eventually, for the purpose of analysis, be re-coded into broad categories (i.e. having achieved the sufficient response vs failed to obtain sufficient response following treatment), it may be preferable to substitute a valid global question for the complex measure at the outset [191].

9.1.3.1 Development of the OxPReSi

Chapter Six of this thesis introduced a new method of assessing outcome following TKR/THR – the Oxford Patient Reported Success Index (OxPReSi), which eliminates the need to fill out any lengthy questionnaires or to calculate for any cut-off scores. The OxPReSi was developed using the novel approach of sequentially combining patients' responses to two global questions – the transition and satisfaction anchor questions. Application of this index directly divides patients into three categorical outcome groups: I) high benefit – significant improvement with satisfaction. II) low benefit – perceptible improvement without satisfaction. III) no benefit – no perceptible improvement. The proportion of patients achieving a high benefit is

83% for TKR and 91% for THR, which is consistent with the results obtained in Chapter Three of this thesis (prevalence of "sufficient response" : between 72% and 90% for TKR, between 77% and 95% for THR).

9.1.3.2 Comparison between OxPReSi and the OKS/OHS

The OxPReSi demonstrated close association with the OKS/OHS – the "high benefit" group had a very large mean ΔOKS/ΔOHS (ΔOKS = 18.5 for TKR patients, and ΔOHS = 22.3 for THR patients), the "low benefit" group showed a mean score change of 9.0 (ΔOKS) and 11.6 (ΔOHS), the "no benefit" group showed a very small but positive change in OKS/OHS. However, the wide variation in ΔOKS/ΔOHS within each outcome group (see Figure 6.3 for detail) suggests that there are fundamental differences between these two patient-reported outcome assessment methods.

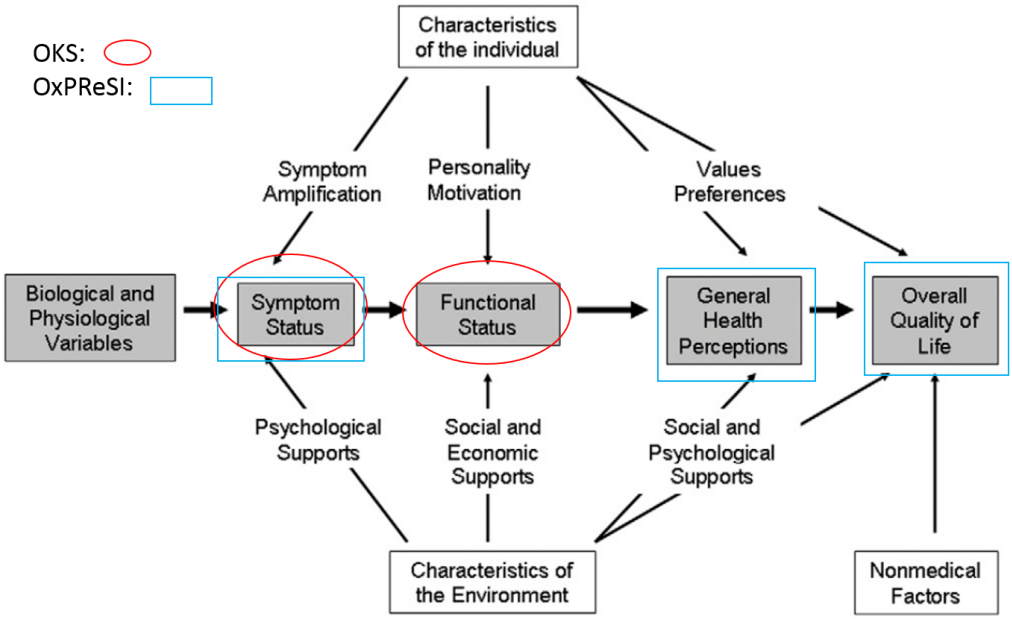


Figure 9.1: The levels measured by the OKS/OHS are marked by red circles, and the levels covered by the OxPReSI method are marked by blue squares. (Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995; 273: 59-65.)

As shown in Figure 9.1, the Wilson and Cleary model illustrates the characteris-

tics of individuals that different outcome measures target. The OKS mainly covers "symptom status" and "functional status", whereas the OxPReSI covers "symptom status", "overall quality of life" and perhaps 'general health perceptions". Both "general health perceptions" and "overall quality of life" are at the more complex end of the spectrum. General health perceptions are shown to be associated with functional status, but can also be affected by a host of social and psychological factors [194]. "Overall quality of life" is the broadest outcome, and can be impacted by a wide range of non-medical factors. Previous research has shown that functional status is not necessarily related to overall happiness. Comparison between OxPReSi and the OKS/OHS in the context of the Wilson-Cleary model indicates that, the OxPReSi method provides more comprehensive assessment of post-operative outcome, while the OKS/OHS measures symptoms and functions in greater detail.

9.1.3.3 Estimation of prior state – response shift or implicit theory of change

Another fundamental difference between traditional multi-item questionnaires used in the field of arthroplasty and OxPReSi lies in the way patients' pre-operative state is estimated. Traditional measures ask patients to offer prospective rating of their symptoms before and after operation. In other words, there are the pre-test and post-test conducted at different time points. The OxPReSi is only administered post-operation, at which point patients are asked to :

- Rate their present state – the "post-test",
- Provide retrospective judgement of their pre-operative state – the "then-test",
- Compare the change in symptom levels from "then-test" to "post-test".

The theory of *response shift* is crucial in understanding the validity of prospective

vs retrospective judgement of pre-operative state (i.e. the validity of pre-test vs then-test). Response shift refers to the fact that patients' perception of their health state at a particular time point (e.g. before undergoing an operation, before taking a new drug) can change over time as a result of acquisition of new standards. With total knee replacement patients, before undergoing operation they may have been accustomed to the pain and functional limitations, so rate their QoL fairly high – this would be their pre-test results. Only after surgery, looking back, will these patients realise how incapacitated they were pre-operation. Consequently, patients' retrospective assessment of their pre-operative QoL would be low – this would be their then-test results. Response shift theory considers the retrospective judgement of pre-operative symptom state, the then-test, as more valid than pre-test. This is because retrospective judgement is made in light of the new (post-operative) standards, which was not obtainable at the time of baseline assessment [12] [195]. The "response shift" theory therefore bolsters the validity of the OxPReSi method in assessing post-operative outcome.

9.1.4 Application of OxPReSi to the Total and Partial Knee Arthroplasty Trial data

Following the development of the OxPReSi, Chapter Seven investigated whether this new anchor-based metric was suitable for measuring and comparing outcomes in comparitive studies such as randomised controlled trials (RCT). The data utilised to test the applicability of OxPReSi was first year data from the Total and Partial Knee Arthroplasty Trial (TOPKAT), which was designed to compare the clinical and cost effectiveness of partial knee replacement (PKR) with that of total knee replacement (TKR). Results from the study suggest that with some modification, the OxPReSi was able to effectively separate participants receiving TKR and PKR

into three clinically meaningful categories – "high benefit", "low benefit" and "no benefit". The proportion with "high benefit" is 8% higher amongst PKR patients than their TKR counterpart.

The results obtained using the OxPReSi were consistent with the results from the original trial (where the average OKS at one-year follow-up was the primary outcome measure) in terms of the overall trend. The outcome of PKR is superior to TKR – higher post-operative OKS and greater proportion of "high benefit" was found in PKR patients compared to TKR. It could be argued that using the OxPReSi as the primary outcome measure leads to an end result that provides a greater level of detail and is easier to interpret. Furthermore, for the patients, the two anchor questions in OxPReSi are far less burdensome to complete than traditional multi-item PROM questionnaires. The clarity and brevity may yield higher response rate, as well as, saving significant time and resources.

Results obtained using other threshold-based methods also showed the same pattern, with the "sufficient response" rate being consistently higher in PKR patients. Those patients (TKR and PKR combined) with different types of discordant results were identified and their pre-operative scores were compared. The results confirmed previous findings about cut-off scores being baseline dependent. Patients with worse baseline conditions would need to improve their PROM score by a greater amount than those with better baseline functions in order to obtain "sufficient response". Currently, it is common practice to derive a *single* cut-off score for a response criterion. Future work dedicated to extracting baseline-specific threshold values would be highly valuable, as universal application of a single cut-off score can potentially lead to incorrect judgement of post-operative outcome.

Lastly, comparison between the NHS PROM data and the TOPKAT data revealed a notable difference between the post-operative outcome from an RCT data and that

from a population/cohort data. Researchers need to be mindful of such difference, which was not detected using other existing threshold-based methods. Lastly, the results also demonstrated that when constructing a rating scale, it is important to keep the response choices *balanced*, meaning presenting equal number of favourable and unfavourable items.

9.1.5 The optimal pre-operative condition for maximum post-operative benefit

The purpose of the study described in Chapter Eight was to examine the effect of pre-operative condition on post-operative outcome, as well as to identify the optimal condition that can maximize patients' chance of obtaining a "sufficient response" (i.e. "high benefit", when assessed by the OxPReSi method). The results indicate that the effect of pre-operative condition (as measured by the OKS/OHS) for TKR patients is distinct from that for THR patients. For TKR patients, the "sufficient response" rate remains above 85% so long as their pre-operative OKS is higher than 17 points. The data reveals that once a patient's pre-operative score drops below 17 points, the likelihood of getting a "sufficient response" decreases rapidly. In comparison, the pre-operative score appeared to have a much more minor effect on post-operative outcome for THR patients, as the "sufficient response" rate remained above 85% regardless of their baseline score.

9.2 Future directions

9.2.1 Defining and quantifying the "sufficient response"

Findings from our study indicate that the key to accurate interpretation of HRQoL & PRO data from TKR/THR patients is an appropriate definition and quantification of the "sufficient response", which can be achieved in three steps: I) Selection of a suitable outcome measure. II) The choice of a set of suitable response criteria. Response criterion should be decided on an individual basis – identify the aspect most important to the patient, whether it be pain, function or satisfaction, and then a subsequent selection of the response criterion that matches patients' priority. Shared decision-making is paramount in patient outcome research. III) Identification of the cut-off score – a number of response criteria and associated cut-off scores have already been identified, but in the case where the cut-off score is not available, an anchor-based approach should be the primary method for estimating the threshold value [154, 196]. Anchor selection is a crucial step as the anchor needs to be highly meaningful and interpretable. There also needs to be a theoretical basis for the relationship between the anchor and the PROM (or the specific domain of a multi-dimensional PROM). For instance, if a TKR patient has identified improving their activity of daily living (ADL) as their primary goal, then the response criterion would be "minimal important improvement in function". Next, the cut-off for the OKS-function subscale should be determined, and ideally the anchor question would explore the degree of change in their ADL function compared to before the operation. An anchor question addressing pain would be inappropriate [191]. An empirical correlation of at least 0.30 between the anchor and the HRQoL has been recommended [191].

Lastly, it would be helpful to use various definitions of "sufficient response" following

TKR/THR in order to obtain a more comprehensive evaluation of outcome. With regard to reporting, each component that describes sufficient response, including the outcome measure, the response criterion, the cut-off score, and the method used to derive the cut-off all need to be specified. Instead of continuing to develop new MID-related categorical thresholds, it would be desirable if all information regarding each available version of "sufficient response" was consolidated into guidelines for measuring patient reported outcome in orthopaedics with regular updates as new discoveries emerge.

9.2.2 Analysis of long-term outcome following TKR

The main limitation of the longitudinal study presented in Chapter Four is that multiple imputation was not used during data analysis; the high proportion of missing data from the original trial could affect the precision of the estimate [197]. Although there is a mixed opinions regarding the application of multiple imputation in clinical data [197, 198], in the future, it would still be worthwhile to perform the same analysis with proper imputation.

Furthermore, to confirm our findings about the fluctuation between "success" and "failure" regarding patients' long-term outcome following TKR, it would be useful to carry out further studies using a different definition for sufficient response – either using a different patient-reported outcome measurement (e.g. the WOMAC), or still using the OKS but a different response criterion and associated cut-off score. It would be reassuring if a similar pattern is replicated.

Moreover, this study demonstrated the fluctuating and dynamic nature of long-term outcome following TKR, and identified sub-groups of patients following different outcome trajectories over time. Future studies investigating characteristics unique to each sub-group would be highly useful in patient education and expectation man-

agement.

Lastly, data were analysed using a purely statistical method in this study. With the rising popularity of health informatics, it would be useful to feed the current data into an appropriate unsupervised learning algorithm and observe how many categories these patients are divided into. More details about the potential usage of machine learning in orthopaedic research can be found later in Section 9.2.4.

9.2.3 Future of the Oxford Patient Reported Success Index

There is still room to make improvement on the OxPReSi prototype in the future. To develop the "ideal" version of OxPReSi, patient and public involvement sessions would be useful in modifying the way patients are divided into "high benefit", "low benefit" or "no benefit" groups. Also, opinions from experts in other disciplines (e.g. linguistics) would be valuable for optimizing the rating scale, making it as balanced and unbiased as possible. Studies aimed to test the measurement properties, including construct validity, reproducibility, responsiveness and response rate of the OxPReSi would be worthwhile doing.

In clinical practice, the OxPReSi can potentially be used as the first step in measuring post-operative outcome. Patients identified as having a "low benefit" or "no benefit" by OxPReSi will need more detailed follow-up assessment. In addition, when using the OxPReSi to compare outcomes from different interventions, it is important to keep the wording of the transition/satisfaction question AND their response choices identical.

Furthermore, for OA patients with medium level of pain and function at the knee awaiting TKR, these patients need to be provided with the intervention in a timely manner before their symptoms become too severe. Also, for patients seen in clinic

presented with severe symptoms (i.e. high pain and low function), pre-operative education and management of expectation need to be carried out, as the chance of them experiencing a successful outcome can be 10-15% lower than other higher functioning TKR patients.

9.2.4 Machine learning for research in patient-reported outcome following total joint arthroplasty

Another area for future work is machine learning. This section provides a brief overview of how machine learning (ML) techniques can potentially be utilised in analysing patient-reported outcome data following total joint arthroplasty; technical details (e.g. formulations, modelling, etc) will not be included.

The "machine" part in machine learning refers to the process where a computer is provided with data or past experience, as well as an algorithm (or model) defined up to some parameters. The "learning" part in machine learning refers to optimisation of the parameters of the existing model by the (pre-programmed) computer using the training data or past experience provided. The end result is a system that is capable of continuous self-improvement and thereby offering increased efficiency and effectiveness. The model may be for descriptive or predictive purpose [199].

9.2.4.1 Data source – electronic health record (EHR)

In the last ten years, there has been a continued increase in using machine learning approaches for clinical data analysis. Within the realm of knee and hip arthroplasty, the type of data available in large retrospective analyses is electronic health record (EHR) data.

In the UK, the NHS Hospital Episode Statistics (HES) contain records of all pa-

tients admitted to NHS hospitals in England. Each HES record contains an array of information about an individual patient admitted to a NHS hospital, including clinical information (e.g. diagnoses, operations etc), demographical information (e.g. age group, gender, ethnicity, etc) as well as geographical information, such as where the patient resides and received treatment [200]. In the case with hip and knee replacement, since the implementation of the national PROM program in 2009, all hip and knee replacement patients are encouraged to complete a pre-operative and a post-operative questionnaire. The PROM questionnaires are then linked to HES record data where possible. Such large-scale population-based EHR data accumulated during the routine care of patients make them ideal for ML-related analysis.

On a conceptual level, there are two areas of arthroplasty-related outcome research, where ML models can potentially provide clinical values.

9.2.4.2 Patient clustering

As shown in Chapter Six and Seven, using the OxPReSi method, knee and hip replacement patients can be categorised into three groups – "high benefit", "low benefit" and "no benefit". Within each group, patients also vary in terms of their age, gender, the severity of their pre-operative symptoms, etc. Therefore, it is likely that amongst all patients with end-stage OA at their hip or knee and having undergone hip and knee replacement, there are distinct subtypes. These subtypes may be related to personal factors or environmental factors . A better understanding of these subtypes will facilitate progress toward "precision medicine" by enabling the delivery of more personalised care throughout the process [201]. A variety of unsupervised ML techniques provide a powerful way to probe large electronic health record data in order to identify subgroups of patients. To date ML approaches have already been utilised in subtyping patients with chronic obstructive pulmonary

disorders and type II diabetes [202, 203].

9.2.4.3 Application of extreme value theory to longitudinal data

Hip and knee replacement are supposed to provide both short- and long-term benefits. Currently, despite a low overall revision and re-operation rate, patients' self-reported post-operative outcome are more heterogeneous. As shown in Chapter Four, patients follow different trajectories in terms of long-term post-TKR outcome – some achieve a sufficient response early on and maintain it, some take more than one year to obtain sufficient response, some keep fluctuating between "success" and "failure".

When monitoring patients over a long period of time, often it is the deviation beyond what is considered a "normal" range that are of interest. In the case of long-term follow up for TKR patients using the PASS-ΔOKS score as the cut-off, it becomes a concern when a patients' ΔOKS stays constantly well below the 11-point threshold. Therefore, it is sensible to focus specifically on modelling the extremes of the distribution of data points. Extreme value theory (EVT) is a branch of statistics that seeks to model the behaviour of the tails of distributions, and therefore is well-suited for problems in which the points of interest are those points positioned in the extremes. EVT has already been used to monitor financial and environmental data for many years, and the EVT model certainly has the potential to be applied to outcome-related research [201].

Bibliography

[1] N. J. Eynon-Lewis, D. Ferry, and Pearse M F, "Themistocles Gluck: an unrecognised genius," *BMJ*, vol. 305, pp. 1534–6, 1992.

[2] R. Trebse, *Infected Total Joint Arthroplasty: The Algorithmic Approach*. Springer-Verlag, 2012.

[3] The NJR Editorial Board, "2017 14th Annual Report," tech. rep., The National Joint Registry, 2017.

[4] P. H. Bullens, C. J. van Loon, M. C. de Waal Malefijt, R. F. Laan, and R. P. Veth, "Patient satisfaction after total knee arthroplasty," *J. Arthroplasty*, vol. 16, pp. 740–747, sep 2001.

[5] R. D. Leffert, "Ernest Amory Codman: The End Result of a Life in Medicine," *Journal of Shoulder and Elbow Surgery*, vol. 9, p. 259, may 2000.

[6] E. Codman, *The shoulder rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa*. N.Y.: C. Miller, 1934.

[7] B. Mallon, *Ernest Amory Codman : the end result of a life in medicine*. Saunders, 2000.

[8] P. Pynsent, J. Fairbank, and A. Carr, *Outcome Measures in Orthopaedics and Orthopaedic trauma*. London: Arnold, a member of the Hodder Headline Group, 2nd ed., 2004.

[9] C. Jenkinson, *Measuring Health And Medical Outcomes*. UCL Press, nov 1994.

[10] S. Wood-Dauphinee, "Assessing Quality of Life in Clinical Research: From Where Have We Come and Where Are We Going?," *Journal of Clinical Epidemiology*, vol. 52, pp. 355–363, apr 1999.

[11] "The International Classification of Functioning, Disability and Health: ICF," tech. rep., World Health Organisation, Geneva, 2001.

[12] D. L. Streiner, G. R. Norman, and J. Cairney, *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press, fifth ed., 2008.

[13] A. Cláudia de Souza -Rua Padre Brito, A. Cláudia de Souza, N. Maria Costa Alexandre, and E. de Brito Guirardello, "Psychometric properties in instruments evaluation of reliability and validity," *Applications of epidemiology Epidemiol. Serv. Saude, Brasília*, vol. 26, no. 3.

[14] D. F. Polit, "Assessing measurement in health: Beyond reliability and validity," *International Journal of Nursing Studies*, vol. 52, pp. 1746–1753, nov 2015.

[15] A. Bowling, *Measuring Disease*. Philadelphia: Open University Press, 2nd ed., 2001.

[16] M. H. Liang, "Evaluating measurement responsiveness.," *The Journal of rheumatology*, vol. 22, pp. 1191–2, jun 1995.

[17] M. Corzillius, P. Fortin, and G. Stucki, "Responsiveness and sensitivity to change of SLE disease activity measures," *Lupus*, vol. 8, pp. 655–659, oct 1999.

[18] K. Knutson and O. Robertsson, "The Swedish Knee Arthroplasty Register – The Inside Story," *Acta Orthopaedica.*, vol. 81, pp. 5–7, feb 2010.

[19] C. Delaunay, "Registries in orthopaedics," *Orthopaedics & Traumatology: Surgery & Research*, vol. 101, pp. S69–S75, 2015.

[20] O. Robertsson, "Knee arthroplasty registers," *Bone Joint J.*, vol. 89-B, no. 1, 2007.

[21] M. Bhandari, J. Smith, L. E. Miller, and J. E. Block, "Clinical and economic burden of revision knee arthroplasty.," *Clin. Med. Insights. Arthritis Musculoskelet. Disord.*, vol. 5, pp. 89–94, 2012.

[22] D. W. Murray, A. J. Carr, and C. Bulstrode, "SURVIVAL ANALYSIS OF JOINT REPLACEMENTS," *J Bone Jt. Surg [Br]*, vol. 75, pp. 697–704, 1993.

[23] A. Darzi, "High Quality Care For All – NHS Next Stage Review Final Report," tech. rep., The Stationery Office, England, 2008.

[24] K. Harris, J. Dawson, E. Gibbons, C. R. Lim, D. J. Beard, R. Fitzpatrick, and A. J. Price, "Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty.," *Patient related outcome measures*, vol. 7, pp. 101–8, 2016.

[25] A. D. Siderowf and R. M. Werner, "The EQ-5D–a generic quality of life measure–is a useful instrument to measure quality of life in patients with Parkinson's disease.," *Journal of neurology, neurosurgery, and psychiatry*, vol. 70, p. 817, jun 2001.

[26] R. L. Mizner, S. C. Petterson, K. E. Clements, J. A. Zeni, J. J. Irrgang, L. Snyder-Mackler, and L. Snyder-Mackler, "Measuring functional improvement after total knee arthroplasty requires both performance-based and

patient-report assessments: a longitudinal analysis of outcomes.," *J. Arthroplasty*, vol. 26, pp. 728–37, aug 2011.

[27] FDA and HHS, "Guidance for Industry Use in Medical Product Development to Support Labeling Claims Guidance for Industry," *Clinical/Medical Federal Register*, no. December, pp. 1–39, 2009.

[28] M. Peters, H. Crocker, C. Jenkinson, H. Doll, and R. Fitzpatrick, "The routine collection of patient-reported outcome measures (PROMs) for long-term conditions in primary care: a cohort survey," *BMJ Open*, vol. 4, p. 3968, 2014.

[29] Porter, Gonçalves-Bradley, Ricci-Cabello, Gibbons, Gangannagaripalli, Fitzpatrick, Black, Greenhalgh, Valderas, and JM, "This is a repository copy of Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities," 2016.

[30] E. F. Juniper, G. H. Guyatt, A. Willan, and L. E. Griffith, "Determining a minimal important change in a disease-specific Quality of Life Questionnaire.," *Journal of clinical epidemiology*, vol. 47, pp. 81–7, jan 1994.

[31] H. J. Schünemann, E. A. Akl, and G. H. Guyatt, "Interpreting the results of patient reported outcome measures in clinical trials: The clinician's perspective," *BMC : Health and Quality of Life Outcomes*, vol. 4, pp. 1–8, sep 2006.

[32] I. R. Gough, C. M. Furnival, L. Schilder, and W. Grove, "Assessment of the quality of life of patients with advanced cancer.," *European journal of cancer & clinical oncology*, vol. 19, pp. 1161–5, aug 1983.

[33] W. O. Spitzer, A. J. Dobson, J. Hall, E. Chesterman, J. Levi, R. Shepherd, R. N. Battista, and B. R. Catchlove, "Measuring the quality of life of can-

cer patients: a concise QL-index for use by physicians.," *Journal of chronic diseases*, vol. 34, no. 12, pp. 585–97, 1981.

[34] J. Wennberg, "Outcomes research, const containment, and the fear of health care rationing," *The New England Journal of Medicine*, vol. 323, no. 17, pp. 1202–1204, 1990.

[35] P. D. Cleary, S. Edgman-Levitan, M. Roberts, T. W. Moloney, W. McMullen, J. D. Walker, and T. L. Delbanco, "Patients evaluate their hospital care: a national survey.," *Health Aff. (Millwood).*, vol. 10, no. 4, pp. 254–67, 1991.

[36] H. Haga, H. T. Suzuki, S. Yasumura, H. Nagai, H. Amano, and H. Shibata, "Factors predisposing to decline of physical function between age 70 and 85 years," in *The XV International Congress of Gerontology*, (Budapest), 1993.

[37] D. M. Levine, L. L. Morlock, A. I. Mushlin, S. Shapiro, and F. E. Malitz, "The role of new health practitioners in a prepaid group practice: provider differences in process and outcomes of medical care.," *Med. Care*, vol. 14, pp. 326–47, apr 1976.

[38] J. W. Bush, J. P. Anderson, R. M. Kaplan, and W. R. Blischke, ""Counterintuitive" preferences in health-related quality-of-life measurement.," *Med. Care*, vol. 20, pp. 516–25, may 1982.

[39] L. Burke, "History of Patient-Reported Outcome Measurement at FDA: My Perspective." `https://mapigroup.com/pro{_}newsletter/history-patient-reported-outcome-measurement-fda-perspective/`, Accessed: 2018-11-12.

[40] I. B. Wilson and P. D. Cleary, "Linking Clinical Variables With Health-Related Quality of Life A Conceptual Model of Patient Outcomes," *JAMA*, vol. 273, no. 1, pp. 59–65, 1995.

[41] J. M. Valderas and J. Alonso, "Patient reported outcome measures: a model-based classification system for research and clinical practice," *Quality of Life Research*, vol. 17, pp. 1125–1135, nov 2008.

[42] P. R. Deshpande, S. Rajan, B. L. Sudeepthi, and C. P. Abdul Nazir, "Patient-reported outcomes: A new era in clinical research.," *Perspectives in clinical research*, vol. 2, pp. 137–44, oct 2011.

[43] N. Black, "Patient reported outcome measures could help transform healthcare What are PROMs?," *BMJ*, vol. 346, 2013.

[44] D. G. Kyte, M. Calvert, P. J. Van Der Wees, R. Ten Hove, S. Tolan, and J. C. Hill, "An introduction to patient-reported outcome measures (PROMs) in physiotherapy," *Physiotherapy*, vol. 101, pp. 119–125, 2015.

[45] N. Forget and J. Higgins, "Comparison of generic patient-reported outcome measures used with upper extremity musculoskeletal disorders: Linking process using the International Classification of Functioning, Disability, and Health (ICF)," *Journal of Rehabilitation Medicine*, vol. 46, no. 4, pp. 327–334, 2014.

[46] R. Fitzpatrick, C. Davey, M. J. Buxton, and D. R. Jones, "Evaluating patient-based outcome measures for use in clinical trials HTA Health Technology Assessment NHS R&D HTA Programme," Tech. Rep. 14, 1998.

[47] S. McKenna, "Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science.," *BMC medicine*, vol. 9, p. 86, jul 2011.

[48] J. E. Ware, B. Gandek, R. Guyer, N. Deng, and N. Deng, "Standardizing disease-specific quality of life measures across multiple chronic conditions: de-

velopment and initial evaluation of the QOL Disease Impact Scale (QDIS®).," *Health Qual. Life Outcomes*, vol. 14, p. 84, jun 2016.

[49] C. J. M. Bachmeier, L. M. March, M. J. Cross, H. M. Lapsley, K. L. Tribe, B. G. Courtenay, and P. M. Brooks, "A comparison of outcomes in osteoarthritis patients undergoing total hip and knee replacement surgery," *Osteoarthritis and Cartilage*, vol. 9, pp. 137–146, 2001.

[50] S. M. McPhail, K. S. Bagraith, M. Schippers, P. J. Wells, and A. Hatton, "Use of Condition-Specific Patient-Reported Outcome Measures in Clinical Trials among Patients with Wrist Osteoarthritis: A Systematic Review," *Advances in Orthopedics*, vol. 2012, pp. 1–10, nov 2012.

[51] M. M. Versteegh, A. Leunis, C. A. Uyl-de Groot, and E. A. Stolk, "Preference-Based Assessments Condition-Specific Preference-Based Measures: Benefit or Burden?," 2012.

[52] A. W. Riley, C. B. Forrest, G. W. Rebok, B. Starfield, B. F. Green, J. A. Robertson, and P. Friello, "The Child Report Form of the CHIP-Child Edition: reliability and validity.," *Med. Care*, vol. 42, pp. 221–31, mar 2004.

[53] R. G. Parrish, "Measuring population health outcomes.," *Preventing chronic disease*, vol. 7, p. A71, jul 2010.

[54] C. O'Boyle, H. McGee, A. Hickey, C. Joyce, J. Browne, K. O'Malley, and B. Hiltbrunner, "The Schedule for the Evaluation of Individual Quality of Life (SEIQoL). Administration Manual.," *Psychology Reports*, jan 1993.

[55] L. W. Chambers, L. A. Macdonald, P. Tugwell, W. W. Buchanan, and G. Kraag, "The McMaster Health Index Questionnaire as a measure of quality of life for patients with rheumatoid disease.," *The Journal of rheumatology*, vol. 9, no. 5, pp. 780–4, 1982.

[56] R. F. Meenan, P. M. Gertman, and J. H. Mason, "Measuring health status in arthritis. The arthritis impact measurement scales.," *Arthritis and rheumatism*, vol. 23, pp. 146–52, feb 1980.

[57] R. Kane, "Commentary: functional assessment questionnaire for geriatric patients – or the Clinical Swiss Army knfit," *J Chronic Dis.*, pp. 40(Suppl.), 95S, 1987.

[58] S. Ziebland, R. Fitzpatrick, C. Jenkinson, A. Mowat, A. Mowat, S. R. Ziebland Fitzpatrick, R. C. Fitzpatrick Jenkinson, and A. Mowat Ann Mowat, "Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ," *Annals ofthe Rheumatic Diseases*, vol. 51, pp. 1202–1205, 1992.

[59] D. Marks and B. Chipperfield, "A health and lifestyle survey of Londoners: preliminary data from the Bloomsbury and Islington health and lifestyle survey.," *Symposium on Health Behaviour VIth European Health Psychology Society Conference.*, 1992.

[60] Y. El Miedany, *Patient reported outcome measures in rheumatic diseases*. Springer, 2016.

[61] A.-K. Nilsdotter, E. M. Roos, J. P. Westerlund, H. P. Roos, and L. S. Lohmander, "Comparative responsiveness of measures of pain and function after total hip replacement," *Arthritis & Rheumatism*, vol. 45, pp. 258–262, jun 2001.

[62] M. J. Dunbar, O. Robertsson, L. Ryd, and L. Lidgren, "Appropriate questionnaires for knee arthroplasty. Results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry.," *The Journal of bone and joint surgery. British volume*, vol. 83, pp. 339–44, apr 2001.

[63] N. Black, "Patient reported outcome measures could help transform healthcare What are PROMs?," *BMJ*, vol. 346, 2013.

[64] N. J. Devlin, J. Appleby, and M. Buxton, "Getting the most out of PROMs: putting health outcomes at the heart of NHS decision-making," tech. rep., Office of Health Economics, London, 2010.

[65] N. E. Anderson, M. Calvert, P. Cockwell, M. Dutton, O. L. Aiyegbusi, and D. Kyte, "Using patient-reported outcome measures (PROMs) to promote quality of care in the management of patients with established kidney disease requiring treatment with haemodialysis in the UK (PROM-HD): a qualitative study protocol.," *BMJ Open*, vol. 8, p. e021532, oct 2018.

[66] A. M. Stiggelbout, T. Van der Weijden, M. P. T. De Wit, D. Frosch, F. Légaré, V. M. Montori, L. Trevena, and G. Elwyn, "Shared decision making: really putting patients at the centre of healthcare.," *BMJ*, vol. 344, p. e256, jan 2012.

[67] A. Coulter, "Do patients want a choice and does it work?," *BMJ*, vol. 341, p. c4989, oct 2010.

[68] P. N. Baker, T. Petheram, S. S. Jameson, P. J. Avery, M. R. Reed, P. J. Gregg, D. J. Deehan, and P. J. Avery, "Comparison of patient-reported outcome measures following total and unicondylar knee replacement," *J Bone Joint Surg Br*, vol. 9494, no. 7, pp. 919–27, 2012.

[69] R. Jaeschke, J. Singer, and G. H. Guyatt, "Measurement of health status," *Controlled Clinical Trials*, vol. 10, pp. 407–415, dec 1989.

[70] C. D. Coon and J. C. Cappelleri, "Interpreting change in scores on patient-reported outcome instruments," *Therapeutic Innovation & Regulatory Science*, vol. 50, no. 1, pp. 22–29, 2016.

[71] FDA, "Health and Quality of Life Outcomes Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance," tech. rep., 2006.

[72] M. Kosinski, S. D. Keller, H. T. Hatoum, S. X. Kong, and J. E. Ware, "The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability.," *Medical care*, vol. 37, pp. MS10–22, may 1999.

[73] N. C. Laucis, R. D. Hays, and T. Bhattacharyya, "Scoring the SF-36 in Orthopaedics: A Brief Guide," *BY THE JOURNAL OF BONE AND JOINT SURGERY*, vol. 97, pp. 1628–1662, 2015.

[74] J. N. Katz, M. G. Larson, C. B. Phillips, A. Fossel, and Liang Matthew H, "Comparative Measurement Sensitivity of Short and Longer Health Status Instruments," *Medical Care*, vol. 30, no. 10, pp. 917–925, 1992.

[75] C. A. Jones, D. C. Voaklander, D. W. Johnston, and M. E. Suarez-Almazor, "Health related quality of life outcomes after total hip and knee arthroplasties in a community based population.," *The Journal of rheumatology*, vol. 27, pp. 1745–52, jul 2000.

[76] J. Ware, M. Kosinski, and S. D. Keller, "A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity.," *Medical care*, vol. 34, pp. 220–33, mar 1996.

[77] C. Jenkinson, R. Layte, D. Jenkinson, K. Lawrence, S. Petersen, C. Paice, and J. Stradling, "A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies?," *Journal of Public Health Medicine*, vol. 19, no. 2, pp. 179–186.

[78] O. Rolfson, A. Rothwell, A. Sedrakyan, K. E. Chenok, E. Bohm, K. J. Bozic, and G. Garellick, "Use of patient-reported outcomes in the context of different levels of data.," *The Journal of bone and joint surgery. American volume*, vol. 93 Suppl 3, pp. 66–71, dec 2011.

[79] O. Rolfson, S. Wissig, L. van Maasakkers, C. Stowell, I. Ackerman, D. Ayers, T. Barber, T. Benzakour, K. Bozic, N. Budhiparama, J. Caillouette, P. G. Conaghan, L. Dahlberg, J. Dunn, J. Grady-Benson, S. A. Ibrahim, S. Lewis, H. Malchau, M. Manzary, L. March, N. Nassif, R. Nelissen, N. Smith, and P. D. Franklin, "Defining an International Standard Set of Outcome Measures for Patients With Hip or Knee Osteoarthritis: Consensus of the International Consortium for Health Outcomes Measurement Hip and Knee Osteoarthritis Working Group.," *Arthritis care & research*, vol. 68, no. 11, pp. 1631–1639, 2016.

[80] N. J. Devlin and R. Brooks, "EQ-5D and the EuroQol Group: Past, Present and Future.," *Applied health economics and health policy*, vol. 15, pp. 127–137, apr 2017.

[81] R. Rabin and F. de Charro, "EQ-SD: a measure of health status from the EuroQol Group," *Annals of Medicine*, vol. 33, pp. 337–343, jan 2001.

[82] C. Jenkinson, R. Fitzpatrick, and M. Argyle, "The Nottingham Health Profile: an anlysisi of its sensitivity in different illness groups," *Sor SCI Med*, vol. 27, pp. 1411–1414, 1988.

[83] S. M. Hunt, J. McEWEN, and M. S. R McKENNA, "Measuring health status: a new tool for clinicians and epidemiologists," *Journal of the Royal College of General Practitioners*, vol. 35, pp. 185–188, 1985.

[84] L. T. Nilsson, H. Frazén, A. S. Carlsson, and R. Onnerfält, "Early radiographic

loosening impairs the function of a total hip replacement. The Nottingham Health Profile of 49 patients at five years.," *J. Bone Joint Surg. Br.*, vol. 76, pp. 235–9, mar 1994.

[85] A. R. Feinstein, "The pre-therapeutic classification of co-morbidity in chronic disease," *J. Chronic Dis.*, vol. 23, pp. 455–68, dec 1970.

[86] J. Dawson, R. Fitzpatrick, D. Murray, and A. Carr, "Questionnaire on the perceptions of patients about total knee replacement," *Journal of Bone & Joint Surgery, British Volume*, vol. 80, no. 1, pp. 63–69, 1998.

[87] R. Fitzpatrick and J. Dawson, "Health-related quality of life and the assessment of outcomes of total hip replacement surgery," *Psychology & Health*, vol. 12, pp. 793–803, dec 1997.

[88] K. K. Harris, A. J. Price, D. J. Beard, R. Fitzpatrick, C. Jenkinson, and J. Dawson, "Can pain and function be distinguished in the Oxford Hip Score in a meaningful way? An exploratory and confirmatory factor analysis," *Bone Joint Res.*, vol. 3, pp. 305–9, nov 2014.

[89] K. K. Harris, J. Dawson, L. D. Jones, D. J. Beard, and A. J. Price, "Extending the use of proms in the nhs—using the oxford knee score in patients undergoing non-operative management for knee osteoarthritis: a validation study," *BMJ Open*, vol. 3, no. 8, 2013.

[90] H. Weel, R. Lindeboom, S. E. Kuipers, and T. M. J. S. Vervest, "Comparison between the Harris- and Oxford Hip Score to evaluate outcomes one-year after total hip arthroplasty," vol. 83, pp. 98–109, 2017.

[91] J.-Y. Jenny, P. Louis, and Y. Diesinger, "High Activity Arthroplasty Score has a Lower Ceiling Effect Than Standard Scores After Knee Arthroplasty," *J. Arthroplasty*, vol. 29, pp. 719–721, 2014.

[92] K. Harris, C. R. Lim, J. Dawson, R. Fitzpatrick, D. J. Beard, and A. J. Price, "The Oxford knee score and its subscales do not exhibit a ceiling or a floor effect in knee arthroplasty patients: an analysis of the National Health Service PROMs data set," *Knee Surgery, Sport. Traumatol. Arthrosc.*, vol. 25, pp. 2736–2742, sep 2017.

[93] D. F. Hamilton, J. V. Lane, P. Gaston, J. T. Patton, D. MacDonald, A. H. R. W. Simpson, and C. R. Howie, "What determines patient satisfaction with surgery? a prospective cohort study of 4709 patients following total joint replacement," *BMJ Open*, vol. 3, no. 4, 2013.

[94] N. Bellamy, W. W. Buchanan, C. H. Goldsmith, J. Campbell, and L. W. Stitt, "Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee.," *J. Rheumatol.*, vol. 15, pp. 1833–40, dec 1988.

[95] S. McConnell, P. Kolopack, and A. M. Davis, "The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties," *Arthritis Rheum.*, vol. 45, pp. 453–461, oct 2001.

[96] O. Rolfson, E. Bohm, P. Franklin, S. Lyman, G. Denissen, J. Dawson, J. Dunn, K. E. Chenok, M. Dunbar, S. Overgaard, and G. Garellick, "Patient-reported outcome measures in arthroplasty registries Report of the Patient-Reported Outcome Measures Working Group of the Inter-national Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis," *Acta Orthop.*, vol. 87, pp. 9–23, 2016.

[97] S. L. Whitehouse, E. A. Lingard, Katz J. N, and Learnmonth I. D, "Development and testing of a reduced WOMAC function scale. - PubMed - NCBI," *J*

Bone Jt. Surg Br, vol. 85, no. 5, pp. 706–11, 2003.

[98] E. M. Roos, H. P. Roos, L. S. Lohmander, C. Ekdahl, and B. D. Beynnon, "Knee Injury and Osteoarthritis Outcome Score (KOOS)-Development of a Self-Administered Outcome Measure," *Journal of Orthopaedic & Sports Physical Therapy*, vol. 28, no. 2, pp. 88–96, 1998.

[99] A. K. Nilsdotter, S. Lohmander, M. Klässbo, and E. M. Roos, "Hip disability and osteoarthritis outcome score (HOOS)-validity and responsiveness in total hip replacement," 2003.

[100] E. M. Roos and S. Toksvig-Larsen, "Health and Quality of Life Outcomes Knee injury and Osteoarthritis Outcome Score (KOOS)-validation and comparison to the WOMAC in total knee replacement," 2003.

[101] O. Rolfson, K. E. Chenok, E. Bohm, A. Lübbeke, G. Denissen, J. Dunn, S. Lyman, P. Franklin, M. Dunbar, S. Overgaard, G. Garellick, and J. Dawson, "Patient-reported outcome measures in arthroplasty registries Report of the Patient-Reported Outcome Measures Working Group of the Inter-national Society of Arthroplasty Registries Part I. Overview and rationale for patient-reported outcome measures," *Acta Orthop.*, vol. 87, pp. 3–8, 2016.

[102] D. J. Beard, K. Harris, J. Dawson, H. Doll, D. W. Murray, A. J. Carr, and A. J. Price, "Meaningful changes for the Oxford hip and knee scores after joint replacement surgery.," *Journal of clinical epidemiology*, vol. 68, pp. 73–9, jan 2015.

[103] C. A. Mancuso, T. P. Sculco, T. L. Wickiewicz, E. C. Jones, L. Robbins, R. F. Warren, and P. Williams-Russo, "Patients' Expectations of Knee Surgery," *J. Bone Jt. Surg.*, vol. 83-A, no. 7, pp. 1005–1012, 2001.

[104] P. C. Noble, M. A. Conditt, K. F. Cook, and K. B. Mathis, "The John Insall

Award: Patient expectations affect satisfaction with total knee arthroplasty.," *Clin. Orthop. Relat. Res.*, vol. 452, pp. 35–43, nov 2006.

[105] R. B. Bourne, B. Chesworth, A. Davis, N. Mahomed, and K. Charron, "Comparing patient outcomes after THA and TKA: is there a difference?," *Clinical orthopaedics and related research*, vol. 468, pp. 542–6, feb 2010.

[106] V. Brander, S. Gondek, E. Martin, and S. D. Stulberg, "THE JOHN INSALL AWARD:," *Clin. Orthop. Relat. Res.*, vol. PAP, pp. 21–6, jun 2007.

[107] "Finalised patient reported outcome measures (proms) in england - april 2013 to march 2014." `http://www.hscic.gov.uk/catalogue/PUB17876`.

[108] P. N. Ramkumar, J. D. Harris, P. C. Noble, P. N. Ramkumar, J. D. Harris, O. Surgeon, and P. C. Noble, "Patient-reported outcome measures after total knee arthroplasty," *Bone Jt. Res*, vol. 444, no. 7, 2015.

[109] A. D. Beswick, V. Wylde, R. Gooberman-Hill, A. Blom, and P. Dieppe, "What proportion of patients report long-term pain after total hip or knee replacement for osteoarthritis? A systematic review of prospective studies in unselected patients," *BMJ Open*, vol. 2, p. e000435, feb 2012.

[110] G. A. Hawker, E. M. Badley, C. M. Borkhoff, R. Croxford, A. M. Davis, S. Dunn, M. A. Gignac, S. B. Jaglal, H. J. Kreder, and J. E. M. Sale, "Which patients are most likely to benefit from total joint arthroplasty?," *Arthritis Rheum.*, vol. 65, pp. 1243–52, may 2013.

[111] A. Escobar, M. Gonzalez, J. M. Quintana, K. Vrotsou, A. Bilbao, C. Herrera-Espiñeira, L. Garcia-Perez, F. Aizpuru, and C. Sarasqueta, "Patient acceptable symptom state and OMERACTeOARSI set of responder criteria in joint replacement. Identification of cut-off values," *Osteoarthr. Cartil.*, vol. 20, pp. 87–92, 2012.

[112] R. Jayadevappa, R. Cook, and S. Chhatre, "Minimal important difference to infer changes in health-related quality of life—a systematic review," *J. Clin. Epidemiol.*, vol. 89, pp. 188–198, sep 2017.

[113] N. Black and Others, "Patient reported outcome measures could help transform healthcare," *BMJ*, vol. 346, p. f167, jan 2013.

[114] G. R. Norman, J. A. Sloan, and K. W. Wyrwich, "Point/Counterpoint Interpretation of Changes in Health-related Quality of Life The Remarkable Universality of Half a Standard Deviation," *Official Journal of the Medical Care Section, American Public Health Association*, vol. 41, no. 5, pp. 582–592, 2003.

[115] A. Judge, N. K. Arden, A. Kiran, A. Price, M. K. Javaid, D. Beard, D. Murray, and R. E. Field, "Interpretation of patient-reported outcomes for hip and knee replacement surgery: identification of thresholds associated with satisfaction with surgery.," *The Journal of bone and joint surgery. British volume*, vol. 94, pp. 412–8, mar 2012.

[116] T. Pham, D. van der Heijde, R. D. Altman, J. J. Anderson, N. Bellamy, M. Hochberg, L. Simon, V. Strand, T. Woodworth, and M. Dougados, "OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited.," *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society*, vol. 12, pp. 389–99, may 2004.

[117] A. Escobar, J. Quintana, A. Bilbao, I. Aróstegui, I. Lafuente, and I. Vidaurreta, "Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement," *Osteoarthritis and Cartilage*, vol. 15, pp. 273–280, mar 2007.

[118] J. Quintana, A. Escobar, A. Bilbao, I. Arostegui, I. Lafuente, and I. Vidau-

rreta, "Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement," *Osteoarthritis and Cartilage*, vol. 13, pp. 1076–1083, dec 2005.

[119] F. A. Frans, P. T. Nieuwkerk, R. Met, S. Bipat, D. A. Legemate, J. A. Reekers, and M. J. W. Koelemay, "Statistical or Clinical Improvement? Determining the Minimally Important Difference for the Vascular Quality of Life Questionnaire in Patients with Critical Limb Ischemia," *European Journal of Vascular and Endovascular Surgery*, vol. 47, no. 2, pp. 180–186, 2012.

[120] W. P. Maksymowych, R. Richardson, C. Mallon, D. van der Heijde, and A. Boonen, "Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing spondylitis," *Arthritis & Rheumatism*, vol. 57, pp. 133–139, feb 2007.

[121] P. Tugwell, M. Boers, P. Brooks, L. Simon, V. Strand, and L. Idzerda, "OMERACT: An international initiative to improve outcome measurement in rheumatology," *Trials Journal*, 2007.

[122] "Mission & Vision — Osteoarthritis Research Society International (OARSI)."

[123] "Finalised Patient Reported Outcome Measures (PROMs) in England: April 2012 to March 2013," tech. rep., aug 2014.

[124] O. Ethgen, O. Bruyère, F. Richy, C. Dardennes, and J.-Y. Reginster, "Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature.," *J. Bone Joint Surg. Am.*, vol. 86-A, pp. 963–74, may 2004.

[125] P. Rissanen, S. Aro, P. Slätis, H. Sintonen, and P. Paavolainen, "Health and quality of life before and after hip or knee arthroplasty.," *J. Arthroplasty*, vol. 10, pp. 169–75, apr 1995.

[126] V. Wylde, P. Dieppe, S. Hewlett, and I. D. Learmonth, "Total knee replacement: is it really an effective procedure for all?," *The Knee*, vol. 14, pp. 417–23, dec 2007.

[127] N. Arden, A. Kiran, A. Judge, L. Biant, M. Javaid, D. Murray, A. Carr, C. Cooper, and R. Field, "What is a good patient reported outcome after total hip replacement?," *Osteoarthritis and Cartilage*, vol. 19, no. 2, pp. 155–162, 2011.

[128] F. Angst, A. Aeschlimann, B. A. Michel, and G. Stucki, "Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities.," *J Rheumatol*, vol. 29, pp. 131–138, jan 2002.

[129] J. M. Giesinger, D. F. Hamilton, B. Jost, H. Behrend, and K. Giesinger, "WOMAC, EQ-5D and Knee Society Score Thresholds for Treatment Success After Total Knee Arthroplasty.," *J. Arthroplasty*, vol. 30, pp. 2154–8, dec 2015.

[130] P. N. Baker, J. H. van der Meulen, J. Lewsey, P. J. Gregg, and National Joint Registry for England and Wales, "The role of pain and function in determining patient satisfaction after total knee replacement. Data from the National Joint Registry for England and Wales.," *The Journal of bone and joint surgery. British volume*, vol. 89, pp. 893–900, jul 2007.

[131] W. G. Harcourt, S. H. White, and P. Jones, "Specificity of the Oxford knee status questionnaire. The effect of disease of the hip or lumbar spine on patients' perception of knee disability.," *The Journal of bone and joint surgery. British volume*, vol. 83, pp. 345–7, apr 2001.

[132] I. Rombach, C. Jenkinson, A. Gray, D. Murray, and O. Rivero-Arias, "Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials," *Patient Related Out-*

come Measures, vol. 9, pp. 197–209, jun 2018.

[133] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.," *BMJ*, vol. 338, p. b2393, jun 2009.

[134] P. Dieppe, K. Lim, and S. Lohmander, "Who should have knee joint replacement surgery for osteoarthritis?," *International Journal of Rheumatic Diseases*, vol. 14, pp. 175–180, may 2011.

[135] T. Lange, E. Rataj, C. Kopkow, J. Lützner, K.-P. Günther, and J. Schmitt, "Outcome Assessment in Total Knee Arthroplasty: A Systematic Review and Critical Appraisal.," *The Journal of arthroplasty*, vol. 32, pp. 653–665.e1, feb 2017.

[136] O. Robertsson, M. Dunbar, T. Pehrsson, K. Knutson, and L. Lidgren, "Patient satisfaction after knee arthroplasty: A report on 27,372 knees operated on between 1981 and 1995 in Sweden," *Acta Orthopaedica Scandinavica*, vol. 71, pp. 262–267, jan 2000.

[137] M. A. Ritter, A. E. Thong, K. E. Davis, M. E. Berend, J. B. Meding, and P. M. Faris, "Long-term deterioration of joint evaluation scores.," *The Journal of bone and joint surgery. British volume*, vol. 86, pp. 438–42, apr 2004.

[138] R. Gandhi, H. Dhotar, F. Razak, P. Tso, J. R. Davey, and N. N. Mahomed, "Predicting the longer term outcomes of total knee arthroplasty," *The Knee*, vol. 17, pp. 15–18.

[139] D. P. Williams, C. M. Blakey, S. G. Hadfield, D. W. Murray, A. J. Price, and R. E. Field, "Long-term trends in the Oxford knee score following total knee replacement.," *The bone & joint journal*, vol. 95-B, pp. 45–51, jan 2013.

[140] A.-K. Nilsdotteryz, S. Toksvig-Larsenz, and E. M. Rooszx, "A 5 year prospective study of patient-relevant outcomes after total knee replacement," *Osteoarthritis and Cartilage*, vol. 17, pp. 601–606.

[141] KAT Trial Group, L. Johnston, G. MacLennan, K. McCormack, C. Ramsay, and A. Walker, "The Knee Arthroplasty Trial (KAT)," *The Journal of Bone and Joint Surgery-American Volume*, vol. 91, pp. 134–141, jan 2009.

[142] D. W. Murray, G. S. MacLennan, S. Breeman, H. A. Dakin, L. Johnston, M. K. Campbell, A. M. Gray, N. Fiddian, R. Fitzpatrick, R. W. Morris, and A. M. Grant, "A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT)," *Health Technology Assessment*, vol. 18, mar 2014.

[143] P. Rissanen, S. Aro, H. Sintonen, Slatis P, and Paavolainen P, "Quality of Life and Functional Ability in Hip and Knee Replacements : A prospective Study," *Quality of life Research*, vol. 5, no. 1, pp. 56–64, 1996.

[144] M.M. Dowsey, A.J. Smith, and P.F.M. Choong, "Latent Class Growth Analysis predicts long term pain and function trajectories in total knee arthroplasty: a study of 689 patients," *Osteoarthritis and Cartilage*, vol. 23, pp. 2141–2149, 2015.

[145] A. M. Garratt, S. Brealey, W. J. Gillespie, and DAMASK Trial Team, "Patient-assessed health instruments for the knee: a structured review," *Rheumatology*, vol. 43, pp. 1414–1423, nov 2004.

[146] P. G. Conaghan, M. Emerton, and A. Tennant, "Internal construct validity of the Oxford Knee Scale: Evidence from Rasch measurement," *Arthritis & Rheumatism*, vol. 57, pp. 1363–1367, dec 2007.

[147] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–592, dec 1976.

[148] B. C. Johnston, S. Ebrahim, A. Carrasco-Labra, T. A. Furukawa, D. L. Patrick, M. W. Crawford, B. R. Hemmelgarn, H. J. Schunemann, G. H. Guyatt, and G. Nesrallah, "Minimally important difference estimates and methods: a protocol.," *BMJ open*, vol. 5, no. 10, 2015.

[149] C. L. Petersen, J. B. Kjærsgaard, N. Kjærgaard, M. U. Jensen, and M. B. Laursen, "Thresholds for Oxford Knee Score after total knee replacement surgery: a novel approach to post-operative evaluation," *J. Orthop. Surg. Res.*, vol. 12, p. 89, dec 2017.

[150] N. D. Clement, D. MacDonald, and A. H. R. W. Simpson, "The minimal clinically important difference in the Oxford knee score and Short Form 12 score after total knee arthroplasty," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 22, pp. 1933–1939, aug 2014.

[151] J. M. Giesinger, D. F. Hamilton, B. Jost, H. Behrend, and K. Giesinger, "WOMAC, EQ-5D and Knee Society Score Thresholds for Treatment Success After Total Knee Arthroplasty," *The Journal of Arthroplasty*, vol. 30, pp. 2154–2158, dec 2015.

[152] J. C. Keurentjes, F. R. Van Tol, M. Fiocco, C. So-Osman, R. Onstenk, A. W. M. M. Koopman-Van Gemert, R. G. Pöll, and R. G. H. H. Nelissen, "Patient acceptable symptom states after totalhip or knee replacement at mid-term follow-up: Thresholds of the Oxford hip and knee scores.," *Bone & joint research*, vol. 3, no. 1, pp. 7–13, 2014.

[153] F. Tubach, P. Ravaud, G. Baron, B. Falissard, I. Logeart, N. Bellamy, C. Bombardier, D. Felson, M. Hochberg, D. van der Heijde, and M. Dougados, "Eval-

uation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement.," *Annals of the rheumatic diseases*, vol. 64, pp. 29–33, jan 2005.

[154] D. Revicki, R. D. Hays, D. Cella, and J. Sloan, "Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes.," *Journal of clinical epidemiology*, vol. 61, pp. 102–9, feb 2008.

[155] N. Carter, D. Bryant-Lukosius, A. DiCenso, J. Blythe, and A. J. Neville, "The Use of Triangulation in Qualitative Research," *Oncology Nursing Forum*, vol. 41, pp. 545–547, sep 2014.

[156] "Kendall's Tau." Statistical Research – Random Statistics and Data Science, `https://statistical-research.com/wp-content/uploads/2012/09/kendall-tau1.pdf`, Accessed: 2018-04-16.

[157] Douglas G. Altman, *Practical Statistics For Medical Research*. London: Chapman & Hall, first ed., 1991.

[158] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research.," *Malawi medical journal : the journal of Medical Association of Malawi*, vol. 24, pp. 69–71, sep 2012.

[159] D. Heeger, "Signal Detection Theory." `http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html`, Accessed: 2018-06-20.

[160] F. Tubach, M. Dougados, B. Falissard, G. Baron, I. Logeart, and P. Ravaud, "Feeling good rather than feeling better matters more to patients," *Arthritis & Rheumatism*, vol. 55, pp. 526–530, aug 2006.

[161] R. L. Barrack, E. L. Ruh, J. Chen, A. V. Lombardi, K. R. Berend, J. Parvizi, C. J. Della Valle, W. G. Hamilton, and R. M. Nunley, "Impact of socioeco-

nomic factors on outcome of total knee arthroplasty," *Clinical Orthopaedics and Related Research®*, vol. 472, no. 1, pp. 86–97, 2014.

[162] E. A. Lingard, C. B. Sledge, I. D. Learmonth, K. O. Group, *et al.*, "Patient expectations regarding total knee arthroplasty: differences among the united states, united kingdom, and australia," *J Bone Joint Surg Am*, vol. 88, no. 6, pp. 1201–1207, 2006.

[163] C. E. H. Scott, C. R. Howie, D. MacDonald, and L. C. Biant, "Predicting dissatisfaction following total knee replacement," *Bone & Joint Journal*, vol. 92-B, no. 9, pp. 1253–1258, 2010.

[164] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaume Associates, second ed., 1988.

[165] A. J. Pearse, G. J. Hooper, A. Rothwell, C. Frampton, A. J. Pearse, G. J. Hooper, A. Rothwell, E. Professor, and C. Frampton, "Survival and functional outcome after revision of a unicompartmental to a total knee replacement THE NEW ZEALAND NATIONAL JOINT REGISTRY," *J Bone Joint Surg [Br]*, vol. 92, pp. 508–12, 2010.

[166] A. J. Price, J. Webb, F. H. Topf, C. A. F Dodd, J. W. Goodfellow, D. W. Murray, the Oxford Hip, and K. Group, "Rapid Recovery After Oxford Unicompartmental Arthroplasty Through a Short Incision,"

[167] N. M. Brown, N. P. Sheth, K. Davis, M. E. Berend, A. V. Lombardi, K. R. Berend, and C. J. Della Valle, "Total Knee Arthroplasty Has Higher Postoperative Morbidity Than Unicompartmental Knee Arthroplasty: A Multicenter Analysis," *The Journal of Arthroplasty*, vol. 27, pp. 86–90, sep 2012.

[168] C. A. Willis-Owen, K. Brust, H. Alsop, M. Miraldo, and J. P. Cobb, "Unicondylar knee arthroplasty in the UK National Health Service: An analysis

of candidacy, outcome and cost efficacy," *The Knee*, vol. 16, pp. 473–478, dec 2009.

[169] G. R. Scuderi and A. J. Tria, *MIS of the Hip and the Knee: A clinical Perspective*. New York: Springer-Verlag, 2004.

[170] T. Borus and T. Thornhill, "Unicompartmental knee arthroplasty.," *The Journal of the American Academy of Orthopaedic Surgeons*, vol. 16, pp. 9–18, jan 2008.

[171] D. Bruni, F. Iacono, I. Akkawi, M. Gagliardi, S. Zaffagnini, and M. Marcacci, "Unicompartmental knee replacement: a historical overview.," *Joints*, vol. 1, no. 2, pp. 45–7, 2013.

[172] K. A. N. Saldanha, G. W. Keys, U. C. G. Svard, S. H. White, and C. Rao, "Revision of Oxford medial unicompartmental knee arthroplasty to total knee arthroplasty — results of a multicentre study," *The Knee*, vol. 14, no. 4, pp. 275–279, 2007.

[173] A. Arirachakaran, P. Choowit, C. Putananon, S. Muangsiri, and J. Kongtharvonskul, "Is unicompartmental knee arthroplasty (UKA) superior to total knee arthroplasty (TKA)? A systematic review and meta-analysis of randomized controlled trial," *European Journal of Orthopaedic Surgery & Traumatology*, vol. 25, pp. 799–806, jul 2015.

[174] S. Johnson, P. Jones, and J. H. Newman, "The survivorship and results of total knee replacements converted from unicompartmental knee replacements," *The Knee*, vol. 14, pp. 154–157, mar 2007.

[175] D. A. Fisher, M. Watts, and K. E. Davis, "Implant Position in Knee Surgery and Total Knee Arthroplasty," *The Journal of Arthroplasty*, vol. 18, no. 7 Suppl 1, pp. 2–8, 2003.

[176] D. Beard, A. Price, J. Cook, R. Fitzpatrick, A. Carr, M. Campbell, H. Doll, H. Campbell, N. Arden, C. Cooper, L. Davies, and D. Murray, "Total or Partial Knee Arthroplasty Trial - TOPKAT: study protocol for a randomised controlled trial," *Trials*, vol. 14, p. 292, sep 2013.

[177] M. Baker, "Statisticians issue warning over misuse of P values," *Nature*, vol. 531, pp. 151–151, mar 2016.

[178] H. H. Friedman and T. Amoo, "Rating the Rating Scales," *Friedman, Hershey H. and Amoo, Taiwo (1999)." Rating the Rating Scales." Journal of Marketing Management, Winter*, pp. 114–123, jan 1999.

[179] C. Lavernia, M. D'Apuzzo, M. D. Rossi, and D. Lee, "Is Postoperative Function After Hip or Knee Arthroplasty Influenced by Preoperative Functional Levels?," *The Journal of Arthroplasty*, vol. 24, pp. 1033–1043, oct 2009.

[180] P. R. Fortin, J. R. Penrod, A. E. Clarke, Y. St-Pierre, L. Joseph, P. Bélisle, M. H. Liang, D. Ferland, C. B. Phillips, N. Mahomed, M. Tanzer, C. Sledge, A. H. Fossel, and J. N. Katz, "Timing of total joint replacement affects clinical outcomes among patients with osteoarthritis of the hip or knee," *Arthritis & Rheumatism*, vol. 46, pp. 3327–3330, dec 2002.

[181] A. Judge, N. K. Arden, C. Cooper, M. Kassim Javaid, A. J. Carr, R. E. Field, and P. A. Dieppe, "Predictors of outcomes of total knee replacement surgery.," *Rheumatology (Oxford, England)*, vol. 51, pp. 1804–13, oct 2012.

[182] R. K. Elmallah, J. J. Cherian, K. Robinson, S. F. Harwin, and M. A. Mont, "No Title," *Journal of Knee Surgery*, vol. 28, no. 05, pp. 411–416, 2015.

[183] P. R. Fortin, A. E. Clarke, L. Joseph, M. H. Liang, M. Tanzer, D. Ferland, C. Phillips, A. J. Partridge, P. Belisle, A. H. Fossel, N. Mahomed, C. B. Sledge, and J. N. Katz, "Outcomes of total hip and knee replacement: Preoperative

functional status predicts outcomes at six months after surgery," *Arthritis & Rheumatism*, vol. 42, pp. 1722–1728, aug 1999.

[184] M. E. Fernandez-Cuadros, O. S. Perez-Moro, and J. A. Miron-Canelo, "Age and Comorbidities Affect Quality of Life in Patients With Osteoarthrtitis and Knee Replacement," *Middle East Journal of Rehabilitation and Health*, vol. 3, aug 2016.

[185] F. L. Loth, J. M. Giesinger, K. Giesinger, D. J. MacDonald, A. H. R. Simpson, C. R. Howie, and D. F. Hamilton, "Impact of Comorbidities on Outcome After Total Hip Arthroplasty," *The Journal of Arthroplasty*, vol. 32, pp. 2755–2761, sep 2017.

[186] C. A. Jones, D. C. Voaklander, and M. E. Suarez-Alma, "Determinants of function after total knee arthroplasty.," *Physical therapy*, vol. 83, pp. 696–706, aug 2003.

[187] S. V. Onsem, C. Van Der Straeten, N. Arnout, P. Deprez, G. V. Damme, and J. Victor, "A New Prediction Model for Patient Satisfaction After Total Knee Arthroplasty," *The Journal of Arthroplasty*, vol. 31, no. 12, pp. 2660 – 2667, 2016.

[188] Y.-J. Choi and H. J. Ra, "Patient Satisfaction after Total Knee Arthroplasty.," *Knee surgery & related research*, vol. 28, pp. 1–15, mar 2016.

[189] J. B. Mistry, M. Chughtai, R. K. Elmallah, S. Le, P. M. Bonutti, R. E. Delanois, and M. A. Mont, "What Influences How Patients Rate Their Hospital After Total Hip Arthroplasty?," *J. Arthroplasty*, 2016.

[190] H. Beckerman, M. E. Roebroeck, G. J. Lankhorst, J. G. Becher, P. D. Bezemer, and A. L. Verbeek, "Smallest real difference, a link between reproducibility and responsiveness.," *Quality of life research : an international journal of quality*

of life aspects of treatment, care and rehabilitation, vol. 10, no. 7, pp. 571–8, 2001.

[191] M. T. King, "A point of minimal important difference (MID): a critique of terminology and methods," *PharmacoEconomics Outcome Research*, vol. 11, no. 2, pp. 171–184, 2011.

[192] H. C. de Vet, C. B. Terwee, R. W. Ostelo, H. Beckerman, D. L. Knol, and L. M. Bouter, "Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change.," *Health and quality of life outcomes*, vol. 4, p. 54, aug 2006.

[193] A. Bowling, "Just one question: If one question works, why ask several?," *Journal of epidemiology and community health*, vol. 59, pp. 342–5, may 2005.

[194] R. A. Keith, "Functional Status and Health Status," vol. 75, pp. 478–483, 1994.

[195] G. Norman, "Hi! How are you? Response shift, implicit theories and differing epistemologies," *Qual. Life Res.*, vol. 12, no. 3, pp. 239–249, 2003.

[196] R. D. Hays, S. S. Farivar, and H. Liu, "Approaches and recommendations for estimating minimally important differences for health-related quality of life measures.," *COPD*, vol. 2, pp. 63–7, mar 2005.

[197] H. Kang, "The prevention and handling of the missing data," *Korean J Anesthesiol*, no. 5, pp. 402–406, 2013.

[198] N. J. Horton and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models.," *The American statistician*, vol. 61, pp. 79–90, feb 2007.

[199] E. Alpaydin, *Introduction to machine learning.* Cambridge, Massachusetts: The MIT Press, second ed., 2010.

[200] N. Graves, C. Wloch, J. Wilson, A. Barnett, A. Sutton, N. Cooper, K. Merollini, V. McCreanor, Q. Cheng, E. Burn, T. Lamagni, and A. Charlett, *A cost-effectiveness modelling study of strategies to reduce risk of infection following primary hip replacement based on a systematic review.*, vol. 20. NIHR Journals Library, 2016.

[201] D. A. Clifton, *Machine Learning for Healthcare Technologies.* Institution of Engineering and Technology, 2016.

[202] M. Pikoula, J. Quint, K. Direk, F. Nissen, A. Gonzalez-lzquierdo, H. Hemingway, L. Smeeth, and S. Denaxas, "Discovery of novel COPD subtypes via cluster analysis of clinical features extracted from electronic health records," *American Journal of Respiratory and Critical Care Medicine*, vol. 197, 2018.

[203] R. Kagawa, Y. Kawazoe, Y. Ida, E. Shinohara, K. Tanaka, T. Imai, and K. Ohe, "Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach.," *Journal of diabetes science and technology*, vol. 11, no. 4, pp. 791–799, 2017.

Appendix A

Chapter Four supplementary material

A.1 Ethics approval of the knee arthroplasty trial (KAT) study

Approval for the KAT study was obtained from the Multi-Research Ethics Committee for Scotland (research protocol MREC/98/0/100 – November 1998, ISRCTN45837371) and from the Local Research Ethics Committees at each study site recruiting participants. All the participants gave written informed consent prior to being included in the study. No additional ethics approval was required for conducting the study described in Chapter Four.

A.2 Analysis of data containing missing values

As explained in Chapter Three, missing data is a common problem in clinical research, and the three main types of missing data include missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In the case with MCAR, removing missing values and analysing the observed data will produce unbiased results. Unbiased results can sometimes be obtained for MAR data. For MNAR data, excluding missing data in analysis always produce biased results. It is difficult to determine the underlying mechanism of missing data by using observed data only. In the case with KAT data, it is reasonable to assume the nature of missing data to be MAR [132].

A.2.1 Analytical methods used in this study

- **Analysis of cases with incomplete data** – Data from patients with at least three years of OKS follow-up data were included in the analysis (n=1359); results are shown in Chapter Four.
- **Complete-case data analysis** – To verify the results obtained from incomplete-case analysis, all longitudinal analyses were repeated using a subset of patients (n=449), for whom all tens years of OKS follow-up data were available.

A.2.2 Results from complete-case analysis

A.2.2.1 Fluctuation in long-term post-TKR outcome

Complete-case analysis was performed in order to verify the results regarding long-term outcome of patients who achieved a **successful** outcome (ΔOKS $\geqslant$11) at

one-year follow up (shown in Figure 4.4). The results are presented in Figure A.1, which shows the exact same pattern as Figure 4.4.

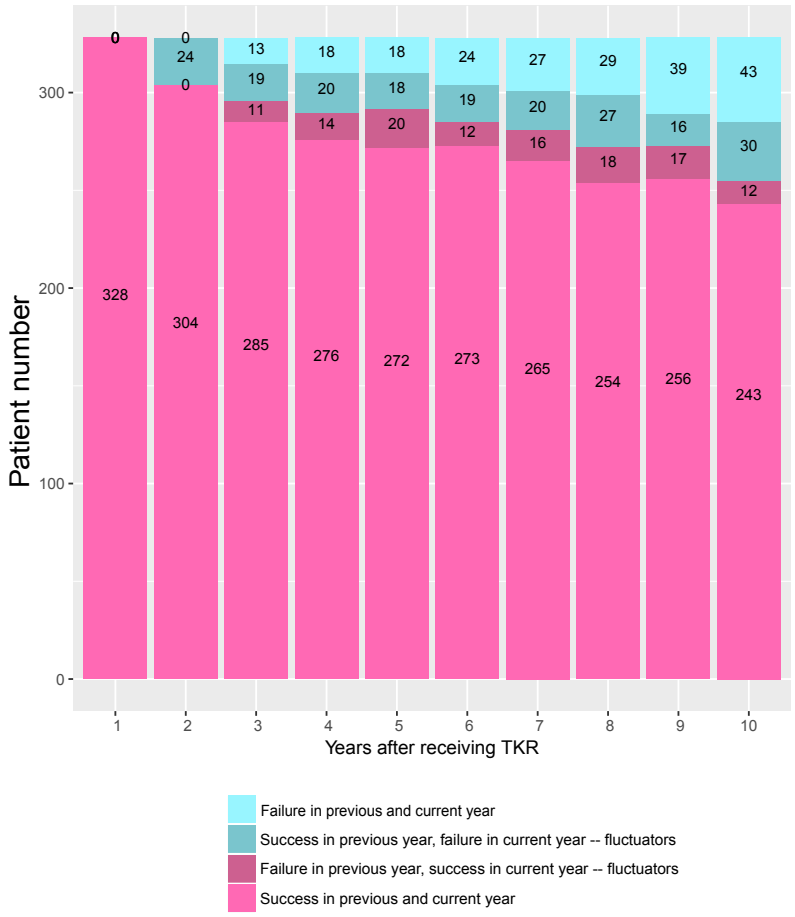


Figure A.1: Bar chart showing the long-term outcome of the patients who achieved a successful outcome at one-year follow up. (complete-case analysis, n=328)

Complete-case analysis was also performed to verify the results regarding long-term outcome of patients who **failed** to achieve a successful outcome at one-year follow up (shown in Figure 4.5). The results are presented in Figure A.2, which bears resemblance to Figure 4.5.

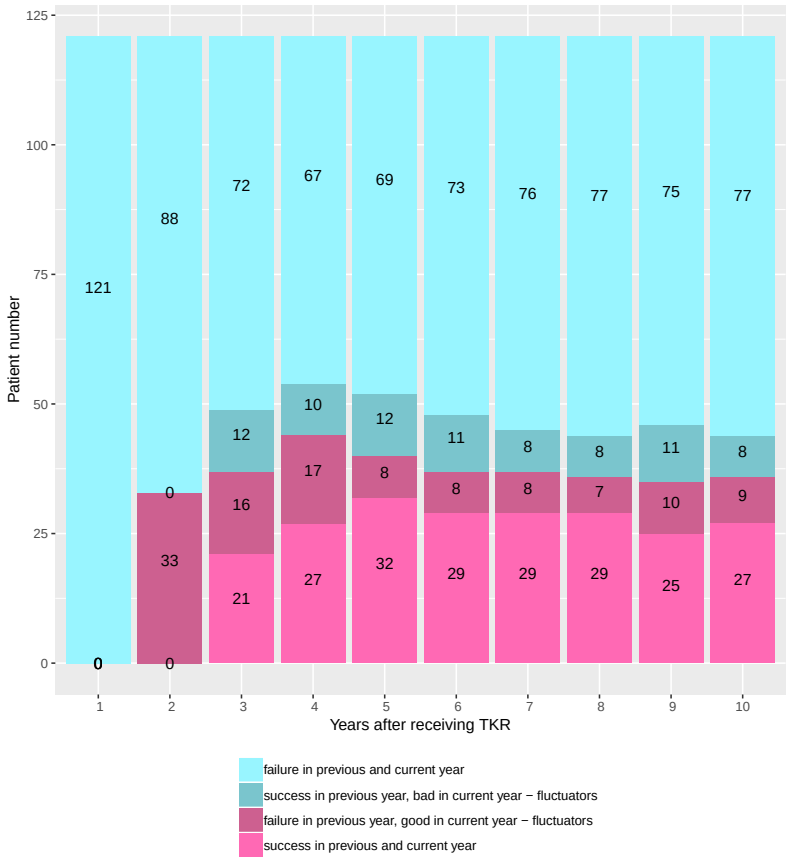


Figure A.2: Bar chart showing the long-term outcome of patients who had a "failed" outcome at one-year post-TKR. (complete-case analysis (n=121)

A.2.2.2 When do patients achieve sufficient response following TKR

To verify the results regarding the distribution of the earliest time patients achieve a sufficient response (shown in figure 4.6), the same analysis was repeated with a subset of 402 patients – those with a complete set (all ten years) of follow-up data. Results obtained from complete-case analysis are highly similar to the incomplete-case analysis, details are shown in Table A.1.

Total sample size	Years after TKR									
	1	2	3	4	5	6	7	8	9	10
n=1205	81.1%	8.5%	3.8%	2.3%	1.4%	0.8%	1.1%	0.4%	0.4%	0.1%
n=402	81.6%	8.2%	4.0%	2.5%	0.7%	0.5%	1.2%	0.5%	0.5%	0.2%

Table A.1: Proportion achieving a sufficient response for the first time each year (amongst those who did meet the PASS criterion some time between year-one and year-ten). Results from complete-case analysis (bottom row, n=402) and from incomplete case analysis (n=1205) are highly similar.

A.2.3 Conclusion

The high level of similarity between results obtained via incomplete-case anlysis and complete-case analysis indicate that patients with complete follow-up data do not behave significantly differently from those with incomplete follow-up data.

Appendix B

Chapter 7 supplementary material

This section presents the following documents regarding the total and partial knee arthroplasty trial (TOPKAT):

- Data request form
- Baseline questionnaire
- One year follow-up questionnaire

Data Request Form

Name of study from which data is requested:

Total or Partial Knee Arthroplasty Trial (TOPKAT)

Person requesting data (including contact details):

Anqi Gao

PhD student under the supervision of Prof David Beard and Prof Andrew Price

NDORMS, University of Oxford,

anqi.gao@ndorms.ox.ac.uk

Wolfson College

OX2 6UD

Reason for request (including background, aims, rationale etc.):

Background:

In order to assess patients' outcome following knee replacement surgery, I have developed a composite metric, which involves combining patients' self-reported transition and satisfaction rating sequentially. This composite metric divides patients into three categorical groups regarding their outcome — "good", "intermediate" and "bad". Using the NHS PROMs datasets (2012-2013, 2013-2014, 2014-2015), our analysis results have consistently shown that patients who are classified into a better outcome group using the composite metric also have greater improvement in their Oxford Knee Score (OKS).

Aims:

1. NHS PROMs datasets are population-based including all TKR patients across England and Wales. However, there may be some inherent difference between patients who agree to participate in clinical trials and those who refuse to. With TOPKAT data, I would be able to validate the usage of this aforementioned composite metric in a clinical trial-based population.
2. The composite metric can be used as an additional outcome measure to assess the clinical effectiveness of TKR in comparison to UKR.

3. Recently, using information from the Knee Arthroplasty Trial (KAT) as the data source, the OKS as the primary outcome measure, and patient acceptable symptom state (PASS) score as the criterion for a "good outcome", we have found that 31% of patients classified as having a "poor outcome" in year-one post-TKR improve significantly and cross-over to the "good outcome" group by the end of the 2nd year. Also, about 10% of patients classified as having a "good outcome" in year-one, become significantly worse by the end of the 2nd year post-op, and cross-over to join the "poor outcome" group. The OKS-based PASS score mainly assesses pain and function, whilst my composite metric addresses patients' satisfaction and self-perceived global improvement. I would like to investigate whether the cross-over phenomenon at the end of the 2nd year post-op still exists when using my composite metric as the criterion for a good outcome, instead of the PASS score. I cannot do this analysis using KAT data, because satisfaction data were not collected during KAT. TOPKAT data would allow me to conduct such an analysis.

Legislative approvals (details of any approvals required):

N/A

Data (details of what data will you need access to, including data items, time-points):

I would like to have all complete, pre- and post-operative data to-date. The specific outcome data I am looking for include: OKS, answers to questions "How satisfied are you with your knee" and "How are the problems related to your knee now, compared to before your knee surgery" (in the "Patient Satisfaction" section), EQ-5D and UCLA-activity score.

Data fields:

Following data fields: **Baseline and 1 year only**

- OKS,
- Questions Patient Satisfaction" section

"How satisfied are you with your knee"

"How are the problems related to your knee now, compared to before your knee surgery"

- EQ-5D
- UCLA-activity score.

Demographic data:

Date of birth

Gender

Time of Onset of OA in knee

All of medical history

Previous treatment on study knee.

Programming implications (for example data extraction, manipulation):

N/A

Proposed statistical analysis (indicate your analysis plans, and who will undertake the analysis):

Anqi Gao will undertake the analysis. Below is the plan:

- Separate patients according to the type of arthroplasty they have received (i.e. TKR vs UKR), then apply the composite metric (explained in the previous sections) to all patients' 1st year post-TKR/UKR data -- calculate the proportion of TKR and UKR patients in each outcome group (namely "good", "intermediate" or "bad"). Compare the prevalence of each outcome category in patients who received TKR versus those who underwent UKR.
- Longitudinally track patients in each outcome group (which was divided based on their 1st year post-op outcome assessed by the composite metric) on a yearly basis, and look for the presence/absence of the cross-over phenomenon (explained in the previous sections).

Publication policy (include details of any planned publication, including proposed authorship, timing of publication, and how collection of the original data will be acknowledged):

Any publication or presentation resulting from the use of data must be sent to the Chief investigator/Trial Management Group prior to submission for their approval.

If things go to plan, I would like to incorporate the analysis results from TOPKAT data into two papers potentially. In that case, I will make sure to acknowledge people involved in TOPKAT.

Proposed start and end date: May 2017- July 2017
Data request was discussed at a Project Management Group Meeting held on: 8 May 2017. As the OKS APQ data is only collected in the year 4 ad 5 yr questionnaires, it was considered too early to access the APQ data. It was agreed that access to the year 1 data would be fine. It was agreed that this data requests did not have to be reviewed by the TSC.
[redacted]

BASELINE QUESTIONNAIRE

Study Number

TOPKAT

NHS National Institute for Health Research

Total Or Partial Knee Arthroplasty Trial

NIHR Health Technology Assessment

CHaRT Centre for Healthcare Randomised Trials

University of Oxford

CONFIDENTIAL

Baseline Questionnaire

ISRCTN03013488

The following questionnaire is broken down into 8 sections. Please work through all sections from start to finish.

Each section asks you to indicate your answers to the question by placing a tick (✓) in the appropriate box. Please read the questions carefully and answer each one as accurately as you can.

There are no right or wrong answers.

Please try to complete the whole questionnaire even though some questions may appear similar.

The answers will be treated with complete confidentiality.

Thank you for your time in completing this questionnaire.

Please start here.

Date questionnaire filled in:

D	D	/	M	M	/	Y	Y	Y	Y

Study Number | | | | | |

Section 1 – Oxford Knee Score

PROBLEMS WITH YOUR KNEE

Please tick ONE box for EVERY question

1. **During the past 4 weeks…**
   How would you describe the pain you <u>usually</u> have from your knee?

None	Very mild	Mild	Moderate	Severe
☐	☐	☐	☐	☐

2. **During the past 4 weeks…**
   Have you had any trouble washing and drying yourself (all over) <u>because of your knee</u>?

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

3. **During the past 4 weeks…**
   Have you had any trouble getting in and out of a car or using public transport <u>because of your knee</u>?(whichever you would tend to use)

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

4. **During the past 4 weeks…**
   For how long have you been able to walk before <u>pain from your knee</u> becomes severe? (with or without a stick)

No pain /More than 30 minutes	16 to 30 minutes	5 to 15 minutes	Around the house only	Not at all /pain severe when walking
☐	☐	☐	☐	☐

5. **During the past 4 weeks…**
   After a meal (sat at a table), how painful has it been for you to stand up from a chair <u>because of your knee</u>?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
☐	☐	☐	☐	☐

6. **During the past 4 weeks…**
   Have you been limping when walking, <u>because of your knee</u>?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

7. During the past 4 weeks…
Could you kneel down and get up again afterwards?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

8. During the past 4 weeks…
Have you been troubled by pain from your knee in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
☐	☐	☐	☐	☐

9. During the past 4 weeks…
How much has pain from your knee interfered with your usual work (including housework)?

Not at all	A little bit	Moderately	Greatly	Totally
☐	☐	☐	☐	☐

10. During the past 4 weeks…
Have you felt that your knee might suddenly 'give way' or let you down?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

11. During the past 4 weeks…
Could you do the household shopping on your own?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

12. During the past 4 weeks…
Could you walk down one flight of stairs?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Section 2 – UCLA Activity Score

Please tick one box that best describes your current activity level

1. Wholly inactive, dependent on others, and cannot leave residence ☐
2. Mostly inactive or restricted to minimum activities of daily living ☐
3. Sometimes participates in mild activities, such as walking, limited housework and limited shopping ☐
4. Regularly participates in mild activities ☐
5. Sometimes participates in moderate activities such as swimming or could do unlimited housework or shopping ☐
6. Regularly participates in moderate activities ☐
7. Regularly participates in active events such as bicycling ☐
8. Regularly participates in active events, such as golf or bowling ☐
9. Sometimes participates in impact sports such as jogging, tennis, skiing, acrobatics, ballet, heavy work or backpacking ☐
10. Regularly participates in impact sports ☐

Section 3 – High Activity Score

Please tick the box that best describes your highest level of function in each of the four categories. Only one box per category can be ticked.

1. Walking

- Over rough ground > 1 hour ☐
- Unlimited on flat, rough ground with difficulty ☐
- Unlimited on flat, no rough ground ☐
- On flat at least 30 minutes ☐
- Short distances unassisted (up to 20m) ☐
- Using walking aids for short distances or worse ☐

2. Running

- More than 5km ☐
- Jog slowly up to 5km ☐
- Run easily across the road ☐
- Run a few steps to avoid traffic if necessary ☐
- Cannot run ☐

3. Stair climbing

- Climb stairs 2 at a time ☐
- Climb without handrail ☐
- Climb with handrail or stick ☐
- Cannot climb stairs ☐

4. Activity Level

- Competitive sports e.g. singles tennis, running >10km, cycling >80km ☐
- Social sports e.g. doubles tennis, skiing, jogging <10km, high impact aerobics ☐
- Vigorous recreational activities e.g. hill-walking, low impact aerobics, heavy gardening or manual work/farming ☐
- Moderate recreational activities e.g. golf, light gardening, light working activities ☐
- Light recreational activities e.g. short walks, lawn bowls ☐
- Required outdoor activities only e.g. short distance to shop ☐
- Housebound without assistance ☐

Section 4 – General Health

1. In general, would you say your health is (*please tick one box only*)

- Excellent ☐
- Very good ☐
- Good ☐
- Fair ☐
- Poor ☐

2. Compared to one year ago, how would you rate your health in general now? (*please tick one box only*)

- Much better now than one year ago ☐
- Somewhat better now than one year ago ☐
- About the same ☐
- Somewhat worse now than one year ago ☐
- Much worse now than one year ago ☐

Section 5 – General Health EUROQOL EQ-5D

For each of the five sets of statements below, please tick the one box that best describes your own health state today.

1. Mobility

I have no problems in walking about ☐

I have some problems in walking about ☐

I am confined to bed ☐

2. Self care

I have no problems with self care ☐

I have some washing or dressing myself ☐

I am unable to wash or dress myself ☐

3. Usual activities (e.g. work, study, housework, family or leisure activities)

I have no problems with performing my usual activities ☐

I have some problems with performing my usual activities ☐

I am unable to perform my usual activities ☐

4. Pain/discomfort

I have no pain or discomfort ☐

I have moderate pain or discomfort ☐

I have extreme pain or discomfort ☐

5. Anxiety/depression

I am not anxious or depressed ☐

I am moderately anxious or depressed ☐

I am extremely anxious or depressed ☐

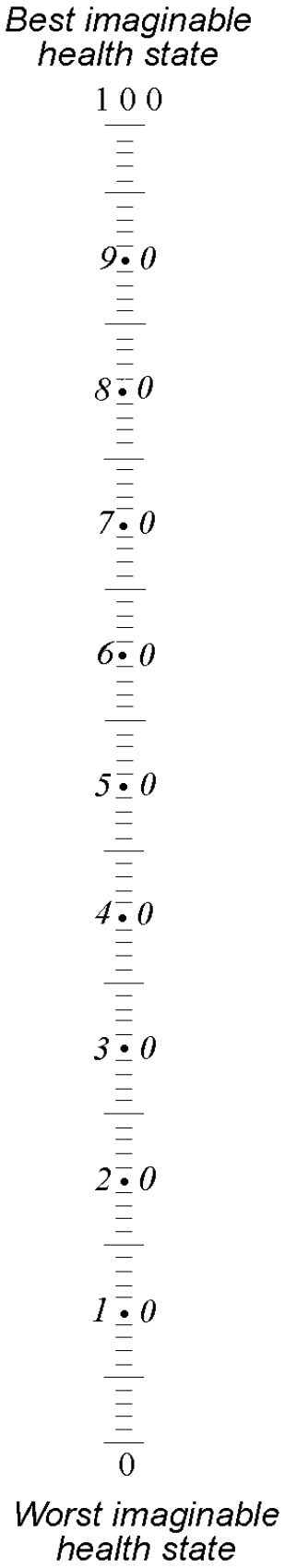
Section 6 – Health State Scale EUROQOL EQ-5D

To help people say how good or bad their health is, we have drawn a scale (rather like a thermometer) on which the **best health state** you can imagine is marked **100** the **worst health state** you can imagine is marked **0.**

We would like you to indicate on this scale how good or bad you think your own health is **today.**

Please do this by drawing a line **from the box below** to whichever point on the scale indicates how good or bad your health state is **today.**

Your own health state today



Section 7 – Health Services (Health Economics)

We would like to know how much contact you have had with the health service **over the last 12 months**. If you are not exactly sure, we would rather have your best guess than no information at all. Please answer every question, even if the answer is "0". For each question we are asking about the knee that is about to be, or has been, operated on in the study.

Please fill in both boxes, for example | 0 | 3 | **if seen three times.**

In the **last 12 months**, how many times have you

1. Seen your GP about your knee? | | |
2. Seen a practice nurse about your knee? | | |
3. Seen a physio or occupational therapist about your knee? | | |
4. Visited a hospital out-patient clinic about your knee? | | |
5. Been in hospital overnight because of your study knee? | | |

5a. If you were in hospital overnight, because of your study knee, please provide details below of each admission to hospital.

Number of nights	Name of hospital	Reason for admission

6. Visited a private practitioner such as an osteopath or chiropractor about your knee? | | |

7. In the **last 12 months** how many medicines or creams have you been prescribed because of your knee? | | |

In the **last 12 months**, approximately how much (to the nearest £) did the following items cost you? If there was no cost, please write "0".

8. Buying painkillers, creams and lotions, dressings or supports as a result of your knee. £ _____

9. Transport, parking, or other costs of visiting the GP or physio, attending exercise clinics, or other health service visits about your knee. £ _____

10. Paying to see private practitioners such as osteopaths or chiropractors about your knee. £ _____

11. If there are any other costs that you have had to pay as a result of your knee in the **last 12 months**, then please use the space below to provide us with details and an approximate figure (to the nearest £).

£ _____

Section 8 – Employment

At present, would you describe yourself as (*please tick one box only*)

Retired ☐

Not in employment ☐

In paid employment ☐

If you are in paid employment, please could you tell us:

What is your job title? _____

How many hours per week do you work? _____

Are you working your usual hours? Yes ☐ No ☐

Are you working reduced hours? Yes ☐ No ☐

Are you off sick? Yes ☐ No ☐

THANK YOU

Thank you very much for your time and patience in filling in this questionnaire. Please hand the questionnaire back to the doctor/nurse who handed it to you.

The information you have given us will be extremely useful in helping us carry out research. It will be treated with the strictest confidence and kept securely.

Thank you again for your help.

If you would like any further information or have any queries about the study, please contact:

The TOPKAT Trial Office in Oxford

Tel: 01865 737210

E-mail: topkat@ndorms.ox.ac.uk

This study is taking place in centres across the UK but the questionnaires are being processed in Aberdeen at the Centre for Healthcare Randomised Trials (CHaRT), Health Services Research Unit, University of Aberdeen, Health Sciences Building, Foresterhill, Aberdeen AB25 2ZD.

12 MONTH FOLLOW-UP QUESTIONNAIRE

Study Number

TOPKAT

NHS National Institute for Health Research

Total Or Partial Knee Arthroplasty Trial

NIHR Health Technology Assessment

CHaRT Centre for Healthcare Randomised Trials

University of Oxford

CONFIDENTIAL

12 Month Follow-up Questionnaire

The following questionnaire is broken down into 9 sections. Please work through all sections from start to finish.

Each section asks you to indicate your answers to the question by placing a tick (✓) in the appropriate box. Please read the questions carefully and answer each one as accurately as you can.

There are no right or wrong answers.

Please try to complete the whole questionnaire even though some questions may appear similar.

The answers will be treated with complete confidentiality.

Thank you for your time in completing this questionnaire.

Please start here.

Date questionnaire filled in:

D	D	/	M	M	/	Y	Y	Y	Y

Section 1 – Oxford Knee Score

PROBLEMS WITH YOUR KNEE

Please tick ONE box for EVERY question

1. During the past 4 weeks…
How would you describe the pain you usually have from your knee?

None	Very mild	Mild	Moderate	Severe
☐	☐	☐	☐	☐

2. During the past 4 weeks…
Have you had any trouble washing and drying yourself (all over) because of your knee?

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

3. During the past 4 weeks…
Have you had any trouble getting in and out of a car or using public transport because of your knee?(whichever you would tend to use)

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
☐	☐	☐	☐	☐

4. During the past 4 weeks…
For how long have you been able to walk before pain from your knee becomes severe? (with or without a stick)

No pain /More than 30 minutes	16 to 30 minutes	5 to 15 minutes	Around the house only	Not at all /pain severe when walking
☐	☐	☐	☐	☐

5. During the past 4 weeks…
After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your knee?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
☐	☐	☐	☐	☐

6. During the past 4 weeks…
Have you been limping when walking, because of your knee?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

7. During the past 4 weeks…
Could you kneel down and get up again afterwards?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

8. During the past 4 weeks…
Have you been troubled by <u>pain from your knee</u> in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
☐	☐	☐	☐	☐

9. During the past 4 weeks…
How much has <u>pain from your knee</u> interfered with your usual work (including housework)?

Not at all	A little bit	Moderately	Greatly	Totally
☐	☐	☐	☐	☐

10. During the past 4 weeks…
Have you felt that your knee might suddenly 'give way' or let you down?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
☐	☐	☐	☐	☐

11. During the past 4 weeks…
Could you do the household shopping <u>on your own</u>?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

12. During the past 4 weeks…
Could you walk down one flight of stairs?

Yes, easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, impossible
☐	☐	☐	☐	☐

Section 2 – UCLA Activity Score

Please tick one box that best describes your current activity level

1. Wholly inactive, dependent on others, and cannot leave residence ☐
2. Mostly inactive or restricted to minimum activities of daily living ☐
3. Sometimes participates in mild activities, such as walking, limited housework and limited shopping ☐
4. Regularly participates in mild activities ☐
5. Sometimes participates in moderate activities such as swimming or could do unlimited housework or shopping ☐
6. Regularly participates in moderate activities ☐
7. Regularly participates in active events such as bicycling ☐
8. Regularly participates in active events, such as golf or bowling ☐
9. Sometimes participates in impact sports such as jogging, tennis, skiing, acrobatics, ballet, heavy work or backpacking ☐
10. Regularly participates in impact sports ☐

Section 3 – High Activity Score

Please tick the box that best describes your highest level of function in each of the four categories. Please tick only one box per category.

1. Walking

- Over rough ground > 1 hour ☐
- Unlimited on flat, rough ground with difficulty ☐
- Unlimited on flat, no rough ground ☐
- On flat at least 30 minutes ☐
- Short distances unassisted (up to 20m) ☐
- Using walking aids for short distances or worse ☐

2. Running

- More than 5km ☐
- Jog slowly up to 5km ☐
- Run easily across the road ☐
- Run a few steps to avoid traffic if necessary ☐
- Cannot run ☐

3. Stair climbing

- Climb stairs 2 at a time ☐
- Climb without handrail ☐
- Climb with handrail or stick ☐
- Cannot climb stairs ☐

4. Activity Level

- Competitive sports e.g. singles tennis, running >10km, cycling >80km ☐
- Social sports e.g. doubles tennis, skiing, jogging <10km, high impact aerobics ☐
- Vigorous recreational activities e.g. hill-walking, low impact aerobics, heavy gardening or manual work/farming ☐
- Moderate recreational activities e.g. golf, light gardening, light working activities ☐
- Light recreational activities e.g. short walks, lawn bowls ☐
- Required outdoor activities only e.g. short distance to shop ☐
- Housebound without assistance ☐

Section 4 – Patient Satisfaction

Side assessed Right ☐ Left ☐

How satisfied are you with your knee? *(please tick only one box)*

Dissatisfied	Uncertain	Satisfied	Very satisfied
☐	☐	☐	☐

How are the problems related to your knee now, compared with before your knee surgery? *(please tick only one box)*

No problems at all now	Much better	Slightly better	No change	Slightly worse	Much worse
☐	☐	☐	☐	☐	☐

If you could go back in time, would you still choose to have the knee operation? *(please tick only one box)*

Yes ☐ No ☐ Not sure ☐

Section 5 – General Health

1. In general, would you say your health is (*please tick one box only*)

Excellent ☐
Very good ☐
Good ☐
Fair ☐
Poor ☐

2. Compared to one year ago, how would you rate your health in general now? (*please tick one box only*)

Much better now than one year ago ☐
Somewhat better now than one year ago ☐
About the same ☐
Somewhat worse now than one year ago ☐
Much worse now than one year ago ☐

Section 6 – General Health EUROQOL EQ-5D

For each of the five sets of statements below, please tick the one box that best describes your own health state today (*please tick one box only*).

1. Mobility

I have no problems in walking about ☐

I have some problems in walking about ☐

I am confined to bed ☐

2. Self care

I have no problems with self care ☐

I have some washing or dressing myself ☐

I am unable to wash or dress myself ☐

3. Usual activities (e.g. work, study, housework, family or leisure activities)

I have no problems with performing my usual activities ☐

I have some problems with performing my usual activities ☐

I am unable to perform my usual activities ☐

4. Pain/discomfort

I have no pain or discomfort ☐

I have moderate pain or discomfort ☐

I have extreme pain or discomfort ☐

5. Anxiety/depression

I am not anxious or depressed ☐

I am moderately anxious or depressed ☐

I am extremely anxious or depressed ☐

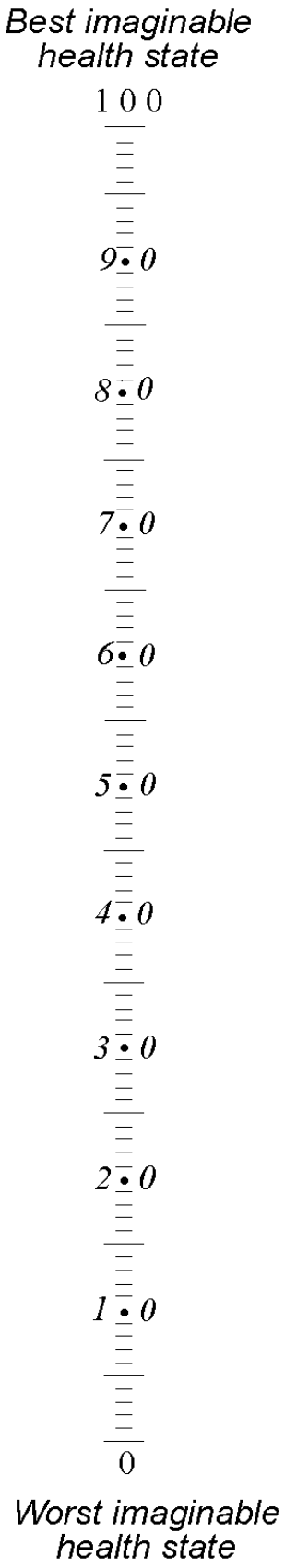
Section 7 – Health State Scale EUROQOL EQ-5D

To help people say how good or bad their health is, we have drawn a scale (rather like a thermometer) on which the **best health state** you can imagine is marked **100** the **worst health state** you can imagine is marked **0.**

We would like you to indicate on this scale how good or bad you think your own health is **today.**

Please do this by drawing a line **from the box below** to whichever point on the scale indicates how good or bad your health state is **today.**

Your own health state today



Section 8 – Health Services (Health Economics)

We would like to know how much contact you have had with the health service **over the last 10 months**. If you are not exactly sure, we would rather have your best guess than no information at all. Please answer every question, even if the answer is "0". For each question we are asking about the knee that is about to be, or has been, operated on in the study.

Please fill in both boxes, for example | 0 | 3 | **if seen three times.**

In the **last 10 months**, how many times have you

1. Seen your GP about your knee? | | |
2. Seen a practice nurse about your knee? | | |
3. Seen a physio or occupational therapist about your knee? | | |
4. Visited a hospital out-patient clinic about your knee? | | |
5. Been in hospital overnight because of your study knee? | | |

5a. If you were in hospital overnight, because of your study knee, please provide details below of each admission to hospital.

Number of nights	Name of hospital	Reason for admission

6. Visited a private practitioner such as an osteopath or chiropractor about your knee? | | |

7. In the **last 10 months**, how many medicines or creams have you been prescribed because of your knee? | | |

In the **last 10 months**, approximately how much (to the nearest £) did the following items cost you? If there was no cost, please write "0".

8. Buying painkillers, creams and lotions, dressings or supports as a result of your knee. £ _____

9. Transport, parking, or other costs of visiting the GP or physio, attending exercise clinics, or other health service visits about your knee. £ _____

10. Paying to see private practitioners such as osteopaths or chiropractors about your knee. £ _____

11. If there are any other costs that you have had to pay as a result of your knee in the **last 10 months**, then please use the space below to provide us with details and an approximate figure (to the nearest £).

£ _____

Section 9 – Employment

At present, would you describe yourself as (*please tick one box only*)

Retired ☐

Not in employment ☐

In paid employment ☐

If you are in paid employment, please could you tell us:

What is your job title? _____

How many hours per week do you work? _____

Are you working your usual hours? Yes ☐ No ☐

Are you working reduced hours? Yes ☐ No ☐

Are you off sick? Yes ☐ No ☐

THANK YOU

Thank you very much for your time and patience in filling in this questionnaire. Please return this questionnaire in the pre-paid envelope provided.

The information you have given us will be extremely useful in helping us carry out research. It will be treated with the strictest confidence and kept securely.

Thank you again for your help.

If you would like any further information or have any queries about the study, please contact:

The TOPKAT Trial Office in Oxford

Tel: 01865 737210

E-mail: topkat@ndorms.ox.ac.uk

This study is taking place in centres across the UK but the questionnaires are being processed in Aberdeen at the Centre for Healthcare Randomised Trials (CHaRT), Health Services Research Unit, University of Aberdeen, Health Sciences Building, Foresterhill, Aberdeen AB25 2ZD.

Appendix C

Instruments for collection of patient-reported outcome data in orthopaedics (hip and knee)

This section presents some of the most widely-distributed patient-reported outcome measures in orthopaedics, including:

- Oxford Hip Score (OHS)
- Western Ontario and McMaster Universities Arthritis Index (WOMAC)
- Knee Injury and Osteoarthritis Outcome Score (KOOS)
- Hip Injury and Osteoarthritis Outcome Score (HOOS)

The Oxford knee score and EuroQol 5-dimension (EQ-5D) were presented in the previous section as part of the TOPKAT One year follow-up questionnaire.

Problems with your hip

During the past 4 weeks..

✓tick <u>one</u> box for <u>every</u> question.

1. ***During the past 4 weeks****.....*

How would you describe the pain you <u>usually</u> had from your hip?

None	Very mild	Mild	Moderate	Severe
❑	❑	❑	❑	❑

2. ***During the past 4 weeks****.....*

Have you had any trouble with washing and drying yourself (all over) <u>because of your hip</u>?

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

3. ***During the past 4 weeks****.....*

Have you had any trouble getting in and out of a car or using public transport <u>because of your hip</u>? *(whichever you tend to use)*

No trouble at all	Very little trouble	Moderate trouble	Extreme difficulty	Impossible to do
❑	❑	❑	❑	❑

4. ***During the past 4 weeks****.....*

Have you been able to put on a pair of socks, stockings or tights?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

5. ***During the past 4 weeks****.....*

Could you do the household shopping <u>on your own</u>?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

6. ***During the past 4 weeks****.....*

For how long have you been able to walk before <u>pain from your hip</u> becomes severe? (*with or without a stick)*

No pain/ More than 30 minutes	16 to 30 minutes	5 to 15 minutes	Around the house <u>only</u>	Not at all -pain severe on walking
❑	❑	❑	❑	❑

The Oxford Hip Score©Department of Public Health, University of Oxford, Old Road Campus, Oxford OX3 7LF , UK.

P.T.O./

During the past 4 weeks...

✓**tick <u>one</u> box for <u>every</u> question**

7

During the past 4 weeks*.....*

Have you been able to climb a flight of stairs?

Yes, Easily	With little difficulty	With moderate difficulty	With extreme difficulty	No, Impossible
❑	❑	❑	❑	❑

8

During the past 4 weeks*.....*

After a meal (sat at a table), how painful has it been for you to stand up from a chair <u>because of your hip</u>?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
❑	❑	❑	❑	❑

9

During the past 4 weeks*.....*

Have you been limping when walking, <u>because of your hip</u>?

Rarely/ never	Sometimes, or just at first	Often, not just at first	Most of the time	All of the time
❑	❑	❑	❑	❑

10

During the past 4 weeks*.....*

Have you had any sudden, <u>severe</u> pain - 'shooting', 'stabbing' or 'spasms' - <u>from the affected hip</u>?

No days	Only 1 or 2 days	Some days	Most days	Every day
❑	❑	❑	❑	❑

11

During the past 4 weeks*.....*

How much has <u>pain from your hip</u> interfered with your usual work *(including housework)*?

Not at all	A little bit	Moderately	Greatly	Totally
❑	❑	❑	❑	❑

12

During the past 4 weeks*.....*

Have you been troubled by <u>pain from your hip</u> in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
❑	❑	❑	❑	❑

©Department of Public Health, University of Oxford, Old Road Campus, Oxford OX3 7LF , UK.

Your Full Name: _____ **Today's Date:**
_____/_____/_____

Month Day Year

WOMAC OSTEOARTHRITIS INDEX

1. The following questions concern the amount of pain you are currently experiencing in your knees. For each situation, please enter the amount of pain you have experienced in the past 48 hours.

		None	mild	moderate	severe	extreme
A. Walking on a flat surface	A.	☐	☐	☐	☐	☐
B. Going up or down stairs	B.	☐	☐	☐	☐	☐
C. At night while in bed	C.	☐	☐	☐	☐	☐
D. Sitting or lying	D.	☐	☐	☐	☐	☐
E. Standing upright	E.	☐	☐	☐	☐	☐

2. Please describe the level of pain you have experienced in the past 48 hours for each one of your knees.

		None	mild	moderate	severe	extreme
A. Right knee	A.	☐	☐	☐	☐	☐
B. Left knee	B.	☐	☐	☐	☐	☐

3. How severe is your stiffness after first awakening in the morning?

None	mild	moderate	severe	extreme
☐	☐	☐	☐	☐

4. How severe is your stiffness after sitting, lying, or resting later in the day?

None	mild	moderate	severe	extreme
☐	☐	☐	☐	☐

5. The following questions concern your physical function. By this we mean your ability to move around and to look after yourself. For each of the following activities, please indicate the degree of difficulty you have experienced in the last 48 hours, in your knees.

What degree of difficulty do you have with:

		None	mild	moderate	severe	extreme
A. Descending (going down) stairs	A.	☐	☐	☐	☐	☐
B. Ascending (going up) stairs	B.	☐	☐	☐	☐	☐
C. Rising from sitting	C.	☐	☐	☐	☐	☐
D. Standing	D.	☐	☐	☐	☐	☐
E. Bending to floor	E.	☐	☐	☐	☐	☐
F. Walking on a flat surface	F.	☐	☐	☐	☐	☐
G. Getting in/out of car	G.	☐	☐	☐	☐	☐
H. Going shopping	H.	☐	☐	☐	☐	☐
I. Putting on socks/stockings	I.	☐	☐	☐	☐	☐
J. Rising from bed	J.	☐	☐	☐	☐	☐
K. Taking off socks/stockings	K.	☐	☐	☐	☐	☐
L. Lying in bed	L.	☐	☐	☐	☐	☐
M. Getting in/out of bath	M.	☐	☐	☐	☐	☐
N. Sitting	N.	☐	☐	☐	☐	☐
O. Getting on/off toile	O.	☐	☐	☐	☐	☐
P. Heavy domestic duties (mowing the lawn, lifting heavy grocery bags)	P.	☐	☐	☐	☐	☐
Q. Light domestic duties (such as tidying a room, dusting, cooking)	Q.	☐	☐	☐	☐	☐

KOOS KNEE SURVEY

Today's date: _____/_____/_____ Date of birth: _____/_____/_____

Name: _____

INSTRUCTIONS: This survey asks for your view about your knee. This information will help us keep track of how you feel about your knee and how well you are able to perform your usual activities.
Answer every question by ticking the appropriate box, only one box for each question. If you are unsure about how to answer a question, please give the best answer you can.

Symptoms

These questions should be answered thinking of your knee symptoms during the **last week**.

S1. Do you have swelling in your knee?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S2. Do you feel grinding, hear clicking or any other type of noise when your knee moves?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S3. Does your knee catch or hang up when moving?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S4. Can you straighten your knee fully?

Always	Often	Sometimes	Rarely	Never
☐	☐	☐	☐	☐

S5. Can you bend your knee fully?

Always	Often	Sometimes	Rarely	Never
☐	☐	☐	☐	☐

Stiffness

The following questions concern the amount of joint stiffness you have experienced during the **last week** in your knee. Stiffness is a sensation of restriction or slowness in the ease with which you move your knee joint.

S6. How severe is your knee joint stiffness after first wakening in the morning?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

S7. How severe is your knee stiffness after sitting, lying or resting **later in the day**?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Pain

P1. How often do you experience knee pain?

Never	Monthly	Weekly	Daily	Always
☐	☐	☐	☐	☐

What amount of knee pain have you experienced the **last week** during the following activities?

P2. Twisting/pivoting on your knee

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P3. Straightening knee fully

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P4. Bending knee fully

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P5. Walking on flat surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P6. Going up or down stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P7. At night while in bed

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P8. Sitting or lying

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P9. Standing upright

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Function, daily living

The following questions concern your physical function. By this we mean your ability to move around and to look after yourself. For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your knee.

A1. Descending stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A2. Ascending stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your knee.

A3. Rising from sitting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A4. Standing

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A5. Bending to floor/pick up an object

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A6. Walking on flat surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A7. Getting in/out of car

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A8. Going shopping

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A9. Putting on socks/stockings

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A10. Rising from bed

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A11. Taking off socks/stockings

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A12. Lying in bed (turning over, maintaining knee position)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A13. Getting in/out of bath

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A14. Sitting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A15. Getting on/off toilet

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your knee.

A16. Heavy domestic duties (moving heavy boxes, scrubbing floors, etc)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A17. Light domestic duties (cooking, dusting, etc)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Function, sports and recreational activities

The following questions concern your physical function when being active on a higher level. The questions should be answered thinking of what degree of difficulty you have experienced during the **last week** due to your knee.

SP1. Squatting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP2. Running

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP3. Jumping

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP4. Twisting/pivoting on your injured knee

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP5. Kneeling

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Quality of Life

Q1. How often are you aware of your knee problem?

Never	Monthly	Weekly	Daily	Constantly
☐	☐	☐	☐	☐

Q2. Have you modified your life style to avoid potentially damaging activities to your knee?

Not at all	Mildly	Moderately	Severely	Totally
☐	☐	☐	☐	☐

Q3. How much are you troubled with lack of confidence in your knee?

Not at all	Mildly	Moderately	Severely	Extremely
☐	☐	☐	☐	☐

Q4. In general, how much difficulty do you have with your knee?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Thank you very much for completing all the questions in this questionnaire.

HOOS HIP SURVEY

Today's date: _____/_____/_____ Date of birth: _____/_____/_____

Name: _____

INSTRUCTIONS: This survey asks for your view about your hip. This information will help us keep track of how you feel about your hip and how well you are able to do your usual activities.
Answer every question by ticking the appropriate box, only one box for each question. If you are uncertain about how to answer a question, please give the best answer you can.

Symptoms

These questions should be answered thinking of your hip symptoms and difficulties during the **last week**.

S1. Do you feel grinding, hear clicking or any other type of noise from your hip?

Never	Rarely	Sometimes	Often	Always
☐	☐	☐	☐	☐

S2. Difficulties spreading legs wide apart

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

S3. Difficulties to stride out when walking

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Stiffness

The following questions concern the amount of joint stiffness you have experienced during the **last week** in your hip. Stiffness is a sensation of restriction or slowness in the ease with which you move your hip joint.

S4. How severe is your hip joint stiffness after first wakening in the morning?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

S5. How severe is your hip stiffness after sitting, lying or resting **later in the day**?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Pain

P1. How often is your hip painful?

Never	Monthly	Weekly	Daily	Always
☐	☐	☐	☐	☐

What amount of hip pain have you experienced the **last week** during the following activities?

P2. Straightening your hip fully

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

What amount of hip pain have you experienced the **last week** during the following activities?

P3. Bending your hip fully

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P4. Walking on a flat surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P5. Going up or down stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P6. At night while in bed

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P7. Sitting or lying

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P8. Standing upright

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P9. Walking on a hard surface (asphalt, concrete, etc.)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

P10. Walking on an uneven surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Function, daily living

The following questions concern your physical function. By this we mean your ability to move around and to look after yourself. For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your hip.

A1. Descending stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A2. Ascending stairs

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A3. Rising from sitting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A4. Standing

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

For each of the following activities please indicate the degree of difficulty you have experienced in the **last week** due to your hip.

A5. Bending to the floor/pick up an object

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A6. Walking on a flat surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A7. Getting in/out of car

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A8. Going shopping

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A9. Putting on socks/stockings

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A10. Rising from bed

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A11. Taking off socks/stockings

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A12. Lying in bed (turning over, maintaining hip position)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A13. Getting in/out of bath

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A14. Sitting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A15. Getting on/off toilet

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A16. Heavy domestic duties (moving heavy boxes, scrubbing floors, etc)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

A17. Light domestic duties (cooking, dusting, etc)

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Function, sports and recreational activities

The following questions concern your physical function when being active on a higher level. The questions should be answered thinking of what degree of difficulty you have experienced during the **last week** due to your hip.

SP1. Squatting

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP2. Running

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP3. Twisting/pivoting on loaded leg

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

SP4. Walking on uneven surface

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Quality of Life

Q1. How often are you aware of your hip problem?

Never	Monthly	Weekly	Daily	Constantly
☐	☐	☐	☐	☐

Q2. Have you modified your life style to avoid activities potentially damaging to your hip?

Not at all	Mildly	Moderately	Severely	Totally
☐	☐	☐	☐	☐

Q3. How much are you troubled with lack of confidence in your hip?

Not at all	Mildly	Moderately	Severely	Extremely
☐	☐	☐	☐	☐

Q4. In general, how much difficulty do you have with your hip?

None	Mild	Moderate	Severe	Extreme
☐	☐	☐	☐	☐

Thank you very much for completing all the questions in this questionnaire.

Appendix D

Chapter 8 supplementary material

Ordinal regression has been used to model nominal outcome variables, in which the log odds of the outcomes are modelled as a linear combination of the predictor variables. Examples of outcome suitable for using ordinal regression are occupational choices, food choices, exercise choices, etc. The predictor variables can be either categorical or continuous.

In addition to multiple logistic regression, an ordinal regression model, using the "low benefit" group as baseline, was also fitted, results are shown in Tables D.1 and D.2.

Variables	TKR		THR	
	OxPReSI-good	OxPReSI-bad	OxPReSI-good	OxPReSI-bad
	Coefficient (OR, good or bad/intermediate)		Coefficient (OR, good or bad/intermediate)	
Gender				
-Female	0	0	0	0
-Male	0.037(1.04)	0.157(1.17)**	0.038(1.04)	-0.095(0.91)
Age				
-Under 60yo	0	0	0	0
-60-69yo	0.187(1.21)**	-0.13(0.88)*	0.187(1.21)*	0.43(1.04)
-70-79yo	0.345(1.41)***	-0.20 (0.82)*	0.067(1.07)	0.008(1.007)
-Over 80yo	0.421(1.52)***	-0.26(0.77)*	0.152(1.16)	0.136(1.15)
Baseline OKS/OHS				
-0-4	-0.36(0.70)*	-0.05(0.95)	-0.543(0.58)***	-0.234(0.79)
-5-8	-0.36(0.70)***	0.04(1.04)	-0.545(0.58)***	-0.463(0.63)***
-9-12	-0.25(0.78)***	0.04(1.04)	-0.267(0.77)**	-0.157(0.85)
-13-16	0.097(0.91)	0.033(1.03)	-0.14(0.87)	0.025(0.97)
-17-20	0	0	0	0
-21-24	-0.088(1.09)	0.052(1.05)	0.088(1.09)	0.103(1.11)
-25-28	0.157(1.17)	-0.013(0.987)	0.093(1.10)	0.278(1.32)
-29-32	0.218(1.24)*	0.22 (1.25)	0.082(1.08)	0.372(1.45)*
-33-36	0.359(1.43)*	0.24 (1.27)	0.358(1.43)	0.73(2.07)**
Symptom Period				
-$\leqslant 1yr$	-	-	0	0
-1-5yrs	-	-	-0.18(0.84)*	-0.036(0.96)
-6-10yrs	-	-	0.206(0.81)	0.295(1.34)*
-$\geqslant 10yrs$	-	-	0.28(0.75)*	0.12(1.12)

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table D.1: Multinomial logistic regression analysis

Variables	TKR		THR	
	OxPReSI-good	OxPReSI-bad	OxPReSI-good	OxPReSI-bad
	Coefficient (OR, good or bad/intermediate)		Coefficient (OR, good or bad/intermediate)	
Depression				
-No	0	0	0	0
-Yes	-0.16(0.86)*	0.12(1.13)	-0.45(0.64)***	0.15(1.17)
Previous Sx				
-No	0	0	0	0
-Yes	-0.34(0.71)***	0.28(1.32)***	-0.82(0.44)*	0.33(1.39)***
Pre-op disabilities				
-No	0	0	0	0
-Yes	-0.24(0.78)***	0.01(1.01)	-0.30(0.74)***	0.09(1.09)
Heart disease				
-No	0	0	0	0
-Yes	-0.18(0.83)**	0.05(1.05)	-0.29(0.75)***	0.03(1.03)*
Lung disease				
-No	0	0	-	-
-Yes	-0.20(0.82)**	0.13(1.14)	-	-
Circulation problem				
-No	0	0	0	0
-Yes	-0.43(0.65)***	0.14(1.14)	-0.32(0.72)***	0.28(1.32)*
Diabetes				
-No	0	0	0	0
-Yes	-0.25(0.78)***	0.02(1.02)	0.03(1.03)	0.39(1.47)*

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table D.2: Table D.2 above and combined with D.1 before show results from ordinal regression analysis, the purpose of which was to determine the effects of different factors on the odds of achieving a good or bad outcome comparing to the intermediate outcome, using the OxCom method.

Results from ordinal regression model are mostly consistent with results from logistic regression model shown in Chapter Eight. The odds of obtaining "high benefit" as opposed to "low benefit" post operation is significantly reduced when pre-operative score is below 17; the negative effect of pre-operative score is greater for TKR patients. Other co-morbidities, including lung disease, diabetes, circulation problems and heart disease all significantly decrease patients' odds of success.

Comparing the choice of ordinal regression with logistic regression, from a purely statistical point of view, an ordinal regression model would be a more superior choice. The outcome variable consists of three ordered categories, "high benefit", "low benefit" and "no benefit", no manipulation of original data was needed in order to fit an ordinal regression model. Fitting a logistic regression required collapsing of "low benefit" and "no benefit" into one level, resulting in a loss of information and statistical power. However, from a clinical point of view, multiple logistic regression is more appropriate. Even though being able to draw distinction between the "low benefit" and "no benefit" group can potentially be useful in clinical practice, the overarching theme of this thesis is about defining and measuring "successful" outcome, and the main research question of this chapter is "what is the effect of pre-operative score on the prevalence of *successful* outcome post-operatively?" Results from logistic regression address the research question more directly and are easier to interpret.