UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Primary Care and Population Sciences

Total Hip Replacement in the UK: Cost-Effectiveness of a Prediction Tool and Outcomes Mapping

by

Rafael A. Pinedo Villanueva

Thesis for the degree of

Doctor of Philosophy

October 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Primary Care and Population Sciences

Doctor of Philosophy

TOTAL HIP REPLACEMENT IN THE UK: COST-EFFECTIVENESS OF A PREDICTION TOOL AND OUTCOMES MAPPING

by Rafael A. Pinedo Villanueva

Total hip replacements (THRs) have been found to be highly cost-effective. For an important number of patients, however, results are not satisfactory. An outcome prediction tool has been developed to identify, preoperatively, *poor* outcome patients after THR and this study assessed whether its implementation would be cost-effective.

Most published evaluations of THRs have focused on assessing their cost-effectiveness against other surgical procedures or different prostheses, but no study has assessed a tool aimed at predicting *poor* outcomes. To that aim, we developed a lifetime Markov model featuring two unique elements: it starts at the orthopaedic surgeon's assessment and it distinguishes between two outcome categories after primary and revision procedures.

To facilitate populating this and other economic models with health utility estimates, we compared the performance of several econometric models mapping Oxford Hip Score onto the EQ-5D index. All models reported high predictive power. Transition probabilities for the model were obtained from expert elicitation, the NHS PROMs initiative and the EPOS study. Both PROMs and EPOS were also used to estimate health utilities. Procedure as well as primary-care costs were obtained from NHS and CPRD data, respectively. An important contribution of this research was the estimation of a model predicting surgery outcome category based on resource use.

For men and women aged 45 or more, implementation of the prediction tool was associated with savings of £1,000 to £1,800 and a reduction of 1.4 to 3.7 QALYs compared to current practice, with most variation due to age. Results indicate that the health utility improvement resulting from THRs, even for *poor* outcomes, means that a tool rationing the operation would produce significantly less net benefits than current practice. Finally, we found that the model structure and data analyses employed for this assessment would be highly applicable to other interventions.

Contents

List of tables

List of figures

DECLARATION OF AUTHORSHIP

I, Rafael A. Pinedo Villanueva

declare that the thesis entitled

Total Hip Replacement in the UK: Cost-Effectiveness of a Prediction Tool and Outcomes Mapping

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- part of this work has been published as: Pinedo-Villanueva, R., Turner, D., Judge, A., Raftery, J., & Arden, N. (2013). Mapping the Oxford hip score onto the EQ-5D utility index. Quality of Life Research, 22(3), 665-675.

Signed: ……..

Date:……..

Acknowledgements

Whilst conducting the research reported in this thesis I was fortunate to count on the splendid support of many people. My supervisors Prof James Raftery and David Turner gave me the opportunity and supported me from the very beginning. Together, they provided guidance and feedback that helped me take this journey to completion. My family, and in particular my parents, brother and sister were a solid rock of support for me, at every step of the way. My daughter's patience and pride for me, even though I chose to dedicate so much of my time and energy to this long project, were a source of inspiration and kept me going day after day. And crucially during the last year, my partner's love and support were there for me to lean on every single day.

I owe thanks to many friends and colleagues who were there for me and supported me personally and professionally in key moments during the past four years. I am grateful to Dr Keith Cooper for opening his house and sharing his wisdom with me. My NETSCC colleagues whilst I was based at Southampton permanently encouraged me. When it came time to commuting, Diana and Salome were a marvellous safety net. I am extremely grateful to my friends in York who not only hosted me every time I visited but also often took the time to discuss specific aspects of my work during these years.

I was very fortunate to conduct this research as part of a joint project with the Musculoskeletal Epidemiology Unit of the Department of Orthopaedics at the University of Oxford. Everyone there, and particularly Prof Nigel Arden, was permanently supportive, happy to engage in constructive discussions about my work and enthusiastic about the progress of my research. That made a huge difference for me.

With the oversight of my main supervisor, editorial advice was sought from Ria Skelton, to whom I am deeply grateful for the kind support she provided in the challenging final stages. No changes of intellectual content were made as a result of this advice and all faults and imperfections in the final text remain my exclusive responsibility.

The list of friends, family and colleagues who have been supportive of me all along the road ending with this doctoral thesis is very long and would not fit on this page, and neither would my words reflect what their love and support means to me. For that, I feel immensely fortunate and grateful.

Definitions and Abbreviations

ACR	American College of Rheumatology
ASA	American Society of Anesthesiologists
BMI	Body-mass index
BNF	British National Formulary
CEA	Cost-effectiveness analysis
CEAC	Cost-effectiveness acceptability curves
CI	Confidence interval
COASt	Clinical Outcomes in Arthroplasty Study
CPRD	Clinical Practice Research Datalink
CRD	Centre for Reviews and Dissemination
CUA	Cost-utility analysis
DARE	Database of Abstracts of Reviews of Effects
DH	Department of Health
DVT	Deep vein thrombosis
EPOS	Exeter Primary Outcomes Study
EUROHIP	Health Technology Assessment of Hip Arthroplasty
EVPI	Expected value of perfect information
GPRD	General Practice Research Database
HES	Hospital Episodes Statistics
HESG	Health Economists' Study Group
HHS	Harris Hip Score
HRG	Healthcare Resource Group
HRQL	Health-related quality of life
HTA	Health Technology Assessment
ICER	Incremental cost-effectiveness ratio
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
MAE	Mean absolute error
MSE	Mean square error
NHS	National Health Service
NHS EED	NHS Economic Evaluation Database
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NJR	National Joint Registry of England and Wales
NMB	Net monetary benefit

NSAID	Non-steroidal anti-inflammatory drug
OA	Osteoarthritis
OHS	Oxford Hip Score
OLS	Ordinary least squares
ONS	UK Office of National Statistics
OPCS	Operating Procedure Codes
PbR	Payment by Results
PCT	Primary Care Trust
PE	Pulmonary embolism
PROMs	Patient-reported outcome measures
PSA	Probabilistic sensitivity analysis
PSSRU	Personal Social Services Research Unit
QALY	Quality-adjusted life year
RMSE	Root mean square error
ROC	Receiver Operating Characteristic
SD	Standard deviation
SMDM	Society for Medical Decision Making
SWLEOC	South West London Elective Orthopaedics Centre
THR	Total hip replacement
TKR	Total knee replacement
TTU	Transfer to utility
VIF	Variance inflation factor
WOMAC	Western Ontario and McMaster Universities Arthritis index

1 Introduction

Revolutionary approaches to treating diseases of the joints can be traced as far back as the 18th Century, as a letter written in 1782 by Mr H. Park of Liverpool demonstrates [1]. Focusing on the treatment of knee and elbow joints, Mr Park spoke of an alternative to the amputation of the limb, which was the treatment of choice at the time for surgeons facing patients with severe joint problems. Park's letter to his teacher Mr Pott is a comprehensive account of the outcome of his first attempt at removing a knee from a patient in 1781. He also performed an operation to remove an elbow, but that was described in much less detail. Mr Park fully extirpated the articulations and fixed the bones previously linked by the joint by way of allowing a callus to form. The patients were left with extremities that could not be bent, yet he argued that there was much value in keeping the limbs, and so did his patients [1].

The approach to handling hip problems has come a long way since then. Amputation led to excision arthroplasty (removal of joint surface and bone), then osteotomy (bone cut) and later interpositional arthroplasty (inserting tissue between bones), each with different flaws such as lack of stability, mobility, effectiveness, and low predictability of outcome [2]. Interpositional arthroplasty was progressively improved by incorporating implants of such different materials as glass, celluloid, Bakelite, Pyrex, and even rubber or ivory, until an alloy called Vitallium showed the most promising results [2]. Vitallium was extensively used until contributions by Sir John Charnley in the 1960s made hip arthroplasties the successful intervention they are today [2]. Charnley considered the biomechanics of the hip joint and essentially changed the way arthroplasties were done by implanting, and fixing using cement, both a femoral component with a small-diameter head in the femur and its counterpart, an acetabular component attached to the hip bone, thus achieving a low friction articulated joint [3]. Charnley implants are still used for what are known as total hip replacements (THRs) half a century after his breakthrough.

Although a very successful operation, not all patients have an excellent outcome after surgery. In this chapter we present a general overview of THRs in Section 1.1, followed in Section 1.2 by a review of the findings and reasons that make them such extraordinary interventions, overshadowing the shortcomings. In Section 1.3 we describe the umbrella project under which our work is framed, with specific PhD research aims detailed in Section 1.4. We close this first chapter with an outline of the structure of the entire thesis in Section 1.5.

1.1 Total hip replacements

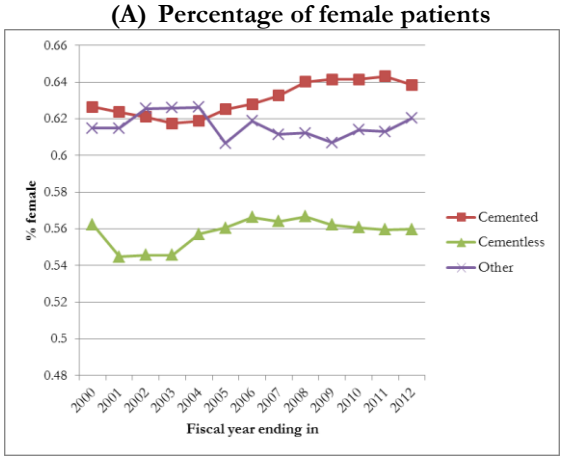
Hip replacements are performed on patients with pain or functional hip problems due mainly, though not exclusively, to osteoarthritis (OA). OA is considered the most common of all chronic joint diseases. Prevalence is rising given population ageing and the obesity epidemic [4] because age and body-mass index (BMI) as well as physical activity and injury are amongst its main risk factors [5]. OA is now understood to be a disorder affecting the whole joint and not only the cartilage[6], where attention used to be focused because as the cartilage becomes damaged and the natural process of repair fails [5], patients are left with bones rubbing against each other thereby causing significant pain, discomfort and limitations. Although different diseases affecting the hip have different pathologies, such as rheumatoid arthritis which is a chronic disease causing inflammation but also capable of destroying the joints [7], THRs can generally help these patients by entirely replacing the hip joint with a prosthesis that can often last for 10 or 20 years.

According to Frankel and colleagues [8], criteria for hip replacement during the 1990s included the National Institutes for Health consensus conference [9, 10] and the New Zealand priority criteria [11]. The former established that patients with moderate to severe persistent pain or disability that was not extensively reduced by non-surgical treatment should be candidates for hip replacement. For the latter, a summary score of various indicators of pain and functional limitations was built and thresholds for moderate and severe disease selected. The score and classification were then considered when making the decision to indicate hip replacement [8]. In the UK, the care and management of OA patients has been clearly delineated by a clinical guideline indicating that patients must first be treated with analgesics such as paracetamol, non-steroidal anti-inflammatory drugs (NSAIDs), COX-2 inhibitors or opioids, and only when these do not alleviate the pain and reduced function should they be referred for hip replacement surgery [12].

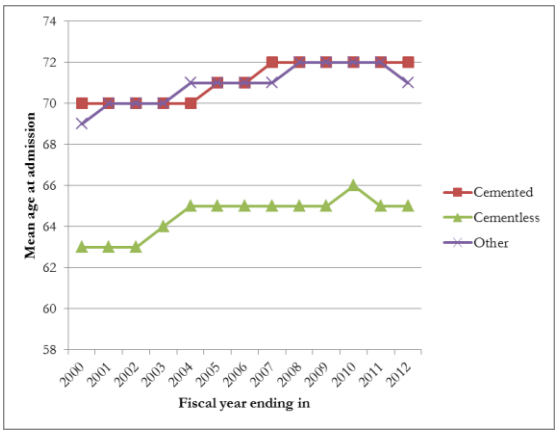
THRs are performed in increasing numbers in the UK. In English NHS hospitals alone, during the fiscal year 1999-2000 some 45,600 hip replacement admissions were recorded, growing to slightly over 70,000 by 2011-2012. Most THR patients are women and the mean age of patients undergoing the operation, around 70 years of age, is slowly increasing over time [13]. Figures 1.1 (A) and (B) show the trend of the percentage of female patients and the mean age of patients admitted for THRs in English hospitals during the last 12 years and separately by fixation type.

Figure 1.1
Hip replacement NHS admissions in England

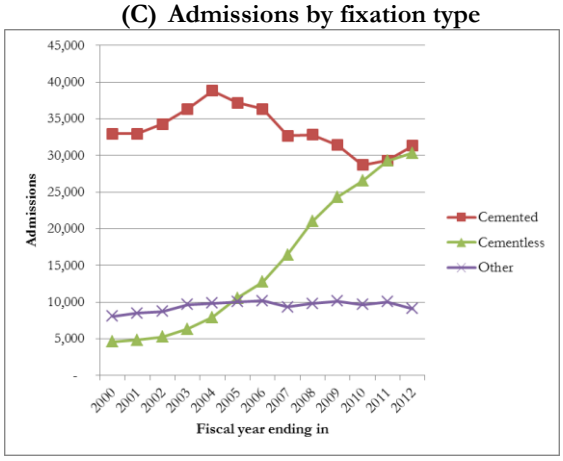
(A) Percentage of female patients



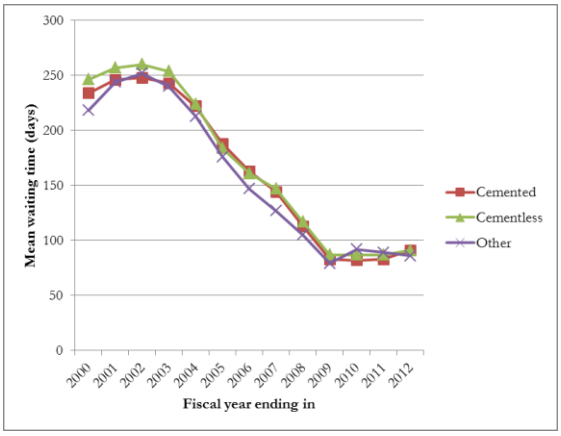
(B) Age at admission



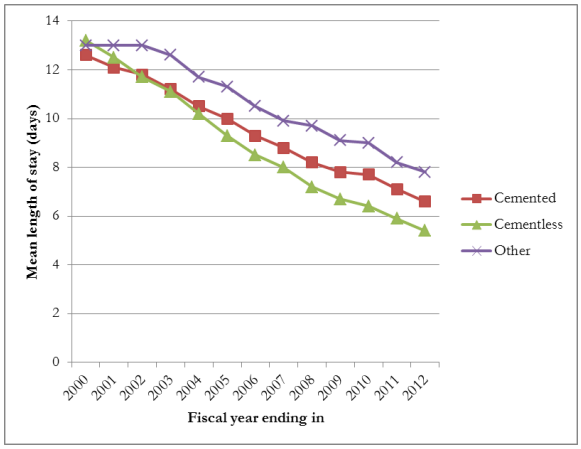
(C) Admissions by fixation type



(D) Waiting time



(E) Length of stay



Source of data: Hospital Episode Statistics

Ever since Charnley revolutionised the approach to hip arthroplasties, prostheses have been fixed to the bones using cement. The more expensive cementless prostheses were developed aiming to increase implant survival, considering that people live longer and hence so must implants [14]. These cementless prostheses have become increasingly popular in many countries [15] and England is no exception. As Figure 1.1 (C) shows,

admissions for cementless procedures in English NHS hospitals have increased substantially since the beginning of the century to the point that today they have become as common as the traditional cemented prostheses [13]. Mean waiting time for THRs has dropped from over 200 days in 1999-2000 to around three months in 2011-2012 regardless of fixation type, a similar reduction to that observed for the mean length of stay of patients. This fell from 13 to around seven days in the same period [13], as Figures 1.1 (D) and (E), respectively, show.

1.2 The problem: beware of miracles, even real ones

Total hip arthroplasties have been regarded as one of the most successful interventions in orthopaedics because they have proved able to take patients from a state of pain and limitation in moving and performing usual activities to a high level of functional ability, reduction of pain and significant improvement in quality of life [16, 17]. Moreover, these benefits are produced at a cost that makes the intervention largely worthwhile. In fact, the high level of effectiveness achieved by THRs considered alongside their cost has meant that the operation is regarded as highly cost-effective [8], not only in the United Kingdom but also in many other countries [14, 18]. Even outside of orthopaedics, THRs have been found to be more cost-effective than interventions such as bypass surgery and dialysis [19]. According to an article published by The Lancet in 2007, their ability to absorb advances in technology and the significant benefits that they can and often provide to patients have earned THRs the title of *operation of the century* [20].

But such high success has come at a price. The widespread regard for THR as one of the greatest interventions in medicine, albeit founded, has been in some cases mistakenly taken to mean that the operation will work wonders for every patient, which, simply, is not true. Although procedures have become increasingly advanced [20] there is evidence showing that a portion of the group of patients undergoing THRs achieves little or no improvement in terms of mobility or is not satisfied with the results [21-23]. In a study based on 1,100 randomly selected THR patients from five different regions in the UK dating back to 2002, 11% of patients were found to be dissatisfied with the procedure at one year whilst only 2.6% had had a revision replacement by then [24], indicating that need for revision is not an indicator of patient satisfaction.

If potential poor or unsatisfactory outcomes following THR could be predicted, then these patients could be treated in some other way that benefited them most, without the

health care system having to incur costs that could otherwise serve those same patients or others more efficiently.

1.3 The COASt project

Despite all the advances that have led to the success of THRs, there are still relevant unanswered questions about the effectiveness and potential cost-effectiveness of this extraordinary procedure. First, not much attention has been paid to the predictors of outcome for THR patients. There has been work linking age, gender, marital status, co-morbidity and the physical status American Society of Anesthesiologists (ASA) score to THR outcome [21], whilst anxiety/depression [22] and socioeconomic factors such as education and employment [23] have also been found to be associated to patient's outcomes and satisfaction. No comprehensive tool predicting the outcome of THR patients has, however, been produced yet.

If such a tool were to be developed, it would be equally important to ascertain whether implementing it would be worth it in terms of its additional costs and benefits. It is possible that an outcome prediction tool could effectively identify patients who would not have a satisfactory or very good outcome after surgery, yet its potential higher benefits may be lower than the health benefits displaced elsewhere in the system by directing resources to implement it. In short, whether an outcome prediction tool would be a cost-effective use of resources for the UK health care system is another question not yet addressed by the research community.

The contribution that the cost-effectiveness analysis of an outcome prediction tool would produce could be expanded further with results by specific patient subgroups. As with any other economic evaluation, aggregate results may hide significantly different effects whether in costs, effectiveness or both, for specific subgroups with potentially different prognosis and costs as they undergo a THR.

Finally, the economic evaluation of a prediction tool is no different from that of any other health care intervention in that measures of both costs and effectiveness are required. Regarding the latter, outcomes of THR have typically been measured in terms of prostheses survival although there is an increasing tendency to focus instead on patient-reported outcome measures (PROMs). The decision by the NHS to collect data on the condition-specific Oxford Hip Score (OHS) as well as on health-related quality of life

(HRQL) through the EQ-5D questionnaire on all publicly-funded THRs starting in 2009 is a clear sign of this [25]. However, longitudinal data is essential for the economic evaluation of a procedure like this because the cost-effectiveness of THRs has proven to be highly sensitive to revision rates [26]. Long-term data on HRQL is hence necessary yet not regularly available. Because the OHS has been widely used since its introduction in the mid-1990s [27], a new methodological challenge arises bearing the question of whether HRQL measures can be estimated based on responses to the OHS. Cross-walked or mapped EQ-5D scores, for example, would allow incorporation of estimated utility values into economic evaluations which would have otherwise lacked such critical input for the assessment of the prediction tool or of THRs themselves.

All the above questions, summarised in Box 1.1, are currently being addressed by the Clinical Outcomes in Arthroplasty Study (COASt) funded by the National Institute for Health Research (NIHR) through its Programme Grants for Applied Research. The project began in September 2008 and is expected to be completed by mid-2014. Led by Prof Nigel Arden, the project brings together researchers from the Universities of Oxford and Southampton as well as from the Medical Research Council with the aim of designing an outcome prediction tool for lower limb arthroplasty that could be implemented by the NHS. The project has four specific objectives organised in an equal number of work packages: 1) to analyse current and future demand for total replacement of hip and knee (TKR) in the UK; 2) to build an outcome prediction tool based on major risk factors associated to THR and TKR; 3) to perform an economic evaluation using modelling techniques and decision analysis of the prediction tools; and 4) to test the tool on a prospective cohort study.

Box 1.1

A set of unanswered questions around THRs

1. Can the outcome of THRs be predicted and, if so, how?
2. Would the implementation of a THR outcome prediction tool be cost-effective?
3. How would this cost-effectiveness change if different patient subgroups were considered separately?
4. Could HRQL measures be estimated based on responses to the OHS questionnaire?

Of the four work packages included in COASt, the third encompasses an economic evaluation that addresses questions two, three and four in Box 1.1. Our research focused on answering those questions.

1.4 PhD research objectives

The aim of the research presented in this thesis was to conduct a comprehensive economic evaluation of the outcome prediction tool for THRs developed under COASt. In order to do so, our research was directed at achieving six intermediate objectives which are summarised in Box 1.2.

Box 1.2

PhD intermediate objectives
1. To review the published literature on economic evaluations of THRs
2. To propose a model structure for the cost-effectiveness analysis of an outcome prediction tool of THR
3. To estimate all possible model parameters from patient-level data representative of UK current practice
4. To estimate and compare the performance of several econometric methods for the mapping of OHS onto EQ-5D
5. To assess the cost-effectiveness of the outcome prediction tool for THRs developed under COASt, conduct relevant sensitivity analysis, and present deterministic and probabilistic results for a set of patient subgroups
6. To discuss results and implications for THR practice and research

As mentioned in Section 1.2, there have been a number of economic evaluations of THRs that have found the intervention largely cost-effective. In order to appropriately frame our own economic evaluation, albeit not assessing the operation but an outcome prediction tool for it, we conducted a systematic review of the literature to explore the different approaches previously employed by researchers when assessing the cost-effectiveness of THRs. Based on findings from this review and on consultations with health care professionals, we structured an economic model incorporating key unique features that allowed it to adapt to the potential implementation and consequences of using an outcome prediction tool. Because no prediction tool has been developed for THRs before, there have been no economic models to assess their cost-effectiveness, making the

one proposed here a unique contribution to the expanding field of economic evaluation of orthopaedic interventions.

Data sources about the outcomes of THRs for UK patients such as the PROMs, together with other large patient-level data sets such as the Clinical Practice Research Datalink (CPRD) made available by COASt, offered our research a unique opportunity to estimate the majority of the model's input parameters from patient-level data. Most economic evaluations have to rely on data from small randomised controlled trials or observational cohorts, and more often than not some input parameters are also obtained from previously published sources. We strived to make this economic evaluation one populated mainly with data obtained from large representative patient-level data sets about the current practice of THRs in the UK. Results would therefore benefit from the highest levels of confidence as well as extraordinary validity for UK decision makers, health care professionals and patients.

Some of the data available about the outcome of THRs are completed OHS questionnaires, whilst data on health utility is normally not as readily available. Estimating econometric models capable of mapping OHS onto EQ-5D would provide a means for this economic evaluation to have access to much needed estimates of HRQL. But the benefits of a mapping algorithm would reach far beyond this particular evaluation. Longitudinal data is most relevant to the follow-up of patients who undergo a THR and with a mapping model, if EQ-5D was not collected, historical data on OHS could serve as a predictor for health utilities and the wealth of disease-specific data could provide an estimate for HRQL indices.

Finally, with all the parameter estimates obtained from the best available sources populating the newly structured economic model, we assessed the cost-effectiveness of the outcome prediction tool for THRs in the UK. We present deterministic and probabilistic results for the selected patient subgroups. This is an assessment that has not been done before. Much is known about the cost-effectiveness of THRs and very little about predicting unsatisfactory outcomes, but the knowledge gap about the cost-effectiveness of one such prediction tool is only beginning to close with the findings from this research. There are many implications of our results, some policy relevant, some about future research.

We trust that this work is a relevant contribution both to the methods required to conducting an economic evaluation of an outcome prediction tool; to the knowledge base helping decision makers direct health care resources in the most efficient manner; and, through these, to patients' quality of life.

1.5 Thesis structure

The thesis is presented in six chapters following this introduction, one for each of the intermediate objectives shown in Box 1.2. Chapter 2 presents a systematic literature review of economic evaluations of THRs using models. This chapter identifies publications using economic models to assess hip replacements following the guidelines of systematic literature reviews. It provides a characterisation of the type of economic analyses performed with the models and discusses their findings, but concentrates specifically on the models' features, their implications and limitations. The review in Chapter 2 served as key background to develop the economic model described later in Chapter 4, to assess the cost-effectiveness of the outcome prediction tool for THRs.

In Chapter 3, different models were estimated to map OHS onto EQ-5D. Data from a London orthopaedic centre was used to develop four different algorithms that produce EQ-5D estimates based on responses to the OHS questionnaire. The comparative performance of the models is assessed based primarily on their ability to predict the observed mean EQ-5D summary score. The chapter also reports on the estimation of an additional two models predicting the change in EQ-5D based on the change in OHS after THR. Both of these sets of models were validated internally and on an out-of-sample dataset. The association between a disease and joint-specific outcome measure such as the OHS and the preference-based generic EQ-5D is an important contribution for researchers aiming to populate economic models on THR with estimates for utility values, when EQ-5D is not collected but the OHS is.

Chapter 4 presents the economic model developed for the cost-effectiveness analysis of the outcome prediction tool for THRs. We first review the prediction of outcome and then explain the distinction between outcome categories which makes our economic model unique. The reasons for choosing a Markov type model and the overall patient care pathway for patients with hip pain are also discussed, before explaining the model in extensive detail. Emphasis was placed on the two characteristic features of the model: the assessment starts at the point where patients are referred either to THR or to a non-

surgical alternative; and, for those who undergo the operation, the outcome can be *good* or *poor*. The fact that this evaluation compares current practice to a hypothetical scenario where the outcome prediction tool would be used is also covered in this chapter, together with the perspective of the analysis and the patient subgroups whose parameters populated the model.

Obtaining model parameter values to populate the economic model deserved a separate chapter. We start Chapter 5 reporting on an extensive expert elicitation exercise conducted to estimate preoperative transition probabilities for the model. The rest of the transition probabilities were estimated using primarily data from the PROMs dataset, which was also key to the following section describing HRQL data used to derive health utility estimates for each model state by patient subgroup. Because PROMs data only covers one pre and one postoperative measure, we used the first five years of follow-up data from the Exeter Primary Outcomes Study (EPOS) to help produce estimates of health utility for the years following the operation. Next we report on an extended analysis of resource use data obtained from the CPRD to estimate the primary care costs of THR patients before and after their operation. Since the first sections of the chapter describe the model parameter values under current practice, we dedicate a section of Chapter 5 to estimate which parameters would change with the introduction of the outcome prediction tool and how. Assumptions made by the model and the data populating it are reported at the end of the chapter.

In Chapter 6 we detail how the model was mathematically structured in Excel and then we present deterministic results for all patient subgroups. We conducted sensitivity analysis on key model parameters and present those results in the following section. Fully probabilistic sensitivity analysis was also performed for each patient subgroup and results are presented in the usual cost-effectiveness acceptability curves (CEAC).

Finally, Chapter 7 is used to present a general discussion about the findings of our research. We discuss how our cost-effectiveness analysis of the outcome prediction tool for THR not only starts to fill a void where nothing had been done before, but it also produces results that are relevant for the current policy about THR referrals in the UK. The extensive efforts put into deriving estimates for each model parameter based mostly on patient-level data allowed us to identify areas for improvement that can positively impact research and THR practice, not only in the UK but also elsewhere. We stop to

consider the potential implications of a novel approach developed during the research for the estimation of surgery outcome based on resource use, and go on to reflect about where we believe that the research about cost-effectiveness of prediction tools might move towards next. We close the chapter with some of the strengths and limitations identified about our work and a final research recommendation that could potentially change one aspect of the current policy about referral of patients for a THR in several regions of the UK.

1.6 Contributions

At the end of each chapter, a separate section details the contributions received from other researchers, if any, for the development of the work.

This chapter was entirely developed by Rafael Pinedo.

2 Economic models in the literature: a systematic review

2.1 Introduction

Despite the clinical success and established value-for-money of THRs, the National Institute for Health and Care Excellence (NICE) continues to explore the clinical and cost-effectiveness of these interventions. In August 2011, NICE announced a combined review of two existing technology appraisals [28] issued on surgical interventions for patients with hip disease, one on replacement prostheses published in 2000 and a second on metal-on-metal hip resurfacing issued three years later. The revised appraisal is expected by December 2013.

The justification for the revised technology appraisal comes from the original guidance, which called for the collection and analysis of long term outcomes following the hip arthroplasties [29]. More than ten years after issuing the technology appraisal on replacement prostheses, much data covering postoperative outcomes is now available. In addition, the Medicines and Healthcare Regulatory Agency issued an alert in 2012 about possible soft-tissue damage from metal-on-metal implants, presumably resulting from interaction with debris as the implants wear [29]. This provided additional justification for the new and revised technology appraisal on surgical interventions.

Given the large and growing number of hip procedures being performed, the important portion of the health budget that they consume, and the increasing quantity of data available to assess their clinical and cost-effectiveness, it is expected that economic evaluations of these procedures will continue to be performed. Future assessments can be expected to explore cost-effectiveness at higher levels of detail, such as the impact of different elements of the surgical process (e.g. fixation type, prosthesis brand and model) or cost-effectiveness stratified by patient subgroups (e.g. according to gender, age, BMI, and comorbidities).

For economic evaluations of interventions such as joint replacements which have a long-term effect on patients' health, decision analytic models, which are key to inform decision-making based on cost-effectiveness [30], would be a natural choice to synthesise data from various sources and over a long period of time. However, we did not find any systematic review in the published literature that focused on the economic models used in lower limb arthoplasty as opposed to the results of their cost-effective analyses. Searches

in the Centre for Reviews and Dissemination's (CRD) Database of Abstracts of Reviews of Effects (DARE) as well as in the NHS Economic Evaluation Database (NHS EED) and the Health Technology Assessment (HTA) Database show mostly reviews on specific evidence, or on the results of cost-effectiveness analyses of total joint replacements or prophylaxis, but not about the economic models used.

The aim of this chapter is therefore to conduct a systematic review of the economic models reported in the literature on economic evaluations of THR. This review offers an updated reference of the key features of decision models used to perform economic evaluations of THR. Findings from this chapter serve as a reference for the development of the economic model applied in the assessment of the cost-effectiveness of the outcome prediction tool for THRs, as presented in Chapter 4. The results of this review are also valuable in the context of the upcoming technology appraisal and new cost-effectiveness studies of THRs being commissioned.

The following section reports on the methods used for the systematic review, with emphasis on the protocol followed, inclusion and exclusion criteria; and the databases in which the searches were performed. Section 2.3 shows the results of the review. Here, our findings are outlined with details on how studies met or did not meet the inclusion criteria, and the description of those selected. A key part of the section is the description of the features of the models contained in the selected studies. In Section 2.4 we look at the complete body of studies included in the review and discuss their similarities and differences, strengths and weaknesses, and then close the section with a set of recommendations for the development of an economic model for the cost-effectiveness analysis of THRs. Finally, Section 2.5 reports on the contribution by other researchers into the development of this literature review.

2.2 Methods

The review was performed following the CRD's guide on health care literature reviews [31] as well as the PRISMA guidelines on systematic reviews [32]. As recommended in both documents, the question under research for this study was structured using the PICOS system by describing population, interventions, comparators, outcomes and study design. The review covers all published studies regardless of the participants involved, as long as interventions included a primary or revision hip replacement procedure. Comparators had to include replacement surgery itself, a particular prosthesis or a more

specific procedure type. There were no specifications in terms of the outcomes considered so long as the study design was a full economic evaluation (i.e. including both costs and effects) making use of a modelling technique.

The main search was performed in EMBASE and MEDLINE databases through OVID in August 2009 with two updates up to September 2010. The NHS EED and HTA databases (accessed through the CRD) as well as the National Institute for Health Research's (NIHR) HTA programme database were also searched up to September 2010. The search strategy run via OVID was conducted using search criteria combining numerous terms to identify publications matching arthroplasty or replacement procedures of the hip, which also reported on costs or were part of economic evaluations using models. For the CRD search only the term "hip replacement" was used, whilst for the NIHR's HTA programme any document with "hip" in it was reviewed. All searches were made irrespective of the year of publication and only records in English were retrieved.

The search was limited to published studies. Although this may lead to a risk of publication bias, it is unlikely that any major modelling development has occurred on hip replacement that is not represented in the published literature. In fact, most of the studies identified made references to economic models reported on other previously published papers [19, 26, 33-39].

The inclusion criteria specified studies that make use of a modelling technique as part of a full economic evaluation of a hip replacement procedure. Both primary or revision surgeries were acceptable and comparators could include the procedure itself or a specific prosthesis, for example. Studies covering only costs or effectiveness were excluded, as were evaluations of prophylactic interventions because their structure and outcomes of interest were not applicable to a model assessing long term costs and benefits of THR as guided by a prediction tool.

In the first stage of the review, publications were identified through the various searches and by back-tracing relevant references. These publications were then screened to identify and exclude duplicates and to validate whether they met the inclusion criteria (based on title and abstract). An eligibility check was then performed by reading the full text of all screened publications, at which point the inclusion criteria were applied again to arrive at the final list of publications included in the review.

Data extracted from selected publications included authors, year of publication, country, type of evaluation and type of model. We also extracted the time horizon of the evaluation, outcome measures, sensitivity analyses performed, the main research question, and sources of evidence for effectiveness and cost data. Finally, we obtained the reference currency and year used in the model, the discount rate, the stage within the clinical pathway at which the model starts, and a summary of the study's findings.

2.3 Results

2.3.1 Selection of studies

Using the same structure reported in the PRISM guidelines [32], Figure 2.1 shows the flow of information through the different phases of the review. A total of 501 publications were identified through the searches in OVID, CRD and HTA databases, and an additional 25 were included by back-tracing relevant references cited in those. After removing duplicates, 279 articles were left to be screened. Titles and abstracts were reviewed and 223 publications were excluded because they were not economic evaluations or did not involve a THR. The full texts of the remaining 56 studies were read to check their eligibility based on the inclusion criteria and 17 were finally selected for the systematic review.

Of the 39 ineligible studies, 26 reported on economic evaluations of hip replacements but specifically assessed different prophylactic interventions: one on surgical and one on anticoagulant prophylaxis, two on the prevention of surgical infections, and the rest on the prevention of thromboembolism (deep venous thrombosis). Five additional studies used hip replacements only circumstantially to explore the cost-effectiveness of blood donation (2), autologous transfusions (2) or a blood conservation technique (1). Four studies were excluded because they only evaluated the costs or present value of the THR, and another one because it more broadly explored the economic impact of the operation. Finally, of the remaining three publications that did not meet the inclusion criteria, one was an economic appraisal performed in the absence of empirical data, the second a literature review without producing an economic evaluation, and the third used THRs for the description of uncertainties using a Bayesian approach to cost-effectiveness. Details of the 39 studies not meeting eligibility criteria are shown in Appendix 1.

2.3.1 Description of studies

A total of 17 studies matched the inclusion criteria and were included in the qualitative analysis. One of them involved not only THRs but also TKRs as well as several other surgical and non-surgical alternatives for the treatment of osteoarthritis. This study was part of a health-sector wide model to identify the most effective and cost-effective interventions at reducing the burden of the disease [40]. The authors found THRs together with TKRs to be the most effective of the treatment options considered, including primary prevention, management, pharmacotherapy, and complementary medicine methods. The study was retained for analysis, summarised along with all remaining selected publications in Table 2.1.

Figure 2.1
Flow of information for the literature review

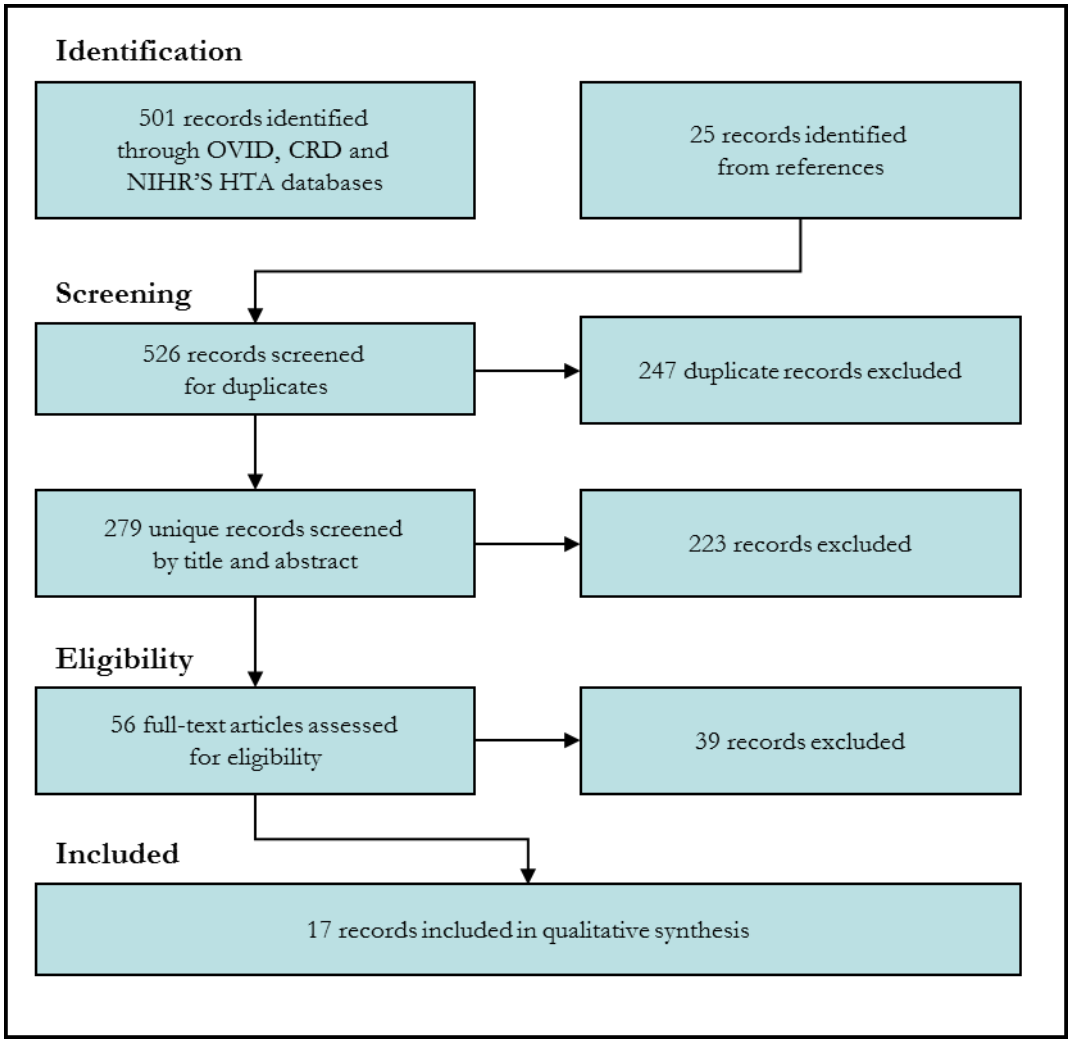


Table 2.1
Summary of studies included in the review

First author	Year	Country*	Evaluation type **	Comparators	Model type	Time horizon	Sensitivity analysis	Perspective	Ref
Chang	1996	US	CEA, CUA	Primary THR vs. non-surgical	Factor stochastic tree	Unspecified	Multi-way deterministic	Societal	19
Saleh	1997	CA	CEA	Revision THR vs. delayed surgery	Decision tree	2 years	Multi-way deterministic	Unspecified	43
Briggs	1998	UK	CUA	Different types of prostheses	Markov	60 years	One-way deterministic	Health care system	26
Faulkner	1998	UK	CEA	Different types of prostheses	Mathematical simulation	20 years	Multi-way deterministic	Unspecified	41
Fitzpatrick	1998	UK	CUA	Different types of prostheses	Markov	60 years	One-way deterministic	Health care system	42
Fisman	2001	US	CEA, CUA	Revision THR vs. other surgical	Markov	Unspecified	One-way deterministic	Societal	44
Vale	2002	UK	CUA	Primary THR vs. other surgical	Markov	20 years	One-way deterministic	Health care system	46
Briggs	2003	UK	CUA	Different types of prostheses	Markov	60 years	Fully probabilistic	Health care system	37
McKenzie	2003	UK	CUA	Primary THR vs. other surgical	Markov	20 years	One-way deterministic	Health care system	36
Briggs	2004	UK	CUA	Different types of prostheses	Markov	60 years	Fully probabilistic	Health care system	45
Segal	2004	AU	CUA	19 primary prevention and patient management interventions	Health-sector wide disease	15 years	One-way deterministic	Societal	40
Bozic	2006	US	CUA	Different bearings on THR	Markov	Lifetime	Multivariate	Unspecified	50
Marinelli	2008	IT	CUA	Different fixation types on THR	Markov	Unspecified	Fully probabilistic	Payer	51
Sharifi	2008	US	CUA	Primary THR vs. other surgical	Decision tree	30 years	Multivariate	Societal	39
de Verteuil	2008	UK	CUA	Primary THR vs. other surgical	Markov	40 years	Deterministic and probabilistic	Unspecified	47
Slover	2009	US	CUA	Primary THR vs. other surgical	Markov	20 years	Two-way	Unspecified	48
Bozic	2010	US	CUA	Primary THR vs. other surgical	Markov	30 years	Deterministic and probabilistic	Health care system	49

* Country: United Kingdom (UK), United States (US), Canada (CA), Australia (AU), Italy (IT) ** Evaluation type: cost-effectiveness analysis (CEA), cost-utility analysis (CUA)

The first reported model to be used for the economic evaluation of hip replacements was published by Chang *et al* in the United States (US) during 1996 [19]. Shortly after that, in 1998, three studies based in the UK were published by Briggs *et al* [26], Faulkner *et al* [41] and Fitzpatrick *et al* [42]. Since then, models of different kinds have continued to be used to assess the cost-effectiveness of THRs until the last year included in the review (2010), including two publications in 2003 and three during 2008.

All but three of the publications identified reporting on economic models assessing THRs came from the UK (8 studies) and the US (6), with Canada, Italy and Australia contributing one each. In terms of the type of economic evaluation performed, 15 of the 17 selected studies were cost-utility analyses (CUA). We considered all those studies measuring outcomes in terms of HRQL such as the quality-adjusted life year (QALY) to be CUA. The two remaining publications carried out a cost-effectiveness analysis (CEA), one using the Harris Hip Score (HHS) as an outcome measure, which reports the level of hip function and symptoms [43], whilst the other used revision rates [41]. Additionally, two of the studies employing QALYs as outcome measure also reported results of a CEA, the first by including in their analyses infection [19] as a secondary measure, and the second relapse-free survival [44].

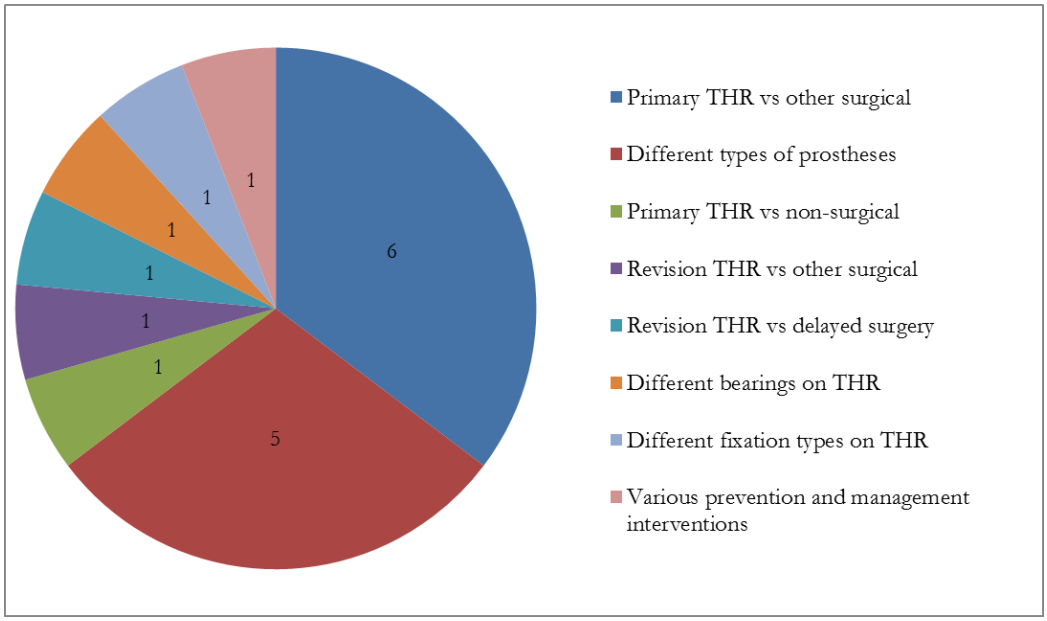
2.3.2 Studies' findings

Most studies reported on assessments aimed at estimating the cost-effectiveness of THR against an alternative intervention, whilst others compared different key components of the surgery such as the prosthesis type. Figure 2.2 shows the breakdown of comparators used in the selected publications.

More than two thirds of the studies assessed either different types of prostheses or compared THR against a different kind of surgical intervention. More specifically, five (29%) of the publications used the economic model to compare the cost-effectiveness of THRs using different prostheses. Of these, two are essentially the same analysis comparing Charnley to Spectron prostheses published first as a report [37] and then as a journal article [45] by the same authors. The other three studies compared Charnley against a newer hypothetical prosthesis [26], Charnely against Stanmore together with Exeter against Muller stems [41], and standard care (Charnley and Stanmore) against another hypothetical newer prosthesis [42]. The two studies (reported in three publications) comparing Charnley directly to a different prosthesis found the former less

cost-effective. Faulkner *et al* [41] found the Stanmore prosthesis to be more cost-effective in an analysis covering up to 20 years after surgery, whilst Briggs *et al* [37, 45] reported that the Spectron prosthesis generated more QALYs than the Charnley, dominating in the case of younger patients and producing an incremental cost-effectiveness ratio (ICER) under £20,000 per QALY for the older subgroups. The two studies comparing the Charnley prosthesis against hypothetical ones, one by Briggs *et al* [26] and the other by Fitzpatrick *et al* [42], identified cost and surgery outcome thresholds that the alternative would need to achieve in order to be more cost-effective than a THR using Charnley.

Figure 2.2
Comparators in THR models



Comparing THR against other surgical procedures was the subject of six (35%) of the selected publications. Two of these were performed before 2008, the remainder between then and 2010. In 2002, Vale *et al* [46] compared metal-on-metal hip resurfacing arthroplasty against THR and two non-surgical interventions. Since we are only interested in comparisons with THRs, we focused our analysis on the assessment of the two surgical procedures, which was possible because the authors performed their evaluations separately. They found that THR dominated metal-on-metal hip resurfacing in patients both under and above 65 years of age, being both cheaper and producing better outcomes. Their analysis covered up to 20 years after the interventions. The following year, McKenzie et al [36] performed a very similar study comparing metal-on-metal hip resurfacing to THR and also to watchful waiting (followed by THR), so again we concentrated on the first of those alternatives. The authors found that THR dominated

metal-on-metal hip resurfacing for both age-groups considered (entering the model at 45-50 or 65-70 years of age), under an analysis span of 20 years.

In 2008, Sharifi *et al* [39] found periacetabular osteotomy to be more cost-effective than THR for the less advanced levels of arthrosis of the hip, whereas THR dominated when it was further advanced. This study was performed on patients younger than 45 years of age with symptoms of dysplasia of the hip. In that same year, minimal-incision THR was estimated to dominate standard THR by de Verteuil *et al* [47] in a study looking at both one-year and 40-year time horizons. Minimal-incision THR seems to have slight perioperative advantages over standard THR, which at a similar cost drives cost-effectiveness in its favour. An important limitation of the study was the lack of data on long-term outcomes after the newer procedure, but the authors claim that these should be similar to those of standard THR and hence advocate its use in the NHS. Slover *et al* [48] later compared THR against hemiarthroplasty as treatment for fractures of displaced femoral necks in patients over 70 years of age and found THR to be more cost-effective. Finally, the costs and benefits of THR were again compared to those of metal-on-metal hip resurfacing by Bozic *et al* [49] in 2010. They found metal-on-metal hip resurfacing to be associated with better clinical outcomes and cost-effectiveness compared to THR for men younger than 65 years of age and women before they reach 55, in an analysis covering 30 years after the operations.

Other than comparing THR against alternative surgical interventions or assessing different prostheses types, the publications included in this analysis reported on a variety of different studies. The seminal work by Chang *et al* [19] compared primary THR against non-surgical management and found that THR produced an ICER of $4,600 per QALY for men age 85, lower than that of interventions such as coronary artery bypass surgery or renal dialysis. Furthermore, they reported that the ICER increased with age and was higher for men than for women [19]. Fisman *et al* [44] compared revision THR to open debridement with retention of the prosthesis for the case of infected primaries in patients of 65 and 80 years of age. The authors concluded that debridement is more costly but produces more QALYs for both age groups and genders, with ICERs around $20,000 per QALY for the 65-year olds, and much lower for the 80-year olds. Revision THR has also been compared to a period of watchful waiting before the intervention was performed; Sale *et al* [43] estimated that immediate surgery would be around $9,000 cheaper than the alternative over a period of analysis of two years. Although the authors included the HHS

as a measure of outcome, it was only used to develop surgical outcome groups and was not incorporated into an ICER.

Finally, economic models were also used for the assessment of the relative cost-effectiveness of different bearings and fixation types on THRs. Bozic *et al* [50] developed a model looking at conventional bearings on THR and identified the thresholds of cost and failure rate that would make hypothetical alternative bearings cost-saving compared to conventional ones. In contrast, Marinelli *et al* [51] compared cemented against cementless fixation in THRs for 70 year-old patients with femoral neck fracture or arthritis of the hip. The authors concluded that cementless prostheses are associated with lower costs than cemented over five years for their selected cohort, but that the difference is not enough to make them more cost-effective at a threshold of €50,000 per QALY [51].

The study by Segal *et al* [40] looking simultaneously at THRs and TKRs along with 17 other preventive and management interventions found the former two to be the most effective interventions for the treatment of osteoarthritis. Both joint replacements were compared to four different primary prevention programmes (three on weight loss and one on surgery for obese people); seven management programmes including two educational, two home-based exercise, two clinic-based exercise and a knee brace; two pharmacotherapies based on NSAIDs; three complementary medicines (glucosamine, avocado/soy unsaponifiables, and capsaicin); and a further surgical intervention, knee arthroscopy with lavage. Using a health-sector wide model, THRs and TKRs were found to be highly cost-effective compared to all of the above.

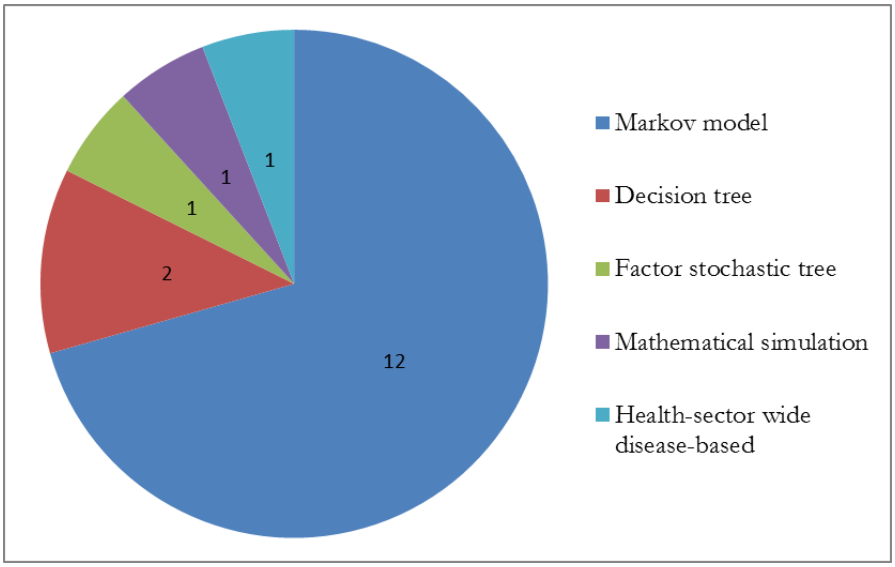
2.3.3 Characteristics of selected models

The types of models used to perform the economic assessments of THRs are shown in Figure 2.3. The most common model type was the Markov, employed by 12 of the 17 studies (71%). When a Markov model was not used, authors applied a factor stochastic tree method [19], a decision tree [39, 43], a mathematical simulation [41], or a disease model based on the entire health sector [40].

All models started their analysis of cost-effectiveness at the point of surgery. Three of them compared immediate THR against delayed THR, therefore in those models the procedure was also assessed at a starting point immediately prior to surgery. This was done by Saleh *et al* (1997), using a decision tree to compare THR against watchful waiting;

by Vale *et al* (2002), comparing THR against metal-on-metal resurfacing, a bone-conserving alternative treatment and watchful waiting; and by McKenzie *et al* (2003), performing the same previous comparison but excluding bone-conserving treatments. Saleh *et al* assumed a 12-month delay before surgery for the watchful waiting group [43], whilst both Vale *et al* and McKenzie *et al*, who used a Markov model, applied the same annual probability of 0.083 for the transition between the watchful waiting state and the THR [36, 46].

Figure 2.3
Types of economic model assessing THRs



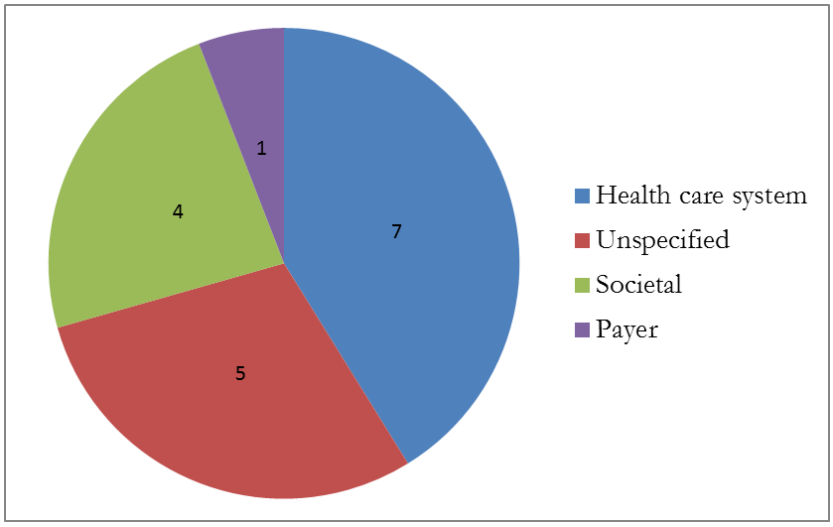
The time horizon used for the analyses varied between studies, as shown in Table 2.1. Four (24%) models used 20 years and another four performed the analysis over 60 years. Thirty years was the choice for two studies (12%), one used 40 years and the decision tree presented by Saleh *et al* only looked at a two-year time horizon. Bozic *et al* (2006) presented a lifetime model of its 50 year-old cohort, whilst three models did not specify their time frame.

The perspective of the health care system was the most common, chosen in seven (41%) of the studies. Three performed the analysis from a broader societal perspective and one from that of the payer. Five of the selected publications (31%) did not explicitly specify the perspective of the model. Figure 2.4 presents these results graphically.

Most authors went to the published literature to obtain effectiveness data to populate their models. In several cases, however, the Swedish Arthroplasty Registry was used to

validate implant survival rates [26] or to obtain survival estimates for primaries [42] or revisions [47]. Cost data came from a much richer list of sources including not only previously published studies but also hospital data, Medicare administrative data in the US, NHS reference costs in the UK, and even from manufacturers of prostheses.

Figure 2.4
Perspective considered by THR models



Populated models produced results that were in all cases further explored by sensitivity analyses, most commonly in its deterministic form with one or more variables involved. Probabilistic sensitivity analysis was only performed by five (29%) of the selected studies. This is largely explained by the fact that THR models started being published in 1996, when probabilistic sensitivity analysis was not common practice. Of the ten THR studies published during or after 2003, half performed probabilistic sensitivity analysis. Discount rates were applied in all analyses, and to both costs and health outcomes with the exception of two studies that only applied them to costs (Saleh *et al* in 1997 and Faulkner *et al* in 1998).

2.4 Discussion

The systematic review of the literature found 17 publications reporting the use of a model in a full economic evaluation of THRs. Most studies identified in the first stage of the review were excluded either because they were not full economic evaluations or because, rather than evaluating hip replacements, they conducted an assessment of associated complications such as thromboembolism.

The majority of economic evaluations were conducted to compare THRs to alternative surgical interventions or to assess the use of different prostheses. THRs were found to be more cost-effective than metal-on-metal hip resurfacing arthroplasties [36, 46], but for younger patients when the time horizon was increased from 20 to 30 years the conclusion was the opposite [52]. It is likely that metal-on-metal hip resurfacing produced such favourable results because the procedure preserves bone from the femoral head by reshaping it to fit the metal cap, instead of removing it altogether, as is done in a THR. Younger patients with a THR are therefore more likely to experience wear of their prostheses and to have revisions, which can be picked up by a life-time or long-term economic model, decreasing the cost-effectiveness of THRs compared to hip resurfacing. THRs were also found to be less cost-effective than periacetabular osteotomy [39], another bone-preserving technique, but designed to treat patients with acetabular dysplasia.

Using a bone-preserving procedure appears to offer younger patients greater benefits in the long-term compared to THRs, as the latter tend to require a revision after full prostheses have been used for many years. The superiority of bone-preserving procedures is therefore likely linked to the severity of the disease and the ability of patients to remain in relative good health for a long time, without requiring later a complete replacement of their hip. If a procedure such as hip resurfacing is performed on a patient in whom deterioration of the remaining bones continues after the operation, they are likely to be a candidate for a THR thereby decreasing its value for money. All three studies comparing THR to metal-on-metal hip resurfacing considered the scenario of hip resurfacing followed by THR in their models.

Economic evaluations comparing THRs with different prostheses were the second most common purpose of the studies. All of those selected for this review used the Charnley prosthesis as a comparator, which is to be expected given that the Charnley was the only prosthesis to have more than 20% of the market of cemented implants in the UK by 1995 [53]. The three studies comparing THRs using a Charnley to THRs using another specific prosthesis (as opposed to a hypothetical one) found the alternatives to be more cost-effective than the Charnley. This superiority in value-for-money is reflected in current figures of clinical practice regarding the choice of prosthesis used in THRs, in the UK at least. According to figures from the National Joint Registry of England and Wales (NJR), by 2011 the Charnley had only about 4% of the market of cemented hip stems in the UK.

This figure is much lower than the 64% recorded by the Exeter V40 and three other prostheses capturing between 5% and 15% of the English and Welsh THR prosthesis market [54].

The models employed to conduct the economic evaluations were mostly Markov models, a natural choice given the chronic nature of the disease and the ease of structuring the care pathway of arthritis patients after a surgical intervention into separate health states. All models started at the point of surgery or immediately prior to it and, although only one presented a lifetime model, the majority covered between 20 to 60 years after the operation. It is unlikely that the choice of time horizon within the above range would make a significant difference in results given that most THRs are performed on patients over 60 years of age. During 2011-2012, mean age of THR patients was around 70 years. However, nearly 20% of THRs in England are performed on patients younger than 60 years of age [55]. For younger patients, a shorter time framework of analysis may bias results by excluding long terms benefits and failing to consider likely higher rates of revision due to the longer survival. In fact, as reported above, two cost-effectiveness studies with different time horizons favoured different alternatives when comparing THR against metal-on-metal hip resurfacing. Although time was not the only difference between the two studies, varying the time horizon clearly has an effect on the relative cumulative costs and HRQL of the alternatives and hence on their cost-effectiveness.

The most common perspective employed for the analyses was that of the health care system. The primary concern of this review is surveying the models used to evaluate THRs within the health policy context. It follows that an assessment of THRs as compared to other surgical or non-surgical interventions, when performed under the perspective of the health care system or that of the broader society, should begin at a point previous to the operation. This is because, whereas at a first level it is important to identify the most cost-effective intervention for specific patient groups, at a broader level all or many of the treatment alternatives will still be performed for some patients. Hence, there is a treatment choice that is not being captured by the structure of the models comparing THRs against other alternatives in one integrated model. If models were integrated, a policy affecting the referral of patients to one or another treatment could truly be assessed from a societal or health-care-system perspective. Otherwise, this perspective is only guiding the costs and outcomes being considered, but not the broader policy implications of the assessment. All models selected for this review which compared

THRs against other alternatives not only started at the point of surgery or immediately prior to that, but they also analysed the model for each alternative separately. As a result, they help inform a decision making process of treatment for certain patient groups but are not able to answer broader questions regarding the clinical pathway of the patient population, from the perspective of the health care system as a whole.

None of the models reviewed considered in its structure or discussion the potential application of an outcome prediction tool. In general, most assessments made it explicit in their models that the possible health states following a THR were a successful operation, revision surgery or death. Only two studies allowed for varying health states after surgery depending on outcome [19, 43], but the distinction was only used to populate the models with different HRQL estimates. The possibility of predicting THR outcome before surgery, and more specifically assessing the cost-effectiveness of such a predictive tool, is completely absent in the published literature of economic evaluations of THRs. The study presented in the following chapters is therefore the first of its kind within the scientific literature.

This review benefited from its systematic protocol, by searching the most relevant databases of publications in the field, and by specifying inclusion criteria that allowed for the identification of any evaluation employing an economic model (regardless of the country where they were based, the year of publication, and the comparators chosen) as long as a THR was included. The findings are limited by the fact that the search was conducted on published literature only; however, it seems unlikely that any unpublished economic model would be significantly different to the ones identified here. As a systematic literature review, it ideally would have included a second reviewer for eligibility checks and data extraction, but being an integral part of a PhD thesis this is considered to be an acceptable limitation. The analysis could also be improved by adding an assessment of the models' conceptualisation, structure and analysis. Several guidelines have been developed for this, including the most recent published jointly by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDM) Modeling Good Research Practices Task Force [56]. These guidelines, however, consist of a list of criteria whose assessment is often subjective and which in this case (considering our interest in describing models as opposed to using their results) would have added little value.

In conclusion, this review offers a description of the economic models that have been used and published for the assessment of THRs. Based on the results, it seems appropriate to use the models to conduct CUAs since PROMs are now widely used, systematically collected and reported in countries such as the UK, where costs per QALY can inform the decision-making process at a sector-wide level. A Markov model appears to be the most appropriate type, given the chronic nature of arthritis, the long-term effects of the interventions, and the ability to structure a system of health states closely associated to the natural history of the disease. The time horizon, when possible, should be the lifetime of the patient group, and the start point that of the surgery or before. If a health care system or societal perspective is chosen, and more than one intervention is likely to be applied to different subgroups of the patient cohort of interest, then it would be useful to start the model at a point where there is a choice between treatment alternatives. If so, the results of separate evaluations such as the ones reviewed here could be used to inform the distribution of patients amongst alternatives. Such criteria could then be assessed against others to find the most cost-effective integrated care pathway. Sensitivity analysis should be performed, both probabilistic given the high level of uncertainty of important parameters such as long term revision rates, and deterministic to identify the most critical drivers of cost-effectiveness.

2.5 Contributions

The database search criteria were established by David Turner (DT) with the support of an information specialist, Alison Price. Rafael Pinedo (RP) joined the project at this point and developed the protocol for the review. The first literature search was performed by DT, the two updates by RP. Inclusion criteria were defined by RP. The first screening was done by RP and DT, and publications progressed to the next phase (eligibility checks) if either of the two selected the article. Eligibility was then performed only by RP. All analyses of results and discussion were produced and written by RP.

3 Utility outcomes for hip interventions

3.1 Introduction

When the Department of Health (DH) requests NICE to conduct appraisals of new or established technologies so that they can make recommendations about implementation, the appraisal involves producing an estimate of the cost-effectiveness of the technology. NICE's *Guide to the methods of technology appraisal 2013* [57] establishes that health effects are to be measured in QALYs, a composite of length of life weighted by quality, the latter being a measure of HRQL. As HRQL measures are considered a proxy for utility, cost-effectiveness studies using QALYs as a measure of health outcome are often termed cost-utility analyses. Economic evaluations of hip interventions in the UK aimed at impacting health policy therefore require utility data. The preferred method of HRQL in adults within the technology appraisal programme is the EQ-5D. NICE's guide also indicates that the valuation of health states should be based on public preferences from a representative sample of the UK and critically that, if these are not available, methods to estimate EQ-5D utility data can be used as long as the functions are estimated and validated on empirical data. The methods to estimate utility scores based on other similar measures are known as 'cross-walking' or 'mapping' methods.

Since April of 2009, NHS providers performing unilateral hip replacements have been required to collect both EQ-5D scores and the Oxford Hip Score (OHS), a condition-specific outcome measure [25]. Prior to this, however, EQ-5D was not routinely collected from THR patients, whereas the OHS questionnaire was commonly collected and regarded as an important indicator of success for THR [27].

This chapter assesses the performance of different mapping methods in estimating the mean EQ-5D score from responses to the OHS questionnaire. The ability to estimate EQ-5D scores based on the OHS would enable estimation of utility data for older datasets where OHS was collected but EQ-5D was not. Older datasets are of key importance given the need for long-term follow up of hip replacement patients whose prostheses, in most cases, last for many years without need for replacement.

We describe the mapping methods used in the following section and in Section 3.3 we report the results of the mapping exercise, including an internal validation. Section 3.4 reports on an application of the mapping methods to predict change in EQ-5D, which is

discussed together with the original results and the strengths and limitations of the study in Section 3.5. Finally, Section 3.6 explains the various contributions made by other researchers to this chapter.

3.2 Methods to map OHS onto EQ-5D

3.2.1 OHS and EQ-5D

The OHS was first reported in 1996 as a consistent and useful measure of outcome for THR [58]. Subsequent studies have assessed it as a valid outcome measure for THR patients [59, 60] and, though considered fairly easy to complete, its results have been shown to be significantly correlated to those of the more complex and physician-assessed HHS [61]. The OHS consists of twelve patient-completed statements covering pain, mobility and ability to carry out regular tasks. Each statement has five categories of response ranging from having no trouble to absolute inability to perform tasks. A copy of the OHS questionnaire is included in Appendix 2. The current scoring system assigns values between zero and four to each item: higher scores corresponding to better outcomes. Individual scores are summed, giving a total score ranging from zero (worst) to 48 (best) [27].

The EQ-5D is a widely used generic measure of health outcomes. It produces a summary index for each of the 243 descriptive health states by applying a preference-based valuation derived from a sample of the general population [62]. A copy of the UK version of the EQ-5D questionnaire [63] is given in Appendix 3 and a more detailed explanation of how summary scores are produced and interpreted is provided in Appendix 4.

A recently published systematic review of mapping methods between non-preference based and preference-based measures found thirty studies covering a wide variety of start and target measures [64]. Of those, two papers considered orthopaedics-related condition-specific measures [65, 66] but neither used the OHS. We found only one reference of mapping OHS onto EQ-5D as part of a report on Patient Reported Outcomes to the DH [67], simply indicating the degree of association between the two scores as produced by a linear regression.

In order to estimate the summary EQ-5D index from OHS responses, we employed two known conversion algorithm techniques, namely transfer to utility regression (TTU) and Response Mapping.

3.2.2 Transfer to utility

The TTU approach uses regression equations to predict the values of one outcome measure, using scores from a second measure as regressor(s) [68]. Methods of this kind require a dataset containing both outcome measures from the same patients. The regression coefficients comprise the mapping algorithm. We used three different TTU methods in this modelling exercise: two are variations of the linear regression model, and the third is a two-part model combining a binary outcome and a linear regression model.

The first model regressed total OHS on the EQ-5D summary index using ordinary least squares (OLS). This model is described by Equation 3.1, where $\hat{E}$ is the expected EQ-5D summary score:

(Equation 3.1)

$$\widehat{E}_i = \beta_0 + \beta_1 \cdot OHS_i$$

Although total OHS is an aggregation of twelve categorical responses, we treated it as a continuous variable under the assumption that it indicates levels of severity of hip arthritis[1].

The second method employed responses to all 12 questions of the OHS questionnaire as categorical regressors and is shown in Equation 3.2:

(Equation 3.2)

$$\hat{E}_i = \beta_0 + \sum_{j=1}^{12} \beta_j \cdot OHS_{ij}$$

where *j* is each of the 12 questions in the OHS questionnaire. One area of concern when including each of the twelve questions of the OHS as regressors is that some of them may be highly correlated, in which case there would be an effect over the variance of coefficients[22]. In order to explore the presence of multiple collinearity between OHS questions, Stata's *collin* command was used on each pair of questions. Results for the variance inflation factor (VIF) and R^2 are shown in Appendix 5. In both cases, the higher the values the greater the collinearity, with VIFs above 10 and R^2 close to 1 being reasons of concern. The highest VIF reported was between the questions on *Description of pain* and *Pain interfering with work* with a factor of 2.92, also showing the highest R^2 at 0.66. Even though the *collin* command runs a simple correlation and ignores the fact that variables are

[1] In www.orthopaedicscore.com ranges of OHS scores are associated to different severity levels of the disease.

categorical, running the correlation accounting for the categorical regressor hardly changes the value of the R^2. Results suggest that none of the correlations between OHS questions are close enough to being nearly perfectly linear to cause concern when fitting a model that includes them all. Mean VIF when including all twelve OHS questions was 2.96.

The third TTU method used was a two-part Logit-OLS model. Many patients report having no problem in the five dimensions included in the EQ-5D after hip replacement, hence a high proportion of postoperative responses have scores of one (full health). Since OLS would not predict a discrete score of one, we formulated this two-part model in order to be able to predict full-health states.

The first part employed a binary outcome logistic model to predict which patients were expected to have EQ-5D scores of one, as shown in Equation 3.3:

$$y_i^* = \beta_0 + \sum_{j=1}^{12} \beta_j \cdot OHS_{ij}$$

$$p_i(EQ5D = 1) = \frac{\exp(y_i^*)}{1 + \exp(y_i^*)} \qquad \text{(Equation 3.3)}$$

$$\hat{E}_i \begin{cases} = 1 & if\ p_i(EQ5D = 1) \geq 0.5 \\ < 1 & if\ p_i(EQ5D = 1) < 0.5 \end{cases}$$

where y* is an unobserved latent variable indicating the log of odds of EQ-5D being equal to one. We then converted this value into a probability using the exponential function, which determined if a one was to be recorded as the expected EQ-5D summary score for the selected observation. Part two used linear OLS regression to estimate EQ-5D values for those patients not predicted to score one.

The underlying assumptions of the linear regression model were checked. Although there seemed to be evidence of heteroskedasticity, linear association between OHS and EQ-5D was confirmed by a fractional polynomial plot and residuals were approximately normally distributed. For the Categorical OLS and Two-part models, different variations were estimated and compared by excluding some or all response categories of certain OHS questions. The best or more efficient variations of each class of model were assessed by internal and external validation.

Potential limitations with the TTU approach have been documented [69]. Firstly, predicted values may fall outside the range of possible EQ5D scores (-0.594 to 1). Secondly, the actual values are unlikely to be matched by a linear regression. Thirdly, regression methods have assumptions that need to hold for a model's estimations to be efficient, or at least unbiased, and these may not always be met.

3.2.3 Response Mapping approach

Response Mapping seeks to predict the responses to each of the five individual EQ-5D questions instead of predicting the summary score directly [69]. A logistic regression model can then be used to estimate probabilities that each set of OHS responses would correspond to a response level of each EQ-5D question. The next step would be to use a Monte Carlo simulation to assign response levels to each EQ-5D question by comparing random numbers to these probabilities. In the original work by Gray *et al* [69] they rightly used the simulation procedure to generate a distribution and then assign the corresponding category, but reported only a single simulation because, given their large sample size, differences were very small. Based on our sample size we also chose to assign health categories after one iteration only. The final index was then computed using the UK's EQ-5D tariff. However, this comes at a cost as assigning a wrong predicted response in just one of the EQ-5D dimensions would result in a significantly different fitted summary EQ-5D score [68].

Responses to EQ-5D questions are ordered, which intuitively implies that the ordered logistic model would be the most appropriate method to use. However, this requires the parallel regression assumption to hold. A likelihood ratio test was used to assess whether this assumption held but it did not, therefore a multinomial logistic model was applied. Equation 3.4 was calculated for two of the three response categories of each EQ-5D dimension, and the third was the reference case against which these probabilities were calculated:

(Equation 3.4)

$$p_{ik} = \frac{\exp(\alpha_k + \sum_{j=1}^{12} \beta_{kj} \cdot X_{ij})}{1 + \sum_{h=2}^{3} \exp(\alpha_h + \sum_{j=1}^{12} \beta_{hj} \cdot X_{ij})}$$

Here, p_{ik} is the probability that respondent *i* will be assigned response category *k* (1, 2 or 3) for the two non-reference categories (*h*). For the reference category, the numerator in Equation 3.4 becomes one.

For all TTU and Response Mapping models we also ran variations that included additional regressors (gender, age, age squared, and deprivation converted into a categorical variable). As none of these variations offered improved performance over the basic models, their results are not reported here.

3.2.4 Data

Data were obtained from the South West London Elective Orthopaedics Centre (SWLEOC) database. The Centre performs hip and knee replacement surgeries for four acute NHS Trusts in South West London. The full data comprised 3,504 hip replacements each with preoperative and/or six-month postoperative responses to the OHS and EQ-5D questionnaires, plus basic demographic, socio-economic and clinical information. All except two operations were performed between 2006 and 2008. All models were estimated on 1,759 operations for which we had data on both pre- and post-op OHS and EQ-5D scores, sex, age and deprivation. Since we were interested in cross-sectional mapping, we pooled pre-op and post-op records together, providing 3,518 outcome observations.

We included primary and revision surgeries, as well as uni- and bi-lateral procedures. Multiple records for the same patient were allowed as long as each record described a separate procedure. As we had at least two observations per patient (pre- and post-op) our dataset was clustered. We allowed for this using Stata's *robust cluster* command during model estimation to show robust standard errors.

We treated the functional relationship between the OHS and EQ-5D as being essentially the same regardless of circumstances and timing of data collection. Even though there could exist such a difference, we considered it would not significantly affect the estimation of the mean score of the group. The data were analysed using Stata/IC 11 statistical software.

3.2.5 Performance assessment and validation

All models were assessed according to their predictive power of the group's mean EQ-5D summary index in the internal validation, i.e. after fitting the models to the same estimation sample. We recorded the range of fitted EQ-5D scores as an indicator of dispersion, and used the percentage of cases for which the estimated score fell within 0.1 of the observed health-state utility estimate as a measure of precision. We estimated the linear correlation between observed and fitted EQ-5D scores and reported their R^2 and root mean square error. In order to assess how well the models performed at estimating EQ-5D scores across the entire range, we calibrated them by calculating the mean absolute error (MAE) by deciles of fitted EQ-5D. Most of these measures have been used in other mapping studies [64] and although we are primarily interested in the models' predictive power, the remaining metrics are also important as they measure different aspects of prediction.

Given the lack of other datasets recording both OHS and EQ-5D, we performed the external validation on 1,616 observations from the subset of the original cohort of 3,504 hip replacements that had not been selected for the estimation sample. The validation sample comprised records with OHS and EQ-5D responses for either the pre- or post-op period, but not for both.

3.3 Mapping results

3.3.1 Descriptive statistics

Table 3.1 shows summary statistics for both selected and excluded observations. There is no statistically significant difference between the groups in terms of age, gender, proportion of primary THR or side of the hip on which the operation was performed.

Figures 3.1 and 3.2 show OHS and EQ-5D scores, respectively, for the selected and excluded groups both before and after the intervention. OHS was near-normally distributed preoperatively, whilst postoperative scores were negatively skewed with a mean of 38 (both groups).

In the case of EQ-5D, an originally bimodal distribution of scores before the operation also became negatively skewed, albeit with a gap in scores between 0.883 and one. The gap is explained by the fact that the highest possible score of non-perfect health using the

UK EQ-5D tariff is 0.883. Of the 3,518 observations included in the estimation dataset, 650 had EQ-5D score of one (perfect health); 1,971 had scores lower than one but greater than 0.5; another 478 scored between zero and 0.5; and 419 observations reported negative EQ-5D scores.

Table 3.1
Summary characteristics of selected and excluded samples

Attribute	Selected (n=1,759)	Excluded (n=1,736)
Mean age at operation in years (SD)	70 (11)	69 (12) †
Female	64%	64% †
Primary THR	96%	95%
Left side	45%	46%
Diagnosis of Primary Osteoarthritis *	62% ‡	56% §

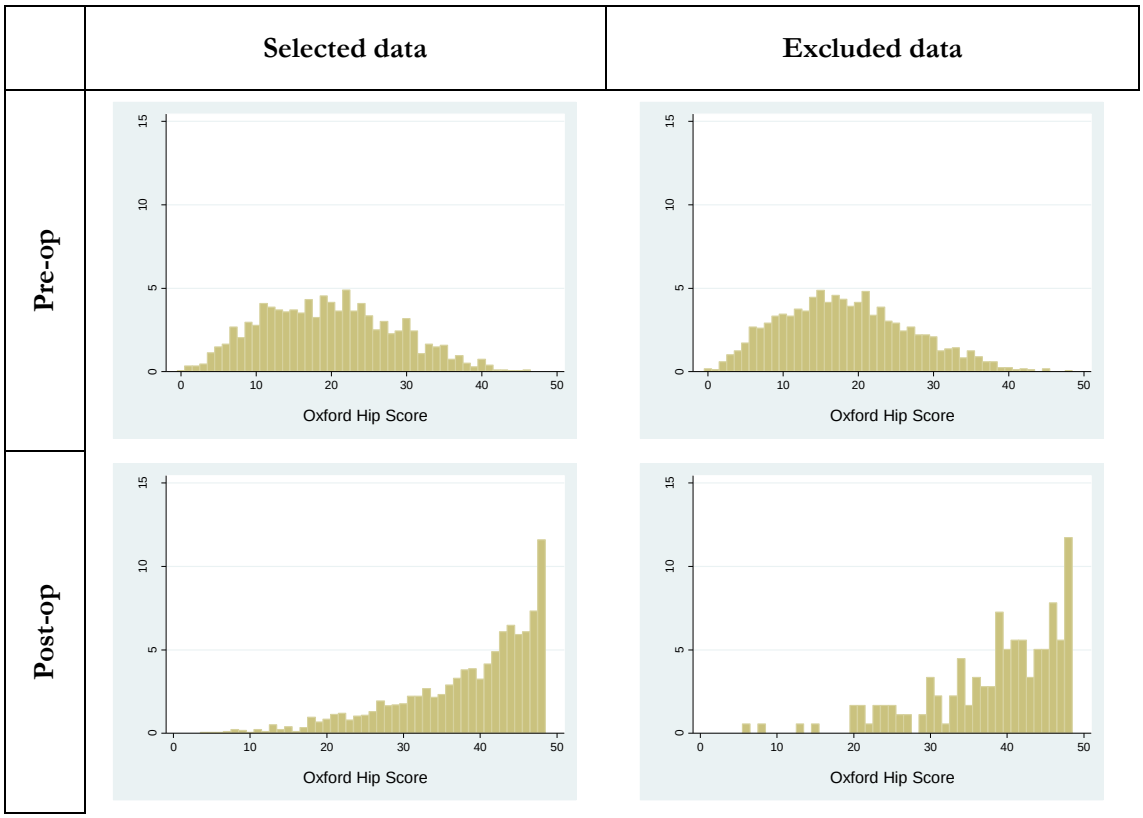
* Two-group mean comparison test, difference of means ≠ 0 statistically significant at 0.05 level

† Less than 1% missing values

‡ 27% missing values

§ 35% missing values

Figure 3.1
Distribution of OHS for selected and excluded data *



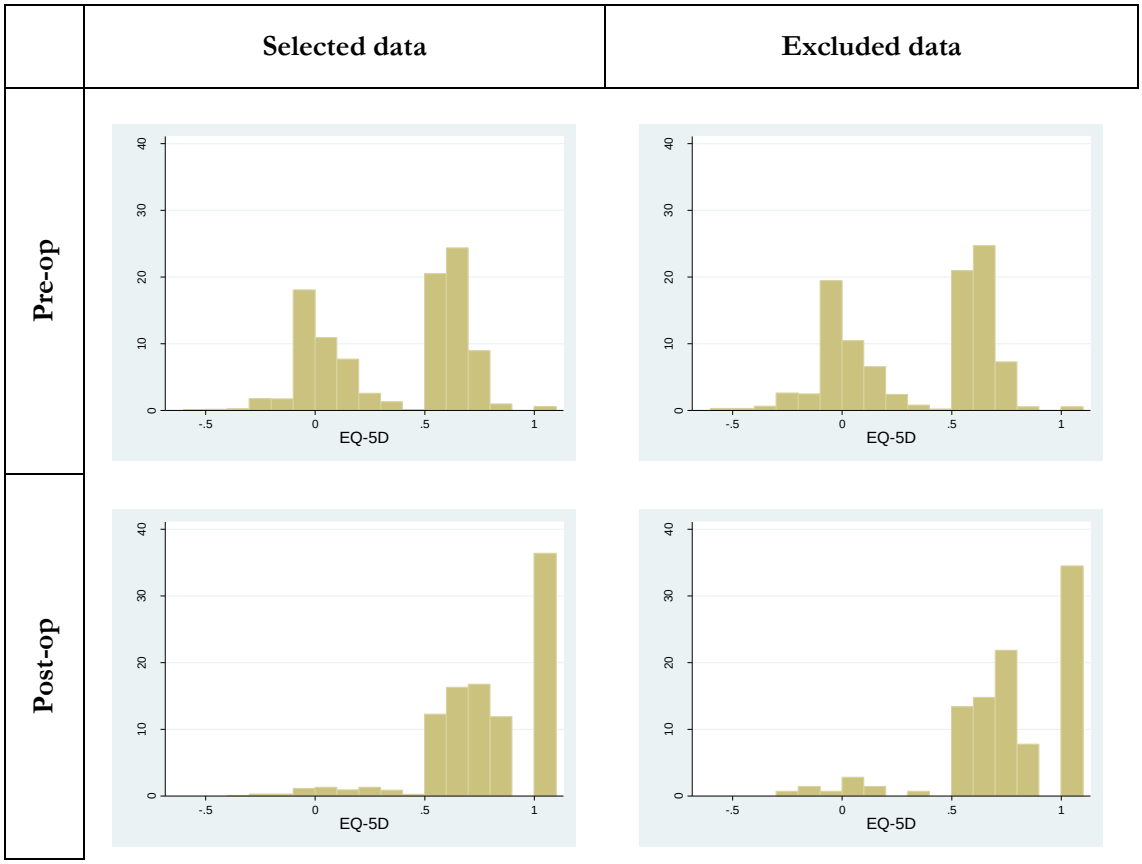
* Number of observations:
Selected data: Pre-op = 1,759. Post-op = 1,759.
Excluded data: Pre-op = 1,682. Post-op = 179

3.3.2 Model performance

The simplest model, Continuous OLS with total OHS as the only regressor, was statistically significant with residuals approximately normally distributed. Model coefficients are shown in Table 3.2.

The categorical version of the linear regression model including all OHS separate questions also reported residuals nearly normally distributed. However, it produced some coefficients that were both statistically not significant and inconsistent with the positive relationship between OHS and EQ-5D, i.e. they were either negative or did not follow an increasing progression within the same question. We tested removing questions which had at least one response level statistically not significant at the 0.05 level, by performing a likelihood-ratio test against the original model. We found no statistically significant evidence suggesting that we should drop any of those questions, hence they were all included in the model.

Figure 3.2
Distribution of EQ-5D scores for selected and excluded data *



* Number of observations: selected pre-op=1,759; post-op=1,759; excluded pre-op=1,502; post-op=142.

In order to produce a more consistent model in which statistical insignificance and inconsistency in the progression of coefficients did not both happen, we collapsed response levels for OHS questions on *Climbing stairs*, *Car and public transport* and *Washing and drying*. Merging item response levels has been applied in previous mapping studies [64]. The new restricted model performed no worse than the full one, hence it was used for external validation. Table 3.3 shows coefficients for the selected model, which varied greatly by question and response level.

Table 3.2
Linear regression output from OLS Continuous model *

Dependent variable: EQ-5D summary score

Independent variables	Coefficient	Robust Standard Error †	95% confidence interval	
Total OHS score	0.0222	0.000	0.021	0.023
Constant	-0.0697	0.010	-0.088	-0.051

* Number of observations = 3,518
F(1,1685) = 7704.93
Prob . F = 0.000
R-squared = 0.672
Root mean square error = 0.200
† Standard error adjusted for 1,686 clusters of patients

For the two-part approach, we estimated the first (logistic) part of the model and found that only 13 of the 48 regressors were statistically significant. According to repetitive likelihood-ratio tests, questions on *Sudden pain*, *Walking duration*, *Pain from standing up from a chair* and *Night pain* did not make a statistically significant difference in predicting whether a patient scores one or not in the EQ-5D so these were excluded from the model. We also combined response levels until coefficients within the same question were either all significant or positive with increasing progression.

For the second part (categorical OLS) we estimated the model on observed EQ-5D scores lower than one and included all OHS questions, since none could be excluded based on repetitive likelihood-ratio tests. We collapsed response levels using the same methods as with the Categorical OLS model. Coefficients are shown in Table 3.3. Again, we found that residuals were approximately normally distributed but with a high peak at zero from perfectly fitted cases of observed EQ-5D equal to one.

Table 3.3

Coefficients and p-values for OLS Categorical and Two-part models

OHS question: response level	Coefficient	p > |t|	Coefficient	p > |t|	Coefficient	p > |t|
	OLS Categorical		Two-part: logit (first stage)		Two-part: OLS (second stage)	
Description of pain: 0	Base case		Base case		Base case	
Description of pain: 1	0.171	0.000	0.339	0.764	0.171	0.000
Description of pain: 2	0.146	0.000	0.460	0.669	0.150	0.000
Description of pain: 3	0.174	0.000	1.866	0.072	0.158	0.000
Description of pain: 4	0.212	0.000	3.035	0.003	0.162	0.000
Night pain: 0	Base case		Excluded		Base case	
Night pain: 1	0.036	0.012			0.038	0.008
Night pain: 2	0.037	0.009			0.039	0.005
Night pain: 3	0.037	0.028			0.040	0.017
Night pain: 4	0.047	0.002			0.049	0.001
Sudden pain: 0	Base case		Excluded		Base case	
Sudden pain: 1	0.004	0.837			0.013	0.456
Sudden pain: 2	0.027	0.089			0.039	0.014
Sudden pain: 3	0.034	0.071			0.052	0.007
Sudden pain: 4	0.044	0.011			0.052	0.003
Limping: 0	Base case		Base case		Base case	
Limping: 1	0.045	0.000				
Limping: 2	0.046	0.001	0.500	0.245	0.014	0.145
Limping: 3	0.045	0.001	0.929	0.003		
Limping: 4	0.055	0.000	1.449	0.000		
Walking duration: 0	Base case		Excluded		Base case	
Walking duration: 1	0.006	0.738			0.009	0.624
Walking duration: 2	0.008	0.618			0.017	0.294
Walking duration: 3	0.031	0.050			0.048	0.004
Walking duration: 4	0.038	0.017			0.050	0.003
Climbing stairs: 0	Base case		Base case		Base case	
Climbing stairs: 1					0.005	0.844
Climbing stairs: 2	0.039	0.009			0.046	0.063
Climbing stairs: 3	0.058	0.001	0.738	0.006	0.073	0.006
Climbing stairs: 4	0.085	0.000			0.072	0.008
Socks and stockings: 0	Base case		Base case		Base case	
Socks and stockings: 1	0.042	0.005				
Socks and stockings: 2	0.039	0.010	0.425	0.267	0.011	0.324
Socks and stockings: 3	0.055	0.001	0.946	0.005	0.018	0.143
Socks and stockings: 4	0.087	0.000	1.516	0.000	0.041	0.009
Pain from standing up from chair: 0	Base case		Excluded		Base case	
Pain from standing up from chair: 1	0.072	0.004			0.076	0.001
Pain from standing up from chair: 2	0.101	0.000			0.107	0.000
Pain from standing up from chair: 3	0.117	0.000			0.128	0.000
Pain from standing up from chair: 4	0.118	0.000			0.127	0.000
Car and public transport: 0	Base case		Base case		Base case	
Car and public transport: 1						
Car and public transport: 2	0.034	0.018			0.037	0.011
Car and public transport: 3						
Car and public transport: 4	0.044	0.014	0.934	0.000		
Washing and drying: 0	Base case		Base case		Base case	
Washing and drying: 1						
Washing and drying: 2	0.018	0.256			0.019	0.223
Washing and drying: 3	0.049	0.005	1.105	0.036	0.051	0.003
Washing and drying: 4	0.063	0.001			0.059	0.001
House shopping: 0	Base case		Base case		Base case	
House shopping: 1	0.001	0.967				
House shopping: 2	0.036	0.014	0.981	0.003	0.035	0.007
House shopping: 3	0.065	0.000			0.057	0.000
House shopping: 4	0.102	0.000			0.074	0.000
Pain interfering work: 0	Base case		Base case		Base case	
Pain interfering work: 1	0.097	0.000			0.103	0.000
Pain interfering work: 2	0.166	0.000			0.180	0.000
Pain interfering work: 3	0.174	0.000	1.715	0.000	0.194	0.000
Pain interfering work: 4	0.236	0.000			0.206	0.000
Constant	-0.165	0.000	-9.816	0.000	-0.154	0.000

For the Response Mapping approach we found that all five multinomial models (one for each EQ-5D question) were statistically significant ($p<0.001$), however many of the individual regressors were not. We built an alternative model, removing entire questions from each equation based on likelihood-ratio tests, and combining response levels with the criteria employed thus far. Performance in all indicators worsened; hence, we retained the full model for all dimensions. Distribution of residuals was largely similar to that of the Two-part model. Coefficients for each of the five multinomial models are shown in Appendix 6.

3.3.3 Validation

All selected variations of each model were internally validated. Table 3.4 shows summary performance indicators for each model, revealing high predictive power of the mean EQ-5D score by all models. Predictive power was highest for both OLS models whilst the Two-part approach and Response Mapping underestimated the mean by less than 0.002.

Table 3.4
Performance of models and internal validation *

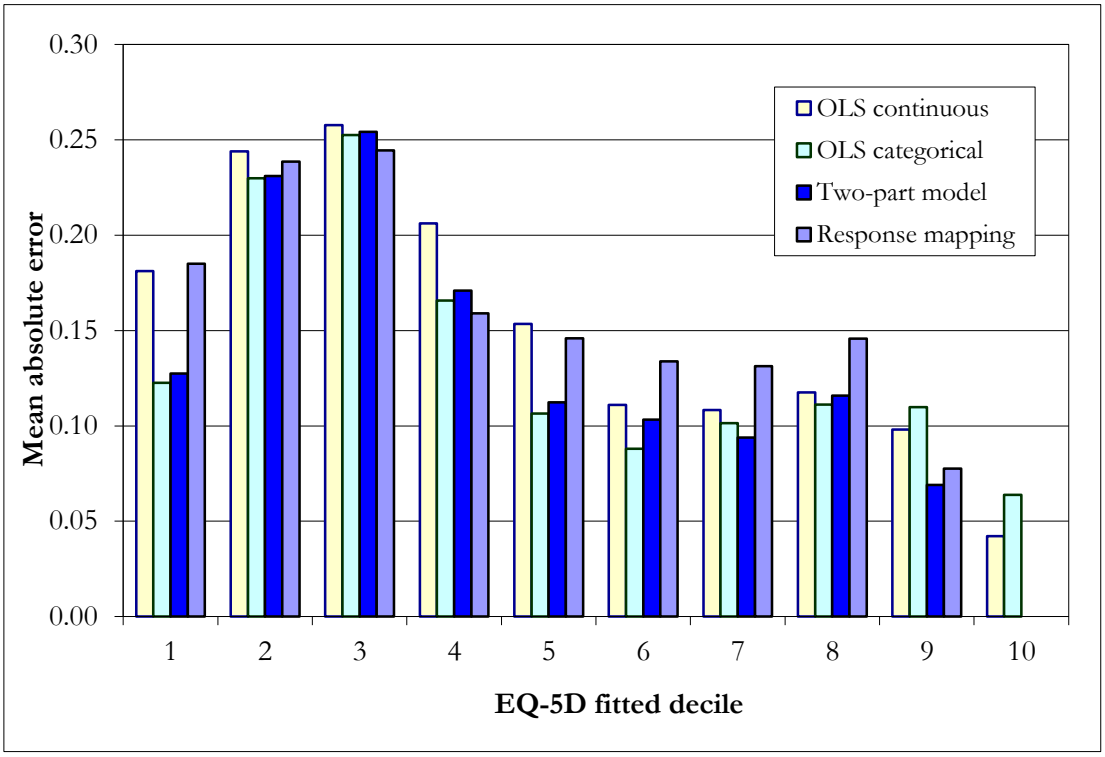
Model: regressors	Mean fitted EQ-5D	Difference of means (observed - fitted)	Range of fitted EQ-5D	Range of residuals	% within 0.10 utility	R2 EQ-5D observed vs. fitted	RMSE† EQ-5D observed vs. fitted
Continuous OLS: Total OHS	0.5750	0.0000	-0.070 , 0.995	-0.91 , 0.76	41.6%	0.67	0.20
Categorical OLS: All OHS questions	0.5750	0.0000	-0.165 , 0.967	-0.91 , 0.78	52.0%	0.72	0.19
Two-part Logit-OLS ‡	0.5735	0.0015	-0.154 , 1.000	-1.11 , 0.82	51.5%	0.70	0.19
Response Mapping: All OHS questions	0.5737	0.0013	-0.484 , 1.000	-0.98 , 1.03	49.0%	0.57	0.23

* For the estimation sample: n = 3,518. Mean observed EQ-5D score = 0.575. Range of observed EQ-5D scores = [-0.594 , 1]
† Root mean square error of the regression
‡ Logit stage used 8 OHS questions, OLS used all OHS questions

As Table 3.4 shows, the Continuous OLS model was, however, a poor predictor of negative EQ-5D indices. Only Response Mapping was able to predict scores largely into the negative range and, together with the Two-part model, up to and including one. All models but the Continuous OLS achieved nearly half of individual estimations within 0.1 of the observed value, whilst all but Response Mapping attained notably high correlations between observed and fitted scores with a R^2 of around 0.7.

In addition to assessing the models' ability to predict the observed mean EQ-5D score, we also calibrated them by recording prediction errors through the range of values of the dependent variable. Figure 3.3 shows the MAE across deciles of the respective model's fitted scores. Although all models reported a high predictive power of the aggregate mean, the level of precision was not uniform across the full range of scores. All models were better predictors at the upper end of the fitted value scale. In fact, for predicted EQ-5D scores under 0.5 all four models had MAEs between 0.20 and 0.23, whereas for predicted values equal or above 0.5 MAEs were half of that (between 0.10 and 0.13). Although there was not much difference amongst the models, the OLS Categorical reported the lowest MAE in more deciles than any other (5) and also had the lowest difference between errors for the groups above and below 0.5 (0.101 and 0.198, respectively).

Figure 3.3
Mean absolute error for all models by EQ-5D decile *

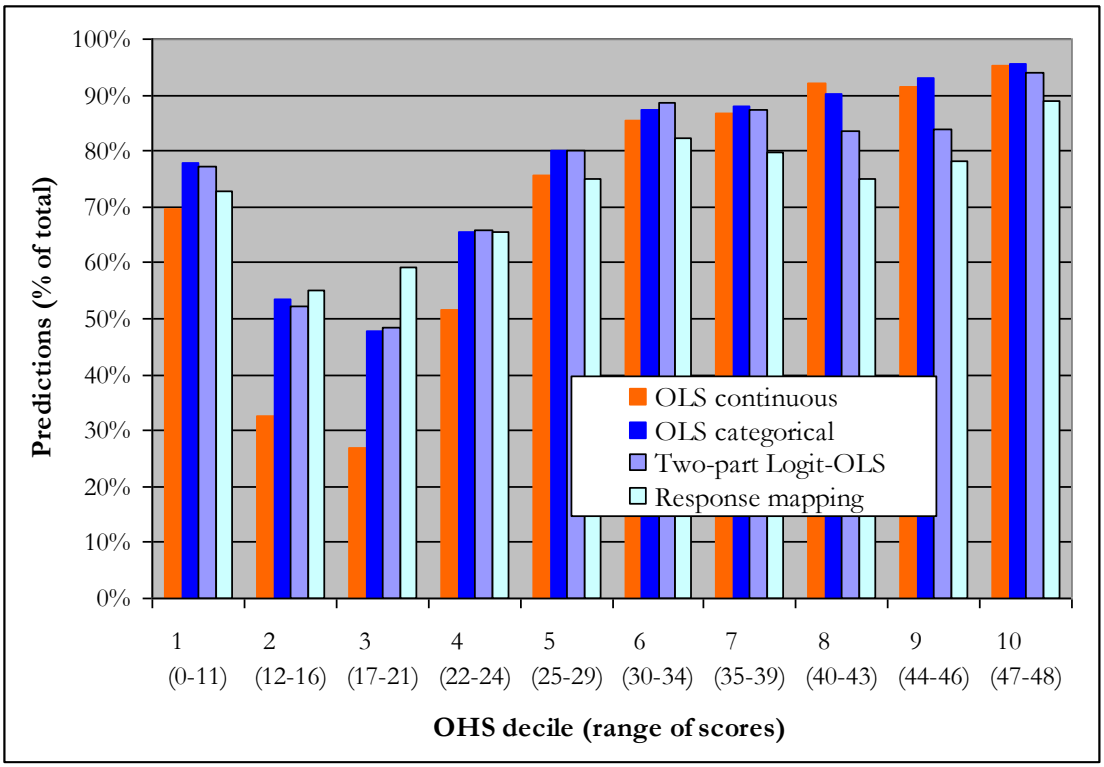


* When generating deciles of fitted EQ-5D scores, Stata forced equal values into the same group. As a result, the two models predicting scores of one (two-part model and response mapping) had a very large ninth decile and an empty tenth decile. Only the OLS categorical model had the same number of cases in each decile.

These results were consistent with the proportion of fitted individual scores estimated within 0.1 of the observed value by decile, since closer predictions were also more common for healthier patients (higher observed OHS and EQ-5D). As Figure 3.4 shows, for the five top deciles of observed OHS with values ranging from 30 to 48, all models

except Response Mapping estimated more than 80% of the observations within 0.1 of the observed EQ-5D score. In fact, for the highest scores of OHS (40 or above) the two OLS methods predicted more than 90% of all observations with this level of accuracy. This contrasts with observations where OHS was between 17 and 21 (the third decile), where OLS Continuous, for example, would only predict values with a residual lower than 0.1 for less than 30% of cases. If OHS was lowest (less than 12), however, predictive power improved so that predicted EQ-5D scores (also very low) fell within 0.1 of a health-state utility estimate of the observed value in 70% to 80% of cases.

Figure 3.4
Highly accurate EQ-5D predictions by observed OHS decile *



* "Highly accurate" is used to mean predicted observations where the absolute value of the residual < 0.1. Values in brackets under the decile number indicate the range of observed OHS scores included in the decile.

Table 3.5 shows the performance of the four models when fitted to the validation sample. They were all able to estimate the mean EQ-5D score with less than 0.005 of a health-state utility estimate difference from the observed mean, with Continuous OLS achieving the closest estimation. Once again, only the Response Mapping approach was able to fit scores well into the negative range. The mean for pre-op records, computed on 1,478 observations, was predicted with lower error than that of postoperative observations, based on 138 cases only. Prediction error of the mean EQ-5D score for pre-op records was between -0.0002 and +0.007, whilst for post-op observations it was -0.03 for all

models. The four models showed an approximately normal distribution of residuals, largely spread out over values of observed EQ-5D.

Table 3.5
Performance of models and external validation *

Model: regressors	Mean fitted EQ-5D	Difference of means (observed - fitted)	Range of fitted EQ-5D	Range of residuals	% within 0.10 utility	R2 EQ-5D observed vs. fitted	RMSE† EQ-5D observed vs. fitted
Continuous OLS: Total OHS	0.3805	-0.0005	-0.070 , 0.995	-0.78 , 0.67	25.4%	0.56	0.23
Categorical OLS: All OHS questions	0.3845	-0.0045	-0.165 , 0.967	-0.75 , 0.77	42.2%	0.63	0.21
Two-part Logit-OLS‡	0.3820	-0.0020	-0.154 , 1.000	-0.83 , 0.77	42.0%	0.64	0.21
Response mapping: All OHS questions	0.3758	0.0042	-0.429 , 1.000	-0.91 , 1.07	44.4%	0.45	0.26

* For the validation sample: n=1,616. Mean observed EQ-5D score=0.38, with range [-0.594, 1]
† Root mean square error of the regression

‡ Logit stage uses 8 OHS questions, OLS uses all OHS questions

3.4 Predicting EQ-5D change after THR

3.4.1 Justification

For many economic evaluations, the change in EQ-5D after THR may be more relevant than the specific pre and postoperative scores. In cases where EQ-5D data were not collected but OHS were, it would be possible to predict the change in EQ-5D based on the OHS data available by making use of the algorithms developed in the previous section. By applying those mapping algorithms, estimates for the pre and postoperative EQ-5D scores can now be derived and a predicted change calculated. It is not clear, however, whether the prediction errors reported in the previous section would result in a higher predictive power of the change, compared to the alternative of predicting the change in EQ-5D directly based on the observed OHS scores.

As shown in Section 3.3, models predicting the mean EQ-5D score for a group of observations report high predictive power, although errors vary with severity of symptoms as reflected in observed OHS scores (see Figure 3.4) and therefore with the stage with respect to surgery (pre-op or post-op). If the sources of mapping predictive errors are the same for both the pre-op and post-op groups (for example, an inability of the algorithms to map the anxiety/depression dimension of the EQ-5D) and the resulting

prediction bias is in the same direction in both cases, then estimating the change in EQ-5D via individual predictions of the pre- and post-op scores would further increase the error in said direction. This could be avoided, or at least minimised, by estimating the change in EQ-5D based directly on the observed OHS scores instead.

In this section we compare the predictive power of models estimating the change in EQ-5D scores by separately estimating the pre- and post-op scores, using the methods assessed in the previous section, against a direct prediction of the change based on observed OHS scores.

3.4.2 Data, methods and validation

The data used for this analysis are the same as that described in Section 3.2. The change in scores (both for the EQ-5D and OHS) was considered to be the scores reported six months after THR, minus the pre-op scores.

We used two OLS models for the direct estimation of change in EQ-5D scores. The first specified the change in total OHS scores whilst the second used the change in each individual OHS question as explanatory variables. We used Stata's *cluster* option to control for multiple operations on the same patients and tested for the significance of age, gender, deprivation, baseline BMI and baseline OHS effects on the prediction of change. Main assumptions of OLS were verified as in the previous section. For both models, a manual backward stepwise regression method was applied so that variables not statistically significant at predicting change in EQ-5D in the multivariate equation were progressively dropped. As comparators, we included all four models explored in Section 3.2.

Performance at estimating change in EQ-5D was evaluated by looking at the difference in mean score change, the range of fitted values compared to the range of those observed, residuals, the percentage of fitted variations that fell within 0.1 of a utility value of the observed change, R^2 of the regression between observed and fitted score changes, MAE, and mean square error (MSE).

Validation followed essentially the same protocol explained in Section 3.2. Given the lack of a properly external dataset with the required scores, we used an out-of-sample subset comprised of those records excluded from the estimation data set but for which we had both the pre-op and post-op scores. Since this subset was comprised of only 30

observations, we extended this validation by performing a bootstrapping in order to obtain confidence intervals for the predicted EQ-5D changes. Bootstrapping as a form of validation has been employed before in at least one other mapping exercise [70].

3.4.3 Results

3.4.3.1 Descriptive statistics

This analysis was based on the same observations selected for the previous section on mapping OHS onto EQ-5D. Selected data are compared to excluded records in Table 3.1 under Section 3.3.1. It is important to note, however, that for the analysis in this section not all records in the excluded category can be effectively compared with the selected portion of the dataset, because in the excluded group there are some observations with missing pre- or post-op OHS or EQ-5D. Therefore, of all excluded observations we could only calculate change in OHS and change in EQ-5D for some and not for all.

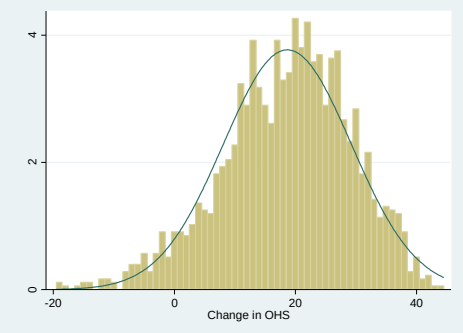
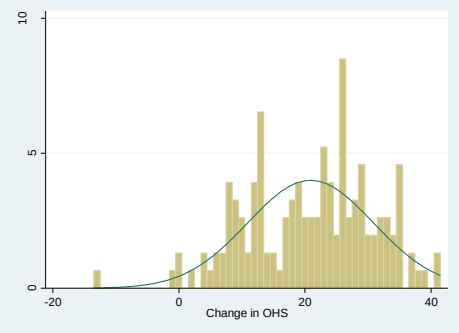
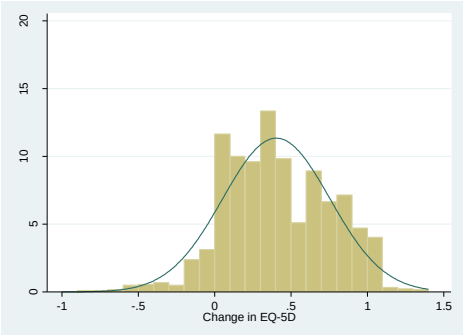
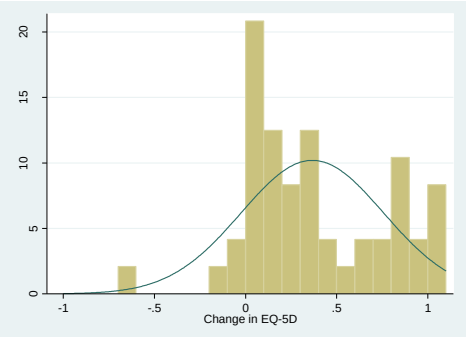
Bearing this in mind, we compared the shape of the distributions of change in OHS and EQ-5D of the selected group to those whose scores were available from the unselected, shown in Figure 3.5. Variation in OHS was approximately normally distributed for the group of 1,759 selected observations which recorded a mean change of 18.7. For the 153 excluded observations the distribution was scattered but with a mean change not far from the latter value at 20.9.

In the case of EQ-5D the distribution does not appear to follow any specific form, although it tends to a symmetric shape centred on its mean at 0.404. There were only 48 excluded observations for which a change in EQ-5D could be calculated, with a distribution very much scattered and different from that of the selected cases, although once again the mean change in EQ-5D was not too distant at 0.366. Because the mean score changes between the two groups are not far from each other and it is their distributions that set them apart, it is likely that the differences are due to the significantly lower number of observations in the excluded group. We are therefore left with no reason to believe that we are using a biased sample of the dataset.

Since one of the methods we want to compare in this section is the direct estimation of EQ-5D change based on observed OHS scores, a visual exploration of the association between both score changes is in order. Figure 3.6 shows a scatterplot of these values, for the selected observations, indicating a clear positive relationship between the two. This

visual association persists when producing the same graph by gender, side of the hip where the replacement was performed, age or BMI group. Graphs for these are shown in Appendix 7.

Figure 3.5
Distribution of OHS and EQ-5D change for selected and excluded data *

	Selected data	Excluded data
OHS change		
EQ-5D change		

* Number of observations:
Selected data: Change in OHS = 1,759. Change in EQ-5D = 1,759.
Excluded data: Change in OHS = 153. Change in EQ-5D = 48.

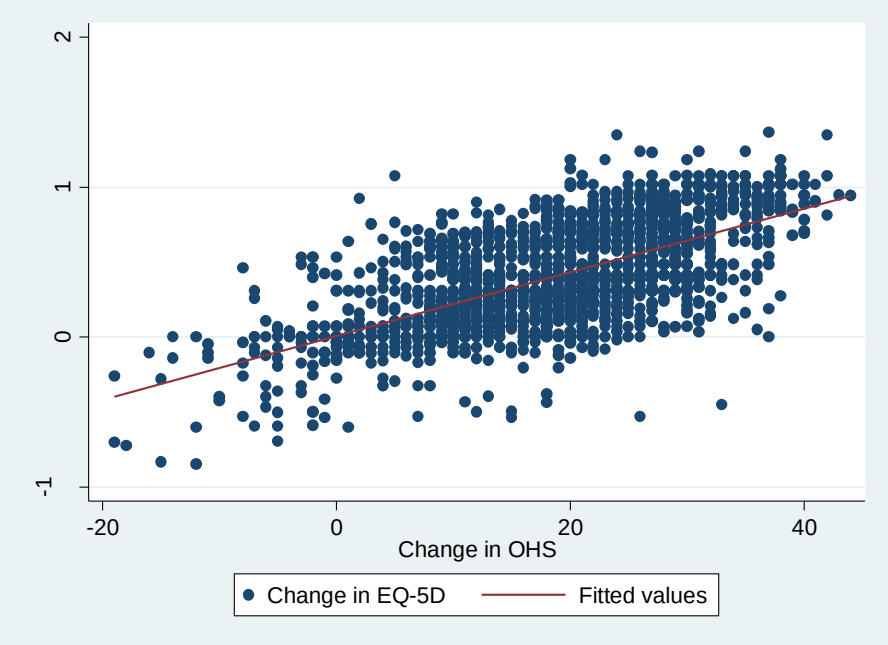
3.4.3.2 Model performance

For the first OLS model using total change in OHS as the main regressor (Model A), OHS change expectedly proved to be strongly and consistently significant together with the baseline OHS score, whilst gender and age (and age^2) were quickly dismissed as significant explanatory variables.

BMI was tested in the model both as a continuous score (linearly and squared) and as a categorical variable following the World Health Organisation's criterion [71]. As categories, all groups were not statistically significant with p-values ranging from 0.266 to 0.653. The squared BMI score was significant, however, both with all original covariables (p=0.013) and with only baseline and total change in OHS (p=0.004) in the right-hand

side of the equation. These results suggest that BMI may play a role in predicting the change in EQ-5D based on the variation on the OHS total score; however, given the low number of observations with recorded height and weight included in the dataset (*n* drops from 1,759 to 361), and that this group may be a biased sample of the original dataset, we decided to exclude BMI from the remainder of the analysis in order to capture and fully explore the variations found in a larger portion of the sample.

Figure 3.6
Change in OHS by change in EQ-5D



When the model was estimated excluding BMI, deprivation category 2 was statistically significant at the 0.05 level. Deprivation was included as a categorical variable by quintiles of the *Index of Multiple Deprivation 2004* at the Lower Layer Super Output Area level (with 1 corresponding to the most deprived quintile and 5 the least deprived). After excluding age and gender, categories 1 and 3 of deprivation also became statistically significant at the 0.05 level. However, the value of coefficients for the different levels did not follow a consistent progression, thus making the apparent effect of deprivation on the prediction of EQ-5D change unclear. In addition, data on deprivation is seldom available hence including this variable in the model is likely to limit its potential use by researchers. We therefore chose to explore the effects of deprivation only up to this point and to proceed with the two variables that were consistently significant as explanatory variables in all cases: baseline and change in OHS.

Table 3.5 shows the output of the linear OLS regression where baseline OHS and change in OHS after THR explain 43% of the variation in changes in the EQ-5D score. The positive coefficient associated to the change in OHS is consistent with the positive relationship observed in Figure 3.6, and the negative coefficient on its baseline score indicates that a given positive change in OHS would be associated to a lower EQ-5D effect for patients with higher pre-op OHS than those with lower baseline scores. This seems reasonable since it is perfectly plausible that patients with worse symptoms before the operation attain greater changes in their HRQL after the procedure.

Table 3.5
Linear regression output from OLS Model A *
Dependent variable: EQ-5D summary score change

Independent variables	Coefficient	Robust Standard Error †	95% confidence interval	
Baseline OHS	-0.0067	0.001	-0.008	-0.005
Change in OHS	0.0180	0.001	0.016	0.020
Constant	0.2011	0.033	0.137	0.266

* Number of observations = 1,759
F(2,1685) = 684.37
Prob . F = 0.000
R-squared = 0.429
Root mean square error = 0.266
† Standard error adjusted for 1,686 clusters of patients

The second OLS model considered for the analysis used the change in each individual OHS question as continuous explanatory variables (Model B). After results from the previous model specification, we found no basis for the inclusion of gender, age, BMI or deprivation, but we did keep baseline OHS as a regressor. The first model estimation identified half the questions as significant and the other half not, with baseline OHS again being significant. We tested the removal of the six questions that were not significant by performing likelihood-ratio test after estimation, having to drop the *robust* option to be able to do this. In all cases but one, the test indicated that the question did not make a statistically significant difference and could be removed from the model. The question on *Climbing stairs* was therefore reintroduced into Model B, now specified with seven questions and baseline total OHS. The output for this model is shown in Table 3.6.

Model B also reports pre-op OHS with a negative coefficient quite close to that of Model A, whilst the seven questions included in the model have all positive coefficients. This is logical given that any increases in OHS questions would naturally be associated with

improvement in the patient's health state, reflected in a positive EQ-5D summary score change. Coefficient values vary amongst questions, as was the case with the mapping results in the previous section, in this case with *Pain interfering with work* and *Description of pain* reporting the highest effect on the change of EQ-5D score. This model, incorporating changes in the specific OHS questions, explains 44% of the variation of EQ-5D changes, roughly the same as Model A. Both Models A and B produced residuals that were highly normally distributed and centred on zero.

Table 3.6
Linear regression output from OLS Model B *
Dependent variable: EQ-5D summary score change

Independent variables	Coefficient	Robust Standard Error †	95% confidence interval	
Baseline OHS	-0.0082	0.001	-0.010	-0.006
Δ Description of pain	0.0398	0.005	0.004	0.024
Δ Night pain	0.0141	0.005	0.004	0.024
Δ Limping	0.0169	0.005	0.006	0.028
Δ Climbing stairs	0.0179	0.008	0.003	0.033
Δ Putting on socks	0.0224	0.006	0.011	0.034
Δ Stand from chair	0.0360	0.008	0.020	0.052
Δ Work interference	0.0500	0.008	0.035	0.065
Constant	0.2105	0.032	0.148	0.273

* Number of observations = 1,759
F(8,1685) = 176.25
Prob . F = 0.000
R-squared = 0.443
Root mean square error = 0.263
† Standard error adjusted for 1,686 clusters of patients

3.4.3.3 Validation

Internal validation was performed by fitting both models to the same estimation dataset. Models A and B are OLS models, as are two of the four estimated in Section 3.2, therefore we did not expect the prediction of mean EQ-5D change to have much informative value given the nature of the OLS method to minimise mean residuals. However, the rest of the indicators provide a more objective picture of model performance. Table 3.7 shows how all models explored in this chapter compare to each other in predicting the change in EQ-5D after THR.

As expected, Models A and B produced a perfect estimation of the mean, whereas the four models that predicted the pre- and post-op scores separately, before calculating the

difference, led to mean score changes with a margin of error between 0.004 (OLS Categorical) and 0.045 (Response Mapping). The latter models were more able to predict large changes (both decreases and increases) in the EQ-5D summary score approaching the observed range, with Response mapping achieving the largest range even beyond the observed maximum decrease. Residuals were similar amongst all models and the percentage of predicted changes falling within 0.1 of an estimated utility value varied between 25% (Continuous OLS) and 34% (Categorical OLS). R^2 of the linear association between observed and predicted values was between 0.4 and 0.5, with the exception of Response Mapping which only reached 0.31. Categorical OLS reported the highest association, as well as the lowest measures of error. Results shown in Table 3.7 are consistent with the graphical representation of observed versus predicted changes in EQ-5D for Models A and B (shown in Figure 3.7).

Table 3.7
Performance of models and internal validation

Model: regressors	Mean EQ-5D change	Range of EQ-5D change	Range of residuals	% within 0.10 utility	R2 EQ-5D observed vs. fitted	MAE	MSE
Observed	0.404	-0.847, 1.367					
Models predicting change of EQ-5D directly							
Model A: Δ total OHS	0.404	-0.376, 0.965	-1.164, 0.845	29%	0.428	0.212	0.071
Model B: Δ individual questions	0.404	-0.407, 0.930	-1.100, 0.794	29%	0.443	0.210	0.069
Models predicting pre- and post-op EQ-5D scores separately							
Continuous OLS: Total OHS	0.414	-0.422, 0.977	-1.184, 0.963	25%	0.410	0.220	0.073
Categorical OLS: All OHS questions	0.400	-0.543 , 1.042	-1.018, 0.758	34%	0.494	0.194	0.063
Two-part Logit-OLS *	0.403	-0.565 , 1.154	-1.074, 0.844	33%	0.460	0.203	0.068
Response mapping: All OHS questions	0.359	-1.016 , 1.239	-1.040, 1.035	30%	0.312	0.250	0.110

* Logit stage uses 8 OHS questions, OLS uses all OHS questions

Validation based on fitting models to the same dataset that was used to estimate them is not ideal. As explained in Sections 3.2.5 and 3.3.3, given the lack of an external dataset, we identified those observations with pre- and post-op scores from the subset originally excluded from this study to build an out-of-sample validation dataset. In this case, only the 30 observations excluded on the basis of missing gender, age or deprivation qualified.

Table 3.8 compares the performance of all six models after being fitted to this out-of-sample validation dataset.

Figure 3.7
Observed versus predicted change in EQ-5D

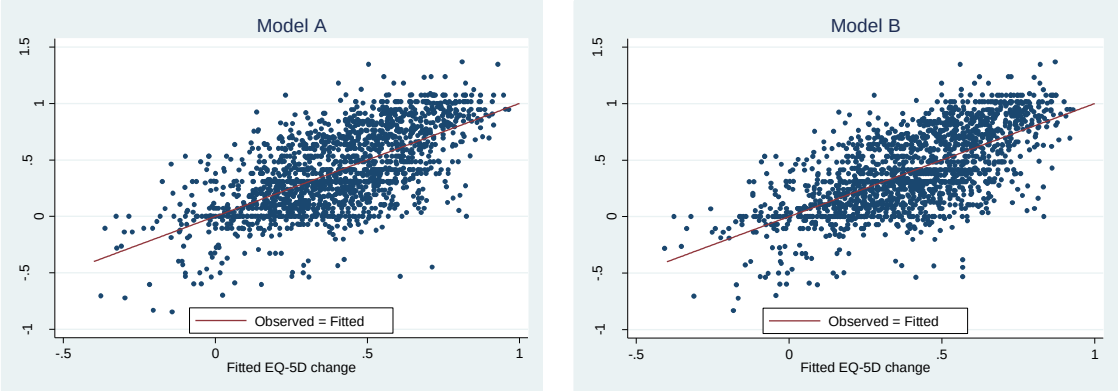


Table 3.8
Performance of models and out-of-sample validation

Model: regressors	Mean EQ-5D change	Range of EQ-5D change	Range of residuals	% within 0.10 utility	R2 EQ-5D observed vs. fitted	MAE	MSE
Observed	0.357	-0.603, 1.074					
Models predicting change of EQ-5D directly							
Model A: Δ total OHS	0.443	-0.174, 0.847	-0.596, 0.464	23%	0.552	0.256	0.095
Model B: Δ individual questions	0.448	-0.119, 0.839	-0.556, 0.425	20%	0.590	0.254	0.091
Models predicting pre- and post-op EQ-5D scores separately							
Continuous OLS: Total OHS	0.438	-0.289, 0.844	-0.622, 0.439	23%	0.604	0.235	0.083
Categorical OLS: All OHS questions	0.454	-0.297, 0.970	-0.583, 0.342	40%	0.675	0.207	0.072
Two-part Logit-OLS *	0.448	-0.299, 1.024	-0.549, 0.445	40%	0.662	0.211	0.072
Response Mapping: All OHS questions	0.396	-0.090, 0.924	-0.603, 1.021	40%	0.364	0.244	0.122

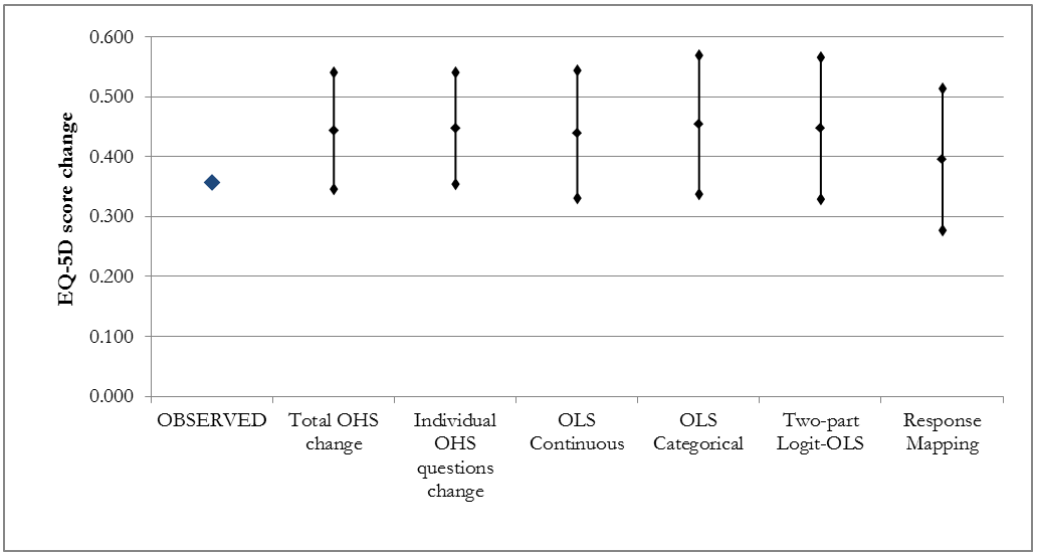
* Logit stage used 8 OHS questions, OLS used all OHS questions

All models overestimated the mean EQ-5D change. Response Mapping achieved the closest estimate with an error of 0.039, whereas the estimated mean by Categorical OLS was the highest (at nearly 0.1 of a utility estimate over the observed mean change). The predicted means produced by Models A and B were approximately 0.09 higher than the observed. Considering individual estimations, the Two-part Logit-OLS model achieved

the widest range of estimated EQ-5D change and the one closest to the observed spread. The Two-part model, Categorical OLS and Response Mapping achieved the highest proportion of estimations within 0.1 of an estimated utility at 40%. Categorical OLS reported the highest R^2 between observed and fitted values as well as the lowest errors.

Due to the low number of observations, we produced 95% confidence intervals of the predicted EQ-5D changes after fitting each model by taking 1,000 bootstrap samples of 30 observations with replacement from the out-of-sample validation dataset. Figure 3.8 shows how the observed EQ-5D change falls inside the 95% confidence interval of all six models. Results suggest that, even for relatively small groups, the two direct methods of EQ-5D change estimation as well as all four mapping algorithms developed in this chapter report similarly high predictive power of the mean EQ-5D change after THR.

Figure 3.8
Confidence intervals of predicted mean EQ-5D change by bootstrapping on out-of-sample validation



We did not explore whether predictive power varied with severity of the disease. It is unlikely, however, that pre-op OHS levels would have much impact on predictive power, since most patients tend to be at similar levels in the lower end of the OHS scale before surgery.

3.5 Discussion

Preference-based utility measures such as the EQ-5D should ideally be collected directly from patients, but in certain cases this is not possible. The present work shows that

models estimated here have a high predictive power when mapping OHS responses onto the summary EQ-5D score, and OHS changes onto EQ-5D change. Furthermore, it demonstrates that all models employed here for score mapping are able to estimate the mean EQ-5D index with a high level of precision. The simplest OLS Continuous model achieved the closest estimation of the mean EQ-5D score, whilst Response Mapping proved to be the only approach capable of estimating individual scores well into the negative range and up to full health. An additional benefit of Response Mapping is that it allows for the estimation of mean EQ-5D scores using different valuation tariffs. For all models, predictive power varied considerably across the range of fitted EQ-5D scores with MAE for predicted low EQ-5D scores doubling that of higher fitted values, a tendency also found in a previously published cross-walking study linking a condition-specific measure to a generic one [72]. The OLS Categorical model reported lower predictive errors across the range of scores than the other models. Overall performance of the four models was within range of other reported mapping studies, based on their root mean square errors of around 0.20 [64].

Results of the Continuous OLS model indicate that, based on the data used, 67% of the variation of hip patients' EQ-5D scores is explained by their OHS score. In other words, most of the variability in their HRQL, as measured by EQ-5D, is associated with the impact their hip problem has on the pain and limitations they experience. In the sample analysed by Brown and colleagues [67] (the only other publication to date reporting the relationship between OHS and EQ-5D, the portion of the EQ-5D variance explained by their model was 42%), a coefficient of 0.02 in the regression of the total OHS score on EQ-5D was reported, the same value we obtained in our OLS Continuous model.

We found an association between severity of health problems and models' predictive power of individual scores so that, in general, better health leads to lower predictive errors of EQ-5D score mapping. This tendency, though explored by only a few authors in the past [64], has already been found in studies cross-walking from disease-specific [72] and generic measures [69] onto EQ-5D. Based on this, and on our own calibration, mapping from OHS to EQ-5D should produce estimates with lower predictive errors after surgery. Our external validation, nevertheless, reported a high predictive power of the preoperative mean with models reporting errors that did not exceed 0.001, whilst the post-op mean score was overestimated by 0.03 by all models. Postoperative observations, however, were only 138 (compared with 1,478 before surgery). This suggests not only that the number of

observations affects predictive power, but also that a similar number of postoperative observations might have produced even lower predictive errors than those obtained preoperatively. This result sustains our assumption of a valid single functional relationship between the OHS and EQ-5D regardless of the timing of data collection.

In addition to predicting EQ-5D scores, the categorical OLS regression model also provided information on the relationship between each OHS question and EQ-5D. There were wide variations in coefficient values for different OHS questions and their response levels. Even though the OHS scoring system assigns equal weight to each question, the results presented here suggest that each question may have very different impacts on the HRQL of patients. OHS questions associated with pain (e.g. *Pain interfering with work*, *Description of pain* and *Pain from standing up from a chair*) have greater effect on EQ-5D than questions about mobility or daily routine.

The mapping exercise benefited from pooling together pre and postoperative responses to the questionnaires, hence providing good power and the full range of scores for model estimation. We also found a number of similarities between EQ-5D and OHS; for example, both ask about pain, mobility, and ability to perform tasks and functions. We felt that this was an extremely important factor in the good performance of the mapping algorithms. Similar mapping exercises are likely to be sensitive to similarities between instruments and it is very likely that mapping would perform poorly in cases where instruments are very different.

The mapping was performed using regression techniques which are very widely used and well understood, which facilitated analysis and interpretation of results. There are some limitations, however, that should be borne in mind when interpreting results. Although there is a substantial overlap between OHS and EQ-5D questions, there is one exception. One of EQ-5D's dimensions explores anxiety/depression, which is not covered in the OHS questionnaire; this limits the ability of the disease-specific measure to predict the scores of the generic one. Also, we would ideally like to have used a completely different dataset for external validation from that used for estimation. Our estimation and validation datasets are bound to have shared many characteristics; nevertheless, the large sample size and wide distribution of scores support the reliability of results. Although both the estimation and validation subsamples came from the same cohort, their method of selection made the validation process more robust than if they had been selected

randomly. In most mapping studies, validation samples are built by randomly selecting cases from the same estimation dataset [73]. By doing so, the validation may simply confirm that the selection was truly random instead of actually testing whether results would vary on different data. Using a non-randomly selected validation sample, we were able to test the validity of the mapping methods whilst controlling for the equivalence effect of randomisation.

Estimating direct mapping functions for cross-walking changes in OHS onto changes in EQ-5D after THR allowed us to test whether the more direct mapping of the change might overcome the addition of errors in mapping the individual pre- and postoperative scores. The lack of OHS questions directly linked to the depression/anxiety dimension in the EQ-5D could explain much of the individual score prediction errors. Mapping individual scores and then calculating the change might therefore produce larger errors than applying a direct mapping algorithm for the change in EQ-5D. This would be the case, for example, if the effects of the missing questions on depression/anxiety were of similar magnitude and in the same direction pre- and postoperatively, thus cancelling each other out. Models A and B estimated in Section 3.4 performed no better at estimating the change in EQ-5D after THR than any of the four models that mapped individual scores. This suggests that individual score prediction errors are not cancelling each other out, and indicates that the mean change in EQ-5D of a group of patients can be predicted just as closely to the true values from change in total OHS and in seven of its individual questions, as it can from estimating each score individually and then calculating the difference.

Descriptive analysis of the change in OHS and EQ-5D after THR also served to show that, despite the significant benefits of this operation to most patients (reflected in a mean increase in total OHS score of 19 points and an impressive 0.4 in the EQ-5D summary score) for some patients the outcome is in fact negative. For 5% of patients in the estimating dataset their total OHS score actually decreased, and for 8% of patients the HRQL measure also dropped at six months after they had the operation. These are some of the poor outcomes for patients that the outcome prediction tool evaluated in this thesis would hope to avoid.

Descriptive statistics in Section 3.4 also showed a positive relationship between change in OHS and change in EQ-5D. The large overlap of both outcome measures discussed

above is thus manifested in a positive relationship between the indices and also between their changes after THR. We also found that BMI may play a role in predicting the change in EQ-5D. The coefficient on the statistically significant squared BMI score was positive (0.0001) which suggests that, *ceteris paribus*, the same change in overall OHS would translate into greater change in EQ-5D for patients with larger BMI. For example, in a patient with normal BMI (BMI score=21) the effect of a given change in OHS would result in a change in EQ-5D plus 0.05. For an obese patient (BMI score=35), the change in EQ-5D would be increased by 0.14. With access to a larger dataset including patients' BMI scores, it would be worthwhile exploring this effect more comprehensively.

Of the two models estimating the change in EQ-5D, Model B, which used the change in individual OHS questions as explanatory variables, showed, again, that not all questions have the same predictive power over the summary EQ-5D score (and in this case, over its change either). Consistent with the results shown by the OLS categorical model when mapping scores, OHS questions related to pain (*Pain interfering with work, Description of pain* and *Pain from standing from a chair*) also reported the highest coefficients and hence are the most powerful predictors of change in EQ-5D based on change in OHS questions after THR.

Finally, internal and out-of-sample validation of the models directly mapping change compared to the four models mapping individual scores showed that they perform quite similarly at estimating EQ-5D change after THR. The OLS Categorical achieved slightly better results in the internal validation and, together with the Two-part and Response Mapping, also reported higher predictive power when tested on the out-of-sample dataset. Models overestimated the observed mean change in EQ-5D in the out-of-sample validation by 0.04 in the best case (Response Mapping) to 0.1 estimated utilities. One of the most important limitations of the validation was the size of the out-of-sample dataset, and that it was not strictly external. Future research should validate these models in much larger and completely independent datasets; however, even with the 30 observations used here, confidence intervals produced by bootstrapping included the observed change in EQ-5D for all six models. Given these results and based on its simplicity, the algorithm mapping the change in EQ-5D from the change in total OHS score appears to be the most convenient for researchers to apply when intending to estimate EQ-5D changes after THR based on OHS scores.

To conclude, the mapping methods tested here enable researchers, clinicians and policymakers to obtain reliable estimates of mean EQ-5D scores and mean changes thereof after THR when these are not directly collected but responses to the OHS questionnaire are available. In Chapter 5, we report on the use of the mapping methods developed in this chapter to produce utility estimates based on OHS measures collected in the absence of EQ-5D data in order to populate our cost-effectiveness analysis. The models presented in this chapter report high predictive power. It is important to stress that, if mapped scores are to be used as part of economic evaluations, the uncertainty added by the mapping process must be properly incorporated into the analysis.

3.6 Contributions

Sections 3.2 and 3.3 are based upon a paper published in *Quality of Life Research* in March 2012: *Mapping the Oxford hip score onto the EQ-5D utility index* [74]. Rafael Pinedo (RP) is the first and corresponding author of the publication, with David Turner (DT), Andrew Judge (AJ), Prof. James Raftery (JR), and Prof. Nigel Arden (NA) as co-authors. RP was responsible for all sections including managing the dataset, writing the Stata code, running all econometric analyses and writing all sections of the manuscript. RP discussed with DT, AJ, JR and NA the results of the study throughout the process and received advice from AJ on specific elements of the statistical analysis. All co-authors read the final draft of the manuscript and provided comments. As an aid to researchers interested in applying the mapping algorithms to their own data, the published paper included a step-by-step guide on how to conduct the mapping following each of the four algorithms for an individual score. Section 3.3 includes comments and a figure on the predictive power of EQ-5D summary scores by observed OHS decile that were not included in the published paper, and were fully developed by RP. Section 3.4 on the estimation of EQ-5D change after THR was not part of the published paper and was also fully developed by RP.

As stated in the acknowledgements section of the published manuscript, we are grateful to the COASt team at the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences at the University of Oxford, in particular Dr Kassim Javaid and Dr Amit Kiran, for the helpful discussions of earlier versions of this paper. We also thank Mr Richard Field for allowing the use of the SWLEOC data as well as Dr Oliver Rivero-Arias from the Health Economics Research Centre at the University of Oxford for clarifications on the response mapping approach. An earlier version of this work was presented at the Winter 2011 conference of the Health Economists' Study Group

(HESG) at the University of York, where we received helpful comments. Earlier versions of this study were also presented at the University of Southampton's Medicine and Biological Sciences Postgraduate Conference (June 2011, poster) and the International Society for Pharmacoeconomics and Outcomes Research 3rd Latin America Conference in Mexico (September 2011, oral). An earlier version of the work on mapping the change in EQ-5D from the change in OHS (Section 3.4) was presented at the Summer Conference of the HESG in Wales (July 2011, poster).

4 A model for the cost-effectiveness of a THR outcome prediction tool

4.1 Introduction

As discussed in Chapter 2, most economic evaluations of THR have assessed the cost-effectiveness of different prostheses or compared THR against other surgical interventions. They conducted the analysis starting at the point of surgery or immediately prior to that and, although most studies reported using a societal perspective or that of the health care system, they all assessed the costs and benefits of the alternatives separately. This chapter presents an economic model designed to assess the cost-effectiveness of the implementation of an outcome prediction tool for THR. Hence, the model presented here does not focus on the costs and effects of the operation, but instead it evaluates the costs and effects that may result from using the outcome prediction tool to guide treatment decisions compared to current practice. The intervention being assessed in this model is the use of the outcome prediction tool. The model therefore starts at a point previous to surgery since the tool is expected to inform and improve the referral of patients to THR based on predicted outcome. The perspective is that of the health care system but, in comparison to previous studies, the model presented here includes both surgical and non-surgical alternatives in the same model. This allows more accurate representations of clinical practice with and without the outcome prediction tool to be compared.

This chapter explores the prediction of THR outcomes and the grouping of these into two categories before explaining in detail the model design developed for this study. Section 4.2 explains the context of outcome prediction tools in general and describes the one developed by COASt and assessed with the economic model presented in this chapter. We then make a distinction between *Good* and *Poor* outcomes after THR with the classification system explained in Section 4.3. In Section 4.4 we justify the model type chosen. In Section 4.5 we describe the care pathway of patients commonly referred to THR, upon which the model design was based. Section 4.6 describes the model structure in detail and Section 4.7 briefly specifies the comparators and perspective employed for the analysis. In Section 4.8 we define the patient subgroups for which the analyses were performed and in Section 4.9 we clarify the contributions of other researchers to this chapter. The model input parameters used for the analysis are presented separately, in Chapter 5.

4.2 Predicting outcome after THR

The development of a prediction tool for hip arthroplasty is not unprecedented. Judge *et al* [75] created a clinical risk prediction tool to identify patients most likely to obtain an improvement in physical function eight years after surgery according to the physical functioning scale of the SF-36 instrument [76]. The tool produces a score based on the patient's gender; age; SF-36 physical function baseline score; radiographic grade of the hip; presence of previous injury to the hip; and number of painful joints in hands, shoulders and knees. The score is then compared to a threshold above which the patient would be deemed most likely to observe a 30 point improvement in the SF-36 physical function section at eight years after the operation. This tool performed well when assessed based on calibration and discriminatory power, yet the model was estimated based on a relatively small number of patients (249) from Portsmouth and North Staffordshire, and the authors were not able to externally validate results. Furthermore, the tool was not based on a measure specifically designed for hip replacement patients but instead on one of the scales of a generic health measure such as the SF-36. This, together with the complexity of the logistic model employed to produce the final risk score, may limit the ability of health care professionals and patients to take full advantage of the tool to help make treatment decisions.

For the development of the outcome prediction tool within COASt, a combined analysis of two different population-based cohort studies was undertaken. The project team identified the main predictors of outcome for a primary THR based on postoperative OHS. Data from 1,589 patients of EPOS and 908 from the Health Technology Assessment of Hip Arthroplasty project in Europe (EUROHIP) were pooled together to explore their predictive power of hip surgery outcome.

The two cohort studies were dissimilar in many respects. In terms of physical outcome measures, for example, EPOS recorded OHS and the SF-36 Physical functioning dimension, whilst EUROHIP used the Western Ontario and McMaster Universities Arthritis index (WOMAC) adding the OHS for its English cohort. In terms of explanatory variables, the two cohorts recorded a wide and different array of variables. Risk factors considered for the analysis included preoperative OHS, age, gender, and BMI. The studies also collected information on whether patients were employed, their use of medication and the type of prosthesis used. Many variables, however, were collected by either of the two studies, but not both. For example, EUROHIP collected data on

whether patients lived alone, details of their employment, whether they cared for someone else, education, number of years with hip pain, ASA grade (a measure of fitness for anaesthesia and surgery of patients before an operation), and the number of preoperative expectations (number of thematic groups, out of a total of 15, into which the patient's expectations, of what they would need to do for them to consider the operation successful, were classified), among other factors. EPOS recorded occupation, the SF-36 mental health score, the number of comorbidities prior to surgery, surgical approach, grade of the operator, and the sizes of the different parts of the prosthesis (head and stem), among other potential explanatory variables for primary THR outcome.

The combined dataset consisted of data for nearly 2,500 osteoarthritis patients including demographic and surgical variables. The development of the prediction tool was based on records from OA patients who reported both the preoperative and at least one postoperative OHS measure. Multiple imputation helped produce a complete dataset by imputing missing values when they were not collected or available for any of the explanatory variables. A linear regression model was then estimated based on data for the 2,092 patients meeting the inclusion criteria with postoperative OHS at 12 months as the dependent variable, adjusting for baseline OHS and controlling for centre effect as well as patient, radiographic and surgical variables. Backward selection using Wald Test was applied to fit the model repeated times via a bootstrapping process to identify the strength of the evidence. Variables kept in the model as significant in at least 70% of the 100 bootstrapped iterations were selected for the final model.

The resulting linear model predicts THR outcome, in the form of a postoperative OHS, based on preoperative demographic and surgical variables. Baseline OHS had the largest effect on the outcome variable, followed by posterior (as opposed to anterolateral) surgical approach and having attended college or university (as opposed to not having done so). Being older than 70 years of age and high BMI were negatively associated to surgery outcome, as were the number of joints with osteoarthritis and a lateral (as opposed to supine) position of the patient during surgery. Mental health and the number of expectations were both found to positively impact postoperative OHS in the sample. Table 4.1 shows the variables included in the final model specification, the percentage of iterations in which they were retained after the repeated backward selections, and the final coefficients. The model explained 25% of the variance of postoperative OHS in the sample.

The above model constitutes the outcome prediction tool whose cost-effectiveness is assessed using the economic model presented in this chapter. The tool could be estimated as a logistic model directly predicting *Good* or *Poor* outcome, but since that was not available at the time this thesis was written, the predictive model of the continuous postoperative OHS as a measure of primary THR outcome as presented in Table 4.1 was used as the intervention assessed. Classifying the predicted postoperative OHS into a category of *Good* or *Poor* outcome is discussed in the following section.

Table 4.1
Linear regression model for postoperative OHS after primary THR

Explanatory variables	Inclusion rate	Coefficient*
Preoperative OHS	100%	2.16
Age ≥ 70	71%	-1.12
BMI	100%	-1.02
Education: College/University	100%	2.04
Preoperative SF-36 mental score	100%	0.67
Number of expectations	97%	0.63
Pattern of OA: superomedial/medial/concentric	97%	-1.53
Number of joints with OA	100%	-1.16
Surgical approach: posterior	98%	2.05
Patient's position: lateral	82%	-1.41
Stem size (mm offset)	93%	0.21

* All coefficients statistically significant at 0.05 level with the exception of Patient's position: lateral

4.3 Surgery outcome categories

As discussed in the introduction, the THR outcome prediction tool is justified by the fact that many otherwise successful arthroplasties leave patients with no or little improvement in their pain and mobility, and an important proportion of them are dissatisfied with the results of the operation. This supports an assessment of hip replacements based on a measure of results as experienced by patients. As explained in Chapter 3, the OHS has been largely accepted as a valid instrument to measure pain and mobility for hip patients. It can therefore be used as a valid and reliable measure of outcome after THR and to distinguish between *good* and *poor* outcomes, between satisfactory and unsatisfactory results.

Introducing the element of satisfaction, however, can be controversial as it is closely linked with patients' expectations. On the one hand, what patients expect to achieve and their level of satisfaction after surgery may be entirely subjective. Some patients may hope to simply be able to carry out basic physical functions whereas others may have

irrationally high expectations once their joint is replaced. These expectations and related level of satisfaction may also vary considerably across gender, age and socioeconomic groups, among other factors. For example, older patients may be satisfied if they simply regain the ability to walk short distances without pain, whilst younger ones may consider it utterly disappointing if they are not able to practice sports. On the other hand, expectations seem to be directly linked to objective measures of surgical outcome. As coefficients of the model presented in the previous section show, the number of expectations helps explain outcome after THR. This is not unexpected because what patients expect to be able to do after the operation can be an important internal driving force for rehabilitation immediately after surgery and it can therefore impact on outcome at 12 months. Expectations may also be a valid indicator of the patients' assessment of their own potential for improvement, more accurate than certain assessments by health professionals and likely to be reasonable in most cases, leaving the irrationally high or low expectations to the odd few.

A classification of outcome categories based on a measure of pain and mobility after surgery such as the OHS combined with a measure of satisfaction would therefore allow the model to be highly sensitive to what currently happens with arthroplasty patients: most of them do well and are satisfied after surgery, but many feel unwell afterwards and are not satisfied. In fact, data from EPOS, one of the observational cohort studies used to build the outcome prediction tool for THR, shows that using satisfaction or prosthesis survival produces staggeringly different results of surgery outcome. Of the 1,589 patients taking part in the study, 1,053 reported their satisfaction with surgery results two years after the operation. Of these, 57 patients (slightly over 5%) reported being either *somewhat* or *very dissatisfied* with surgery results two years after their operation, yet none of them had a revision within the first four years and two had a revision at five years. It is clear, then, that a model for the assessment of hip replacements using prosthesis survivorship as a measure of outcome would overlook the reality of an important number of patients who are not satisfied with the result of their operation but would be considered a *successful* intervention as they did not require replacement surgery. Considering that 78,999 primary THRs were performed in England and Wales during fiscal year 2011-12 [54], assuming a 5% patient dissatisfaction rate at two years means that some 3,950 patients operated on that year alone would be expected to be dissatisfied with surgery results, but only a few of them would require a revision THR. The inability of surgery outcomes based on prosthesis survival to capture such basic factors as the state of pain and mobility for hip

patients favours the use of patient-reported outcomes such as the OHS and patient satisfaction with surgery, though the latter would have to be used with care given its subjective nature.

Of the models reviewed in Chapter 2, only two classified patients after surgery based on outcome measures other than prosthesis survivorship. Chang *et al* [19] used the American College of Rheumatology (ACR) Functional Classification for Hip Osteoarthritis published in 1949 [77], making a distinction between four groups termed *classes* according to the patient's ability to perform usual activities. Saleh *et al* [43] employed the HHS [78], a consistent and commonly reported outcome measure in the second part of the 1990's when their study was published, consisting of an index assessing pain, function, absence of deformity and range of motion. They created three outcome categories: improvement (HHS 80-100), fair (HHS 70-79) and poor (HHS 0-69). Although not used in an economic model according to our review of the literature, a classification of THR outcome based on the patients' postoperative OHS was also proposed by Kalairajah *et al* [61] distinguishing amongst excellent, good, fair and poor outcomes. This classification was obtained simply based on the cumulative frequency distribution of OHS with a measure of intra-category agreement, all from a sample of 115 patients.

For our model assessing the cost-effectiveness of the outcome prediction tool for THR, we chose a method of outcome classification that represents an improvement on the systems described above by combining the OHS with patient satisfaction after surgery. The method was developed by Arden *et al* [79] and it used data from 799 patients who received a THR at St. Helier Hospital in London to determine cut-off points on postoperative OHS associated to patient satisfaction after surgery. Receiver Operating Characteristic (ROC) curve analysis was used to identify 12 and 24-month postoperative OHS of 38 and 33, respectively, that maximised sensitivity and specificity at predicting satisfaction. Results varied by gender, age and BMI groups. The analysis was also performed to identify cut-off points in the change of pre- to postoperative OHS but results were very different for low- compared to high-scoring groups, an expected limitation given OHS's ceiling effect.

Although the authors did not report the number of patients in their dataset who scored below and above the thresholds [79], given the reported satisfaction rate of 91.9% at 12 months and assuming that their 24-month reported sensitivity and specificity of 89.7%

and 86.7%, respectively, also applied to 12-month results, Table 4.2 shows the expected percentage of patients in each outcome category. According to these figures, around 17% of patients would be expected to score under 38 at 12 months after THR.

Table 4.2
Expected relative size of surgical outcome categories at 12 months

		Satisfaction with surgery		
		Dissatisfied	Satisfied	Total
OHS	Poor (<38)	7.0	9.5	16.5
	Good (≥38)	1.1	82.4	83.5
	Total	8.1	91.9	100.0

The above method not only combines the OHS and patient satisfaction, which we have argued are reasonable measures for THR outcome as experienced by patients, but it also brings them together as they have been found to be correlated with each other. In other words, by using this method to determine a cut-off point, a line may be drawn in the OHS scale separating those patients who still have pain and mobility problems and are likely to be unsatisfied, from those who score well in their pain and function and are more likely to be satisfied with the operation. Compared to the classification systems used in earlier studies and reported above, it is evident that, by using the OHS, this new approach incorporates the main aspects of the ACR functional classes employed by Chang *et al* (ability to carry out regular activities) [19], and goes further by adding measures of pain and mobility. The method is also more robust than that used by Saleh *et al* [43] as the latter employed discretionary thresholds simply because they had been commonly reported classified in that way, whereas the method by Arden *et al* [79] identified cut-off points according to an anchoring satisfaction question completed by the patient. Regarding Kalairajah *et al* [61], compared to a method that uses a much larger sample size and which links patients' PROMs to their satisfaction through Receiver Operating Characteristic (ROC) curve analysis, their use of simple cumulative frequencies and the lower number of observations in their study make it an inferior classification system for the purpose of this study.

Hence, by using the outcome classification system proposed by Arden *et al* [79], we will be able to structure a model dividing surgery outcomes into two categories, G*ood* (postoperative OHS above the threshold) and P*oor* (below the threshold). This adds a unique feature to the cost-effectiveness model inasmuch as it will incorporate a classification system that combines a PROM with patient-reported satisfaction.

4.4 Model type

Most published economic evaluations of THR have used Markov models which, as discussed in Chapter 2, are the most appropriate model type given the characteristics of THRs. Although the purpose of this work is not to assess THRs but to evaluate the use of an outcome prediction tool for them, a Markov model is also appropriate in this case since the tool precedes the operation but the care pathway of the patient after the operation continues to be the same.

Following guidelines on the selection of modelling techniques [80, 81], we discarded the more complex discrete event simulation methods to build our model because interaction amongst patients or with the environment is not relevant in the case of THRs. Although decision trees could be used and in fact have been employed for the modelling of THR [39, 43], they are not appropriate either as trees would pose limitations in capturing important features of the disease and interventions in the model. THRs are performed to alleviate pain and disability that are chronic, and in some cases the intervention needs to be performed more than once, which makes a model with a lifelong time horizon more appropriate. These complexities are better handled by a Markov framework [82], hence our model for the cost-effectiveness of the outcome prediction tool for THR is developed as a Markov model.

It is worth noting that the one presented here is the only known Markov model incorporating different categories of outcome after THR since the other two studies distinguishing between surgery outcomes (Chang *et al* and Saleh *et al*) were structured as decision trees [19, 43].

4.5 Care pathway

We designed a model that faithfully reflected the care of hip patients regardless of whether they go on to have a THR or not. Evaluating the impact of an outcome prediction tool using a modelling technique means that the model must start at the point where the tool would be implemented, or before. The outcome prediction tool for THR has been designed to help orthopaedic surgeons to direct patients more efficiently to surgery or to alternative modes of care. The model, as a result, cannot start at the point of surgery as all previous economic models of THR have done (see Chapter 2 for details). We therefore explored the care pathway from the moment that a patient experiences hip pain to a

consultation with an orthopaedic surgeon as a framework of reference about the point at which the prediction tool may be used. This pathway places the economic model appropriately into the context of patient care.

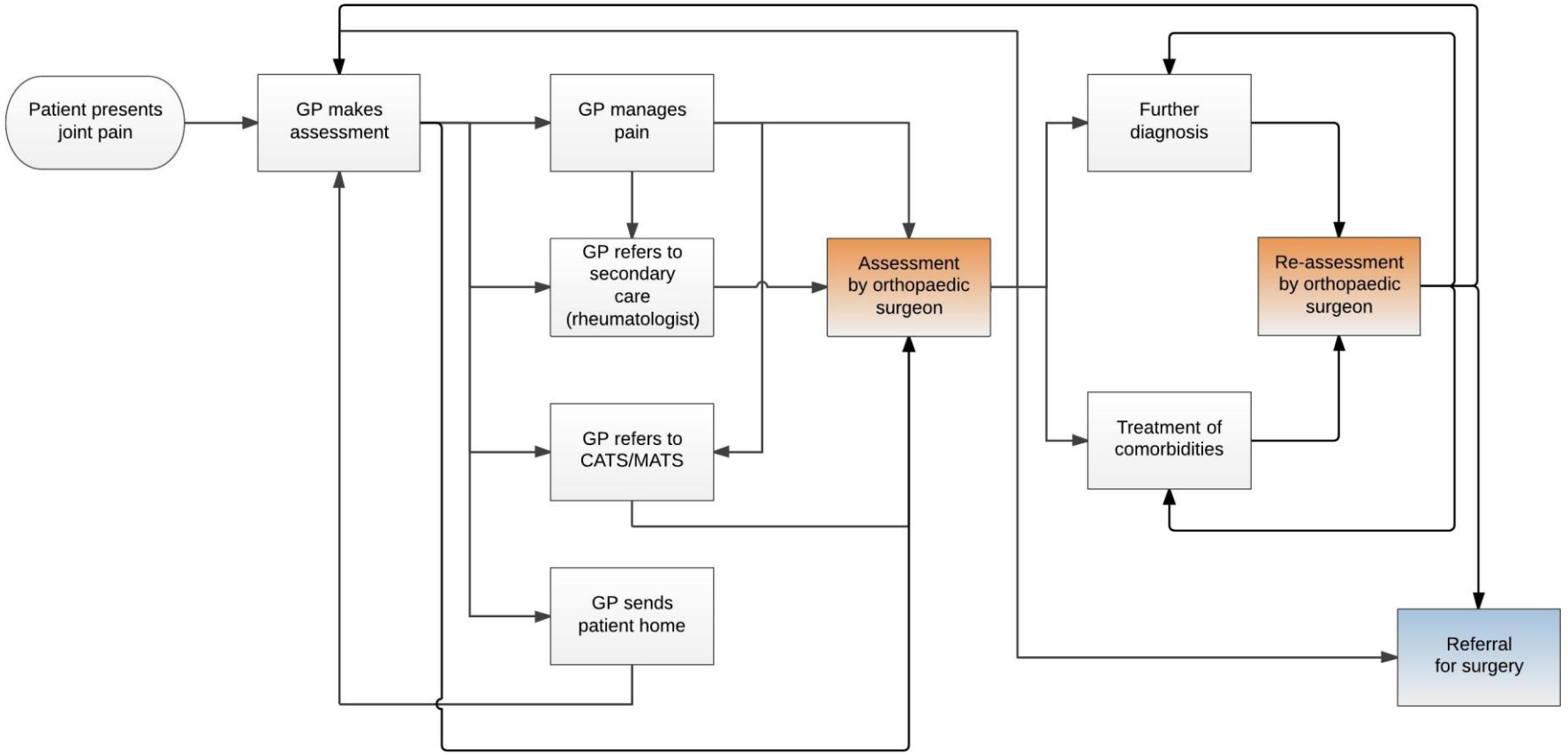
Figure 4.1 shows the care pathway of patients who experience hip pain up to and including the point where they may be seen by an orthopaedic surgeon. This pathway was put together based on discussions with experts including physiotherapists, rheumatologists and orthopaedic surgeons as nothing similar was found in the reviewed literature.

Patients with hip pain would first consult their GP about their condition. The GP then would take one of five actions: 1) to treat the patient with pain relief medication; 2) to refer the patient to a secondary care specialist, commonly a physiotherapist or rheumatologist who can help to alleviate the pain and to restore mobility; 3) to refer the patient to secondary care specialised centres, which are often independent sector treatment centres[2]; 4) to refer the patient directly to an orthopaedic surgeon; or 5) to send the patient home without treatment, a seemingly more common option for elderly patients as GPs may consider pain as merely part of the multidimensional process of aging.

Many patients will respond to any of the first three alternatives (pain relief medication, a secondary care specialist, or a specialised centre) and be able to manage their pain and limitations. Other patients, however, will not respond so favourably and will be referred from any of the above three providers for a consultation with a surgeon to determine if they are candidates for a hip replacement. For those patients sent back home, many are likely to return to the GP seeking help for worsening pain and increasing disability; and if they are suitable candidates for a THR then they would (hopefully) be identified as such on a subsequent round through the care pathway, and referred for a consultation with the orthopaedic surgeon.

[2] Examples of these are the musculoskeletal services provided by Care UK Clinical Assessment & Treatment Services (http://www.careuk.com/content/cats_patients) and the Musculoskeletal Assessment and Treatment Services operated by Bristol Community Health (http://www.briscomhealth.nhs.uk/our-services/item/289).

Figure 4.1
Hip patient care pathway



CATS: Clinical Assessment and Treatment Services; MATS: Musculoskeletal Assessment and Treatment Services

After being assessed by the surgeon, some patients may be referred for further diagnostic tests to confirm if an arthroplasty is the appropriate course of action. The patient may also be sent for treatment of comorbidities such as high blood pressure or excessive weight and only when he or she meets the respective Primary Care Trust's (PCT) criteria for the replacement, and is willing to go through the procedure, will the patient be put on the waiting list for a THR. If the criteria are not met or the patient does not want to have the operation, he or she would be referred back to the GP for further management of the condition and follow up.

Given this care pathway, a model for the evaluation of the THR outcome prediction tool would need to include the assessment by the orthopaedic surgeon, because the tool would be employed at that point to guide patients more efficiently to the different treatment alternatives. Hence, we made the first assessment by the surgeon the starting point of our Markov model.

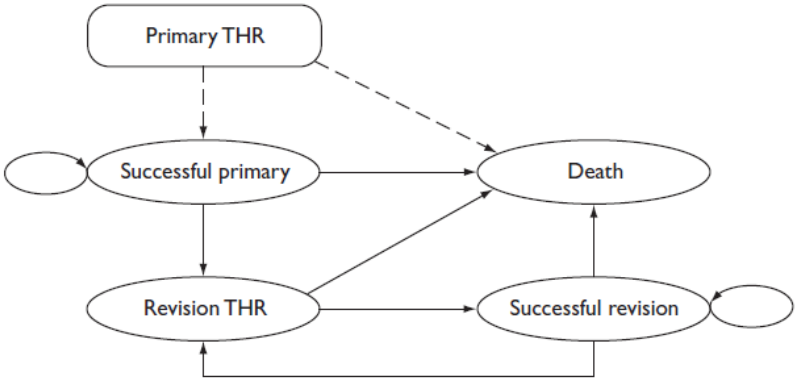
4.6 Economic model

The model for the economic evaluation of the THR outcome prediction tool was developed based on the specific aims of the COASt project. The literature review reported in Chapter 2 provided a framework that helped guide the model design, particularly considering the main two features that set this model apart from those previously published: it starts at the point of consultation with the orthopaedic surgeon and it distinguishes between outcome categories in a Markov model. Initial drafts of the model schema benefited from numerous and extensive discussions with clinical experts and a draft of the final version was validated with a group of orthopaedic surgeons.

4.6.1 Other models

The literature review presented in Chapter 2 showed that most THR economic evaluations have been performed based on Markov models. In fact, one specific model structure was used in six of the 12 studies employing Markov models [26, 36, 37, 42, 45, 46]. This popular model is shown in Figure 4.2 and was first published by Briggs *et al* [26] in 1998. One other study made a slight change by keeping the same structure shown in the figure but substituting "Death" for "Non-operative management", allowing transitions to death from all states [47].

Figure 4.2
Hip replacement model by Briggs and colleagues



Model schema used by six of the 12 economic evaluations assessing THRs identified in the literature review (see Chapter 2) and first reported by Briggs *et al* [26].

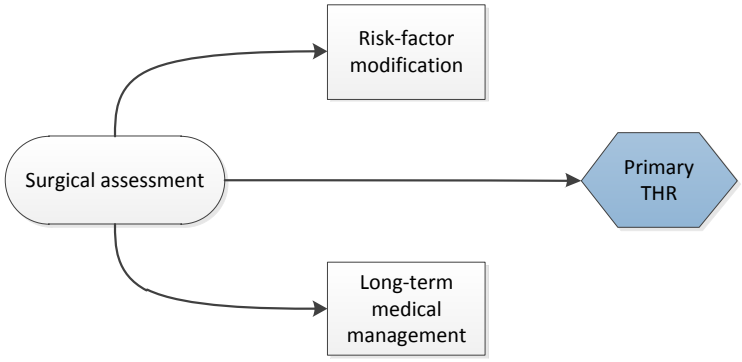
As in standard Markov models, the box and ovals in Figure 4.2 represent health states where patients remain during a given cycle, and the arrows represent the possible transitions amongst the states; that is, the allowed routes for patients to move at each iteration. Relevant features of the THR model in Figure 4.2 include the fact that it starts with the primary operation and that the only possible outcomes after the primary or a revision surgery are a successful procedure or death. These characteristics make this commonly-applied model both simple and effective at capturing the broad results of a very effective intervention such as a THR. Our approach, however, offers a new perspective on post-operative outcomes by distinguishing them between *Good* and *Poor* whilst it also expands the analysis to include what happens to patients before surgery when they are not referred for a THR.

4.6.2 Starting point

As shown in the previous section, at the first points of care, GPs, and then rheumatologists or physiotherapists, most commonly, follow a protocol to treat patients with hip pain. Non-surgical treatment would be effective for some patients and not for others. According to NICE's national clinical guideline for care and management of osteoarthritis (Guideline 59), a referral for consideration of joint replacement should only occur after the patient has been offered the core non-surgical treatment options and these prove ineffective at successfully treating the symptoms as well as reducing their impact on the patient's quality of life [12]. This will be the case with a number of patients who are referred for an assessment with an orthopaedic surgeon.

Our economic model begins with this surgical assessment. From there, patients may be referred not only to the waiting list for a primary THR, but also to other non-surgical health states. Figure 4.3 shows how the model allows for patients to transition into the waiting list for the THR, into a risk factor modification programme or into long-term medical management.

Figure 4.3
Preoperative states



The description of the patient care pathway in the previous section showed that patients referred for a consultation with an orthopaedic surgeon would be likely candidates for arthroplasty because they would have been checked and perhaps even treated by other health care professionals before they were seen by the surgeon. Some of those patients consulted by the surgeon would incidentally present conditions that could compromise the outcome of the replacement surgery. Examples of these conditions are excessive weight, high blood pressure, new onset diabetes, and chronic varicose vein. When any of these or other relevant risk factors are present, they would need to be dealt with before the patient could be put on the waiting list for the THR. These patients would therefore be referred to the appropriate risk factor modification programme, where they would remain until they were found to be fit for surgery at a later re-assessment. Meanwhile, they would also have their hip pain treated, commonly with pain killers and/or physiotherapy.

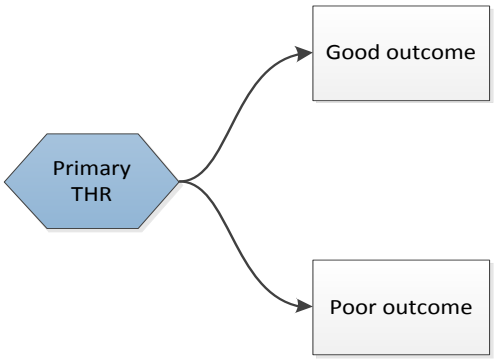
Some other patients may have been referred for the surgical consultation only for the surgeon to diagnose that their hip problem, for example, was not related to the hip (problems with the spine, for instance, are known to cause pain in the hip region), or was not an orthopaedic problem. These, as well as those patients who are found by the orthopaedic surgeon not to be candidates for a THR for any other reason, and the patients who despite being candidates decide that they are not willing to go through surgery, would be referred back to primary or secondary care for long-term medical

management of their condition. After being re-assessed, some of these patients may eventually be found fit and willing to receive a THR.

4.6.3 Following a THR

Since the second defining feature of this Markov model is the distinction between surgery outcomes, the states that patients may find themselves in after a primary THR are now categorised according to a combination of a measure of their postoperative pain and mobility functions together with their satisfaction. As described previously, after a primary THR in our model, patients may be in a state of *Good* outcome (where they would be mostly free from pain and satisfied with surgery results) or in a state of *Poor* outcome (where pain and functional limitations persist on patients generally dissatisfied with the results of the operation). Figure 4.4 illustrates this distinction.

Figure 4.4
Surgery outcome categories



With this model structure, patients with unrevised primaries are no longer put together into one single *successful primary* state. The Markov model presented here captures more accurately the fact that, despite not needing revision surgery, patients may feel very differently (both objectively and subjectively) about their pain, their ability to move around and to perform everyday activities. Making such a distinction also allows for presumably different consumption levels of health care resources by the two outcome categories.

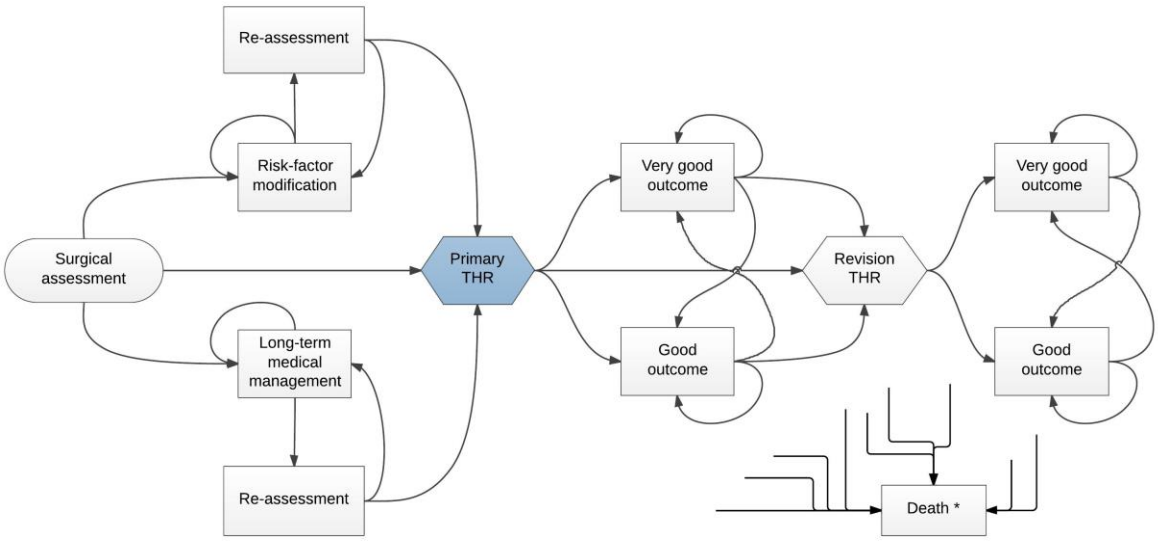
The distinction between *Good* and *Poor* outcomes after surgery is not exclusive to the primary operation. THR patients who require a revision face the same possible outcomes, that is, some will do better than others; some will be satisfied whilst others will not; and those who feel better because pain was reduced and mobility increased are more likely to

be satisfied. The same distinction between outcome categories used after primary THR was therefore applied in the Markov model to revisions, such that a (presumably different) threshold in the OHS can be used to differentiate *Good* from *Poor* outcomes anchored in post-revision satisfaction.

4.6.4 The complete model

Figure 4.5 shows the full cohort Markov model used to assess the cost-effectiveness of the outcome prediction tool for THR. The model combines the pre-surgery states shown in Figure 4.3 and the outcome distinction illustrated by Figure 4.4 and also applied to revision surgeries. Although not shown, a transition to the absorbing state of death is considered from each health state. Previous versions of the model were discussed and modified before arriving at this final and more parsimonious structure. Final adjustments were made based on feedback from a meeting with orthopaedic surgeons, rheumatologists and nurses who found the model to be an accurate representation of reality. The model was conceived to operate with yearly cycles and for as long as patients remain alive.

Figure 4.5
Markov model for the cost-effectiveness of an outcome prediction tool for THR



* Represents transitions from each health state to the absorbing state of *Death*

In the preoperative section of the model, each of the non-surgical states after the first surgical assessment is linked to a separate re-assessment state. Patients in a risk factor modification programme may remain there for a short or long period of time depending on the condition they have and their ability to become fit for surgery. At a certain point

they might be re-assessed by an orthopaedic surgeon who would evaluate them again to ascertain whether they can have their hip replaced. If so, they would be put on the waiting list for the operation, otherwise they would be sent back to the risk factor modification programme. An analogous situation is included in the model for those patients under the long-term medical management state.

Although a yearly cycle was chosen for the model, some of the preoperative transitions may be associated to shorter time spans. For example, the mean waiting time for THR (based on Operating Procedure Codes or OPCS codes W37 and W38 for cemented and uncemented procedures, respectively) was 91 days in fiscal year 2011-12 [55]. Therefore, it is unlikely that a whole year would elapse between a patient's surgical assessment and the time of operation when they are directly referred to the waiting list. Nevertheless, primary care costs and HRQL associated to the surgical assessment state would have been similar if not the same since long before the referral by the orthopaedic surgeon, thus rendering the time difference inconsequential. Similarly, we judged the effects of these time differences unimportant for the remaining preoperative states, including the slight delays in the progress of the cohort and its resulting effects on mortality.

After a primary THR, other than the *Good* and *Poor* outcomes largely discussed above, some patients may require a revision surgery within a year, hence the direct transition between primary and revision THR. Though not common, revisions soon after the primary THR can happen, generally due to infection (as opposed to aseptic loosening, the main cause for revision surgeries after the patient has had the prosthesis implanted for several years) [83]. The model also considers the possibility that patients categorised as *Good* outcomes after primary THR may either remain as *Good* outcome or deteriorate over time and transition to *Poor* state, and original *Poor* outcomes may also remain as such or change states if their pain and mobility improve with time. This explains the transitions between both outcome categories and in both directions, though it is likely that many THR patients would remain in the state they were in after surgery until they die. This section of the model is complemented by the transitions to revision THR in case they require it over time, whether the patient comes from a *Good* or *Poor* outcome, albeit presumably at different rates.

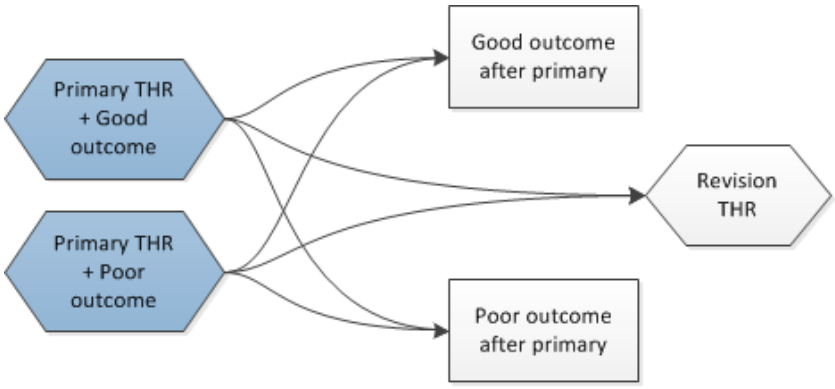
The model is complete once the section on revision THR is included. This part of the model operates in the same way as that of primaries, with *Good* and *Poor* outcomes after

surgery, and transitions between these in both directions. Finally, patients may die at any point in the model, although this absorbing state and the numerous transition arrows are not shown in Figure 4.5 for clarity purposes.

4.6.5 Merging surgery and first year outcome into one state

Because the cycle length was chosen to be one year, and patients do not spend 12 months in either of the surgical procedure states (primary or revision), an adjustment was made to account for this. Length of stay has been reported to average 6.6 days for cemented and 5.4 days for uncemented THRs [55]. A recent study furthermore found that these times vary significantly between NHS centres, independent treatment centres and private treatment centres with the latter two reporting waiting times 18% and 40% lower than NHS public hospitals, respectively [84]. As these times indicate, THR patients spend on average less than a week in the hospital after their operation, therefore we combined the first year of postoperative outcome with the surgical state. This produced two different states where Figure 4.5 shows only one: surgery linked to *Good outcome*, and surgery linked to *Poor outcome* during the first year after the operation. This was applied to both primaries and revisions, and Figure 4.6 illustrates it for the former.

Figure 4.6
THR by first year outcome and transitions



Merging surgery with the first year of postoperative outcome, and leaving the states of *Good* and *Poor* outcome after surgery to capture what happens to patients on year two and afterwards, crucially allowed for a more accurate assignment of parameter values to states. First, probabilities of revision are typically reported for the first year separately from that of the rest, because infections are more common soon after the operation. Second, HRQL would tend to drop immediately after the operation due to the discomfort associated to the surgical procedure and then progressively rise as the patient goes

through rehabilitation. This would not be the case for the following years. And third, costs are likely to be different during the first year after THR compared to the rest as more visits to the GP and physiotherapist, for example, are to be expected.

The disadvantage of separating the primary and revision THR states each into two separate ones by adding the first year of *Good* or *Poor* outcome is the increased number of transitions forced into the model, and by extension its data requirements. By doing this, each of the three transitions leading to the single Primary THR state in Figure 4.5 now points to two possible Primary THR states depending on the outcome category of the first year. This is also the case for revisions. Similarly, each transitions starting from the primary or revision operation in Figure 4.5 became two. The resulting model with the surgical states divided into two and all corresponding transitions, though following the same pathway shown in Figure 4.5, takes away from the parsimony of the latter; hence, we use Figure 4.5 as a faithful graphical representation of the economic model and include the complete one in Appendix 8 for reference.

4.6.6 Modelling a complex reality

As with all models, the one presented here attempts to reflect the true care pathway of patients as they are assessed for a THR, which most undergo, but it necessarily simplifies what in reality is a more complex process: patients' conditions may evolve in ways that have not been simulated in our model; health professionals or patients themselves may make decisions leading to a myriad of health states that are not specifically included in our schema. Modellers face the inevitable trade-off of attempting to capture the complexity of reality vis-à-vis building a manageable and parsimonious model that can be populated with good quality data and produce results that aid the decision-making process. As long as these necessary simplifications do not contradict reality or produce misleading results, then the trade-off can only be expressed and the likely limitations of the simplified models made explicit. The model presented here captures the pathway of THR patients with greater detail and breadth than those used for previously published economic evaluations of THRs; nevertheless, assumptions have necessarily been made and they are discussed in Section 4.9 of the chapter.

4.7 Comparators and perspective

The model described in the previous section was structured considering the current care pathway for patients referred to orthopaedic surgeons as candidates for a hip replacement

operation. To estimate the cost-effectiveness of the outcome prediction tool for THR, we used the model to estimate the costs and benefits (in terms of health utilities) of current practice and compared that to how practice would be if the outcome prediction tool developed under the COASt project were to be implemented. The ratio of the differences in costs and QALYs will produce an ICER which will help assess the tool's potential value for money, compared to the threshold used by NICE as a reference to recommend adoption by the NHS.

If the outcome prediction tool proves effective at identifying poor outcomes before the operation, application of the tool in the UK by the NHS would largely benefit both patients' wellbeing and the health system's efficiency. The analysis presented here is therefore made from the perspective of the UK, and particularly from that of the NHS as payer of health care services. As the model incorporates THR and non-surgical treatment and these are both currently provided by the NHS, a health care system perspective is appropriate. This is additionally supported by the fact that the NHS, through the NIHR, provided the funding for the COASt project and the development of the prediction tool. This shows an interest in the potential application of a tool that could make the significant public investment in THRs through the NHS more efficient.

Furthermore, as reported in Chapter 2, all models evaluating a THR intervention from the perspective of the health system or from that of the broader society, conducted the analysis separately for THR and its alternatives. We offer a different approximation. Our model is based on standard practice starting at a point before surgery where the choice between a THR and non-surgical alternatives is considered, and includes them all in the same model. The analysis is therefore performed jointly and maintains the relationship between the surgical and non-surgical alternatives via the prediction tool: for the alternative to current practice, it will be the tool based on predicted outcomes that determines what proportion of patients found fit for surgery will get a THR, and what proportion won't. This is indeed what happens in the actual care of patients who are deemed candidates for THR at primary care levels: they are referred to an orthopaedic surgeon and then a decision is made about whether to perform a THR or to keep the patient under non-surgical management. We perform an economic evaluation keeping the surgical and non-surgical routes as integral parts of the same model, and produce results that compare current costs and benefits to what they may be if the outcome prediction tool were to be implemented. There is a cost-effectiveness optimum for the referral of

patients to THR and current practice is not at that point, because we know that a non-negligible number of patients perform poorly after surgery. This study shows whether the outcome prediction tool may move clinical practice closer to that optimum.

4.8 Patient subgroups

The economic model was populated with data corresponding to different patient subgroups. Patient cohorts were selected according to gender and age group. Although no discrimination can legally be applied based on age or gender for the provision of health care, justified differential treatment (especially if beneficial to the patient) is perfectly acceptable [85, 86]. The impact of these factors on THR revision rates as well as their proved effect on the likelihood of achieving a clinically significant physical functioning improvement after arthroplasty [75] justify exploring, separately, cost-effectiveness of the prediction tool by these subgroups. Other economic evaluations of THR have also produced results separately for different age and/or gender subgroups as authors have found them to be associated to different model input parameter values [37, 45, 46, 49, 50].

We were also interested in performing the analysis controlling for BMI. BMI thresholds have been applied by some PCTs for patients' assessment and eligibility for joint replacement surgery [87-90]. However, it has been found that BMI does not influence the ability of patients to benefit from THR [75]. This contradiction between the policy being implemented in some parts of England and the evidence already available provides grounds for the inclusion of BMI in this evaluation. BMI was not available, however, in the main sources of data used to populate the model and therefore its impact on the cost-effectiveness of the prediction tool for THR was not analysed here.

Using a combination of gender and four different starting ages (45, 60, 70 and 80) produced a total of eight groups. The starting age of 45 was selected because, even though THRs are sometimes performed on younger patients, it is only after 45 years of age that sufficient patients are found from which to draw reliable data inputs to populate the model. A cohort entering the economic model would be, for example, 45 year-old females, and then a separate analysis would be performed for 60 year-old females. This would also be case for the remaining two subgroups of females and the four equivalent male subgroups. Model input parameters were estimated from data about patients aged 45 to 60 years for the model cohorts with starting age 45, about patients aged 60 to 70 for the model cohorts with starting age 60, and so on. For the purpose of populating the

model, nevertheless, whenever data for an input parameter were not available for a specific subgroup, a common value was applied to several or all subgroups.

As this is a cohort model, the analysis was performed recording direct costs and QALYs accrued by each patient subgroup flowing through the model one yearly cycle at a time. The patients in each state would either remain there (if allowed) or move on to another state in the following iteration according to the probability associated to each transition. Again, although not shown in Figure 4.5, one possible transition from every state in the model is to the absorbing state of death. The parameter inputs for the model are, therefore, direct costs and QALYs associated to every state, and transition probabilities. Data for the model on THR were obtained from various sources, as will be described in detail in the next chapter.

4.9 Assumptions

As highlighted in Section 4.6, our model, as any other, simplifies reality so that we can produce estimates for the cost-effectiveness of the outcome prediction tool. This simplification is achieved by making a number of assumptions that can make the model feasible. It is important to make these assumptions explicit and to consider their possible effects on final results. This section discusses the main assumptions made in the development of the economic model.

Firstly, this model assumes that the outcome prediction tool is capable of identifying potential poor surgical outcomes before patients have the operation. The methods employed to produce the tool are rigorous and appropriate, but they were applied to a set of patients that may or may not be representative of the entire population. We are therefore assuming that the information in the EPOS and EUROHIP datasets are representative of the equivalent characteristics and outcomes in the wider population susceptible of undergoing a THR in the UK. Based on their large number of participants, on the fact that EPOS is a UK-based study, and that EUROHIP is a multicentre study not only in the UK but also in other European countries, we believe that the prediction tool built upon such data is applicable to the wider UK context.

Outcome categories are a key element in this study, hence an important assumption we are making is that the way patients are classified in this model is valid and the most appropriate. We are assuming that all, or most patients who score 38 or above in the OHS

one year after their primary are all free from pain and major mobility limitations as well as satisfied with the operation, and that the opposite is true for those who score less than 38. This may not necessarily be so. First, the method used to identify the cut-off point was anchored on satisfaction, which is a largely subjective concept, as discussed in Section 4.3. And second, satisfaction, and hence the cut-off point for *Good* and *Poor* outcomes, may also vary with gender, age, BMI, expectations, or severity of disease, to name a few. In the study identifying this cut-off point on the postoperative OHS, the authors stratified their results by gender, age and BMI tertiles, and baseline OHS, but differences were not statistically significant from the overall value. They also explored equivalent thresholds using the raw change in OHS after the operation and the percentage of potential improvement achieved as outcome, and in both cases stratifying by the above variables produced results whose difference to the overall values was not statistically significant [79]. We are therefore confident that an overall cut-off point is acceptable as data appears to suggest that the connection between a postoperative OHS score and satisfaction is stable across different groups of patients.

We also assume that all patients found to be candidates for surgery but presenting a risk factor which should be dealt with before the operation, whether it is excessive weight, diabetes, blood pressure or something else, can all be grouped together and therefore the same costs, QALYs and transitions from the *Risk-factor modification* state can be applied. This is likely not the case in real life. However, we are using this health state essentially to introduce a delay into the path towards surgery as attempts to modify risk factors were reported by surgeons to be common when assessing patients considered for an arthroplasty. The *Risk-factor modification* state (where patients would be expected to stay for a short period in most cases) is not intended to reflect the specificities of the risk-factor modification treatment. In fact, as discussed in the following chapter, costs for the modification of the respective risk factors are not included in the costs associated to this state. Also, although HRQL may differ depending on the type of risk-factor patients have, we do not expect variations to be significant as EQ-5D is largely sensitive to mobility, pain and limitation to usual activities, which all patients in the *Risk-factor modification* state would have in common as they have been found to be candidates for a THR. We therefore believe that the heterogeneity of patients diverted to *Risk-factor modification* state reflects clinical practice and that the variation in costs, QALYs and transition probabilities will be appropriately incorporated into results via probabilistic sensitivity analysis (PSA).

We are also grouping a diverse set of patients into the health state of *Long-term medical management*. As above, we have given priority to what these patients have in common, namely their non-surgical treatment, as opposed to their potentially different costs, QALYs and transition probabilities based on what sets them apart. As health care costs are expected to be driven by the non-surgical treatment of their problem, and this will be largely similar for all, bringing such diverse groups of patients together is warranted. QALYs, as explained above, are very much sensitive to hip pain and its consequences, hence however diverse these patients they are all likely to have similar HRQL. Transition probabilities, however, may be different for patients in the *Long-term medical management* state. One of the specific groups of patients that will transit into this state is that comprised of potential candidates for a THR who are not willing to undergo the procedure. These patients, for example, are likely to be much more susceptible to the effects of an outcome prediction tool than patients whose problem is not orthopaedic or hip-related, or simply those found unfit for surgery, all of whom will be in the *Long-term medical management* state. Nevertheless, the distribution of the probability of transition from this health state will capture some of the variation within this group, which through PSA will allow results to incorporate this difference.

Another important assumption is that probabilities of *Good* and *Poor* outcomes are the same in the model whether the patient comes from the risk-factor modification section or from that of long-term medical management. This is a clinically plausible assumption because long-term medical management patients who are ultimately referred for a primary arthroplasty are likely to be very similar to those referred for a THR from the *Risk-factor modification* state in all aspects relevant to surgery outcome.

The model presented here does not allow for multiple revisions. Although there are patients who undergo more than two THRs in their lifetime, not only are they a very small proportion of all patients who receive this operation, but there is no data available about the effect of surgical outcomes on a second or later revision of the prosthesis.

Finally, we are ultimately assuming that the tool will be used by orthopaedic surgeons, when in reality it would be very difficult to know whether the additional information it will provide will be taken into consideration by surgeons, or even patients. It would be unrealistic to think that if the tool predicts that a patient is likely to perform poorly, for example, that this information will supersede the surgeons' criteria when they would

otherwise refer the patient for the operation, or vice versa. We therefore perform the analysis comparing current practice against a hypothetical scenario where the tool will dictate how patients are referred after the surgical assessment as an extreme case. Results will therefore show whether each unit of health benefit brought about by the strict use of this tool would require the NHS to assume additional costs at a rate lower or higher than the opportunity cost within the health system.

4.10 Contributions

The outcome prediction tool described in Section 4.2 was developed by the COASt team in Oxford, led by Prof Nigel Arden and Dr Andrew Judge with Mr Rajbir Batra performing the statistical analyses. This is the tool that the present economic evaluation assessed.

The economic model delineated in Section 4.6 benefited from the contributions of Mr David Turner in its early stages when preoperative states and the differentiation of surgery outcomes were incorporated. All following refinements to the model, the discussion with experts, the validation with orthopaedic surgeons, the merging of the surgical procedure with the first year of postoperative outcome, and the overall final structure of the model were conducted and led by Rafael Pinedo (RP).

Everything else reported in this chapter was the work of RP.

5 Populating the economic model with the best available data: from expert elicitation to patient-level analysis of large administrative datasets

5.1 Introduction

The cost-effectiveness analysis presented here used data from a large variety of sources. In this chapter we report on the data sources used for transition probabilities, including an expert elicitation exercise; for utilities, obtained primarily from the HES-PROMs dataset; and for costs, extracted mainly from the CPRD, previously called General Practice Research Database (GPRD). As with all models, a number of assumptions are made with regards to data in order to conduct the analysis and these are reported at the end of the chapter. Results of the cost-effectiveness analyses performed by populating the model with the data described here are presented in the following chapter.

To assess the cost-effectiveness of the outcome prediction tool, first the model was populated with parameter values relative to current practice. A second set of model input parameters was then estimated to populate the model with data representing the hypothetical scenario of implementing the outcome prediction tool. Costs parameters were the same in both cases, but HRQL values changed for most preoperative states, as explained in Section 5.6. Certain transition probabilities also varied as a result of using the tool. It is these probabilities, and their effects on future costs and outcomes, that drive the cost-effectiveness of implementing the outcome prediction tool. The transition probabilities taking different values in the model for current practice compared to using the prediction tool were the following:

a) Probability of being referred directly to the waiting list for THR (and hence to the two states where Primary THR is associated to *Good* or *Poor* outcome during the first year after surgery)
b) Probability of being referred for Risk factor modification
c) Probability of being referred for Long-term medical management
d) Probability of good outcome after primary (incorporated into (a) above)
e) Probability of poor outcome after primary (incorporated into (a) above)

The first three transition probabilities identified above (*a*, *b* and *c*) capture the direct effects that the outcome prediction tool would have on the decisions made at the surgical assessment stage. The ability of the tool to identify potential poor outcomes would have a

direct impact on the proportion of patients receiving a THR and hence on that of patients continuing with a non-surgical management of their condition. If the tool is able to positively predict good outcomes, then orthopaedic surgeons would more readily suggest a THR and patients would be more willing to accept having the operation. In the case of those patients for whom the tool predicts a poor surgical outcome, the referral rate for THR (and for risk factor management because that model state is intended for patients found fit for surgery) would drop. Likewise, the probability of those patients being referred for long-term medical management would be expected to increase. Hence, the impact of the outcome prediction tool would be primarily reflected on the set of transition probabilities immediately following the surgical assessment.

Consequently, the group referred for a THR if the prediction tool is implemented would be comprised of a higher proportion of patients expected to perform well after the operation than under current practice. The probabilities of having *Good* or *Poor* outcomes after a primary THR would therefore be different between current practice and the application of the prediction tool. These effects are captured by items (d) and (e) in the list above. There are no reasons to think that the remaining transition probabilities would be affected by the outcome prediction tool, hence they remained the same for both sets of model input parameters.

Estimates for the values of transition probabilities, costs and QALYs required to populate the model were obtained from various sources, all with only one exception based on UK patients and practice. Source of data and samples used were associated with varying degrees of uncertainty. We addressed this uncertainty by conducting sensitivity analyses, both one-way on key variables based on their potential effects on results, and a PSA to explore the joint effect of the uncertainty in all input parameters on final results. Data used for transition probabilities are described first and in two separate sections: Section 5.2 dedicated to the preoperative segment of the model and obtained from an expert elicitation exercise and Section 5.3 on transitions after a THR. Parameter values for QALYs are described next in Section 5.4, followed by direct costs associated to each state which are described in Section 5.5. Values used for transition probabilities affected by the use of the outcome prediction tool are presented lastly in Section 5.6. We close the chapter with a section on assumptions made and the usual segment on the statement of contributions received for the development of this chapter.

5.2 Preoperative transitions: obtaining probabilities from expert opinion

The first section of the model covering the states and transitions between the surgical consultation and THR rendered this model not only novel but also contingent on information not systematically collected before. This is because no data were found in the published literature that described referral decisions by orthopaedic surgeons, even if only reflecting the proportion of patients referred for a THR after their initial assessment. In order to obtain estimates for these probabilities we turned to the expert opinion of orthopaedic surgeons and elicited their knowledge in a probabilistic form. We conducted a systematic expert elicitation exercise in order to obtain mean referral rates as well as uncertainty around those values in a way that provided the highest possible level of reliability. Below, we explain the theoretical framework supporting the elicitation exercise, describe the process of conducting the elicitation, and finally present and analyse the results.

5.2.1 Theoretical framework

Elicitation is a process whereby a person's beliefs and judgements about an uncertain quantity or proposition are obtained and represented in the form of a probability distribution [91]. Since our cost-effectiveness model requires estimates for the transition probabilities describing referral decisions by orthopaedic surgeons leading to a THR in the UK, and these are not known, we elicited the opinion of expert orthopaedic surgeons on these values.

This information that we require (and at the onset ignored) is in principle knowable. This kind of uncertainty is known as *epistemic*, as opposed to *aleatory* uncertainty which is induced by randomness or chance [92]. The distinction is important because whereas the notion of probability as the proportion of times that an event occurs fits perfectly those aleatory uncertainties such as the toss of a die, it is only personal or subjective probabilities that can address epistemic uncertainties such as the transition probabilities we require. By personal probabilities we refer to the degree of belief that someone has about an uncertain proposition [92], which will vary between subjects particularly when the proposition at hand is of the epistemic kind, as in our case, hence our choice to elicit those judgements from expert surgeons. Experts' opinion is considered a legitimate source of data for decision-analytic models in health economics as long as the methods used are clearly documented and the selection of experts is performed appropriately [93].

The notion of using personal probabilities to describe an uncertain parameter is supported by the Bayesian approach to statistics, which, in very broad terms, uses data and Bayes theorem to progress from an initial state of knowledge (a prior distribution) to a new one (posterior distribution). In the case of our elicitation exercise, the judgements we collected from each expert allowed us to build prior distributions for the transition probabilities we are interested in. This differs from the frequentist approach, which accepts parameters to be unknown but regards them as fixed, such that no probability distribution can be generated for a given parameter [92].

Expert opinion can be elicited in many different ways. A type of method commonly employed is the Delphi group technique, which brings together a group of experts and aims to have them reach a consensus about specific matters, including a common judgement about current data that is not accurately known or available [94]. This method, however, implies contacting and bringing together a number of experts, which is difficult to achieve and can be expensive. The technique also forces a group dynamic directed at achieving consensus between knowledgeable people who could perfectly disagree, thus potentially creating a challenging task even for the most skilful facilitator. In terms of its outcomes, the Delphi and other consensus techniques have been deemed inappropriate also because they appear to underestimate the uncertainty around the true parameter presented to the group [93] and because forced consensus of several experts seems to produce no clear benefits over methods that do not require interaction among them [95]. We therefore chose an individual, direct method of expert judgement elicitation instead, and more specifically, a mathematical approach that revealed experts' answers as distributions at once.

Although elicitation methods have had limited impact in decision analysis [96], they have been used to obtain parameter estimates and distributions to populate model inputs [97] and even to characterise model structural uncertainty [98]. Of the few expert elicitations applied to economic models, most have focused on deriving mean values of unknown parameters whilst ignoring the uncertainty around those estimates [98].

5.2.2 Method

For our elicitation exercise we individually interviewed experts and presented them with a set of questions about their referral decisions of hypothetical THR patients. More specifically, we were interested in obtaining data describing the transition probabilities

leading from the surgical assessment directly to the waiting list for THR, to risk factor management and to long-term medical management. We also required the probabilities of returning for reassessment from the latter two states, as well as the probabilities of being put on the waiting list for THR after such re-assessments. In total, there were seven probabilities we expected the participating orthopaedic surgeons to provide their expert judgement on. Additionally, since modelling the cost-effectiveness of the prediction tool implied comparing current practice with a scenario where the outcome prediction tool would be implemented, and this had not happened yet, we also asked experts about their opinion regarding the potential effect that such a tool could have on their referrals to the waiting list for THR.

With regards to the seven probabilities we set out to obtain expert opinion on, we required a point estimate of the mean value to populate the deterministic model but we were also interested in obtaining distributions describing the uncertainty around them in order to propagate it through the model as part of the PSA. We believe that eliciting the uncertainty around the parameter values in the form of a distribution would be more comfortable for the experts than asking them to provide a single precise value. We employed, consequently, a method of eliciting experts' judgement that allowed them to express their belief about the seven transition probabilities in the form of a probability distribution.

We adapted to our specific case a method previously used on the elicitation of expert opinion about the probability of sustaining treatment benefits over time, conducted for the economic analysis of enhanced external counterpulsation for the treatment of stable angina and heart failure [97]. For that exercise the authors employed the histogram technique [99], which consists of presenting experts with a frequency chart showing intervals for the range of answers of the question at hand and asking them to specify their relative subjective probabilities for each interval by placing a finite number of crosses throughout the grid.

We used this method for the seven questions about referral decisions. For example, the first question for which we elicited expert opinion stated "What proportion of patients fall in group (a), i.e. those that are referred for a hip replacement?" after having explained the model according to which patients, after a first surgical assessment, would be referred for (a) THR, (b) risk-factor management or (c) long-term medical management. The grid used

to elicit experts' judgements about this question is shown in Appendix 9 and the remaining six questions followed the same format. To explore their beliefs about how the outcome prediction tool would affect their referral decisions, we asked the experts if they thought that the tool would increase or decrease the proportion of patients referred directly for a THR, and then by how much. For this last question, we also used the histogram technique. In all cases, we asked experts questions regardless of patient characteristics, so their responses apply to all patient subgroups defined in Section 4.8.

Although probabilities are a random continuous variable, our elicitation method collected experts' opinion about proportions as if they were discrete instead, by presenting them with a grid showing values that went from 0% to 100% in increments of five points. We chose this method because it provided experts with an easy framework to respond to the questions and because their responses in discrete form could be represented in a probability mass function that could then be approximated to a probability density function associated to the same random variable in continuous form.

5.2.3 The process

We structured the elicitation exercise as face-to-face interviews on the basis of ensuring maximum understanding. An electronic method using the option of simple Microsoft Word or Excel documents was tested as a pilot with one surgeon but instructions were not appropriately followed and responses were inconsistent. A telephone interview was also considered but given how crucial it was that the questions were correctly interpreted and the instructions strictly followed, we decided it was worth assuming the greater costs in terms of time and transportation of the face-to-face interview in order to make the process as clear and easy to understand as possible.

The format of the interview and the choice of questions were aimed to make the elicitation process as natural as possible for expert orthopaedic surgeons to think of the propositions in probabilistic form. This is supported by ecological theories of calibration of subjective probabilistic mental models when making judgements about uncertainty, which argues that people are well adapted to report probabilities based on the relative frequency of relevant cues in their memory [100]. However, reporting probabilities may not involve complex mental calculations but instead various rules of thumb or heuristics, which have been thoroughly studied and reported in the literature [92, 101].

Based on a previous similar exercise conducted by researchers at the University of York [102], we used a series of Microsoft Power Point slides to introduce the background, context, instructions, a dummy elicitation, and finally the specific questions for which experts' judgement was being elicited. We adapted the original slides to fit our study's purpose and specific questions. The set of slides used for our elicitation exercise are shown in full in Appendix 10.

Two pilot interviews were conducted with colleagues from the Wessex Institute at the University of Southampton to assess the clarity of instructions and the timing for the entire session. After slight revisions, a final pilot elicitation interview was conducted with an orthopaedic surgeon before participants were contacted for the interview sessions.

5.2.4 The experts

The elicitation exercise was aimed at obtaining the opinion of experienced orthopaedic surgeons, active at the moment and with many years performing hip replacements. Some names were suggested by the COASt project clinical advisors and those surgeons were contacted by email inviting them to participate in the exercise. Interviews were agreed to be conducted at the convenience of experts, in most cases in the hospitals where they worked. As the session ended surgeons would be asked to suggest the name of other experts to be invited to participate as well.

A total of 28 surgeons were contacted and 10 responded to the invitation. Of these, three had been performing hip replacements for less than five years and during the previous year had only operated on 20 to 30 patients; hence, they were excluded from the sample. Responses from the remaining seven experts were analysed and are reported below. More than half of the selected experts were Consultant Orthopaedic Surgeons, two were Associate Specialists and one an Orthopaedic Hip Fellow. They were based at four different hospitals, namely Southampton General, Bournemouth Royal, Oxford's Nuffield Orthopaedic Centre and Portsmouth's Queen Alexandra. They had all been performing hip replacements for many years, ranging from eight to 20. The selected experts performed 180 hip operations on average during the previous calendar year. The least experienced surgeon had performed about 300 THRs whilst the least number of operations in the period, around 50, was reported by one of the surgeons with most years of experience. Surgeons participating in the study specialised in various surgical

techniques and patient groups (including the elderly, young patients, complex cases, revisions, and sports-related cases).

In terms of the number of experts included in the sample, a diminishing marginal returns has been reported in the literature suggesting that the most benefit is obtained from the first three to five experts [95, 103]. We stopped contacting surgeons after interviewing the seventh since point estimates for mean values varied only slightly with responses from the last two, and the graphical representation of uncertainty was, for most questions, largely unchanged as well.

5.2.5 Calibration, relative weights and consistency checks

Subjective judgements are affected by a random error, even for representative sets of experts' judgements [92]. Two sources of experts' judgements error have been reported, one due to human inconsistency and the other to a sampling error attributable to variability in the environment [104]. We considered the different sources of possible bias for our expert elicitation and designed the exercise in order to minimise them, as they could never be fully eliminated.

An elicitation participant is considered to be well calibrated when the probability he or she reports matches the true frequency [104]. Calibrating the experts we interviewed based on the model questions asked was impossible given that the true values are unknown, and we did not include an additional question for calibration as we assumed all participants to be experts in the field and reflecting a variability that we were, in fact, interested in. It is for this same reason that we did not include a weighting mechanism either, as we believe that the shapes of the individual distributions obtained from experts not only described their subjective uncertainty around the parameter, but they were also a representation of the differences between the clinics, patients and guidelines faced by each surgeon. If we could assume that spread-out responses were a measure of higher uncertainty, then we could use the entropy associated to the probability mass function of the elicited probabilities to identify and assign greater weight to responses from those experts more certain about their answers [101]. However, more spread out answers may also reflect a mixture of the level of uncertainty around the true value of the parameter and the different characteristics of the specific patient subgroups that each expert evaluates.

This makes the present expert elicitation exercise rather unique inasmuch as what is being elicited is not a clinical parameter that may be thought of as having a unique true distribution, presently unknown; this exercise is eliciting judgements about a parameter whose distribution varies as PCT reimbursement rules vary, as surgeons' expertise varies, and very importantly as patient mix varies. For example, as surgeons answered questions based on their own referral decisions, if a surgeon's practice received mainly young, otherwise healthy patients, then answers about referral decisions would be expected to be different from those of surgeons seeing older patients with a myriad of additional health conditions. Calibrating all experts' judgements with one single variable or weighting their responses by any measure would therefore hamper the unavoidably joint elicitation of uncertainty and variability, both relevant to the cost-effectiveness model. We therefore considered all individual judgements equally legitimate and valued them accordingly.

As the degree of uncertainty and the effect of patient/surgeon/PCT heterogeneity cannot be discriminated from the single answers collected from experts, we used the probability distributions resulting from the exercise mainly as a measure of uncertainty. We did, nevertheless, check for consistency of experts' responses to the first three questions as these were supposed to add up to around one. Although we did not expect the mean values of the first three distributions to add exactly to one, we checked that the range of cumulative minimum and maximum values reported by each surgeon included the value of one. In all cases, the uncertainty around the first three transition probabilities elicited included the value of one within the range, which we interpreted as a sign of consistency of experts' responses.

5.2.6 Synthesising individual elicitations

Once all personal judgements were elicited, the next step was to synthesise their knowledge into a single distribution. We did this by mathematical aggregation, which involves adding together the responses from all experts (as opposed to the behavioural aggregation applied in consensus approaches such as the Delphi technique discussed at the beginning of this section). We employed the linear opinion pool method described by Stone [105]:

(Equation 5.1)

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta),$$

where $p(\theta)$ represents the combined probability distribution along all frequency values θ (0%, 5%, …, 100%) presented to the n experts, and w is the weight applied to each, such that the sum of the weights of all experts will equal one. Since we applied the linear opinion pool under the *democratic* condition that all experts' judgements are weighted equally, then for our elicitation process $w = 1/n$ for all i. In essence, this meant that the probability associated to each discrete value presented to the experts was equal to the simple arithmetic mean of the probabilities reported by all respondents.

The graphical equivalent of this aggregation method was a combined histogram where the individual arithmetic means of each frequency value were added together to form a single distribution. Parameters for these distributions were used for the PSA.

5.2.7 Results

Table 5.1 shows the means and standard deviations (SD) of the responses from all seven experts included in the elicitation exercise about their beliefs regarding the proportion of patients referred for *Primary THR*, *Risk-factor modification* and *Long-term medical management*, turned into probabilities. Pooled means and SDs are also reported. Since transitions from the first health state in the model (*Surgical assessment*) lead to the above three alternatives or to death and all of these must add up to one, it is customary for Markov models such as this one to leave one transition probability as the difference between 1 and the sum of all other probabilities. As we obtained all-cause probabilities of death from the UK Office of National Statistics (ONS) Life Tables [106], we chose to select the two transitions reported in Table 5.1 on which experts agreed the most and use their respective values, hence leaving the third to complement up to 1. This implies an upward adjustment to the values reported for the latter because, as Table 5.1 shows, the pooled mean values add up not to 1 but to 0.94.

Values for the SDs reported in Table 5.1 suggest that there was considerably more agreement amongst experts about the proportion of patients referred to Risk-factor modification than to the other two alternatives. This is also confirmed by the spread of curves representing each expert's responses to the question on referrals to this alternative shown in Figure 5.1. Figures 5.2 and 5.3 illustrate experts' responses to the questions on referral to THR and Long-term medical management, respectively. As Figure 5.1 shows, all but one of the surgeons consulted believed that the probability of referral for Risk-factor modification is never more than 25%, whilst one considered 30% to be the most

likely answer but thought that it could be as high as 45%. This particular surgeon (Expert 7) works mostly on elderly patients and children, which may explain the relatively higher referral rate to Risk-factor modification.

Table 5.1
Expert opinion on patient referral from surgical assessment

	Probability of referral from initial surgical assessment: mean (SD)		
	To Waiting list for "THR"	To "Risk-factor modification"	To "Long-term medical management"
Expert 1	0.32 (0.088)	0.04 (0.038)	0.65 (0.060)
Expert 2	0.31 (0.099)	0.12 (0.055)	0.20 (0.072)
Expert 3	0.73 (0.034)	0.15 (0.036)	0.08 (0.025)
Expert 4	0.84 (0.060)	0.02 (0.025)	0.08 (0.039)
Expert 5	0.91 (0.033)	0.06 (0.021)	0.02 (0.024)
Expert 6	0.69 (0.049)	0.12 (0.043)	0.04 (0.027)
Expert 7	0.59 (0.066)	0.31 (0.065)	0.11 (0.059)
Linear pool of experts	**0.63 (0.227)**	**0.14 (0.093)**	**0.17 (0.208)**

Expert surgeons largely disagreed on their beliefs about the proportion of referrals to the waiting list for THR (see Figure 5.2), but they appeared to be in ample agreement regarding referrals to Long-term medical management, as shown in Figure 5.3. Similarly to the question about referrals to *Risk-factor modification*, all but one of the expert surgeons consulted believed that 35% is the largest proportion of patients sent back to *Long-term medical management*, whereas Expert 1 reported that this proportion ranges, instead, between 55% and 75%. We did not explore the justification for the experts' responses and hence we did not explore why Expert 1 in particular had such different belief regarding referrals to *Long-term medical management* in relation to the rest of the experts. It could be due to the fact that Expert 1 was the surgeon with the least number of years performing THRs (eight years, compared to a mean of 16 years amongst the other six surgeons), although the same expert is also the surgeon who reported performing the largest number of operations during the year previous to the interview.

Based on the above results, we chose to take the values about transition probabilities from *Surgical assessment* to *Risk-factor modification* and to *Long-term medical management* obtained from the expert elicitation. We therefore allowed the probability of transition from *Surgical assessment* to the waiting list for a THR to take the value necessary to have all the above and the probability of death from *Surgical assessment* to add up to 1. Since the health state corresponding to the primary THR was divided into two according to the outcome

category for the first year (*Good* or *Poor*) and those probabilities are described in the next section on postoperative transition probabilities, we present parameter values and distributions there.

Figure 5.1
Probability distribution of referral to Risk-factor modification

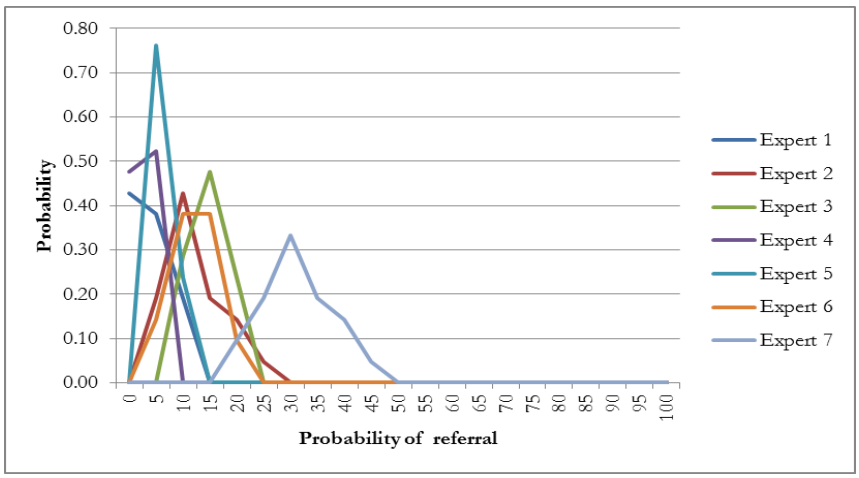


Figure 5.2
Probability distribution of referral to THR

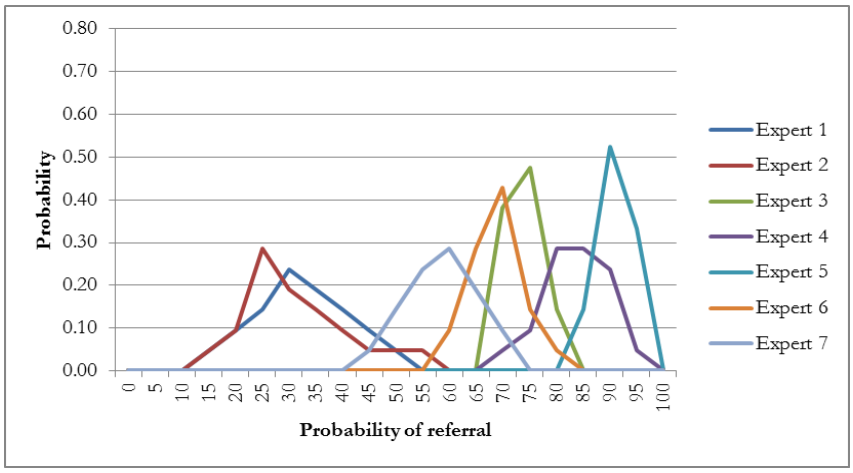
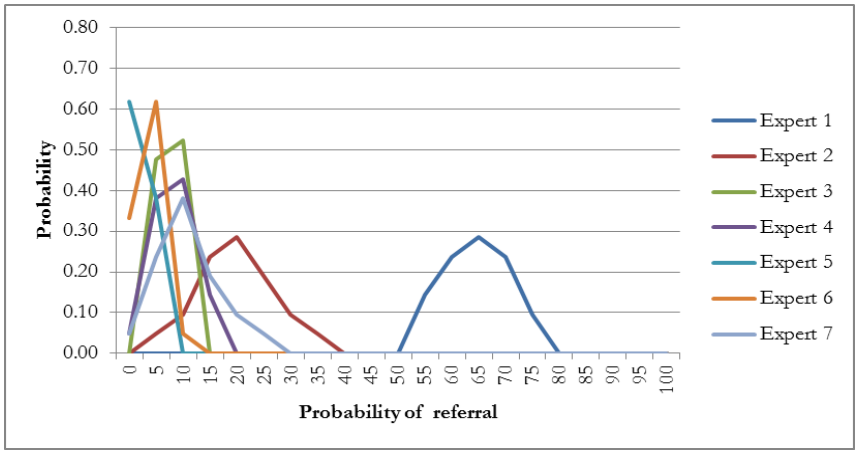


Figure 5.3
Probability distribution of referral to Long-term medical management



With responses from all experts pooled together, Figure 5.4 shows the aggregate distribution of the transition probability from *Surgical consultation* to *Risk-factor modification*. The histograms with the corresponding distributions for transition to *Primary THR* and *Long-term medical management* are shown in Appendix 11.

Once patients transition into the model health states for *Risk-factor modification* or *Long-term medical management* they may either remain in that state, be referred for re-assessment, or die. And if sent for re-assessment, again three alternatives are considered in the model: being found fit for surgery and hence referred to the waiting list for THR, being sent back to the health state they were in before re-assessment, or dying. Since the probability of death is already available, two transition probabilities in each case remained to be populated. We asked experts for their judgment about referral to THR in either case, hence the probability of going back to their previous state was 1 minus the latter minus the probability of dying.

Figure 5.4
Pooled probability distribution of transition probability from Surgical assessment to Risk-factor modification

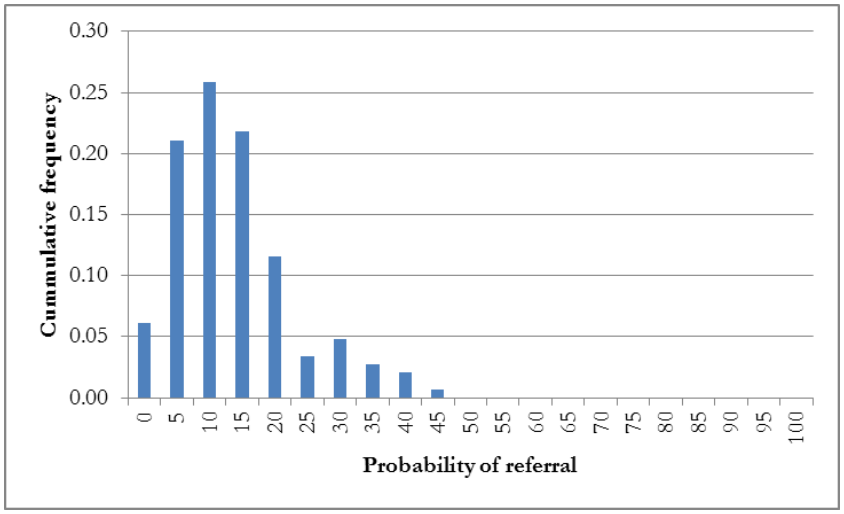


Table 5.2 shows the mean and SD of individual and pooled responses for the probability of referral from *Risk-factor modification* to *Re-assessment*, and specifically for those re-assessed, the probability of being referred to the waiting list for a *Primary THR*.

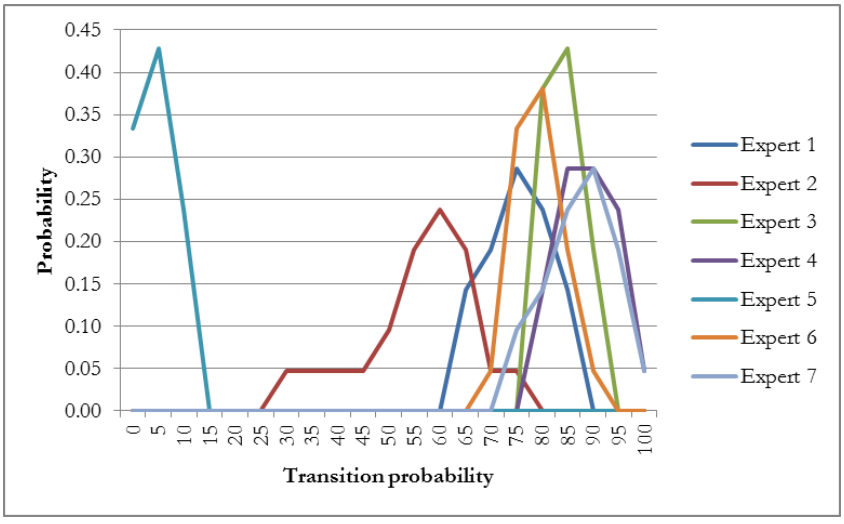
Individual distributions of responses for the transition to Re-assessment after *Risk-factor modification* are shown in Figure 5.5. As shown, five of the seven expert surgeons believed that referral to *Re-assessment* after one year in *Risk-factor modification* can only be higher than

65%, whilst Expert 2 thought that 60% was the most likely rate with a positive chance of it being lower than that as well. Expert 5, on the contrary, thought that this probability is actually rather low, between 0% and 10%. Apart from Expert 5, there seems to be broad agreement that most patients referred to *Risk-factor modification* to care for those factors such as excessive weight or high blood pressure that made them not yet suitable for THR would be re-assessed and considered again for a THR within a year.

Table 5.2
Expert opinion on reassessment after Risk-factor modification

	Probability of referral from risk-factor modification: mean (SD)	
	To re-assessment after one year	**If re-assessed, to THR**
Expert 1	0.75 (0.063)	0.97 (0.025)
Expert 2	0.56 (0.110)	0.85 (0.062)
Expert 3	0.84 (0.037)	0.88 (0.025)
Expert 4	0.89 (0.055)	0.87 (0.057)
Expert 5	0.05 (0.038)	0.63 (0.055)
Expert 6	0.79 (0.047)	0.79 (0.041)
Expert 7	0.87 (0.067)	0.90 (0.051)
Linear pool of experts	**0.68 (0.286)**	**0.84 (0.111)**

Figure 5.5
Probability distribution of referral to Re-assessment after Risk-factor modification



There was much more consensus amongst expert orthopaedic surgeons about the proportion of patients that, once re-assessed after their *Risk-factor modification* programme, would be found suitable and hence referred for a *Primary THR*. As Figure 5.6 shows, all experts considered that most re-assessed patients would be placed in the waiting list for a

THR, with Expert 5 assigning the greatest likelihood to 65% and the rest of the experts recording their belief about this rate between 70% and 100%.

We asked the same set of questions to experts about what they believe happens to patients referred for *Long-term medical management* after the original *Surgical assessment*. Table 5.3 shows the individual and pooled mean and SD of the derived transition probabilities between *Long-term medical management* and *Re-assessment*, and from the latter to a *Primary THR*.

Figure 5.6
Probability distribution of referral to primary THR after Re-assessment post Risk-factor modification

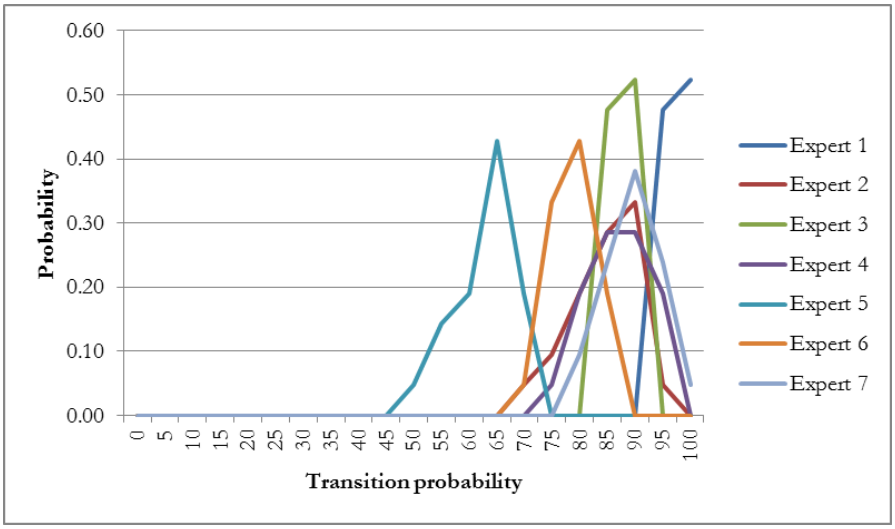


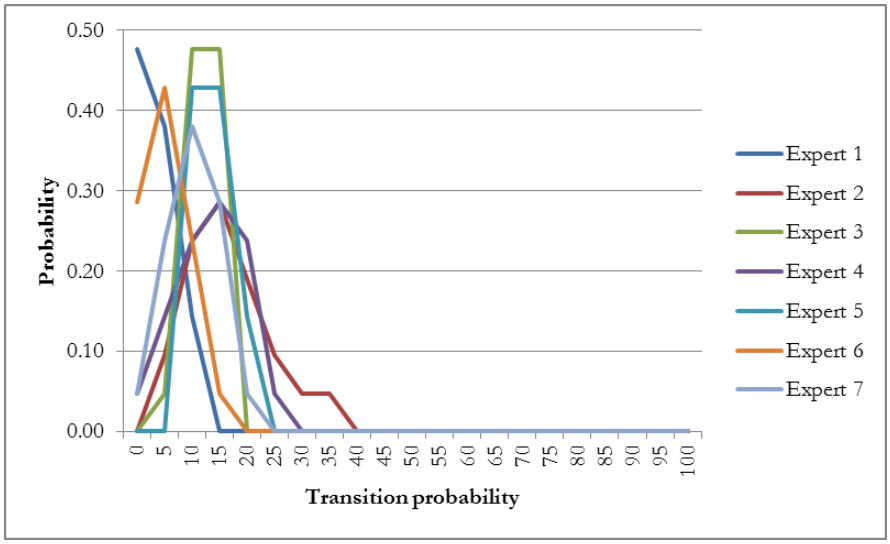
Table 5.3
Expert opinion on reassessment after long-term medical management

	Probability of referral from medical management: mean (SD)	
	To re-assessment after one year	If re-assessed, to THR
Expert 1	0.03 (0.036)	0.50 (0.011)
Expert 2	0.16 (0.076)	0.54 (0.094)
Expert 3	0.12 (0.029)	0.15 (0.039)
Expert 4	0.13 (0.062)	0.07 (0.053)
Expert 5	0.14 (0.035)	0.85 (0.031)
Expert 6	0.05 (0.042)	0.05 (0.034)
Expert 7	0.10 (0.475)	0.05 (0.031)
Linear pool of experts	**0.11 (0.066)**	**0.31 (0.300)**

As SDs in Table 5.3 suggest, there was significantly more agreement amongst experts in relation to how likely it is that patients originally referred to *Long-term medical management*

will be seen again in a surgical *Re-assessment* (SD=0.07), than there was about re-assessed patients being found suitable for a THR (SD=0.30). Patients referred for *Long-term medical management* were assumed to be patients whose pain was found not to be of an orthopaedic nature in the initial *Surgical assessment*, or not related to the hip. They could also be patients found unfit for the operation or otherwise unwilling to go through a major intervention such as a THR. As Figure 5.7 shows, all experts agreed that the probability of such patients being re-assessed within a year was very low, in all cases under 35% but often as low as 0% or 5%.

Figure 5.7
Probability distribution of referral to Re-assessment after Long-term medical management

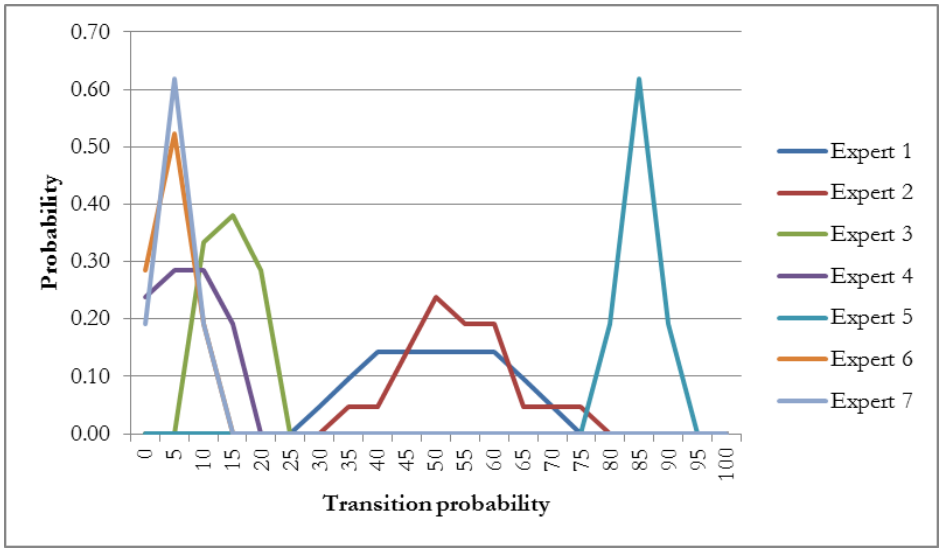


As Figure 5.8 shows, however, there was wide disagreement in the beliefs about re-assessed medical management patients being put on the waiting list for a THR. Although four of the seven experts believed that the probability of a THR referral for these patients is very low (under 20%), two believed that there was roughly a 50/50 chance of finding them fit for surgery. One expert surgeon conversely considered that, once re-assessed, there is a high probability (80% or more) of medical management patients being put on the waiting list for a primary THR.

Given that this is quite a heterogeneous group of patients, based on the fact that they were found unfit for surgery for a variety of reasons in the first place, it is not unreasonable to observe this level of disagreement amongst experts. For example, a patient initially found fit but unwilling to receive the operation may later change his or her opinion and accept to be put on the waiting list for a THR. On the other hand, patients

whose pain persist and do not find medical management satisfactory may keep coming back for a surgical consultation; some only to be sent back to their GPs, others to be successfully referred for a THR. The different views of expert surgeons illustrated in Figure 5.8 seem, therefore, perfectly clinically plausible.

Figure 5.8
Probability distribution of referral to primary THR after Re-assessment post Long-term medical management



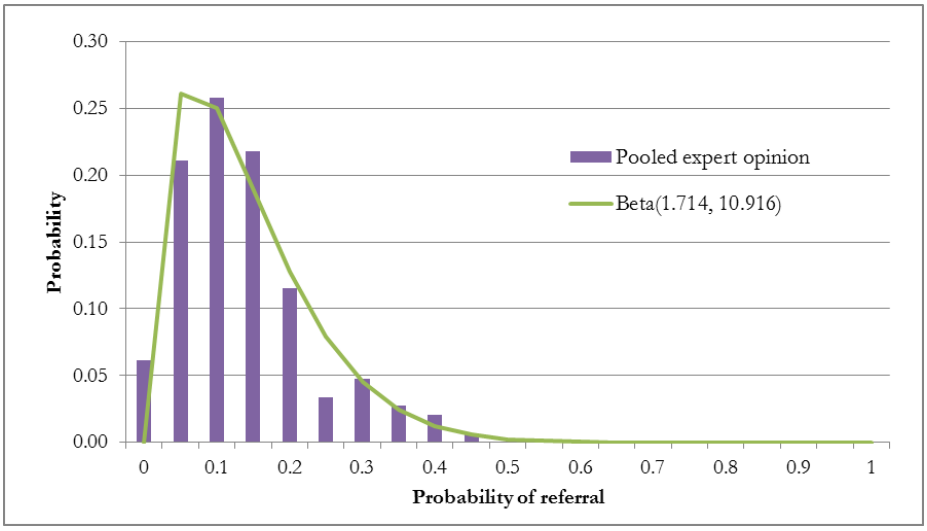
5.2.8 Parameter values

The deterministic model was populated with the mean values obtained from the responses provided in the expert elicitation. We also fitted Beta distributions (by method of moments using the observed mean and SD) as is customary for probabilities since they are bound between 0 and 1 and they are conjugate to binomial data [82]. Questions posed to surgeons referred to scenarios with two alternatives (hence directly associated to binomial data) or three alternatives (also binomial if options are considered conditionally). Although a Dirichlet distribution is sometimes used in these cases, this would have required a common parameter amongst the three transition probabilities that our expert elicitation did not provide. Since patients may die following every model state, age and gender-specific all-cause mortality rates [106] were applied to all preoperative states.

For the PSA, we assigned the corresponding Beta distribution if two conditions were met: first, the resulting probability density function had to appropriately fit the respective pooled probability distribution from experts' responses, as verified by visual exploration; and second, there was no significant difference between the observed mean value and that

generated by the inverse of the cumulative density function evaluated at 0.5 (a difference greater than 0.05 was considered excessive). Figure 5.9 shows how the probability of referral to *Risk-factor modification* is an example of a Beta distribution that fitted the pooled opinion of experts very accurately. The mean probability of referral reported by experts was 0.14 and the estimated mean based on the fitted Beta distribution was 0.12. The Beta distribution in this case was chosen for PSA.

Figure 5.9
Pooled probability distribution of experts' responses and respective fitted probability density function about referral to Risk-factor modification state



Conversely, Figure 5.10 shows the probability of referral to the model state of *Long-term medical management* where the fitted Beta distribution takes the value 0 in a very high proportion of cases, in strong disagreement with the opinion of experts who believe that the true value of this probability lies most likely between 5% and 10%. The fitted distribution thus highly underestimates the mean probability at 0.07, whilst the mean of the experts' elicited opinion was 0.17. In this case, and in all others like this one, we used the empirical distribution, i.e. that built by aggregating the judgements of all experts in its original discrete form, to populate the values for the PSA. Appendix 12 shows the graphical representation of the pooled probability distribution of experts' responses and their respective fitted probability density functions for the remaining transition probabilities.

Figure 5.10
Pooled probability distribution of experts' responses and respective fitted probability density function about referral to Long-term medical management

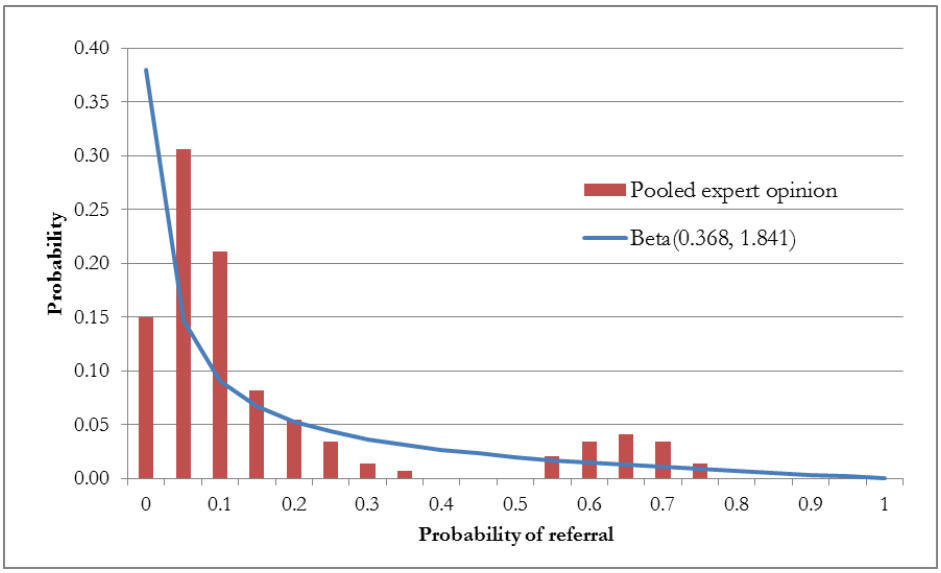


Table 5.4 shows the mean and SD as well as the distribution and its parameters, if applicable, used to populate preoperative transition probabilities for the deterministic and probabilistic economic model.

Table 5.4
Preoperative probabilities: deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.136	0.093	Beta	1.714	10.916
Surgical assessment to Long-term medical management	0.167	0.208	Empirical		
Risk factor modification to Re-assessment	0.679	0.285	Empirical		
Re-assessment after Risk factor modification to THR	0.840	0.111	Beta	8.287	1.581
Long-term medical management to Re-assessment	0.106	0.066	Beta	2.208	18.598
Re-assessment after Long-term medical management to THR	0.315	0.300	Empirical		

5.2.9 Strengths and limitations

In some of the questions presented to expert orthopaedic surgeons participating in this elicitation exercise we found a certain level of disagreement. Given the limited number of surgeons interviewed and the large number of factors that may play a role in their referral patterns, we can only propose plausible explanations for the results obtained from the expert elicitation. We found no correlation between the experts' answers and the hospitals where they work, therefore it is unlikely that differences are due to the varying eligibility criteria for THR amongst PCTs. A plausible explanation may lie in the different type of patients that surgeons have referred to them, and their professional criteria to indicate an

operation. Variation in referral criteria from GPs or treatment centres may also explain the differences, possibly confounded by the surgeons' areas of specialisation or other factors that were not available for the analysis. It is important to emphasize, however, that a larger study involving more surgeons, from other parts of the country and collecting information on determinants or confounding factors of referral decisions would be necessary to further increase the certainty that these differences are not spurious.

The method used for the elicitation of personal judgements also affects the responses obtained. In order to minimise these effects, we followed the practical guidelines suggested by Cooke [101] when eliciting opinions. We made every effort to phrase questions clearly, chose an attractive format for the questions and a graphic one for answers, and performed a 'dry run' before moving on to the actual questions, i.e. surgeons were asked practice questions and they were the same for all surgeons. When experts provided their answers, an analyst was always present to confirm interpretations and clarify any doubts. We offered a brief explanation of the exercise as part of the introduction, a point was made to avoid coaching, and the entire elicitation session did not exceed one hour.

Since we inquired about probabilities and these are by definition bound to the range between zero and one, the elicitation exercise provided important absolute anchors for the experts' judgements in order to minimise relativity effects, which can affect the elicitation of other types of judgements that are relative by nature [92]. Nonetheless, the ones we collected are judgements made by people, hence they will always carry some degree of personal bias.

An alternative mathematical elicitation method commonly employed is the 'quantile' method, whereby experts identify the range of values that parameter can take and then intervals are generated for them to assign probabilities that the value is contained within those 'bins' [96]. This method has been found to be clearer and easier to use than other methods as well as consistent in betting situations, but it also reports higher relative dispersion of the distribution [91, 101]. A tool was recently developed to facilitate the elicitation of expert opinion using the quantile method and with health economic decision-analytic models specifically in mind. The computer-based programme in question would first ask each expert for their lowest, highest and most likely value, based on which it would then provide four probability bins for that expert to report the probability that

the estimated value would lie in such intervals [96]. Although the method and more specifically this tool might have proved useful for our exercise, we found the four probability bins to be too restrictive for our purposes. We considered it important to provide the experts with the option to respond with distributions that could handle many different values with different probabilities. Furthermore, we believe that by asking participants to draw the probability distributions themselves we effectively minimise the gap between the expert's belief and the shape of the elicited distribution, which is the essential measure of a good elicitation method [91]. Under traditional quantile methods, the experts normally would not have the opportunity to see and confirm the probability distribution built based on their responses, but in our case, experts built their own.

Considering that no study has been published providing probabilities of referrals of hip pain patients for a THR, we feel confident that the expert elicitation exercise conducted within this study provides a valid estimate of those probabilities to be used in the economic model.

5.3 Postoperative transition probabilities

Postoperative transition probabilities include the probabilities of *Good* and *Poor* outcomes during the first year after primary THR, probabilities of remaining or moving across outcome categories after the second year post surgery, probabilities of a revision THR, and probabilities of transitioning into each outcome category after the revision as well as moving across them. The main source of data used to obtain estimations for these transition probabilities was a subset of the HES-PROMs dataset acquired by the COASt project. Other data sources used included EPOS (also used for the development of the outcome prediction tool reported in Section 4.2), summary statistics from the NJR, mortality rates from the ONS, and a study about revision rates based on data from the New Zealand Registry. The latter is the only source of data based on patients from outside the UK employed in our economic model.

5.3.1 Outcomes after Primary THR

Since our model combines the surgical procedures (whether primary or revision) with the outcome category during the year immediately following the operation in order to respect the one-year cycles, all transitions to and from each THR state are divided into two. Hence, when a cohort progresses through the model, transitions from preoperative health states to *Primary THR* will immediately be split into *THR+Good outcome* or *THR+Poor*

outcome according to the probabilities of *Good* and *Poor* outcomes at one year after the operation. These are the probabilities discussed in this section.

Data for the breakdown of *Good* and *Poor* outcomes after primary THR were obtained from the NHS' PROMs initiative. This programme, which started in April 2009, requires providers of NHS-funded unilateral hip replacements, unilateral knee replacements, groin hernia surgery and varicose vein surgery to invite patients to participate in the collection of outcomes data to help assess the effectiveness and quality of care provided [25, 107]. The scheme involves the completion by patients of pre- and postoperative questionnaires collecting data on their health status via a condition-specific and a generic, preference based outcomes measure. In the case of hip replacements, postoperative measures are taken at least six months after surgery and the questionnaires patients must complete are the condition-specific OHS and the generic EQ-5D. Other information such as co-morbidities and living arrangements are also collected. The completed preoperative forms are sent to a DH contractor, where NHS numbers are obtained so that forms can be linked to the Hospital Episodes Statistics (HES) database. The DH contractor then sends by post the follow-up questionnaire at the appropriate time to the patient's home address, which, once returned, is linked to the pre-op counterpart [107].

The COASt project obtained from The Health and Social Care Information Centre the non-identifiable HES and PROMs records[3] of all patients who had a hip replacement operation and who accepted to participate in the PROMs initiative. PROMs and HES records were provided as separate data sets, with HES records grouped by fiscal year. Each HES data set contained a link variable unique within each fiscal year and (potentially) linking each record to an entry in the PROMs data set, which included both a variable with the link code and a variable with the HES year corresponding to the PROMs entry. We excluded the preoperative PROMs records from fiscal year 2008-09 because they were used as a pilot for the scheme [107], and those from 2012-13 because HES records were not available for them as the fiscal year had not ended by the time the data were provided. Of a total of 171,881 PROMs records for the three fiscal years between 2009 and 2012, only 128,084 were linkable to their HES counterpart. The remaining 43,797 records were missing the HES year variable necessary to make the link to the HES data set, hence they were excluded from the sample. This is a significant number of

[3] Copyright 2012, used with the permission of The Health and Social Care Information Centre. All rights are reserved.

records (25% of the sample) whose exclusion could potentially bias the analysis; however, since records could not be confidently linked in any other way and the number of PROMs records with a HES year was significantly large, we proceeded with the cleaning protocol regarding the dataset as reliable and the best available source of PROMs data in the country to be used for the model.

Merging the PROMs with the HES data sets was successful for 123,035 of the 128,084 records available for merging (96%). The 3,026 records that could not be matched had a significantly lower proportion of females (26%) than the matched group (60%), a likely cause for the statistically significant differences between their pre- and postoperative OHS and EQ-5D scores (according to results from the non-parametric Wilcoxon-Mann-Whitney mean comparison test). Nevertheless, the number of unmatched records was so low relative to those successfully matched, that we considered it highly unlikely that it would bias results in any relevant manner.

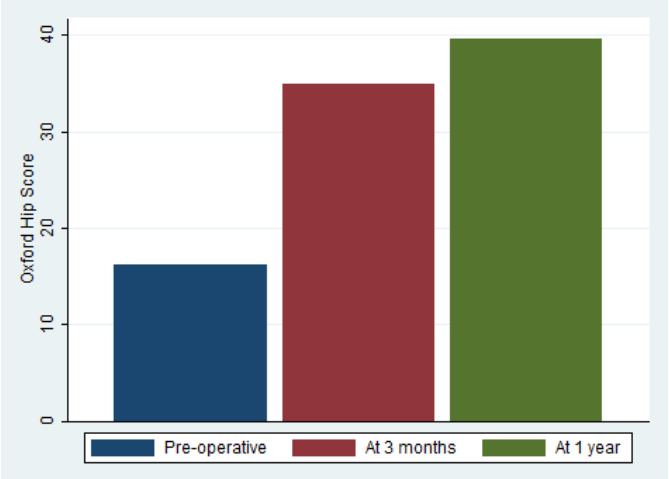
Once merged, we were able to split the sample of PROMs records into primary and revision hip replacements as this information was contained in the HES dataset in the form of OPCS codes, of which each record may have up to 24 starting with the most resource intensive procedure. As we were interested in primary and revision THRs only, we dropped records for other hip replacement procedures such as hybrid prosthetic hip replacement or total prosthetic replacement of the femoral head. THRs accounted for 87% of primaries and 88% of revisions. After removing records from patients younger than 45 years of age and keeping only those with non-missing postoperative OHS score so that their outcome category could be determined, the dataset of primary and revision THR by age and gender group was comprised of 68,156 and 5,320 interventions, respectively. The breakdown of the primary interventions by patient subgroups is shown in Table 5.5.

Table 5.5
Number of matched HES and PROMs records of primary THR patients with non-missing postoperative OHS, by age and gender groups

	Male	Female
45-60 years old	4,801	6,267
60-70 years old	10,103	13,355
70-80 years old	10,102	14,867
80+ years old	2,794	5,867
Total	**27,800**	**40,356**

Since all operations reported in Table 5.5 included a valid postoperative OHS, they could all be classified as *Good* or *Poor* outcomes based on the criteria set in Section 4.3. Patients reporting an OHS score below 38 one year after their primary were accordingly considered *Poor* outcomes. There is, however, a difference between the one year mark at which the postoperative measure was taken from EPOS patients whose answers were used to derive the cut-off point, and the minimum six-month term after which NHS patients may be contacted to provide their postoperative PROMs. Data from the EPOS sample indicate, nevertheless, that most of the improvement captured by the OHS at one year takes place within the first three months after the operation. Of a total of 1,589 EPOS patients with a primary THR, we excluded 43 who were younger than 45 years of age, and identified 1,048 who completed the OHS questionnaire before the operation, at three months and at one year after the primary. Mean OHS values at those three times for these patients were 16.2 (95% confidence interval 15.8-16.7), 35.0 (CI 34.5-35.4) and 39.6 (CI 39.1-40.1), as shown in Figure 5.11. Mean values change less than half of one point if they are calculated based on all patients who answered the questionnaire each time, regardless of whether they completed all three. These statistically significant results clearly indicate that improvement is not linear over time. They show, instead, that most of the progress measured by the OHS takes place in the first few months after surgery. We can therefore expect no significant difference between average postoperative OHS at six months compared to the same measure at one year, and hence that postoperative HES-PROMs records taken at various points at least six months after the operation can appropriately be combined with the cut-off point for surgery outcome categories identified based upon OHS records collected at one year.

Figure 5.11
Mean preoperative and postoperative OHS at three months and one year



The percentage of patients in each outcome category after *Primary THR*, based on postop OHS reported in the HES-PROMs data from 2009 to 2012, allowed for an estimate of the probability of *Poor* outcome by patient group. Table 5.6 shows that, based on this criterion and the cut-off point used, 30% to 50% of patients would be classified as *Poor* outcomes and that it is more likely for women to perform *poorly* than men of similar age. Probabilities reported in Table 5.6 also suggest that it is roughly as likely, for both men and women, to perform *poorly* if they are between 45 and 60 years of age as if they had between 60 and 70. However, after 70 years of age the probability of *Poor* outcome notably increases with age. The younger, much smaller and possibly healthier cohort of patients used by Arden *et al* [79] in their work identifying cut-off points are likely reasons why their expected percentage of *Poor* outcomes (see Table 4.2) is much lower than these.

Table 5.6
Probability of Poor outcome one year after Primary THR

	Male	Female
45-60 years old	0.298	0.359
60-70 years old	0.262	0.329
70-80 years old	0.310	0.410
80+ years old	0.398	0.514

As the split between *Good* and *Poor* outcome can naturally be considered binomial data, we fitted a Beta distribution for the probability of *Poor* outcome immediately after a primary THR based on the counts of *Good* and *Poor* outcomes within each patient subgroup, as reported in Table 5.7. Given the large number of observations, uncertainty around these parameter values was quite narrow. Probabilities of death at one year after THR, whether at *Good* or *Poor* outcome, were obtained from the NJR [54] and are shown in Appendix 13.

Table 5.7
Probability of *Poor* outcome after *Primary THR*: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.298	Beta	1,431	3,370
Males, 60-70 years of age	0.262	Beta	2,647	7,546
Males, 70-80 years of age	0.310	Beta	3,128	6,974
Males, 80+ years of age	0.398	Beta	1,112	1,682
Females, 45-60 years of age	0.359	Beta	2,253	4,014
Females, 60-70 years of age	0.329	Beta	4,399	8,956
Females, 70-80 years of age	0.410	Beta	6,099	8,768
Females, 80+ years of age	0.514	Beta	3,015	2,852

5.3.2 Transitions between outcomes categories after Primary THR

Since our Markov model distinguishes between outcome categories as separate health states both at the first year after the operation (combined with the *Primary THR*) and during the following years as well, we required two sets of transition probabilities. First, we needed estimates for the probabilities of moving from each of the outcome categories in the first year to each outcome category in year two, which also represent outcome categories for the following years in either outcome category. Secondly, we required estimates for the probabilities of moving between health states representing outcome categories during the second and subsequent years after the primary. We used data collected preoperatively and annually during five years after a primary THR by the EPOS group to estimate these probabilities.

As reported in the previous section, the EPOS data available to us on primary THR patients included OHS scores and other demographic information from a total of 1,589 patients. Since we were interested in producing probability estimates for each gender-age patient subgroup, we retained the records of those with non-missing gender and reported age above 45 years, leading to a working total of 1,534 records. Since this number was further reduced because of missing OHS questionnaires, subsequently divided by patient subgroups and finally classified by outcome category, the number of patients transiting between outcome categories from one year to the next within each patient subgroup became too small (under 10 in many cases) to produce reliable estimates. In order to obtain more power, we merged the first and second age groups (45-70 years of age) as well as the third and fourth (70+) respecting the gender differentiation, thereby producing estimates for four patient subgroups instead of eight. The model was populated, therefore, with the same value for each set of two of the original patient subgroups merged into one for this parameter.

Probabilities of transition from *Good* and *Poor* outcome in the first year to *Good* or *Poor* outcome in the second were estimated from the 1,043 patients who completed the OHS questionnaire both at one and two years after the operation. The classification into *Good* or *Poor* was applied using as cut-off points 38 for the first year and 33 for the second, as explained in Section 4.3. Table 5.8 reports the transition probabilities estimated from the sample for each patient subgroup as well as the distribution parameters according to the respective counts. Only one transition from each outcome category is reported as the other will result from calculating 1 minus the probability of death, minus the probability

of revision (reported in the next section), minus the probability reported in Table 5.8. Detailed counts of patients by outcome category in each year are presented in Appendix 14.

As Table 5.8 shows, most patients (between 92% and 99%) classified as *Good* outcomes at the first year following their primary THR will continue to be in a *Good* outcome state at year two. In the model, this is captured by the transition between the *THR+Good outcome* state to that of *Good outcome* after primary. Only a small proportion will go from *Good* to *Poor* outcome between years one and two after their primary, as surgeons consulted during model validation sessions had expected. In the case of those *Poor* outcomes during the first year after the THR, which according to Table 5.7 is between 30% to 40% of patients in all but two of the patient subgroups, about half of them would continue in *Poor* outcome in year two whereas the other half would improve and be classified as *Good* outcomes. This is also clinically feasible as improvements in mobility and pain may take longer than 12 months for many patients who have undergone a major procedure such as a THR.

Table 5.8
Transition probabilities between outcome categories after *Primary THR* from first to second postoperative years: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Good outcome year 1 to *Good* outcome year 2				
Males, 45-70 years of age	0.929	Beta	131	10
Males, 70+ years of age	0.987	Beta	156	2
Females, 45-70 years of age	0.966	Beta	196	7
Females, 70+ years of age	0.920	Beta	230	20
Poor outcome year 1 to *Poor* outcome year 2				
Males, 45-70 years of age	0.444	Beta	24	30
Males, 70+ years of age	0.472	Beta	17	19
Females, 45-70 years of age	0.578	Beta	52	38
Females, 70+ years of age	0.505	Beta	56	55

The pattern of transitions describing the proportion of patients that stay as *Good (Poor)* outcome for the years after the second postoperative year, and those that worsen(improve) and move onto *Poor(Good)*, was obtained from the OHS data by EPOS patients in years two through five after their primary. As Table 5.9 shows, the proportion of patients remaining in *Poor* outcome from year two after their operation onwards increased with respect to the same proportion between years one and two for all patient

subgroups. With the exception of male patients older than 70 between years four and five after the primary THR, for all subgroups and in all yearly transitions between outcome categories, the majority of *Poor* outcome patients stayed as such in the following year. The proportion of *Good* outcomes staying as *Good* outcomes in the three reported yearly transitions remained high for all patient subgroups at levels between 89% and 97%, as shown in Appendix 15.

Since we used a single model state to represent all postoperative years after the second for each outcome category, we estimated an average transition probability that could be applied at each iteration of the Markov model. We considered three different scenarios to extrapolate proportions of *Good* and *Poor* outcomes up to 10 years after the primary THR and compared results to a third clinically plausible alternative. The scenarios included, first, calculating the mean transition probability for remaining in each outcome category between years two and five and applying it to the transitions between years five and ten; second, assigning the transition probabilities between years four and five to the following yearly transitions, calculating an *extrapolated mean* between years two and ten; and finally, applying it to all transitions after year two. We compared these two scenarios to a third plausible alternative whereby transitions up to year five would be as reported by EPOS whilst the last transition, between years four and five, would be maintained over time up to year 10. Figure 5.12 shows the proportion of *Good* and *Poor* outcomes estimated for years two through 10 after the operation according to the above three scenarios for females between 70 and 80 years of age, as they are the largest THR subgroup based on the number of records obtained from the HES-PROMs data set. The proportion of *Good* and *Poor* outcomes at year two were obtained by applying the mean probability reported in Table 5.7 to obtain the split at year one, and then the respective transition probabilities from Table 5.8 as the cohort moved between outcome categories from year one to two. As Figure 5.12 shows, applying a mean probability for all years after the second or using an *extrapolated mean* produced results not only equivalent to one another but also undistinguishable from a progression calculated based on observed yearly transitions up to year five and then maintaining the last transition probabilities over time. This exercise was replicated for all patient subgroups and all three scenarios were almost identical in all cases. Figures showing results for the remaining patient subgroups are shown in Appendix 16.

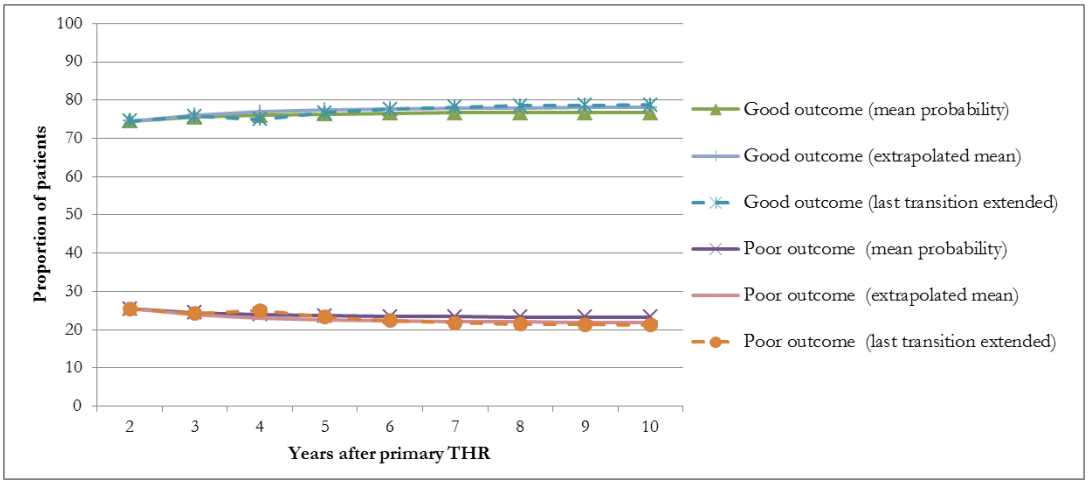
Table 5.9
Observed proportion of EPOS patients remaining in *Poor* outcome category between years two and five after Primary THR

	Years 2-3		Years 3-4		Years 4-5	
	n*	%**	n*	%**	n*	%**
Males, 45-70 years of age	32	65.6	24	70.8	25	68.0
Males, 70+ years of age	17	58.8	16	50.0	18	38.9
Females, 45-70 years of age	54	77.8	53	77.4	49	63.3
Females, 70+ years of age	77	66.2	76	68.4	72	65.3

* number of THR patients classified as *Good* outcomes in year pre transition
** percentage of *n* patients who remained as *Good* outcome in the following year

Figure 5.12 shows that, from the second year onwards after the primary THR, all patient subgroups report about three times as many patients in *Good* outcome compared to *Poor*, and that the proportion of *Good(Poor)* increases(decreases) slightly for the first few years and then plateaus. This is the case even for females over 80 years of age, the majority of whom perform poorly after the operation (see Table 5.7), yet transition probabilities indicate that about half of those *Poor* outcomes at year one reach a level of OHS associated to a *Good* outcome by year two after the primary. Both scenarios of extrapolation considered essentially maintain proportions at year five through year 10.

Figure 5.12
Proportion estimates of *Good* and *Poor* outcomes from three scenarios extrapolating transition probabilities, Females between 70 and 80 years of age



Given the equivalence of the two extrapolating scenarios, we chose to use the one employing the mean of the three yearly transition probabilities because it is simpler and more straight-forward. Table 5.10 shows those mean probabilities and the parameters of

the Beta distributions calculated based on average counts. We confirmed that the Beta distributions thus parameterised produced means that were not further than 0.006 from the observed mean. Uncertainty around the transition leading to remaining in *Poor* outcome for male patients older than 70 years of age produced the widest uncertainty given the low number of cases reported. Finally, yearly mortality rates from both model states were assumed to be the same all-cause gender and age-specific death rates used preoperatively [106].

5.3.3 Revision THR

A patient in *Good* or *Poor* outcome after a primary THR may need a revision of the implanted prosthesis. The Markov model presented here allows for this by introducing transitions that allow patients to remain in the same outcome category after the primary or to move to the other until they die, together with transitions to a *Revision THR* from either the *Good* or *Poor* outcome states after the primary to capture those cases in which patients do have their implants revised. In this section we report on the revision rates employed in the economic model to account for the transitions to a *Revision THR* from either outcome category during the first year after the primary, or from the outcome states from year two onwards.

Table 5.10
Transition probabilities between outcome categories from second year after a primary THR onwards: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Good outcome to Good outcome				
Males, 45-70 years of age	0.958	Beta	157.3	7.0
Males, 70+ years of age	0.919	Beta	144.0	12.7
Females, 45-70 years of age	0.945	Beta	223.7	13.0
Females, 70+ years of age	0.899	Beta	239.7	27.0
Poor outcome to Poor outcome				
Males, 45-70 years of age	0.682	Beta	18.3	8.7
Males, 70+ years of age	0.492	Beta	8.3	8.7
Females, 45-70 years of age	0.728	Beta	38.0	14.0
Females, 70+ years of age	0.666	Beta	50.0	25.0

THR revision rates have traditionally been calculated for the entire sample of primaries, or in some cases they are stratified by fixation and bearing surfaces, revision reason or even by major implant brands, as in the latest annual report of the NJR does [54]. However,

revision rates by categories of outcome based on postoperative OHS have only been reported to date by Rothwell *et al* [108] based on a sample of over 15,000 THRs from the New Zealand Joint Registry. The authors showed that lower postoperative OHS scores were strongly associated with higher revision rates at two years after the primary. Revision rates were calculated for the four groups proposed by Kalairajah [61] using 27, 33 and 41 as cut-off points on postoperative OHS at six months after primary. Table 5.11 reproduces the number of patients who had their primaries revised within two years, the total number of patients by group and the corresponding two-year revision rate, as reported by Rothwell *et al* [108].

We used the figures in Table 5.11 to produce estimated revision rates for our two categories of outcome after THR. Since we are using a postoperative OHS of 38 to classify patients into *Poor* and *Good* outcome one year after the primary, Kalairajah's cut-off points of 34 or 41 could be used to re-categorise the four groups into two. Based on the overall proportion of patients classified as *Poor* and *Good* outcomes in the HES-PROMs dataset (45% and 55%, respectively), we chose to consider the three Kalairajah's groups scoring up to 41 as Poor (42%) and those above 41 as *Good* (58%). *Poor* outcomes (OHS≤41) were hence associated to a two-year revision rate of 2.35% and *Good* outcomes (OHS>41) with 0.48%. The relative risk of revision thus indicates that patients with an OHS lower than or equal to 41 six months after their primary operation are 4.93 times more likely to have a revision in two years than patients scoring above 41.

Table 5.11
Two-year revision rate by Kalairajah OHS classification

Group (OHS)	Patients	Revised	Revision rate
< 27	944	72	7.6%
27 – 33	1,452	32	2.2%
34 - 41	4,170	50	1.2%
> 41	9,257	44	0.5%

For the group with OHS < 27, the authors indicated a revision rate of 7.6% resulting from 72 patients having had a revision out of a total of 874 THRs, which would instead produce a revision rate of 8.2%. We decided to keep the reported revision rate and number of revisions, and to adjust the total number of patients accordingly to produce the reported revision rate of 7.6% at two years.

In order to produce separate revision rates for *Good* and *Poor* outcomes during the first year after the primary, we used the figures in Table 5.11 to produce instantaneous revision

rates (assuming the rate was constant over the two years) and then probabilities of revision at one year for the two outcome categories. The one-year probability of revision for the group scoring less than or equal to 41 was 1.18% and that of the patients scoring above 41 was 0.24%. The relative risk of revision at one year was therefore 4.96, all based on data from the sample of New Zealand THR patients used by Rothwell *et al* [108].

As a similar relative risk of revision by *Poor* to *Good* outcomes is not available for THR patients in the UK, we used this relative risk to produce revision rates stratified by outcome categories based on overall revision rates reported by the NJR and the proportion of *Good* and *Poor* outcomes found in the HES-PROMs dataset. Since NJR does not publicly report revision rates by gender or age groups, we produced overall revision rates and assumed them to be the same for all patient subgroups. In order to obtain the probabilities of revision at one year for each outcome category, proportions of patients revised and not revised in each category had to be estimated whilst at the same time complying with three simultaneous conditions. These conditions are explained below based on the nomenclature shown in Table 5.12, whilst resulting estimates of one-year revision rates by outcome category are shown in Table 5.13. As the latter shows, a one-year revision rate for *Poor* outcomes of 1.29% and *Good* outcomes at 0.26% produce the same relative risk of revision found by Rothwell *et al* [108] whilst maintaining the same proportion of *Poor* to *Good* outcomes found in HES-PROMs and the overall revision rate reported by the NJR [54].

As the economic model requires not only revision rates during the first year after primary but also during subsequent years once patients have moved into the *Poor* or *Good* outcome states, we followed the same procedure explained above to produce revision rates by outcome category after the first year. For this, we first re-categorised the revisions figures reported by Rothwell *et al* [108] at two years into *Good* and *Poor* outcomes using 34 as the cut-off point, almost exactly the same as the score of 33 which, according to Arden *et al* [79], maximises sensitivity and specificity at predicting satisfaction two years after surgery (see Section 4.3). With this reorganisation of groups, the two-year probabilities of revision for *Poor* (<34) and *Good* (≥34) outcomes on the New Zealand sample were 4.34% and 0.70%, respectively, with a relative risk of revision of 6.2. We assumed that this relative risk would hold not only for the second year after the primary but for the following years as well.

Table 5.12
Producing revision rates by outcome categories in the UK based on data from New Zealand

Group	Revised	Not revised	Total	Revision rate
Poor outcome	a	b	c	j
Good outcome	d	e	f	k
Total	**g**	**h**	**i**	**l**

Condition 1: a +b=c=0.4545 and d+e=f=0.5455 as those are the overall proportions of Poor and Good outcomes, respectively, found in HES-PROMs.

Condition 2: where j=a/c and k=d/f, it must hold that j/k=4.96 as that is the relative risk of revision of Poor (OHS≤41) compared to Good outcomes (OHS>41) derived from findings by Rothwell et al [108].

Condition 3: where g=a+d and i=c+f, it must hold that g/i=l=0.0073 as that is the one year overall revision rate reported by the NJR[54].

Table 5.13
Estimated one-year revision rates by outcome category for the UK

Group	Revised	Not revised	Total patients	Revision rate
Poor outcome	0.59%	44.86%	45.45%	1.30%
Good outcome	0.14%	54.41%	54.55%	0.26%
Total	**0.73%**	**99.27%**	**100%**	**0.73%**

We calculated the arithmetic mean of the seven annual overall revision rates reported by the NJR between years two (1.19%) and eight (3.94%) after the primary, obtaining an annual average revision rate of 2.59%. Finally, based on the follow-up data collected under the EPOS study, we know that the proportion of *Good* and *Poor* outcomes varies only slightly between the first and the following years, hence we decided to use the same breakdown of the outcomes categories as used above for the derivation of revision rates during the first year after the primary. Based on the above conditions and solving for the equations as described in Table 5.12, estimated revision rates by outcome category for year two and onwards after the primary are shown in Table 5.14.

Table 5.14
Estimated revision rates by outcome category for the UK at two or more years after the primary

Group	Revised	Not revised	Total patients	Revision rate
Poor outcome	2.17%	43.28%	45.45%	4.77%
Good outcome	0.42%	54.13%	54.55%	0.77%
Total	**2.59%**	**97.41%**	**100%**	**2.59%**

As Table 5.14 shows, revision rates increase for both outcome group categories and especially for the *Poor* outcomes, as expected. The overall mean revision rate of 2.59% is upheld as are the proportions of either outcome category and the relative risk of revision between the groups. THR patients in the UK classified as *Poor* outcomes two or more years after their primary are therefore, according to these estimates, expected to get revisions at an annual rate of 4.8%, whereas *Good* outcomes would only require them in 0.8% of cases every year.

As the data used to derive the revision rates are that of the sample of the New Zealand Joint Registry, we used its number of observations reported in Table 5.11 to produce the parameters for the Beta distributions of the revision rates reported above. For this, we applied the proportions reported in Table 5.13 for the one-year revision rates and Table 5.14 for the rate afterwards to the 15,823 patients followed-up in the study by Rothwell *et al* [108]. Table 5.15 shows the distribution parameters of each revision rate for the PSA, which will be applied to all patient subgroups as the data was not reported by gender and age groups.

Table 5.15
Transition probabilities between outcome categories and *Revision THR*: deterministic and probabilistic parameters

Transition probability	Mean	Distribution	α	β
Good outcome first year to Revision THR	0.0026	Beta	22.2	8609.3
Poor outcome first year to Revision THR	0.0130	Beta	93.4	7098.2
Good outcome year two and onwards to Revision THR	0.0077	Beta	66.5	8565.0
Poor outcome year two and onwards to Revision THR	0.0477	Beta	343.4	6848.2

5.3.4 Outcomes after Revision THR

The previous section described the probabilities of patients in *Good* and *Poor* outcome states undergoing a revision THR. This transition must be combined with the probability of *Good* or *Poor* outcome during the first year after revision because the model consolidated these with *Revision THR* into two health states: revision THR immediately followed by *Good* outcome during the first year, and revision THR followed by *Poor* outcome. In this section, we describe the probabilities of *Good* and *Poor* outcome following a revision of the THR.

As it was done for primary THRs, we used the HES-PROMs data to derive the probabilities of *Good* and *Poor* outcomes after a revision THR. Table 5.16 shows the number of revisions funded by the NHS between fiscal years 2009-2010 and 2011-2012 for whom a postoperative OHS was available through the PROMs Initiative. As the table shows, most THR revisions were performed on women between the ages of 70 and 80, which is consistent with the frequency of primaries reported in Table 5.5 in section 5.3.1.

Table 5.16
Number of matched HES and PROMs records of revision THR patients with non-missing postoperative OHS, by age and gender groups

	Male	Female
45-60 years old	352	500
60-70 years old	702	819
70-80 years old	979	1,093
80+ years old	331	544
Total	**2,364**	**2,956**

As the study identifying cut-off points for outcome categories used data from primary THRs only [79] and it has not been replicated on revision operations, we used the threshold identified for the second year (OHS=33) to classify our HES-PROMs patient records into *Good* or *Poor* outcomes. We chose the lower two-year cut-off point as opposed to that for the first year after the operation because patients undergoing a revision THR would have had problems with their primary prosthesis and are less likely to perform well than the broader spectrum of patients undergoing a THR for the first time. This is confirmed by Figure 5.13 which shows that, whilst the mean postoperative OHS for primary THR patients in the HES-PROMs dataset varies between 35 and 40, following a revision THR these scores drop to values between 31 and 34, around 15% lower for all groups.

Based on of the 33 cut-off point, we used the observed proportion of *Poor* outcomes in each patient subgroup as an estimate of the mean probability of *Poor* outcome after Revision THR. These transition probabilities are shown in Table 5.17 together with the parameters for the Beta distributions assigned to each and taken from the counts for *Good* and *Poor* outcomes. Since mortality rates at one year after THR reported by the NJR were not specific to primary or revision, rates shown in Appendix 13 were also used as death rates after revision THR.

Figure 5.13
Mean OHS after Primary and Revision THR in HES-PROMs, by gender and age groups

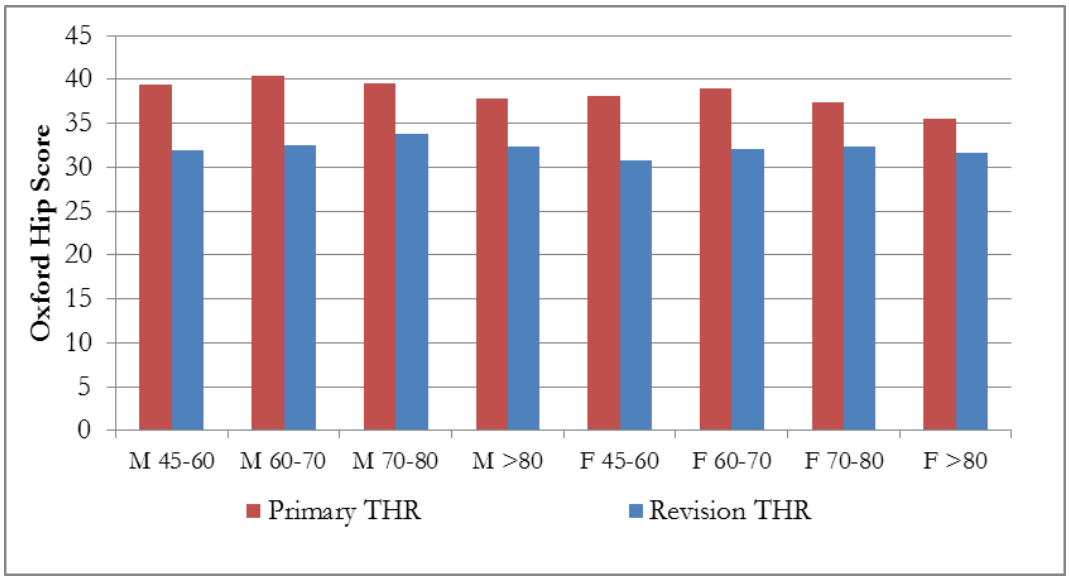


Table 5.17
Probability of *Poor* outcome after Revision THR: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Revision THR				
Males, 45-60 years of age	0.455	Beta	160	192
Males, 60-70 years of age	0.447	Beta	314	388
Males, 70-80 years of age	0.391	Beta	383	596
Males, 80+ years of age	0.474	Beta	157	174
Females, 45-60 years of age	0.510	Beta	255	245
Females, 60-70 years of age	0.476	Beta	390	429
Females, 70-80 years of age	0.450	Beta	492	601
Females, 80+ years of age	0.500	Beta	272	272

5.3.5 Transitions between outcomes categories after Revision THR

As in the case of primary THRs described in Section 5.3.2, our economic model required two sets of transition probabilities as the cohort moves through the Markov model following a revision procedure. First, after their first year in *Good* or *Poor* outcome immediately following the revision, patients who do not die would transit into *Good* or *Poor* outcome at year two, which are modelled as separate health states; and second, patients in either outcome category at year two or onwards may remain in the health state they are in or move to the other one at each iteration. In this section, we present these

two sets of transition probabilities and their distributions as they populated our economic model.

The transition probabilities from outcome categories in the first year to those in the second following a primary THR were estimated based on the 1,043 primary THR patients followed-up in EPOS. Of all patients in the study, a revision surgery was reported only for 25 of them, an insufficient sample size to produce estimates for the transition probabilities we required. Given that no other dataset was available with yearly follow-ups of revision THR patients, we used the same data from EPOS primary THR records to produce estimates for the transition probabilities between outcome categories after a revision. Although primaries and revisions are different in terms of the health condition of patients that undergo the procedure and in the results obtained (see Figure 5.13), we believe that the rates of transition between outcome categories, once the proportions of *Good* and *Poor* outcomes have been determined, are likely to be very similar.

We used 38 and 33 as OHS cut-off points to classify patients into *Good* or *Poor* outcomes for the transition probabilities between years one and two, respectively, after a primary THR. Since we decided to use 33 as the cut-off point for outcome classification at one year after revision THR, as argued in the previous section, we recalculated the probabilities estimated for primary THRs using this threshold for outcome classification for all years in the case of revision THRs. Table 5.18 shows the estimated probabilities of the transition between *Good* or *Poor* outcomes during the first year following the revision to the same outcome category the second year after the procedure. The probability of changing outcome categories is therefore the difference between one and the probabilities indicated in the table.

Table 5.18 shows that, as with primaries, it is most likely that patients will find themselves during the second year in the same outcome category they were in one year after the procedure, and this is especially the case with *Good* outcomes. Almost all of the latter remained as *Good* outcomes in year two, whereas slightly more than half of the *Poor* outcome also stayed in that category, the rest improving and crossing the threshold to become *Good* outcomes by the second year likely due to long rehabilitation periods. As expected, the probabilities of remaining in *Good* outcome are lower in Table 5.18 than in the case of primary THRs because the lower OHS threshold produced a larger group of *Good* outcomes in year one whilst the number of *Good* outcomes in year two remained the

same. In other words, the lower rate is to be expected since the numerator remained fixed and the denominator increased. This also explains the increase in the probabilities of remaining in *Poor* outcome.

Table 5.18
Transition probabilities between outcome categories after *Revision THR* from first to second postoperative years: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Good outcome year 1 to *Good* outcome year 2				
Males, 45-70 years of age	0.902	Beta	148	16
Males, 70+ years of age	0.954	Beta	167	8
Females, 45-70 years of age	0.913	Beta	219	21
Females, 70+ years of age	0.878	Beta	259	36
Poor outcome year 1 to *Poor* outcome year 2				
Males, 45-70 years of age	0.581	Beta	18	13
Males, 70+ years of age	0.579	Beta	11	8
Females, 45-70 years of age	0.717	Beta	38	15
Females, 70+ years of age	0.606	Beta	40	26

Transitions between outcome categories for year two onwards after primary THR were based on a threshold of 33. Hence, since we also applied this cut-off point for the case of revision THRs, the transition probabilities calculated for primaries were also used for revisions. Table 5.19 shows the mean transition probabilities and distribution parameters entered in the model for the transition of patients between outcome categories after a revision procedure. As indicated in the case of primaries and is also shown here in Table 5.18 and Table 5.19, the largest uncertainty around the mean value of transitions is found in the case of males older than 70 years of age starting in the *Poor* outcome state. The low number of such patients reported by EPOS means that PSA will assign values varying widely from the reported mean. Finally, all-cause gender and age-specific mortality rates from the ONS [106] were applied to both model states on outcome categories after revision.

The ones reported here and those in Section 5.2 complete the set of transition probabilities used by the economic model representing current practice. The following two sections describe the QALYs and costs, respectively, accrued by patients as they transit through each health state in the model.

Table 5.19
Transition probabilities between outcome categories after from second year after a revision THR onwards: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Good outcome to *Good* outcome				
Males, 45-70 years of age	0.958	Beta	157.3	7.0
Males, 70+ years of age	0.919	Beta	144.0	12.7
Females, 45-70 years of age	0.945	Beta	223.7	13.0
Females, 70+ years of age	0.899	Beta	239.7	27.0
Poor outcome to *Poor* outcome				
Males, 45-70 years of age	0.682	Beta	18.3	8.7
Males, 70+ years of age	0.492	Beta	8.3	8.7
Females, 45-70 years of age	0.728	Beta	38.0	14.0
Females, 70+ years of age	0.666	Beta	50.0	25.0

5.4 Quality-adjusted life years

As the cohort of patients passes through the different health states in the Markov model according to the transition probabilities described in the previous two sections, different levels of HRQL (measured as QALYs in our model) are assigned to each patient. In this section, we describe the mean values and distribution parameters of the QALYs associated to each health state, organised by sub-sections of the model: preoperative, at and after the primary, and at and after a revision THR.

5.4.1 Preoperative QALYs

The HES-PROMs data we used to estimate the probabilities of *Good* and *Poor* outcome after a primary THR was the most appropriate to estimate the HRQL of patients before they undergo the operation. Since the PROMs initiative collected a preoperative EQ-5D questionnaire, we were able to derive estimated health utilities for each patient subgroup by applying the utility weights obtained from the valuations of a sample of the UK's general population using the time trade-off method [62].

The economic model specified four different health states that patients may find themselves in after a consultation with the orthopaedic surgeon and prior to surgery. The fact that the model starts with this surgical assessment, which patients are referred to after having been evaluated and unsuccessfully treated by other health care professionals (see the hip patient care pathway in Figure 4.1 under Section 4.5) means that all patients seen by the surgeon are likely in a similar general health state. Furthermore, the mapping

exercise reported in Chapter 3 showed that the large overlap between the OHS and EQ-5D questionnaires means that the former is a significant predictor of the latter; or, in other words, that patients with severe and unresolved hip problems will see their preference-based health utility largely and similarly affected by their pain and its immediate consequences. We therefore used the preoperative EQ-5D questionnaires in the HES-PROMs data to populate all preoperative states of the model with a common health utility estimate.

Answers to the descriptive EQ-5D questionnaire produce any of 243 health states, which after applying the UK tariff of utility weights generate an equally finite number of summary scores ranging from -0.594 to one. As Figure 5.14 shows, the distribution of preoperative summary scores was far from uniform, with higher frequency of scores around zero and 0.7, and various ranges of unobserved values, not unexpected given the discrete nature of the EQ-5D index. Distributions by specific patient subgroups were very similar to one another.

Natural boundaries of health utilities between minus infinity (or -0.594 in the case of the EQ-5D-3L for the UK) and 1 suggest calculating disutilities to invert the range and then fit a log normal or gamma distribution to the data. Since mean values of preoperative EQ-5D summary scores for all patient subgroups were far from zero (between 0.3 and 0.4), all with very large sample sizes and hence minimal uncertainty about the true mean value being positive, applying the common pragmatic approach of fitting a Beta distribution directly to utility estimates [82] seemed acceptable. Fitted Beta distributions based on mean values and SDs of the summary EQ-5D scores for each patient subgroup were, nonetheless, U-shaped with high probabilities for values close to zero and 1, an inaccurate fit to the observed health utility data. The orthodox approach of fitting a Gamma distribution to previously calculated disutilities (i.e. what separates someone from perfect health) was followed, using the mean and SDs of the latter to produce distribution parameters. Disutilities were calculated by a simple transformation of 1 minus estimated utility, hence bound between zero and positive infinity allowing the application of a Gamma distribution. Table 5.20 shows the mean health disutility estimates associated to all preoperative health states and the respective Gamma distribution parameters for PSA. Although the Gamma distribution does not exactly match that of the observed data (e.g. it is not bimodal) and estimated distributions could theoretically reach values corresponding to health utilities lower than -0.594, all distributions in Table 5.20 proved to be unbiased

estimators as they generated mean values between 0.04 and 0.06 estimated health utilities away from those observed. For the PSA, random disutility values were converted back into health utility estimates and then multiplied by 1 to obtain QALYs, because the model was designed to run on yearly cycles and patients were expected to maintain their reported health utility throughout the year.

Figure 5.14
Distribution of preoperative EQ-5D summary scores from HES-PROMs data, all patient subgroups

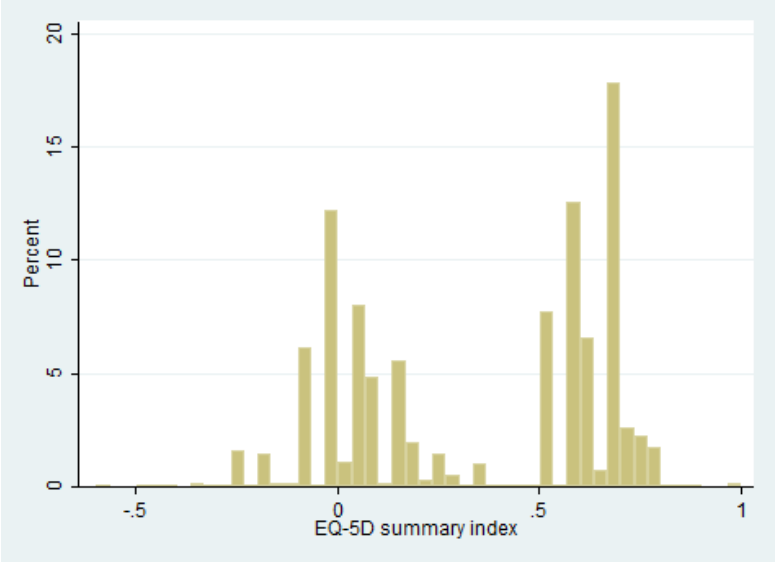


Table 5.20
Disutility associated to preoperative states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
All preoperative states				
Males, 45-60 years of age	0.615	Gamma	3.82	0.161
Males, 60-70 years of age	0.592	Gamma	3.68	0.161
Males, 70-80 years of age	0.597	Gamma	3.74	0.160
Males, 80+ years of age	0.656	Gamma	4.33	0.151
Females, 45-60 years of age	0.694	Gamma	4.61	0.151
Females, 60-70 years of age	0.653	Gamma	4.16	0.157
Females, 70-80 years of age	0.666	Gamma	4.32	0.154
Females, 80+ years of age	0.724	Gamma	5.05	0.143

5.4.2 QALYs after primary THR

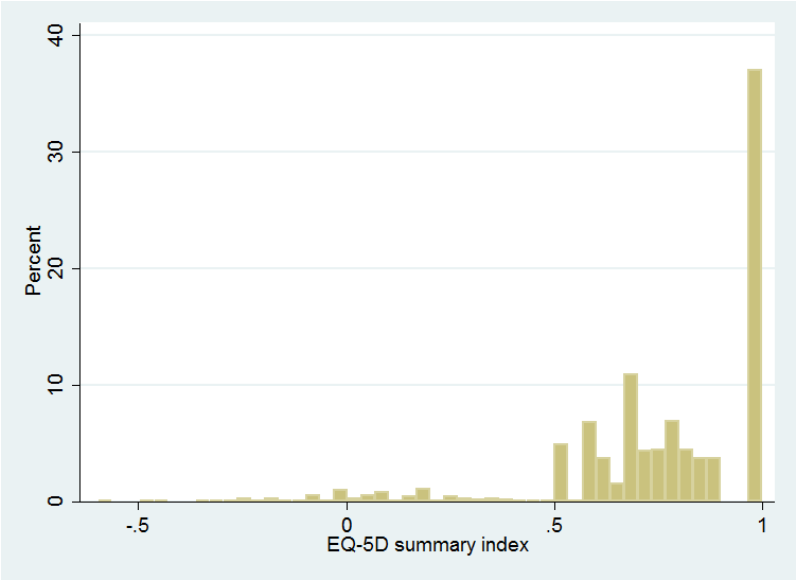
Patients undergoing a primary THR accrued health utilities depending on their outcome category. First, for health utilities associated to model states including the operation, we considered the EQ-5D postoperative summary scores by patient subgroups reported in

the HES-PROMs dataset. We produced QALYs from the latter by incorporating the progression of scores observed in EPOS because the latter reported a measure at three months that helps better understand patients' rate of improvement during the first year after the primary. For the model states representing years two and onwards, we used the expected health utility levels obtained from applying OHS progression rates observed in EPOS to postoperative scores reported in HES-PROMs.

Regarding the use of health utility estimates collected in EPOS, it is worth noting that the study used the SF-36 as a generic measure of health outcome whereas our economic model was consistently populated with health utility estimates derived from responses to the EQ-5D questionnaire. Although data collected from the SF-36 can produce a single preference-based index comparable to the EQ-5D summary score via the SF-6D [109], the two indices have been shown to produce diverging results. Utility estimates generated by the EQ-5D and SF-6D have produced different mean values, varying results across patient subgroups and severity levels, different ranges and variances including the known SF-6D floor effects and EQ-5D ceiling effects in various disease groups, among them osteoarthritis patients [110]. The two measures have been shown to produce final health utility estimates that are not interchangeable [111]. We therefore used the mapping algorithms developed in Chapter 3 to estimate summary EQ-5D scores from the responses to the OHS questionnaires available from EPOS to help understand the progression by *Good* and *Poor* outcome patients in the first year after surgery. We also benefited from the mapping of scores to produce utility estimates for each outcome category at two years and onwards after the operation.

To assign a health utility to the model states combining the primary THR and either outcome category, the main input was the summary scores of postoperative EQ-5D collected by the PROMs initiative. Figure 5.15 shows how, after the operation, the distribution of EQ-5D scores shifts markedly to the right so that most indices surpass 0.5 and over one third of patients report perfect health. Basic descriptive statistics for utility estimates by patient subgroup and outcome category are shown in Table 5.21. Whilst gender and age do not seem to have a significant effect on the variance of health utility levels attained by THR patients, outcome category does. Patients labelled as *Poor* outcomes only achieved health utility levels circa 0.5, whereas patient subgroups categorised as *Good* outcomes reached mean values as high as 0.9.

Figure 5.15
Distribution of postoperative EQ-5D summary scores from HES-PROMs data, all patients



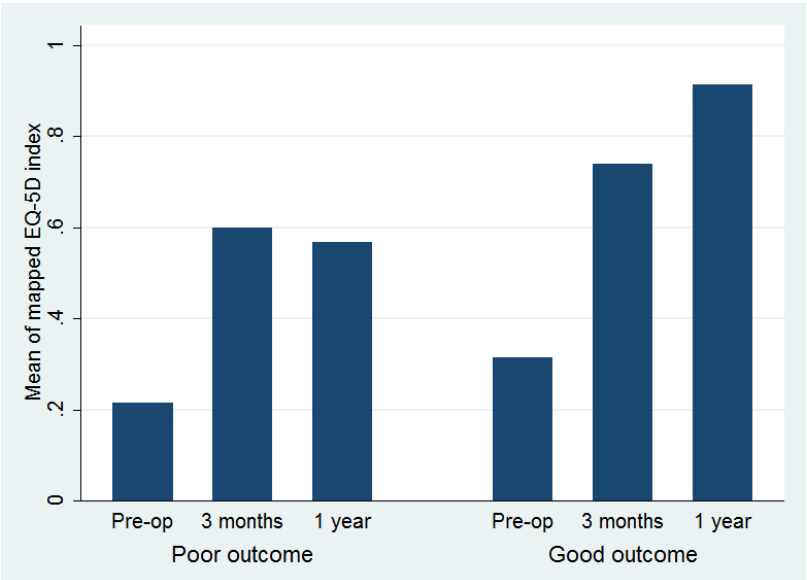
A patient transiting through the health states combining the THR and the first postoperative year, however, would not accrue the full yearly cycle at the above utility levels. They would have arrived at either of these health states at the much lower levels indicated in Table 5.20; and, as reported in Section 5.3.1, we know that improvement, measured by the OHS, is not linear over the first year after the procedure. In fact, after applying the OLS Continuous mapping approach presented in Chapter 3 (as it was the one with the lowest prediction error of the observed EQ-5D mean score) to EPOS records reporting non-missing OHS scores before the operation, at three months and one year after the procedure, we confirmed not only that most of the health utility improvement occurs during the first three months after surgery, but also that this progression is quite different between *Good* and *Poor* outcomes. As Figure 5.16 shows, the mean mapped EQ-5D score of the 310 EPOS patients categorised as *Poor* outcomes almost tripled from the time before surgery (0.215) to three months after (0.599), but then halted and actually registered a slight decrease at one year (0.567). For the 738 *Good* outcomes, however, the significant improvement in the first three months after surgery (from 0.236 to 0.753) reduced its pace but continued until it reached 0.913 at one year after the operation. In other words, according to the data reported by EPOS, *Poor* outcomes reach, at three months, approximately the same level of health utility they will report one year after surgery, whilst *Good* outcomes increase their utility in the first three months as much as three quarters (73%) of the total gain they will see in the full first year after the operation, improving still some more during the following nine months.

Table 5.21
HES-PROMs data: Postoperative EQ-5D summary scores by patient subgroup and outcome category

	Poor outcome		*Good* outcome	
Patient subgroup	**Mean**	**SD**	**Mean**	**SD**
Males, 45-60 years of age	0.490	0.282	0.902	0.153
Males, 60-70 years of age	0.551	0.251	0.903	0.153
Males, 70-80 years of age	0.586	0.231	0.895	0.148
Males, 80+ years of age	0.585	0.222	0.873	0.159
Females, 45-60 years of age	0.505	0.283	0.885	0.170
Females, 60-70 years of age	0.562	0.250	0.889	0.159
Females, 70-80 years of age	0.589	0.228	0.881	0.155
Females, 80+ years of age	0.586	0.238	0.863	0.161

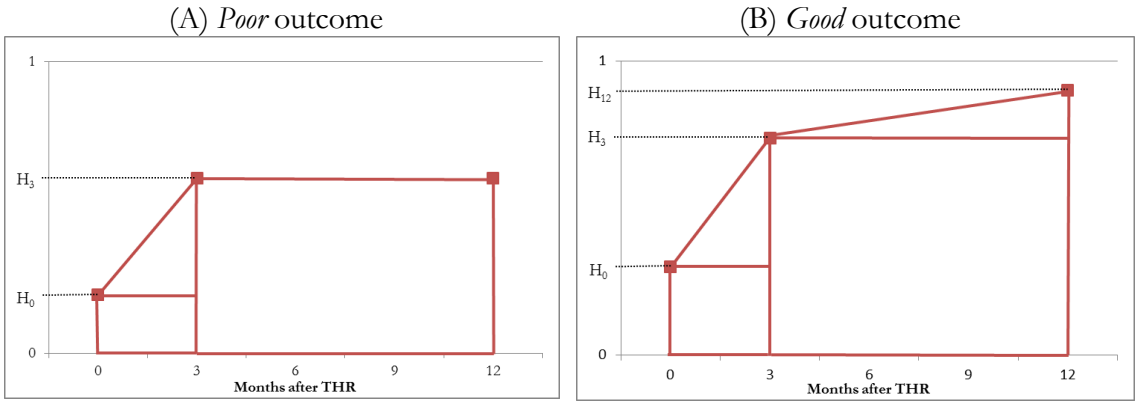
We applied these progression patterns to the data on preoperative and postoperative EQ-5D summary scores reported in the HES-PROMs dataset, to estimate QALYs associated to the first year after primary THR for each patient subgroup by outcome category. We did not include a disutility effect from the pain and discomfort produced by the operation because, as shown in Figure 5.16, even *Poor* outcomes improve significantly in the first 12 weeks after surgery, suggesting it is highly unlikely that such disutility would be of significant size to affect final results. Moreover, such effect would have to vary with the application of the prediction tool or between *Good* and *Poor* outcome patients for it to be relevant for this analysis, and we have no reason to believe either to be the case.

Figure 5.16
Mean EQ-5D summary scores mapped from OHS reported by EPOS patients



The general pattern of improvement is illustrated in Figure 5.17 for *Good* and *Poor* outcomes following results from EPOS. Although shown in the figures as improvement, the change in the summary EQ-5D score may in some cases be negative, as it was in fact for a number of patients in the HES-PROMs. If such drop was substantial and moved the patient from a positive to a negative health utility estimate, then the QALYs associated to the first year after the operation would have been negative if the area below zero was greater than the area in the positive range. A negative health utility during the first year was also negative for cases in which the EQ-5D summary score was negative before the operation and it remained under zero afterwards, or if it improved, it would still be negative in cases for which the change was not sufficient to achieve high enough levels in the positive range to compensate for the negative health utility. We considered all such cases accurate reflections of the negative health utility experienced by a number of patients and reported via the EQ-5D questionnaire under the PROMs initiative.

Figure 5.17
Components of health utility associated to first year after THR, by outcome category



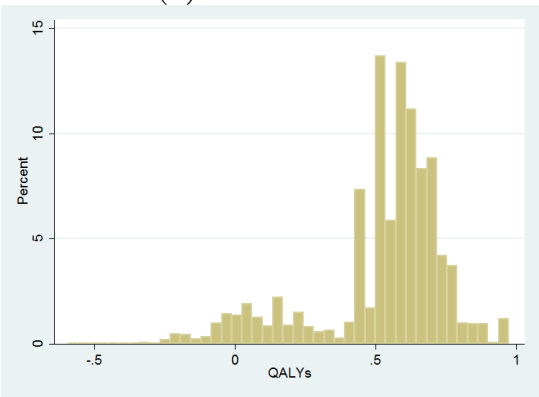
Where H_0 is the preoperative summary EQ-5D score, H_3 the score at 3 months, and H_{12} the score one year after the operation. For *Poor* outcomes, it is assumed that $H_3 = H_{12}$, whereas for *Good* outcomes, $H_3 = H_0+0.73(H_{12}–H_0)$.

After applying the progression patterns described above to HES-PROMs patient records, we obtained distributions of QALYs associated to the first year after THR that varied noticeably between *Good* and *Poor* outcomes. As shown in Figure 5.18, although an estimated 0.5 QALY or more for the first year was the norm for both outcome categories, accruing less than 0.5 of a QALY or even a negative measure was not uncommon for *Poor* outcome patients, whereas *Good* outcomes only experienced such low values in exceptional cases. We fitted Beta distributions to the respective means and SDs by patient subgroups but the resulting functions assigned similar probabilities to a large portion of the full range between 0 and 1. Since this is not consistent with the expected variation of

the mean value based on the observed EPOS data, we calculated disutilities and estimated the parameters for the corresponding Gamma distributions. Table 5.22 shows the parameters for the Gamma distributions describing uncertainty around the mean value of the QALYs associated to the first year after a *Primary THR* for both *Good* and *Poor* outcome patients.

Figure 5.18
Distribution of estimated QALYs during first year after primary THR, by outcome category

(A) *Poor* outcomes



(B) *Good* outcomes

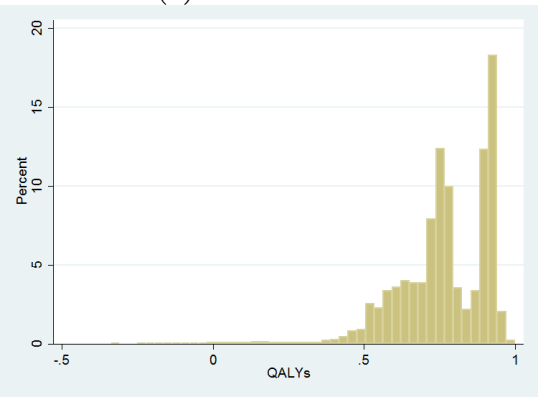


Table 5.22
Disutility associated to *Primary THR* and first postoperative year states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Primary THR + first year in *Poor* outcome				
Males, 45-60 years of age	0.540	Gamma	4.20	0.129
Males, 60-70 years of age	0.485	Gamma	4.29	0.113
Males, 70-80 years of age	0.450	Gamma	4.37	0.103
Males, 80+ years of age	0.457	Gamma	4.87	0.094
Females, 45-60 years of age	0.536	Gamma	4.24	0.126
Females, 60-70 years of age	0.478	Gamma	4.22	0.113
Females, 70-80 years of age	0.454	Gamma	4.60	0.099
Females, 80+ years of age	0.462	Gamma	4.40	0.105
Primary THR + first year in *Good* outcome				
Males, 45-60 years of age	0.217	Gamma	2.14	0.101
Males, 60-70 years of age	0.212	Gamma	2.09	0.101
Males, 70-80 years of age	0.220	Gamma	2.39	0.092
Males, 80+ years of age	0.247	Gamma	2.72	0.091
Females, 45-60 years of age	0.248	Gamma	2.39	0.104
Females, 60-70 years of age	0.239	Gamma	2.45	0.098
Females, 70-80 years of age	0.245	Gamma	2.68	0.091
Females, 80+ years of age	0.270	Gamma	2.98	0.091

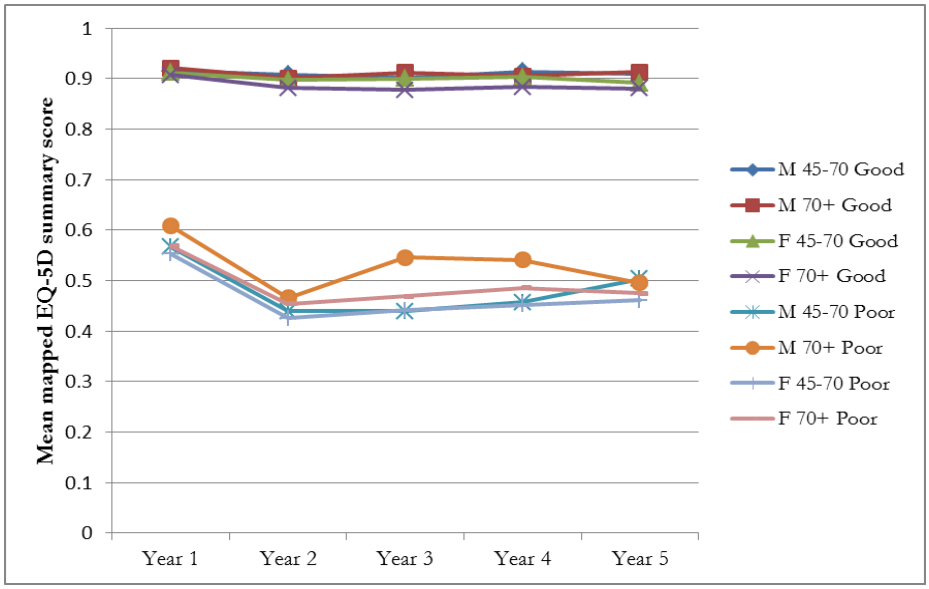
For the health states representing the second and subsequent years after the primary in either outcome category, we also benefited from the follow-up performed under EPOS and the representativeness in the HES-PROMs data when estimating the required QALY values. If data were available and a model was not necessary, the distribution of EQ-5D scores amongst patients in each of the second and subsequent years after surgery for *Good* and *Poor* outcome patients would produce the ideal parameter values. Since an economic model is being used to inform the decision making process, even if the above data were available, the structure of the model requires estimating a unique QALY value that applies to all years starting with the second after the primary. We estimated these values based on the data in EPOS and HES-PROMs by combining the relationships found in the former with the representativeness of the latter.

First, we explored the evolution over time of estimated EQ-5D summary scores mapped from OHS responses by EPOS patients grouped in outcome categories. Patient subgroups were merged as explained in Section 5.3.2. EQ-5D scores were estimated by applying the Continuous OLS mapping method described in Chapter 3, and outcome category groups were determined by the thresholds described in Section 4.3 (i.e. OHS at 38 at year one after the primary and at 33 after that). As Figure 5.19 shows, the mapped EQ-5D scores of those patients classified as *Good* outcomes remained very high and largely unchanged in years one through five after the primary. This, however, was not the case with *Poor* outcomes. The scores of many *Poor* outcome patients in the first year after the operation improved in such a way that, by year two, about half of them became part of the *Good* outcomes category (see transition probabilities reported in Table 5.8, Section 5.3.2). This caused the mean scores of *Poor* outcomes in year two to decrease in magnitudes of about 15% for all patient subgroups.

For the deterministic analysis, we needed a mean QALY estimate that represented most accurately the HRQL experienced by patients each year starting with the second after primary THR. For *Good* outcomes, any one between years two and five was equally appropriate because annual scores during that period were very similar. Scores from *Poor* outcome patients, as mentioned above, varied slightly. We chose year two as the reference year because the HRQL level then was representative of the levels afterwards for both *Good* and *Poor* outcome patients. Although *Poor* outcome males and females older than 70 years of age reported a slight improvement after year two, after that their mapped scores decreased again, approaching the level reported at year two. It is likely, moreover, that

HRQL would continue to decrease beyond that point after the fifth year given the natural progression of the disease, particularly for patients who performed poorly after the operation. For the younger patients, instead, their improvement was maintained beyond one or two years, however it is not unreasonable to expect their EQ-5D scores to eventually drop as well, until they reach and probably go under the values attained two years after the operation. Year two seemed therefore an appropriate choice for a reference year in terms of HRQL for THR patients after their operation.

Figure 5.19
Mean mapped EQ-5D summary scores by yearly outcome category, reported by EPOS patients



In order to obtain a representative measure of the summary EQ-5D score by patient subgroup and outcome category two years after the primary, we estimated models to predict OHS at year two based on EPOS data, applied the models to HES-PROMs records to obtain expected OHS at year two, and then mapped those values onto predicted EQ-5D summary scores using the same mapping algorithm employed thus far.

To predict OHS at year two, we estimated OLS models on EPOS data using OHS at year two as the outcome variable and OHS at year one as the regressor. Alternative models including age, gender, baseline OHS and change in OHS as covariates were also estimated but coefficients were not statistically significant or model performance was not improved. Figure 5.20 shows, nonetheless, that the slope of the curves connecting total OHS between years one and two after the primary varies not only significantly depending on outcome category at year one, but also slightly with age and gender. We therefore

estimated models for each patient subgroup within each outcome category and Table 5.23 shows the respective coefficients and performance indicators. Residuals were mostly centred away from, yet near, zero; their distributions were not symmetric, but as we were mainly concerned with mean values, models seemed appropriate because mean absolute errors were very close to zero in all cases, as shown in Table 5.23. Graphs showing the distributions of residuals by patient subgroup can be found in Appendix 17.

Figure 5.20
Mean OHS by outcome category according to scores at one year after primary THR, reported by EPOS patients

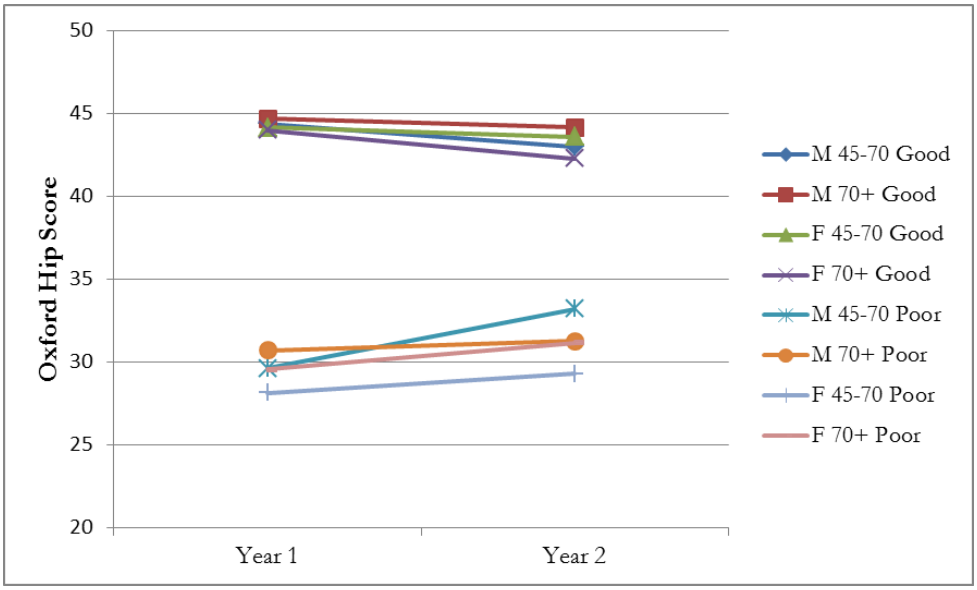


Table 5.23
Models predicting OHS at two years based on OHS one year after primary THR

Outcome category / Patient subgroup	n	Constant	OHS at year 2 Coeff	OHS at year 2 p-value	R^2	RMSE	MAE
Good outcomes							
Males, 45-70 years of age	141	-3.73	1.05	0.000	0.20	6.47	$-5.5e^{-9}$
Males, 70+ years of age	158	19.51	0.55	0.000	0.14	3.93	$5.0e^{-9}$
Females, 45-70 years of age	203	5.70	0.86	0.000	0.21	4.79	$2.1e^{-8}$
Females, 70+ years of age	250	-1.42	0.99	0.000	0.20	6.03	$-9.4e^{-9}$
Poor outcomes							
Males, 45-70 years of age	54	12.83	0.69	0.001	0.18	9.66	$-1.6e^{-8}$
Males, 70+ years of age	36	9.79	0.70	0.018	0.15	8.81	$2.3e^{-8}$
Females, 45-70 years of age	90	10.74	0.66	0.000	0.26	9.03	$-4.6e^{-8}$
Females, 70+ years of age	111	19.08	0.41	0.003	0.08	8.76	$3.0e^{-8}$

Outcome variable = OHS two years after primary THR, RMSE = Root mean square error, MAE = Mean absolute error

The models described in Table 5.23 were used to predict OHS at year two for all records in the HES-PROMs dataset which, together with expected transitions between outcome categories from years one to two after the primary, allowed classifying patients as *Good* or *Poor* outcomes at year two. For the transitions, *Good* outcomes in the first year with the lowest expected OHS at year two were selected to transition into the *Poor* outcome category in year two according to probabilities reported in Table 5.8, Section 5.3.2. Likewise, patients with the highest predicted OHS at year two amongst the *Poor* outcomes at year one were labelled as *Good* outcomes for the second year following the primary. Performing these transitions on the HES-PROMs records was necessary in order to estimate, as accurately as possible, the expected utility scores of primary THR patients two years after primary because these data were not otherwise available. After performing these transitions, the overall split of 65/35 for *Good*/*Poor* outcomes in year one became 79/21 by year two.

Finally, estimated EQ-5D summary scores at year two were multiplied by one to produce associated yearly QALYs as the health utility level was assumed to remain constant over the cycles. Mean overall EQ-5D summary score decreased slightly for *Good* outcome patients between years one and two after the primary, from 0.89 to 0.85, and even less so for *Poor* outcomes, which decreased from 0.57 to 0.56. The slight drop for *Good* outcomes is consistent with EPOS data, but *Poor* outcomes' levels decreasing only slightly is not, as compared to the clear drop shown in Figure 5.19. This is likely due to the combination of several factors, including: predicting OHS at year two with models that did not fit the data perfectly, different mean EQ-5D summary scores by outcome categories between EPOS and HES-PROMs, estimating EQ-5D scores by mapping predicted OHS, and applying patterns observed in EPOS to a HES-PROMs dataset containing a higher proportion of *Poor* outcomes (35% compared to 24% in EPOS). Nevertheless, both EPOS and HES-PROMs report a drop in the proportion of *Poor* outcomes from year one to two (24% to 15% and 35% to 21%, respectively), and HES-PROMs shows resulting mean EQ-5D summary scores that did not change much after transitions, which is a likely scenario in a more representative sample.

Selecting one specific year as a reference for the QALYs associated to years two and onwards after the primary had the added benefit of allowing the use of SDs to estimate parameters for the distributions feeding the PSA. For the probabilistic analysis, we

converted utilities into disutilities to estimate Gamma distribution parameters using the estimated means and SDs. Parameter values are shown in Table 5.24.

Table 5.24
Disutility associated to second and subsequent years after *Primary THR*: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Second and subsequent years after Primary THR in *Poor* outcome				
Males, 45-60 years of age	0.427	Gamma	10.42	0.041
Males, 60-70 years of age	0.398	Gamma	10.23	0.039
Males, 70-80 years of age	0.473	Gamma	15.95	0.030
Males, 80+ years of age	0.471	Gamma	20.02	0.024
Females, 45-60 years of age	0.489	Gamma	14.48	0.034
Females, 60-70 years of age	0.459	Gamma	13.92	0.033
Females, 70-80 years of age	0.405	Gamma	23.18	0.017
Females, 80+ years of age	0.414	Gamma	29.39	0.014
Second and subsequent years after Primary THR *Good* outcome				
Males, 45-60 years of age	0.134	Gamma	1.97	0.068
Males, 60-70 years of age	0.128	Gamma	2.19	0.058
Males, 70-80 years of age	0.139	Gamma	2.12	0.065
Males, 80+ years of age	0.160	Gamma	2.47	0.065
Females, 45-60 years of age	0.139	Gamma	1.75	0.080
Females, 60-70 years of age	0.135	Gamma	1.81	0.074
Females, 70-80 years of age	0.185	Gamma	2.97	0.062
Females, 80+ years of age	0.215	Gamma	4.13	0.052

Given the number of transformations performed on the data, it would have been ideal to incorporate specific parameter uncertainty associated to the mapping exercise. However, this has not been sufficiently studied and no established methodology is available [112]. Siani *et al* have explored applications using analytic and non-parametric bootstrap procedures to incorporate uncertainty originating from the mapping transformations onto resulting confidence intervals of cost-effective and cost-utility analyses [113, 114]. These methods, however, require wider understanding and further validation before wider application can be warranted. The distributions characterised by the parameters in Table 5.24 offer, nonetheless, an important range of variation for health utility estimates associated to the second and following years after *Primary THR*.

5.4.3 QALYs after revision THR

The structure of the economic model after a primary THR is exactly the same as after a revision procedure; therefore, health utility estimates were obtained following an analogous procedure as that followed in the previous section.

For both health states including the revision procedure, QALYs were estimated based on pre- and postoperative EQ-5D summary scores from patients undergoing a revision THR in HES-PROMs combined with the estimated EQ-5D progression by EPOS primary THR patients (reported in the previous section). We used the progression of scores after a primary procedure because no data set was available containing follow-up HRQL measures for revision THR patients before and one year after the operation, as well as at a third point in-between. Table 5.25 shows the means and standard deviations of the EQ-5D summary scores of revision THR patients extracted from the HES-PROMs data set. As with primaries, values do not vary much with age or gender but they show significant differences by outcome category.

Table 5.25
HES-PROMs data: Pre- and postoperative health utility estimates by patient subgroup and outcome category

			Postoperative			
	Preoperative		***Poor* outcome**		***Good* outcome**	
Patient subgroup	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Males, 45-60 years of age	0.338	0.353	0.367	0.284	0.811	0.194
Males, 60-70 years of age	0.355	0.349	0.395	0.290	0.840	0.169
Males, 70-80 years of age	0.406	0.329	0.459	0.276	0.824	0.173
Males, 80+ years of age	0.358	0.319	0.496	0.241	0.790	0.193
Females, 45-60 years of age	0.341	0.347	0.404	0.309	0.798	0.197
Females, 60-70 years of age	0.365	0.343	0.431	0.286	0.819	0.191
Females, 70-80 years of age	0.364	0.329	0.479	0.266	0.804	0.188
Females, 80+ years of age	0.312	0.331	0.490	0.246	0.789	0.185

In order to estimate the QALYs associated to this first year after the revision THR, we connected the start and end points reported in Table 5.25 using the differential progression by outcome category found for primary patients (and illustrated in Figure 5.17 in the previous section). *Poor* outcome patients, therefore, were assumed to reach their postoperative health utility level, as reported by HES-PROMs, by the third month after the operation, whereas *Good* outcomes would attain 73% of their gain by then, and the

rest linearly during the last nine months. We estimated QALYs based on this progression, converted them into disutilities and produced the mean values and Gamma distribution parameters shown in Table 5.26.

Table 5.26
Disutility associated to *Revision THR* and first postoperative year states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Revision THR + first year in *Poor* outcome				
Males, 45-60 years of age	0.655	Gamma	6.24	0.105
Males, 60-70 years of age	0.627	Gamma	5.35	0.117
Males, 70-80 years of age	0.565	Gamma	4.90	0.115
Males, 80+ years of age	0.535	Gamma	5.81	0.092
Females, 45-60 years of age	0.616	Gamma	4.65	0.133
Females, 60-70 years of age	0.592	Gamma	5.06	0.117
Females, 70-80 years of age	0.549	Gamma	4.97	0.110
Females, 80+ years of age	0.543	Gamma	5.75	0.094
Revision THR + first year in *Good* outcome				
Males, 45-60 years of age	0.280	Gamma	2.41	0.116
Males, 60-70 years of age	0.259	Gamma	2.54	0.102
Males, 70-80 years of age	0.262	Gamma	2.50	0.105
Males, 80+ years of age	0.297	Gamma	2.77	0.107
Females, 45-60 years of age	0.294	Gamma	2.51	0.117
Females, 60-70 years of age	0.272	Gamma	2.22	0.122
Females, 70-80 years of age	0.288	Gamma	2.68	0.108
Females, 80+ years of age	0.313	Gamma	3.01	0.104

Once again, we did not consider a utility decrement as a result of the surgery when calculating the above QALY estimates. It is possible that, being a revision procedure, patients suffer greater pain and discomfort after the surgery and that a linear progression may therefore not be an accurate representation of patients' HRQL improvement. *Poor* outcomes see only a slight increase in their EQ-5D summary scores before and after the operation and this may suggest an immediate drop after the intervention and a subsequent improvement only to levels similar to those they had before. Nevertheless, not introducing this hypothetical decrement is supported by the fact that this drop may also be experienced by *Good* outcomes, that patients stay in this health state during only one cycle, and that the possible effects of this unknown decrement are unlikely to alter the overall results of this lifetime model.

The lack of follow-up data on sufficient revision THR patients meant that, for the model states representing the second and subsequent years after their revision operation, we used the patterns of progression observed in patients who underwent a primary. Using postoperative OHS scores reported by HES-PROMs patients who had a revision THR, we estimated OHS scores at year two based on the models described in Table 5.23 of the previous section. Values at two years after the revision were assumed to be representative of all subsequent years for each outcome category. EQ-5D scores were estimated using the same mapping algorithm employed in the previous section, and lowest scoring *Good* outcomes as well as highest scoring *Poor* outcomes transitioned into the other outcome category according to probabilities indicated in Table 5.17 under Section 5.3.4. Estimated EQ-5D summary scores, assumed to remain constant over the year, were converted into disutilities, and their mean values and Gamma distribution parameters estimated. These values are shown in Table 5.27.

Table 5.27
Disutility associated to second and subsequent years after *Revision THR*: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Second and subsequent years after Revision THR in *Poor* outcome				
Males, 45-60 years of age	0.531	Gamma	23.38	0.023
Males, 60-70 years of age	0.517	Gamma	23.20	0.022
Males, 70-80 years of age	0.541	Gamma	15.96	0.034
Males, 80+ years of age	0.529	Gamma	19.80	0.027
Females, 45-60 years of age	0.551	Gamma	20.05	0.027
Females, 60-70 years of age	0.529	Gamma	19.53	0.027
Females, 70-80 years of age	0.462	Gamma	45.08	0.010
Females, 80+ years of age	0.454	Gamma	49.52	0.009
Second and subsequent years after Revision THR *Good* outcome				
Males, 45-60 years of age	0.217	Gamma	3.17	0.069
Males, 60-70 years of age	0.209	Gamma	3.17	0.066
Males, 70-80 years of age	0.190	Gamma	2.40	0.079
Males, 80+ years of age	0.214	Gamma	3.16	0.068
Females, 45-60 years of age	0.209	Gamma	2.49	0.084
Females, 60-70 years of age	0.196	Gamma	2.50	0.079
Females, 70-80 years of age	0.238	Gamma	4.06	0.059
Females, 80+ years of age	0.255	Gamma	5.33	0.048

As with primaries, *Poor* outcomes after revision THR became less common, dropping from 46% one year after surgery to an estimated 33% at year two. Despite the

transformations performed based on evidence from primary procedures, this may well be an accurate reflection of the improvement experienced by patients who take longer a time to recover from the more complex revision surgery. Summary EQ-5D scores for *Good* outcomes dropped from 0.81 to an estimated 0.79, in line with evidence from EPOS primary THR patients shown in Figure 5.19. *Poor* outcomes show an increase from 0.44 to 0.49, contrary to the drop reported by primary EPOS patients, but likely a direct result of fitting a prediction model for OHS at year two based on primary THR patients in EPOS with higher scores than those reported by revision THR patients in HES-PROMs. Appendix 18 reports the mean QALY value associated to each model state. PSA helped incorporate uncertainty in these estimated values into final model results.

5.5 Direct costs

The model was populated with data on the costs associated to each health state by combining measures of resource use with their respective costs. Resource use was obtained primarily from the CPRD. As the analysis was performed from the perspective of the NHS, data on prices are those reported in the most recent Department of Health's publication of reference costs [115] for in-patient events, the Personal Social Services Research Unit's (PSSRU) unit costs of health and social care [116] for primary care, and the British National Formulary (BNF) [117] for drug prices. NHS reference costs correspond to the period 2011-2012 [118]; PSSRU's unit costs are based on the period 2010-2011; and the online version of the BNF was last updated in November 2011. All unit costs are therefore the most recently available and all in 2011 pounds sterling.

The CPRD is a very large database containing primary care data on approximately 4.8 million patients from about 600 GP practices in the UK. It is considered representative of the general population and holds data collected since the late 1980's until present time. GP practices provide the CPRD with anonymized data such as consultations, prescriptions, test results, referrals, measurements of height and weight, and smoking habits [119]. Extracts from this database have been used as the main source of input to estimate primary care costs for many economic evaluations. However, publications rarely explain in detail how consultations, for instance, can be attributed to the medical condition being assessed. Lafuma and Berdeaux [120, 121], for example, looked at glaucoma treatment in the old GPRD in two separate studies selecting patients and consultations based on GP visits with simultaneous glaucoma-related referral, diagnosis or prescription. The authors might have underestimated resource use if patients visited the GP for a glaucoma-related problem without the GP necessarily recording a diagnosis,

making a referral or prescribing a test or drug. Moore *et al* [122] also used the GPRD to estimate not only costs but also transition probabilities associated to herpes zoster and post-herpetic neuralgia. For resource utilisation, many authors seem to only consider related GP consultations and referrals, and although they assume that GPs perform the diagnosis and treatment, they do not specify whether they assume all GP consultations to be associated to the condition, or the criteria to discern this otherwise. Violato *et al* [123] used a different approach by estimating resource use and costs associated to coeliac disease from the GPRD through the comparison of cases and controls. They looked at all consultations and prescriptions recorded for both cases and controls, calculated mean values for each, and subtracted them to produce an estimate of resource use attributable to coeliac disease. The authors reported the difference of the means as well as a confidence interval for it, although no details were provided as to how this confidence interval was calculated without a patient-level analysis.

Although faced with a similar situation as that of Violato *et al* [123] in that we also looked at the comparison between cases and controls to estimate the level of resource use attributable to a specific condition, we followed a different approach guided by our interest in producing a measure of uncertainty based on observed data. We calculated the mean quantity of each resource used by sets of controls and subtracted this from the level reported by corresponding cases. We interpreted this difference as the amount of resources used by each hip pain patient in excess of what their controls, on average, demanded from the health care system. The overall mean of these differences was then an estimate of the resource use attributable to the hip problem. Although not a common approach likely because the weight of each control in the overall mean estimate varies with the number of controls for each case, critically it allowed obtaining an estimate of variability from the observed resource use attributable to hip pain, which the method by Violatto *et al* [123] did not provide. Appendix 19 further discusses the difference between both methods based on comparative results on preoperative consultation data.

The extract of the CPRD data set employed for our analysis identified controls by matching gender, GP practice and age (+/- 5 years) to each case. The specific criteria for cases and controls varied depending on the model stage and are reported in the corresponding section. Matched cases and set of controls were therefore largely equivalent, thus eliminating the need to adjust for confounding factors based on demographic or socio-economic differences. Clinical factors, however, were expected to

vary between cases and controls. We relied on the large number of observations to balance out the differences in comorbidities, so that the effect of controls with more comorbidities and hence greater resource use than cases would be offset by the effect of those with fewer comorbidities and therefore less use of resources. When subtracting the mean resource use of controls from the values reported for their respective cases, individual sets might be biased by the differences in comorbidities but the overall mean of the difference could therefore be considered an unbiased estimator of mean resource use attributable to the hip problem.

There are many benefits of large observational datasets such as the CPRD. The collection of data is generally non-intrusive, it takes place continuously thus providing time-series capabilities, and it is subject to important quality checks. Coverage tends to be high, with considerable potential for linkage and, crucially in our case, both cases and controls can be selected post hoc [124]. However, administrative databases also have their limitations, such as missing data and the fact that data have already been collected. For our analysis, the most relevant limitation is that the CPRD lacks outcomes data, which is a necessary element to be able to estimate resource use by THR patients postoperatively according to their outcome category. To overcome this obstacle, we estimated a model predicting outcome category based on resource use. The model was estimated on the first wave of data from the COASt cohort study and applied to CPRD records in order to estimate their outcome category.

This section is divided, as the previous one, into three. First, we present our findings of resource use and costs associated to all preoperative health states in Section 5.5.1. Under Section 5.5.2 we describe in detail the model predicting outcome category based on resource use and show results of costs associated to health states following a primary THR. Finally, resource use and costs associated to the health care of patients after a revision procedure are reported in Section 5.5.3.

5.5.1 Preoperative costs

For the costs associated to all health states previous to a THR, we used a CPRD extract consisting of the records of all patients with medical diagnosis code for hip arthroplasty performed before 31 December 2006 who were 45 years of age or older at the time of the operation. For each case, the data set included up to five controls without any clinical or referral record for hip arthroplasty, OA or arthritis, ever, and matched to a case by GP

practice, gender and age. The extract was of high quality data as the cohort was limited to patients deemed acceptable based on a standard set of conditions about registration details established by the CPRD.

Records about both consultations and prescriptions were considered. As the CPRD stores a consultation as an event performed by a specific GP practice staff, we included only events performed by health care professionals such as GPs, nurses, physiotherapists or alternative practitioners who would be directly involved in providing care to patients. A complete list of the health care professionals undertaking events included in the analysis, as well as those whose events were excluded, is shown in Appendix 20. But not all events by the above health care professionals were of interest. The data sent by GP practices to the CPRD include events such as "results recording", "administration", and "mail" to or from the patient, for example, which were not relevant for our analysis. "Surgery consultations", "follow-up visits", or even "phone calls" and "night visits" captured, on the other hand, the use of resources we were interested in. Appendix 21 shows the type of events included in the analysis and groups them into "Day visits", "Night visits" and "Telephone calls"; excluded events are also listed.

To capture the use of medication, we searched the CPRD extract for all prescriptions given to patients for drugs related to their hip pain. The list of medication was based on the responses of the first 314 patients in the COASt cohort of THR patients whose one-year follow-up forms had been returned and analysed by the end of 2012. These patients were asked to list the medication they were taking at that moment and the reason for taking them. Based on their answers and after verification by an experienced GP, a final list of 25 different drugs were identified and grouped into six categories: antidepressants, NSAIDs, opioid analgesics, non-opioid analgesics, laxatives, and ulcer prevention medication. Amitriptyline was the only drug in the antidepressants group and was included because it was reported to be used by many patients for pain relief. Paracetamol was the only medication in the non-opioid analgesics category and one of the most commonly reported drugs taken by COASt patients. Laxatives and ulcer prevention drugs were included because they prevent or alleviate side effects directly associated with opioid analgesics and NSAIDs, respectively. The list of specific drugs included in the analysis is shown in Appendix 22. When searching for these drugs in the CPRD extract, we included all available presentations.

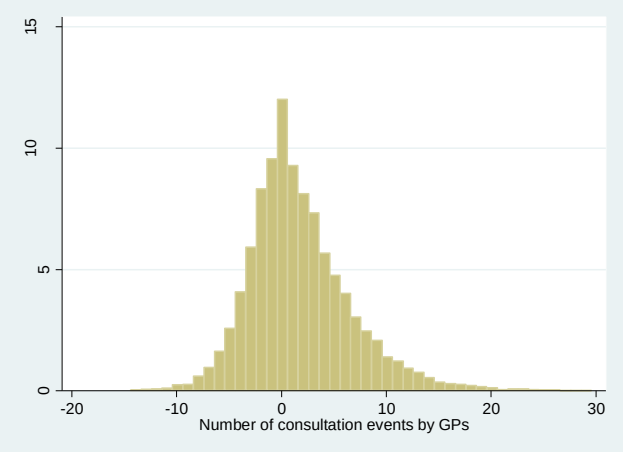
Consultations and prescriptions were considered for patients as long as these were linked to the CPRD for more than six months. Patients enter the database when their GP practice joins the CPRD or when the patient registers with a GP practice that participates in the scheme. They leave the database when they change to a GP practice that is not part of the scheme or when they die. As a consequence, we had data on consultations and prescriptions for periods ranging from as short as a few days to up to 15 years before the THR, although rarely for this long. We added the number of consultations and prescriptions by patient in each year as long as both the case and at least one control were in the database for six months or more during that year. Patients who had more than one primary THR were also excluded because health care resources could have been used for either of the operated hips and there was no way to make the distinction as laterality is not recorded in the CPRD.

The mean number of consultations for each set of controls was calculated and subtracted from the number reported for their respective case. This was done for each health care professional and type of event, and for each year. The difference was assumed to be an estimate of the number of consultations the case had with the specific health care professional due to his or her hip problem. This assumption was based on the fact that controls were matched to cases by GP practice, gender and age and therefore they were expected to have similar socio-economic status, general health and comorbidities. We therefore observed the THR-related and non-THR-related costs of caring for THR patients (cases), and subtracted from that the non-THR-related costs of caring for comparable patients (controls), thus obtaining an estimate of only THR-related costs. Figure 5.21 shows the distribution of the number of day visits to GPs attributable to the hip pain by the 21,572 cases included in the data set during the year prior to their operation.

THR patients used, on average, an estimated 1.8 extra GP day visits during the year immediately prior to their operation compared to similar patients who did not have their hip replaced, suffer from OA or even hip pain. As Figure 5.21 shows, however, for many of these patients the number of extra visits to their GP was much higher, whilst for others it was considerably lower. This was a natural and expected result. Our assumption that comorbidities would balance between the two groups (cases and controls) refers to the aggregate level, given the matching process and the large number of observations. For individual sets of matched patients, however, no such balance was expected, as is

confirmed by our results. The difference between cases and controls was not only the fact that the former had OA and the latter did not; other conditions must have been present in different rates between the two groups. Figure 5.21 reflects this by showing that some future THR patients saw their GP as many as 20 more times than their controls, in average, the year before the operation, likely because they had other comorbidities requiring additional day visits. The same happened with controls who reported seeing their GP many more times than their respective case, presumably about other conditions, thus generating negative additional GP visits by THR patients the year before the operation.

Figure 5.21
Estimated GP day visits due to hip pain by THR patients during the year prior to surgery, patients older than 45 years in the CPRD



A perfect balance of comorbidities in both cases and controls would be represented by a perfectly symmetric distribution. The distribution of additional GP day visits presented in Figure 5.21 is nearly but not perfectly symmetric, confirmed by a measure of skewness of 1.16. A more symmetric distribution would have been generated if we had not reduced the records of the controls for each case to one mean value, but we considered producing a single average control for each case more important than the likely minor effects of slightly less balanced comorbidities.

This analysis was performed for all health care staff and events, and for the 15 years for which we had records of cases and controls. To estimate costs associated to hip pain, we multiplied the estimated number of day visits, night visits and telephone calls made by each health professional and attributable to the condition by its mean unit cost as reported in the PPSRU [116]. Appendix 23 shows the costs associated to each type of event per health care staff.

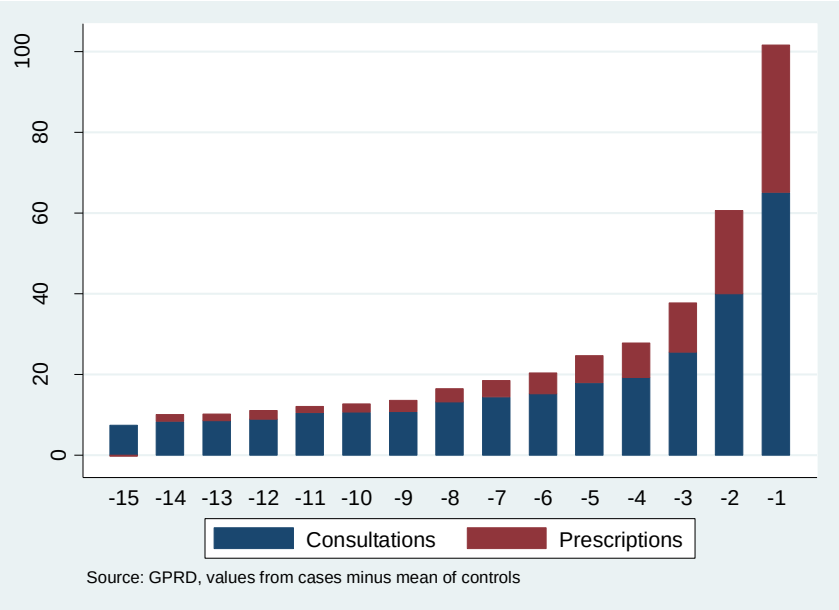
The use of medication associated to hip pain was estimated in the same manner as consultations. The mean number of drug units (tablets, capsules, oral solutions) prescribed to controls was subtracted from the value reported for the respective case. The data on prescriptions found in the CPRD required deeper cleaning as the key piece of data for each record was the number of units prescribed, and this information was sometimes unreasonably high or simply missing. Of a total of 1,034,319 prescription records for all patients and years, only 14 had missing quantity and daily dose which were therefore dropped. Almost 9% of the remaining records had missing daily dose, which was important to determine whether the number of units prescribed would fall under a clinically plausible range. We assigned the mode daily dose indicated for patients who were prescribed the same number of units or the same drug presentation to those missing the daily dose. If neither of these were available, the BNF's recommended daily dose was applied [117]. Around 0.1% of records had daily dose but were missing the quantity prescribed, and for those few we also assigned the mode number of units reported for patients who were prescribed the same drug at the same dose, or the same presentation otherwise. Once all prescription records had an associated quantity and daily dose, we identified 1,541 (0.15%) records associated to a prescription time longer than six months or to more than one litre of oral solution. Those records were dropped.

Mean prescriptions units were subsequently calculated for each set of controls and this value subtracted from that of the case to obtain an estimate of the number of prescriptions associated to hip pain. For most drugs, the resulting difference was zero for the majority of patients in all preoperative years. This was largely due to the fact that, in the aggregate of all years, 54% of the cases with any record of consultation had no counterpart record in the prescriptions database for the 25 drugs we searched the CPRD for. For example, 42% of cases reported no prescriptions for tablets of paracetamol during the year prior to their operation, and neither did their controls. The majority of those cases did not have any prescriptions for other presentations of paracetamol or any the other drugs included in the analysis. All non-zero differences represented therefore the use of medication attributable to the hip pain, and in the case of paracetamol tablets during the year prior to surgery, our 21,572 cases were prescribed a mean of 119 units more than their controls. This distribution was also nearly but not perfectly symmetric with a measure of skewness of 1.5. As in the case of consultations, there were patients in the THR waiting list who were prescribed many more tablets than the mean of 119, and there were also many controls that, on average, were prescribed more tablets than the

patients awaiting the operation. This reflects the wide variation of comorbidities amongst patients included in the sample, which also validates the mean number of prescriptions in this sample as an unbiased estimator of the mean use of this particular resource because of the hip pain experienced by patients who had a THR in the following year.

To estimate medication costs, the number of units attributable to hip pain was estimated as described above for the 272 different CPRD codes associated to the 25 included drugs, and then multiplied by their unit cost as reported in the BNF [117]. Appendix 24 shows the unit cost for all drugs by CPRD code. Consultation and medication costs were then added together and thus the progression of estimated total costs attributable to hip pain used by patients during the 15 years prior to their THR could be produced, which is shown in Figure 5.22. The growing estimated costs confirmed the increasing burden generated by unresolved hip pain and problems experienced by patients who are referred for a THR, who markedly demand many more health care resources during the year immediately prior to their operation. It is also notable that the relative weight of prescription as a portion of total costs increase much more rapidly than that of consultations as patients approach a THR, going from 20% to 27% to 36% at eight, five and one year before surgery, respectively. This indicates that patients are not able to successfully manage their pain through more consultations with health care professionals and have to recourse to more medication until their hip is replaced.

Figure 5.22
Mean cost of consultations and prescriptions due to hip problem, by year before THR



Although the economic model has two distinct branches for patients who are not referred directly for surgery, one leading to *Risk-factor management* and the other to *Long-term medical management*, we used the estimates for the year immediately prior to a THR to populate all preoperative states. Estimates for this year were chosen to represent the costs of patients directed towards *Risk-factor management* because the majority of them would ultimately be referred for surgery (see transition probabilities in Section 5.2.8) and all additional consultations and prescriptions provided in preparation for the operation would apply. The states associated to *Long-term medical management*, however, host patients in the model who are more likely not to be referred for a surgical intervention. All extra consultations and drugs associated to an imminent operation would not apply, yet patients would generally continue to experience pain and difficulties for which they will demand additional consultations with GPs, nurses, physiotherapists or other health care professionals, and even more medication to manage their pain. In the absence of data to reflect such resource use increase, we also applied the estimates for the year immediately prior to a THR to patients in the *Long-term medical management* branch under the assumption that those estimates are an approximation to the primary care cost of hip pain patients who are not referred for a THR over the years.

During the year prior to the operation, costs associated to primary care consultations are categorically dominated by visits to the GP, whereas the greatest portion of prescriptions costs go to cover NSAIDs and paracetamol. Tables 5.28 and 5.29, respectively, illustrate this for the subgroup of females aged between 70 and 80, with relative weights by component being essentially the same for all patient subgroups. Appendix 25 reports on the mean number of day visits, night visits and phone calls by each health care staff found to be attributable to hip pain, shown by patient subgroup. The mean and standard deviations of the costs associated to these consultations events summarised by health care staff and for each patient subgroup are presented in Appendix 26. The cost attributable to hip pain and associated to each individual drug included in the analysis is reported in Appendix 27, also by patient subgroup.

Deterministic analysis was hence based on the above mean values for costs attributable to hip problems. For PSA, distributions reflecting the uncertainty around costs are generally of the Gamma or Log Normal type given that costs are commonly presented as right-skewed non-negative distributions [82]. In our case, however, we have identified not plain costs but the difference in resource use between two groups to ascertain the costs

attributable to hip pain. The resulting distribution is highly normal, as expected given that it is a difference. Figure 5.23 shows the distribution of total costs attributable to hip pain for females between the ages of 70 and 80 as this was the largest patient subgroup in the preoperative CPRD sample (24% of cases), and the normal distribution produced based on the observed mean and SD. Normal distributions were hence used to model the uncertainty around preoperative costs as they were the best fit for the data used to estimate the costs specifically associated to hip problems with an important degree of variation reflecting the wide spectrum of comorbidities of typically elderly patients. Moreover, CEACs ultimately reflect the variation in ICERs given uncertainty in all parameters, and whose reliability is directly linked to the selection of distributions accurately describing the uncertainty around individual parameters, which in the case of preoperative costs, is undoubtedly a normal distribution.

Table 5.28
Components of the mean cost of consultations due to hip problem during the year immediately prior to THR, Females 70-80 years of age

Staff role	Mean	SD	% of total
GPs	£58.20	£173.36	96%
Health visitor	£0.37	£17.86	1%
Physiotherapist	£1.04	£22.17	2%
Practice nurse	£0.62	£47.02	1%
Others	£0.30		0%
TOTAL	**£60.54**		**100%**

Table 5.29
Components of the mean cost of prescriptions due to hip problem during the year immediately prior to THR, Females 70-80 years of age

Medication category	Mean	% of total
NSAIDs	£14.72	40%
Non-opioid (Paracetamol)	£14.71	40%
Opioid	£9.01	24%
Anti-depressants	£0.00	0%
Laxatives	£-0.47	-1%
Ulcer prevention	£-1.14	-3%
TOTAL	**£36.84**	**100%**

Table 5.30 shows the mean and SD of the primary care costs assigned to all preoperative states of the model. For those states involving an assessment by the surgeon, £105 was added as this was the average unit cost reported in the National Schedule of Reference

Costs for year 2011-2012 for an outpatient orthopaedics consultation at NHS Trusts and Foundation Trusts [118].

Figure 5.23
Total cost attributable to hip problem for female patients aged 70 to 80 during the year immediately prior to THR and normal density curve

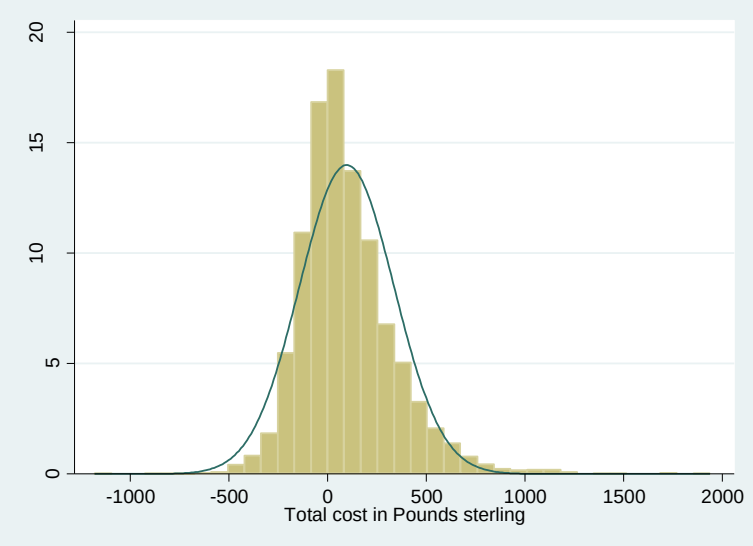


Table 5.30
Primary care costs associated to all preoperative states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	SD	Distribution
Surgical assessment *,			
Risk-factor modification,			
Reassessment after Risk-factor modification *,			
Long-term medical management, and			
Reassessment after Long-term medical management *			
Males, 45-60 years of age	£98.0	£200.35	Normal
Males, 60-70 years of age	£98.2	£209.08	Normal
Males, 70-80 years of age	£87.6	£219.69	Normal
Males, 80+ years of age	£101.1	£222.5	Normal
Females, 45-60 years of age	£121.7	£216.32	Normal
Females, 60-70 years of age	£118.9	£237.33	Normal
Females, 70-80 years of age	£97.4	£240.31	Normal
Females, 80+ years of age	£92.0	£257.74	Normal

* The cost of the consultation with the orthopaedic surgeon must be added (£105) to obtain total cost for the state.

5.5.2 Primary and postoperative costs, and a model mapping resource use to surgery outcome category

If and when patients had one of their hips replaced, a new set of costs applied in the model. A cost was associated to the surgery itself, and then there were postoperative primary care costs that, we argue, vary depending on surgery outcome. In this section, we estimate the costs for the model health states encompassing primary THR followed by a first year either in the *Good* or *Poor* outcome category, as well as the primary care costs associated to *Good* or *Poor* outcomes in the second and subsequent years with the unrevised prosthesis. For the health states including the operation we applied NHS reference costs corresponding to the HRG groups associated to each patient subgroup regardless of outcome category. For postoperative primary care costs, we used CPRD records and reference unit costs as described in the previous section reported by patients after they had a THR. However, since the CPRD does not collect data on outcomes, we developed a model to predict surgery outcome category from the use of health care resources recorded during the first year after the operation based on data collected from the COASt cohort. Using this model, we differentiated patients in the CPRD extract by outcome category which allowed us to provide estimates of primary care costs for *Good* and *Poor* outcome categories separately.

Although we held HES data for all THRs performed in the NHS between 2009 and 2012, this could not be used to estimate the cost of surgery by patient subgroup because reference costs are obtained from Healthcare Resource Groups (HRGs) whereas HES only reports OPCS codes. Instead of calculating a national average cost for all THR patients weighted by the activity numbers reported by the NHS in any particular year for those HRGs apparently associated to a hip replacement, we calculated mean values based on actual HRG assignment and separately for each patient subgroup as defined here. This method was chosen because it considered HRGs actually applied and therefore the corresponding amounts effectively paid by the NHS for THRs, and it also respected the differences amongst costs by patient subgroups according to age and gender. In order to do this, we obtained the table of relative frequencies of HRGs assigned by the Payment by Results (PbR) system in the NHS to each THR reported in HES and eligible for PROMs for fiscal year 2011-2012 [125]. Based on these frequencies and the NHS reference costs by HRGs for the same year [126], we calculated the mean cost for a primary THR by patient subgroup (shown in Table 5.31). Appendix 28 reports the national average unit cost for all relevant HRG groups for the year 2011-2012 paid to

NHS Trusts and Foundation Trusts, and Appendix 29 shows the relative frequency breakdown of HRGs by patient subgroup for HES records in the same fiscal year.

Table 5.31
Cost of a primary THR to the NHS by patient subgroup

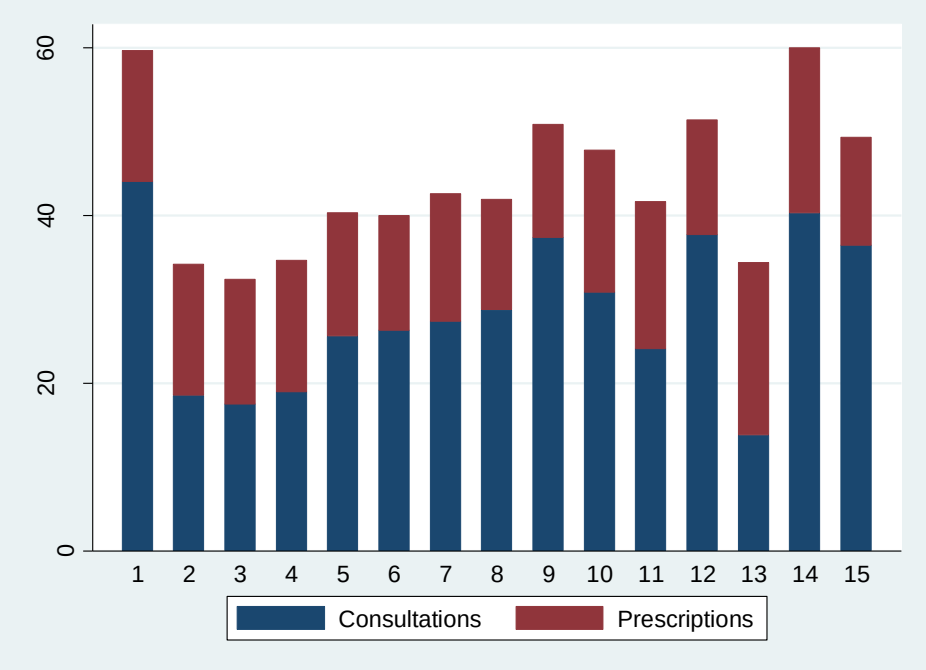
Patient subgroup	n	Mean	SD
Males, 45-60 years of age	4,696	£6,069	£249.4
Males, 60-70 years of age	7,632	£6,102	£233.5
Males, 70-80 years of age	7,948	£6,186	£182.0
Males, 80+ years of age	2,578	£6,352	£143.0
Females, 45-60 years of age	5,121	£6,063	£262.4
Females, 60-70 years of age	10,164	£6,083	£208.6
Females, 70-80 years of age	12,838	£6,139	£111.7
Females, 80+ years of age	6,043	£6,307	£104.9

As Table 5.31 shows, the mean cost of a primary THR to the NHS varies between about £6,000 for those patients aged 45 to 60 years, to around £6,300 for those over 85. We found only a slight variation between genders. For all patient subgroups, the most common HRG assigned to the operation was code HB12C corresponding to "Major hip procedures for non-trauma Category 1 without complications", which was reported in 75% to 90% of cases within patient subgroups. As a result, the standard deviation for surgery costs within subgroups was very small (between 2% and 4% of the mean value), hence we did not model the uncertainty around this parameter value and simply added the reported mean to the first year of postoperative primary care costs to produce the aggregate cost associated to model states combining primary THR and the first 12 months after surgery.

To estimate postoperative primary health care costs, we followed the same process described in the previous section. We used the same high-quality CPRD extract with records of all patients who had a hip arthroplasty before the end of 2006 as well as up to five controls for each who did not have OA or an arthroplasty, but this time we looked at the resource use after the operation. Following the same data cleaning criteria applied to preoperative data, we obtained resource use measures for consultations with the health care professionals listed in Appendix 20 for all relevant events (as reported in Appendix 21), as well as medication use for the 272 presentations of the 25 drugs listed in Appendix 22. We calculated mean values for each set of controls and subtracted this from the values reported by the respective case to produce an estimate of the resource use due to the hip

problem reported by THR patients. We then applied reference unit costs for consultations and drugs and obtained the mean total and component costs attributable to the hip problem shown in Figure 5.24.

Figure 5.24
Mean cost of consultations and prescriptions due to hip problem, CPRD primary THR patients by year after surgery



As Figure 5.24 shows, there is a peak of costs at £60 on average during the year immediately following surgery due to an expected high level of follow-up visits with the GP. This drops to slightly above £30 on average by the second year. The medication component appears relatively stable during the 10 years following surgery whilst consultations, primarily with the GP, drive the variation in total costs over time. It is worth noting that the average primary care costs reported during the 15 years following surgery are lower than the £100 calculated for the first year before surgery and, with the exception of year 14, even lower than the £60 estimated for the second year before the operation.

These cost estimates, however, pool together the use of health care resources by many patients who had a good or excellent outcome and likely did not make many visits to their GPs or take much medication for their hip pain, with the records of those patients who were not satisfied with the results of their surgery, still experienced pain and limitations, and therefore used health care resources a great deal more. Unfortunately, the CPRD does

not record outcome measures, hence patients could not be classified into *Good* or *Poor* outcomes as we did with patients who completed the postoperative OHS questionnaire as part of the HES-PROMs initiative. To produce separate estimates by outcome category, we developed a model to predict surgery outcome based on healthcare resource use during the first year after the primary. The model was estimated on observed data collected from the 314 patients in the COASt cohort who completed the OHS one year after the operation and from whom resource use, including the list of medication used in the cost analysis, was also collected.

As Table 5.32 shows, *Poor* outcome patients in the COASt cohort, who are unsurprisingly less satisfied with the outcome of the operation than *Good* outcome patients, indeed visited their GP, nurse and physiotherapist more than the latter group, and they also took more pain medication. Regarding visits to GPs specifically, Figure 5.25 shows that many *Poor* outcome patients did not visit their GP at all during the 12 months following surgery because of their hip, but this was more markedly the case within the group of patients scoring above 38 in their postoperative OHS. Whilst 40% of *Poor* outcome patients in the COASt cohort did not go to see their GP for problems related to their hip, for *Good* outcomes this proportion was 80%. Also, a few *Poor* outcome patients saw their GP 10 or more times because of their hip during the year following surgery, whereas within the *Good* outcomes no patient visited their GP that many times for problems related to their operated hip.

The relationship between use of resources and surgery outcome is very clear. Patients who still experience pain, limitations and difficulties performing ordinary activities after their THR would naturally visit their GP and possibly other health care professionals, and they would also take more pain medications than patients who have hardly or no pain at all and who regain most or all of their mobility. Dissatisfaction would also be expected to increase the number of visits to the GP. As Table 5.32 shows, whilst almost all *Good* outcome patients in the COASt cohort were satisfied, only half of *Poor* outcome patients were. A model predicting outcome category based on resource use was hence feasible.

We estimated a logit model to predict *Poor* outcome as defined in this study, i.e. scoring less than 38 in the OHS at one year after the primary. All resource use variables in the one-year postoperative follow-up form used in the COASt cohort were originally included in the model, together with age and gender. The latter two as well as nurse and

physiotherapy visits, and drugs other than opioids and paracetamol were not statistically significant predictors of *Poor* outcome. As Table 5.33 shows, a model explaining 24% of the variance of outcome category was estimated from a three-level categorical variable counting GP visits (zero being the base case), whether patients were taking paracetamol or not, and the number of opioid drugs taken. As we lacked an external dataset for validation, we fitted the model to the same estimation dataset obtaining the ROC curve shown in Figure 5.26 which reported an area under the curve of 0.80. At certain cut-off points, the model was able to predict between 70% and 80% of both *Good* and *Poor* outcomes correctly.

Table 5.32
Use of resources by outcome category, COASt cohort patients

	All patients	Good outcome	Poor outcome	Missing OHS
Patients	329 (100%)	276 (84%)	38 (12%)	15 (5%)
Female	198 (60%)	166 (60%)	22 (58%)	10 (71%)
Age: mean (SD)	68 (10.4)	68 (10.3)	71 (8.0)	67 (15.8)
BMI: mean (SD)	28 (4.9)	28 (4.9)	30 (4.7)	27 (4.7)
OHS: mean (SD)	42 (8.0)	44 (4.1)	24 (6.1)	-
EQ-5D: mean (SD)	0.82 (0.253)	0.88 (0.185)	0.40 (0.288)	0.75 (0.238)
Satisfied with outcome	298 (92%)	266 (96%)	19 (50%)	13 (87%)
Do not smoke	307 (94%)	258 (93%)	35 (92%)	14 (93%)
Visits to GP ≥ 2	46 (14%)	28 (10%)	16 (42%)	2 (13%)
Visits to NHS physiotherapist ≥ 1	89 (27%)	66 (24%)	17 (45%)	6 (40%)
Visits to NHS nurse ≥ 1	32 (10%)	24 (9%)	6 (16%)	2 (13%)
Taking any non-opioid drugs	50 (15%)	28 (10%)	17 (45%)	5 (33%)
Taking any NSAIDs	39 (12%)	28 (10%)	8 (21%)	3 (20%)
Taking any opioid drugs	59 (18%)	39 (14%)	15 (39%)	5 (33%)

Since the predictors of *Poor* surgery outcome were measures of resource use also available in the CPRD, we fitted the above model to CPRD's postoperative data to predict the outcome category that patients have most likely have been classified into based on their patterns of resource use. We fitted the model to data from the first postoperative year after the primary because this, combined with the cost of surgery previously reported, produced overall costs associated to each model state covering primary THR and the first year in either outcome category.

Figure 5.25
Number of visits to GPs by outcome category, all COASt patients

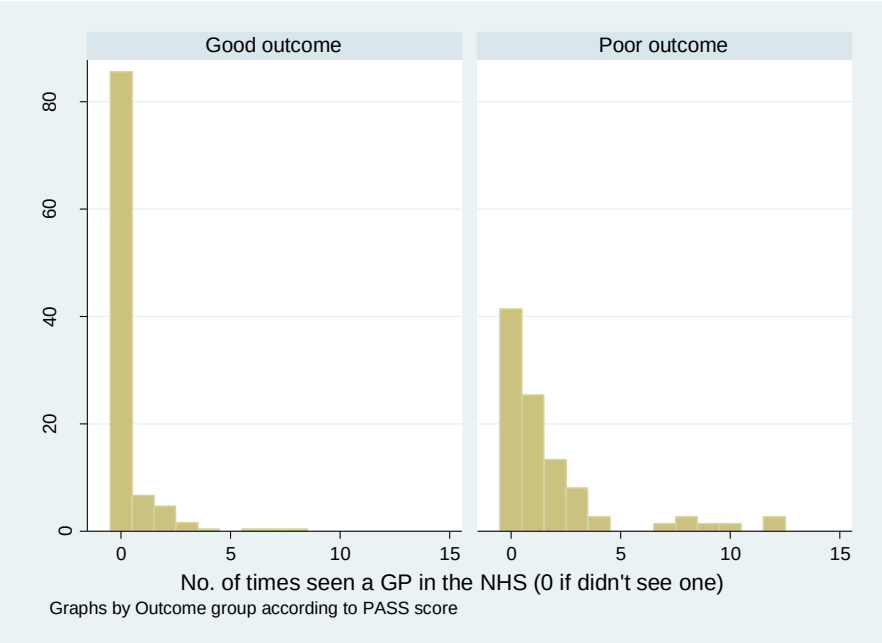


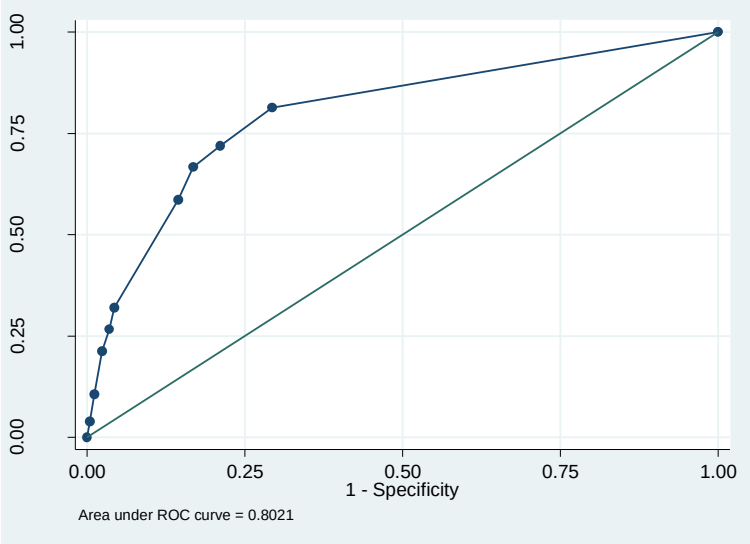
Table 5.33
Surgery outcome predictive model at one year: Logit regression for Poor outcome, estimated on COASt cohort data

Predictor	Coefficient	p-value	95% confidence interval	
Number GP visits = 1 to 4	2.120	0.000	1.446	2.794
Number GP visits = 5 or more	2.468	0.003	0.851	4.085
Paracetamol	1.062	0.010	0.256	1.868
Number of opioid drugs	1.113	0.002	0.421	1.804
Constant	-2.569	0.000	-3.066	-2.071
n = 314				
Pseudo R2 = 0.241				

Predictors in the CPRD were equivalent to those used to estimate the model in COASt. The model was estimated based on the number of visits to the GP specifically because of problems with the hip and, even though the CPRD collects the number of GP visits regardless of reason, we used the reported number of visits after subtracting the mean of controls as an approximation for visits due to hip problems. When fitting the model, we used the combined number of consultations, whether at surgery or night visits. Any presentation of paracetamol was regarded as valid for the binary predictor, and the number of opioid drugs was a straight forward count, also regardless of presentation,

from any of the drugs included in the analysis and classified as opioid (shown in Appendix 22).

Figure 5.26
ROC curve for the model predicting surgery outcome category



Those patients for whom the model estimated a probability greater than 0.4 of being classified as *Poor* outcome were considered *likely Poor* outcomes, the rest were labelled as *likely Good* outcomes. Appendix 30 expands on the rationale for using 0.4 as the probability cut-off point. As a result, we were able to obtain estimates for primary care costs during the first year after a primary THR separately for each outcome category. Whilst Figure 5.24 showed a mean cost of £60 for postoperative primary care of all THR patients during the first year after the operation, the classification of the same patients by outcome group using the model led to an estimated mean of £280 for *likely Poor* outcome patients and £34 cheaper than controls for *likely Good* outcomes, as shown in Figure 5.27.

The distributions of costs for both groups were not strictly symmetric yet they were close to normal with a significant overlap, as shown in Figure 5.28. This overlap, concentrated between negative £200 and positive £500, is important because it confirmed that of all patients with total primary care costs in that range, some were labelled as *likely Good* outcomes and some others as *likely Poor* outcomes. In other words, the classification performed by the model was not equivalent to applying a threshold to costs; it used statistically significant resource use predictors of surgery outcome and assigned probabilities of being *Poor* outcome to patients that, although in most cases reported high costs, in some cases had low costs or even lower than their controls. Appendices 31 through 36 show the number of consultations by staff as well as consultation and

prescription costs by patient subgroup separately for likely *Poor* and *Good* outcomes as predicted by the model.

Figure 5.27
Mean cost of consultations and prescriptions due to hip problem during first year after THR by outcome category, CPRD all patients

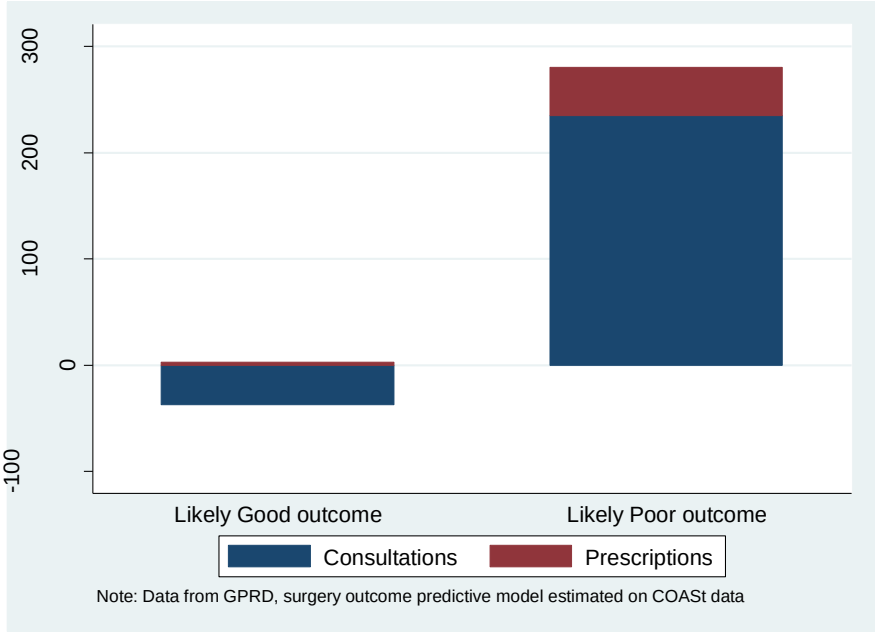
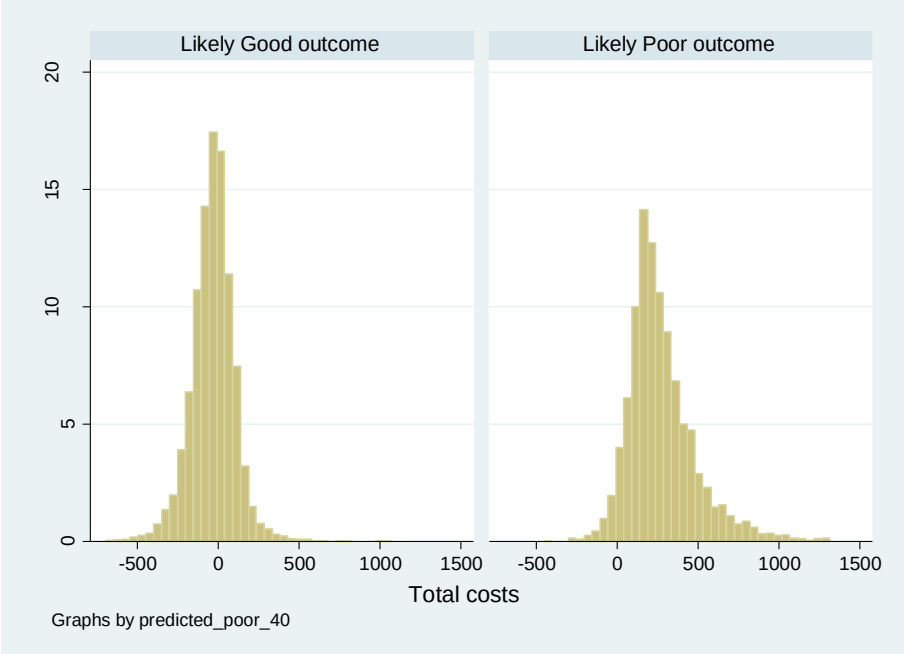


Figure 5.28
Distribution of total primary care costs due to hip problem during first year after THR by outcome category, CPRD all patients



The mean values and distribution of costs for the first year after the operation added to the mean costs of surgery presented in Table 5.31 produced the total costs associated to the model states combining the primary THR and the first postoperative year in *Good* or *Poor* outcome. These are shown in Table 5.34 with parameter distributions set to normal, as with preoperative costs, and SD reflecting only uncertainty around primary care costs because costs of surgery hardly varied at all. Since THR costs were estimated regardless of surgery outcome, the higher overall mean costs of *Poor* outcomes are explained by their higher primary care costs.

Table 5.34
Costs associated to Primary THR followed by first year in either outcome category states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	SD	Distribution
Primary THR + first year in *Good* outcome			
Males, 45-60 years of age	£6,070	£123.5	Normal
Males, 60-70 years of age	£6,082	£120.4	Normal
Males, 70-80 years of age	£6,143	£134.9	Normal
Males, 80+ years of age	£6,320	£149.3	Normal
Females, 45-60 years of age	£6,049	£125.6	Normal
Females, 60-70 years of age	£6,054	£140.1	Normal
Females, 70-80 years of age	£6,096	£139.6	Normal
Females, 80+ years of age	£6,250	£153.3	Normal
Primary THR + first year in *Poor* outcome			
Males, 45-60 years of age	£6,352	£215.2	Normal
Males, 60-70 years of age	£6,379	£227.3	Normal
Males, 70-80 years of age	£6,469	£255.5	Normal
Males, 80+ years of age	£6,637	£222.6	Normal
Females, 45-60 years of age	£6,376	£242.5	Normal
Females, 60-70 years of age	£6,362	£223.0	Normal
Females, 70-80 years of age	£6,421	£228.8	Normal
Females, 80+ years of age	£6,570	£284.8	Normal

Costs for the second and subsequent years in either outcome category were estimated based on CPRD records of resource use, as above, but with the application of an adjusted surgery outcome prediction model. First, the model was re-estimated using 33 as the OHS threshold to distinguish between *Good* and *Poor* outcome patients (notwithstanding COASt resource use data was about the first postoperative year, because no dataset was available with similar information for the second and subsequent years after the primary). Coefficients from this predictive model are shown in Appendix 37. The adjusted model

was then applied to all CPRD records of resource use from cases minus the mean of their respective controls for years two through 10 as after that the number of cases per year dropped rapidly from above 1,000 to under 750. A probability cut-off point of 0.3 was applied considering that at 0.4 (as used for the first year data) the proportion of *likely Poor* outcomes would drop under 10%. This is an unreasonably low percentage even after considering the evidence from EPOS reported in Section 5.3.2 and Appendix 14, indicating that *Poor* outcomes decreased from 27% in the first year to 18% in the year after that. Once CPRD records were labelled as *likely Poor* and *likely Good* outcomes, records for years two through 10 were pooled together and mean values and distributions estimated to represent the yearly cost of primary care for patients two years and onwards after a THR. These values are reported in Table 5.35 by patient subgroup.

Once again, whilst Figure 5.24 showed, regardless of surgery outcome, an overall mean cost of postoperative primary care that increased slowly from around £30 at year two to approximately £50 at year 10, Table 5.35 summarises a very different scenario for *likely Poor* and *likely Good* outcome patients. The mean cost of *Poor* outcomes is now estimated to be between £250 and £300 a year, whilst for *Good* outcomes it stays under £40. Variability, however, is quite significant, as confirmed by the high SDs, which allow PSA to reflect the uncertainty associated to these mean values. As with primary care costs for the first year after the primary, the pooled number of consultations by staff and costs for these as well as for prescriptions are reported in Appendices 38 through 43 by outcome group.

It is important to note that the model used here to assign outcome categories to patients for whom PROMs were not collected needs further validation, which was not done as part of this study because no other dataset was available where both OHS and resource use were systematically collected. Furthermore, using the pattern of resource use observed in COASt to map surgery outcome in the CPRD in order to disaggregate cost parameters by outcome group is potentially a circular process that begins with a cost component (i.e. resource use) and also finishes with a measure of cost. As such, there is a risk that all patients with the highest resource use will be predicted to be poor outcomes, and vice-versa. Nevertheless, the fact that only resource use variables were included in the model to predict surgery outcome in the CPRD was justified by the evidence showing that none of the available demographic variables (gender or age) were statistically significant predictors of THR outcome. The clear causal relationship between (poor) outcome and

(higher) costs produces a strong correlation that was used here in the opposite direction to identify surgery outcome based on resource use, the variable component of cost. Lastly, the fact that the CPRD reports the use of many more resources than those included in the model also means that some cases mapped as being *likely Good* outcomes, for example, may well report costs as high as other *likely Poor* outcomes because of their higher use of any of the many resources not included in the model, as shown in Figure 5.28.

Table 5.35
Costs associated to second and subsequent years after THR in either outcome category states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	SD	Distribution
Second and subsequent year in *Good* outcome			
Males, 45-60 years of age	£21	£183.3	Normal
Males, 60-70 years of age	£19	£191.5	Normal
Males, 70-80 years of age	£5	£213.8	Normal
Males, 80+ years of age	£34	£250.6	Normal
Females, 45-60 years of age	£33	£206.6	Normal
Females, 60-70 years of age	£17	£224.8	Normal
Females, 70-80 years of age	£9	£225.7	Normal
Females, 80+ years of age	-£1	£239.1	Normal
Second and subsequent year in *Poor* outcome			
Males, 45-60 years of age	£316	£304.5	Normal
Males, 60-70 years of age	£237	£264.3	Normal
Males, 70-80 years of age	£241	£245.0	Normal
Males, 80+ years of age	£298	£363.2	Normal
Females, 45-60 years of age	£314	£295.7	Normal
Females, 60-70 years of age	£285	£285.6	Normal
Females, 70-80 years of age	£255	£296.8	Normal
Females, 80+ years of age	£253	£355.9	Normal

5.5.3 Revision THR and postoperative costs

Parameters for the four revision and post-revision model states were estimated following the same protocol described in the previous section and with data from the same sources. First, the costs of revision operations were obtained from relative frequencies of HRGs as reported by the NHS's PbR system on PROMs-eligible revision records for the year 2011-2012 [125]. Table 5.36 reports the mean and SD's of the costs associated to the revision THR operation by patient subgroup produced from the above frequencies and NHS reference costs by HRGs for the same year [126]. Relative HRG frequencies by patient

subgroup are reported in Appendix 44 and national average unit cost for all relevant HRG groups for the year 2011-2012 paid to NHS Trusts and Foundation Trusts are shown in Appendix 28.

Table 5.36 reports higher mean costs for revisions than primaries because the former are generally more complex than the latter. Revision costs are also more dispersed, as confirmed by the higher SDs because, as opposed to primaries, revisions are not assigned primarily to one HRG but instead (in around 65% to 75% of cases) to two: HR04C "Reconstruction procedures category 3 without complications" and HR05Z "Reconstruction procedures category 2", each costing the NHS £8,492 and £7,340, respectively [126]. Although these two HRGs are quite common for revision procedures, a number of operations were assigned other cost groups costing as much as £12,000 or as little as under £1,000. This explains the high SDs which account for as much as 6% to 11% of the mean value, yet since no 'well-behaved' distribution could be accurately fitted to this cost distribution, we only used the mean value reported in Table 5.36 and added the cost of a first year of primary care to compute the costs associated to the two model states combining the revision procedure and the following 12 months. As with primaries, the costs associated to the operation are presented regardless of surgical outcome.

Table 5.36
Cost of a revision THR to the NHS by patient subgroup

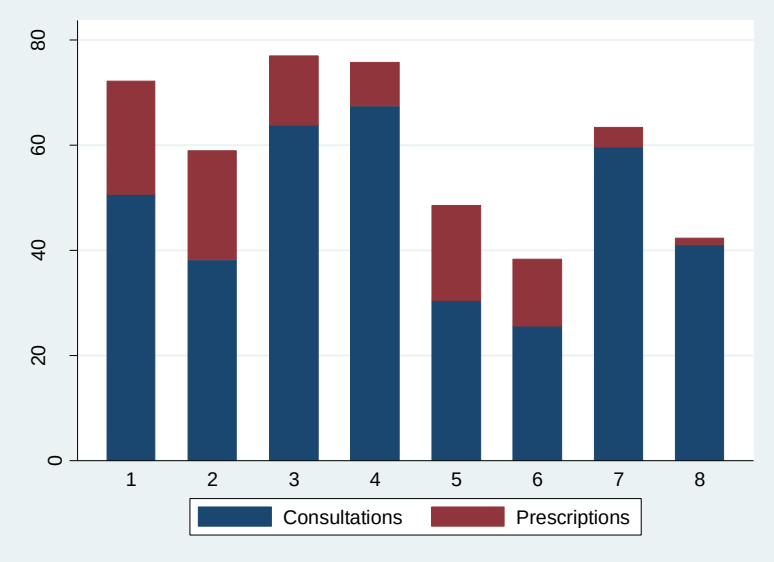
Patient subgroup	n	Mean	SD
Males, 45-60 years of age	499	£7,899	£885.3
Males, 60-70 years of age	835	£8,096	£697.5
Males, 70-80 years of age	1,247	£8,145	£479.6
Males, 80+ years of age	488	£8,191	£941.6
Females, 45-60 years of age	636	£7,733	£777.5
Females, 60-70 years of age	1,080	£7,910	£664.6
Females, 70-80 years of age	1,537	£7,996	£458.9
Females, 80+ years of age	885	£8,001	£574.7

For the costs of primary care provided to revision THR patients, we used CPRD data on patients with a medical diagnosis code for revision hip arthroplasty in their clinical or referral record on or before 31 December 2006, who were at least 18 years of age at index diagnosis. Controls (without record of a THR or even OA) were included and the data on consultations as well as prescriptions processed as described in previous sections. The mean number of consultations and prescriptions by controls were calculated and

subtracted from the figures reported for the respective cases producing, again, a measure of the use of resources attributable to the hip condition and intervention.

Figure 5.29 shows the mean cost of consultations and prescriptions for all patients following a revision THR and up to eight years after the operation. Because revision arthroplasties are rare, the numbers of yearly observations were much lower than those of primaries. In fact, the 497 case records analysed for the first year after the revision quickly dropped to 363 by year two, 130 by year five, and 44 by the eighth year. We considered data only up to year eight because after that the number of cases dropped below 30 and the resulting total yearly costs swung unreasonably from negative to positive values, suggesting that the data might likely be from a biased small group of patients. Total cost of £72 for the first year after revision is slightly higher than costs for the analogous period after the primary, as expected given that revisions are usually more elaborate procedures performed on patients who have had problems with their original replacement. Costs after that do not seem to drop as quickly for revisions as they do for primaries, again a likely consequence of the higher complexity of the intervention.

Figure 5.29
Mean cost of consultations and prescriptions due to hip problem, CPRD revision THR patients by year after surgery

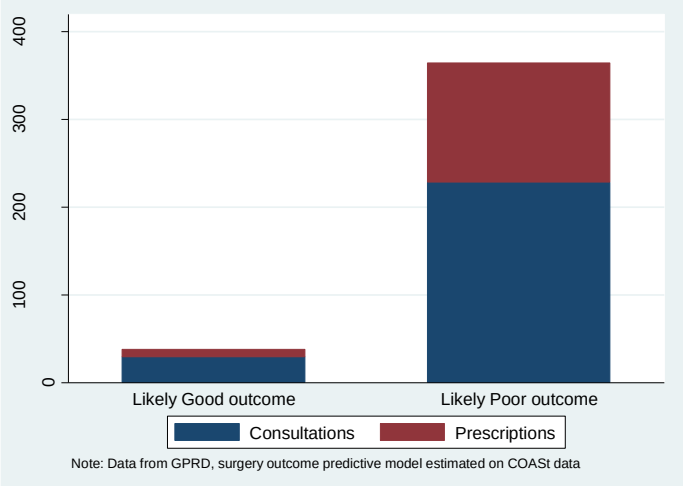


In order to distinguish between primary care costs provided to *Good* and *Poor* outcome patients after a revision THR, we employed the same model introduced in the previous section. Since no threshold had been determined for outcome categories after a revision THR, we used, as argued in Section 5.3.4, a cut-off point of 33. And we used this

threshold not only for the first but for all subsequent years after the procedure because we had no evidence suggesting a drop in the level of OHS associated to satisfaction from the first to the second year after a revision, as it has been found for primaries [79]. Also, the model applied to revisions was based on resource use data from primary THR patients because COASt had collected data about resource use and OHS only from a handful of revision patients. Those numbers were not sufficient to estimate a model predicting revision surgery outcome based on the use of healthcare resources, nonetheless the association between outcome and resource use implicit in the model based on primary THR data suggests that its application on revision patients is acceptable at the least.

The predictive model using 33 as a threshold for outcome categories and resource use data from primary THR patients was the same employed as that employed in the previous section for the second and subsequent years after the primary, and it is described in Appendix 37. The threshold of 0.3 employed for that analysis was also used in this case. Figure 5.30 shows how, after fitting the predictive model and distinguishing between outcome categories, the £72 spent on average on all revision THR patients during the first year immediately following their surgery hides an important gap between an estimated £364 invested on the primary care of *likely Poor* outcome patients, compared to only £38 on patients predicted to be *likely Good* outcomes. The higher relative weight of prescriptions as part of total costs for *likely Poor* outcome revision compared to primary THR patients is also noteworthy. Whilst prescriptions accounted for 16% of *Poor* outcomes' primary care costs during the first year immediately following a primary, this figure was higher at 37% for *Poor* outcome revision patients.

Figure 5.30
Mean cost of consultations and prescriptions due to hip problem during first year after revision THR by outcome category, CPRD all patients



Given the low number of revision cases further divided after fitting the predictive model to the data, results could not be presented by patient subgroups and were instead produced in aggregate form for all patients. Table 5.37 shows the mean cost associated to the model states combining revision THR and the first postoperative year where the differences by outcome categories are explained entirely by primary care costs associated to each group. Variation by patient subgroups is therefore that coming from surgery costs, which were estimated by gender and age groups and regardless of surgery outcome. Uncertainty around costs represented by the common SD within each outcome category is that obtained from the primary care costs analysis. Appendices 45 through 47 report the detailed number of consultations by staff as well as cost estimates for consultations and prescriptions for all patient subgroups.

Table 5.37
Costs associated to Revision THR followed by first year in either outcome category states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	SD	Distribution
Primary THR + first year in *Good* outcome			
Males, 45-60 years of age	£7,937	£38.0	Normal
Males, 60-70 years of age	£8,134	£38.0	Normal
Males, 70-80 years of age	£8,183	£38.0	Normal
Males, 80+ years of age	£8,229	£38.0	Normal
Females, 45-60 years of age	£7,771	£38.0	Normal
Females, 60-70 years of age	£7,948	£38.0	Normal
Females, 70-80 years of age	£8,034	£38.0	Normal
Females, 80+ years of age	£8,039	£38.0	Normal
Primary THR + first year in *Poor* outcome			
Males, 45-60 years of age	£8,264	£337.9	Normal
Males, 60-70 years of age	£8,460	£337.9	Normal
Males, 70-80 years of age	£8,510	£337.9	Normal
Males, 80+ years of age	£8,555	£337.9	Normal
Females, 45-60 years of age	£8,098	£337.9	Normal
Females, 60-70 years of age	£8,275	£337.9	Normal
Females, 70-80 years of age	£8,361	£337.9	Normal
Females, 80+ years of age	£8,365	£337.9	Normal

For the model states representing each outcome category after a revision THR, Table 5.38 indicates the mean and SD of primary care costs that, given the low number of CPRD observations, was applied to all patient subgroups. These were estimated by pooling together all CPRD records for the years two through eight, as explained above. Large SDs

with respect to mean values indicate the wide variability in costs associated to each group, explained by the varying levels of comorbidities amongst patients. Distributions were not as normal as with data from patients following a primary because of the lower number of observations. They were still largely symmetric, however, hence the application of the normal distribution for the PSA. Appendices 48 through 50 show the details of consultations and prescriptions resource use and costs.

Table 5.38
Costs associated to second and subsequent years after THR in either outcome category states: deterministic and probabilistic parameters

State / Patient subgroup	Mean	SD	Distribution
Second and subsequent year in *Good* outcome			
All patient subgroups	£43	£229.6	Normal
Second and subsequent year in *Poor* outcome			
All patient subgroups	£261	£242.2	Normal

5.6 Parameter values when using the prediction tool

The parameter values reported throughout this chapter correspond to current practice and evidence about the transitions, costs and HRQL of patients referred to an orthopaedic surgeon for a THR in the UK. As explained in the introductory section, the cost-effectiveness of applying the outcome prediction tool was estimated by comparing expected lifetime costs and QALYs based on current practice with the costs and QALYs expected to be produced by implementing the tool. This section explains how the outcome prediction tool developed under COASt would affect key model parameters and it presents estimates for the values that those parameters took.

Five transition probabilities and the QALYs associated to several preoperative model states, as indicated at the beginning of this chapter, would take different values if the prediction tool were in use compared to current practice. The prediction tool produces an expected OHS at one year after the primary, if a THR were to take place. We assume that this information can be useful to help surgeons more efficiently direct patients to a THR or to long-term medical management of their problems, such that, if they are expected to perform poorly after the operation, patients would be better off by not going through the procedure. This means that the probability of being referred for a THR would be directly affected by the implementation of the tool, as would the probability of being referred to risk-factor modification because patients referred to that path are in principle considered

suitable for a THR. The probability of a patient going into long-term medical management will necessarily be adjusted to compensate for the changes made to referrals for the previous two surgical pathways. The probabilities of *Good* and *Poor* outcome are the main reason why the tool is considered in the first place, hence they will also be affected by its implementation, by definition leading to a smaller proportion of *Poor* outcomes. And finally, given that those patients having the operation are expected to perform well, and that there exists an association between outcome and preoperative OHS as well as between OHS and EQ-5D, the QALYs associated to all preoperative states but the surgical assessment would also be affected by the shuffling of patients at this stage.

As described under Section 4.2 about the details of the outcome prediction tool, data from 2,092 patients were used to estimate the linear model predicting OHS at one year. We used the estimating dataset to explore the effects of using the prediction tool as the definitive guide to refer patients to a THR (including risk-factor modification), or to the long-term medical management state. We assumed, therefore, that the results of the prediction tool compared to a set threshold would be strictly followed. Although the tool could not replace the knowledge and complex criteria that surgeons rely upon to direct potential THR patients to treatment, this analysis reported on the changes in health benefits expected to be obtained, and its corresponding effect on healthcare direct costs, from applying the tool strictly. Because the tool would guide decisions based on a comparison of predicted OHS against a set threshold, we used 38 as a natural cut-off point since that was the level used to distinguish *Poor* from *Good* outcomes throughout the economic model. Given its key importance, we conducted a sensitivity analysis on the value of the cut-off point and estimated results based on other higher and lower thresholds to guide patients into THR or medical management. The model, nonetheless, was in all cases populated by parameter values based on a definition of *Poor* outcomes as those scoring under 38 in their OHS at one year after a primary THR.

The first immediate effect of applying the prediction tool was hence on the probability of referring patients to a THR or to risk-factor modification. We fitted the prediction model to the same estimating dataset and Table 5.39 shows how the proportion of patients referred for surgery would decrease as the threshold used for patient referral increases. Whilst not applying the tool in fact led to all patients in the dataset being referred for surgery, fitting the model to those same patients indicated that, if the threshold had been

38, for example, only 60% of patients currently receiving a THR would have been referred for the operation (or risk-factor modification) but the remaining 40% would have been sent for long-term medical management, instead.

The application of the outcome prediction tool may guide a patient to a THR because predicted OHS is above the threshold, yet the patient may have a condition that needs to be dealt with before surgery so he or she would be sent for risk-factor modification first. We therefore decreased the transition probabilities leading directly to THR and to risk-factor modification, as if they were one unit but considering their relative weights, to reflect the percentage reduction of patients referred for a THR implicit in Table 5.39. The transition probability to long-term medical management was adjusted upwards accordingly.

Table 5.39
Expected immediate effects of implementing the outcome prediction tool

Threshold OHS point	% referred for THR	% Poor outcomes after THR	Mean OHS gain forgone
32	95.7	28.3	16.6
34	90.1	25.6	18.6
36	78.0	21.4	22.2
38	60.6	16.6	23.9
40	38.8	9.3	24.2
42	20.8	5.8	24.1

As a result, whilst Table 5.4 in Section 5.2.8 reported that, under current practice, 17% of patients are referred for long-term medical management and 83% are considered suitable for a THR, these figures would change considerably if the outcome prediction tool were to be applied. As only predicted *Good* outcome patients would be considered for a THR, applying the tool at a threshold of 38 would reduce the proportion of suitable THR candidates to 60% of the current 83% THR referral rate, hence 50% of patients seen at a surgical assessment would be referred directly for a THR or to risk-factor management, and the other half for long-term medical management. This is shown in Table 5.40, where the original SD obtained from the expert elicitation exercise was kept and the corresponding Beta parameters calculated. The effects over THR and risk-factor modification were apportioned respecting their original relative weights, which produced the values shown in Table 5.40. The distribution of risk-factor modification parameter values for PSA was obtained empirically, i.e. by applying the corresponding reduction to the original distribution of experts' opinions, because a Beta distribution did not fit

properly. Since these transition probabilities were applied after removing deaths according to mortality rates, the values assigned to the direct transition to THR were calculated by subtracting those reported in the table from 1. Appendix 51 shows the corresponding values for the application of the tool at alternative threshold points.

Table 5.40
Preoperative probabilities with the tool: deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to risk factor modification	0.082	0.093	Empirical		
Surgical assessment to long-term medical management	0.495	0.208	Beta	2.365	2.413

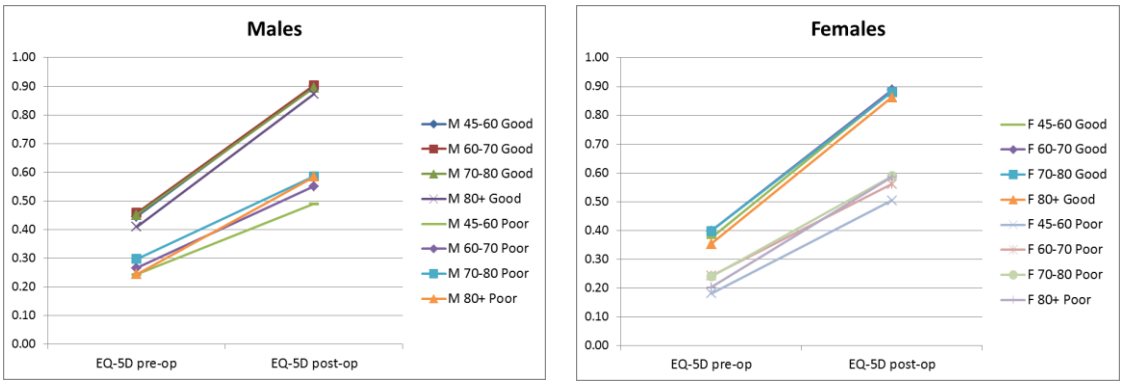
Since the outcome prediction tool produces an expected OHS whose main predictor is its own baseline measure, and we also established that OHS is correlated with HRQL with the work presented in Chapter 3, a change in patient referral patterns based on the tool would change the utilities associated to preoperative states. This would necessarily be true in the estimating dataset, because these data were used to generate the coefficients for the model in which baseline OHS had the greatest impact. This is shown in Table 5.41, which indicates how patients with a predicted OHS lower than 38 had also lower baseline OHS than those whose predicted score was greater than 38.

Table 5.41
Preoperative OHS within estimating dataset of outcome prediction tool

Patient group	Observations	Mean	95 % confidence interval	
All patients	2,092	16.03	15.68	16.39
Predicted OHS < 38	824	9.94	9.59	10.29
Predicted OHS ≥ 38	1,268	19.99	19.58	20.40

Data from HES-PROMs further confirms that the EQ-5D summary scores of patients not reaching an OHS of 38 at one year after the primary were statistically significantly lower than the scores of those who did attain this threshold. Figure 5.31 shows how, for all patient subgroups, *Good* outcomes had higher health utilities before surgery than *Poor* outcomes did. Although the tool is not perfectly sensitive or specific, on average it would therefore be expected to increase the HRQL of patients added to the waiting list for a THR or sent for risk-factor modification, and to lower that of patients referred for long-term medical management with respect to the model where the tool is not used.

Figure 5.31
Pre and postoperative EQ-5D scores of Good and Poor outcome patients by gender, HES-PROMs data



Since the HES-PROMs dataset did not contain all the variables used as predictors for the tool, we estimated a simplified version on the original EPOS-EUROHIP dataset using only preoperative OHS (since age and gender proved not statistically significant) to distinguish patients according to their predicted OHS. Appendix 52 reports on the resulting coefficients of the simplified model. With this approximation of a predicted OHS at one year, we obtained values of health utility estimates from HES-PROMs for the groups that would have been referred for THR or risk-factor modification, or long-term medical management if the outcome prediction tool were in use. Table 5.42 shows the resulting mean disutilities and distribution parameters separately for the two preoperative branches of the economic model. Appendix 18 included QALY values for all health states and Appendix 53 shows disutilities at the alternative threshold points.

Finally, as the tool would identify potential *Poor* outcome patients before they have the operation and refer them to a non-surgical treatment, a smaller proportion of *Poor* outcomes would also be expected from those who ultimately have their hip replaced. As Table 5.39 showed, this was indeed the case. Since all patients in the dataset effectively had surgery, we know that 31% of them scored less than 38 one year after their operation. Keeping this criterion to classify patients as *Poor* outcomes, if the tool had been applied this proportion would have been lower, reaching 17% if the referral threshold had been 38, or only 6% if 52 had been chosen. Table 5.43 shows the probability of *Poor* outcome if the tool had been in place using a threshold of 38 to refer patients. Although the aggregate proportion of *Poor* outcomes in this dataset (31%) was slightly lower than that reported in HES-PROMs (35%) and used when the tool is not implemented, most of the drop is due to the selective referral of patients made by the prediction tool, as indicated by the values reported in Table 5.43. For sensitivity analysis purposes, the corresponding

values for the alternative thresholds applied by the prediction tool are shown in Appendix 54.

Table 5.42
Disutility associated to preoperative states with the tool: deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Risk-factor modification, and				
Reassessment after risk-factor modification				
Males, 45-60 years of age	0.497	Gamma	3.89	0.128
Males, 60-70 years of age	0.484	Gamma	3.76	0.129
Males, 70-80 years of age	0.490	Gamma	3.81	0.128
Males, 80+ years of age	0.518	Gamma	4.01	0.129
Females, 45-60 years of age	0.536	Gamma	3.84	0.139
Females, 60-70 years of age	0.511	Gamma	3.76	0.136
Females, 70-80 years of age	0.522	Gamma	3.87	0.135
Females, 80+ years of age	0.539	Gamma	4.07	0.132
Long-term medical management, and				
Reassessment after long-term medical management				
Males, 45-60 years of age	0.954	Gamma	19.55	0.049
Males, 60-70 years of age	0.942	Gamma	20.99	0.045
Males, 70-80 years of age	0.933	Gamma	18.17	0.051
Males, 80+ years of age	0.948	Gamma	21.10	0.045
Females, 45-60 years of age	0.951	Gamma	19.39	0.049
Females, 60-70 years of age	0.946	Gamma	20.60	0.046
Females, 70-80 years of age	0.943	Gamma	18.44	0.051
Females, 80+ years of age	0.957	Gamma	19.51	0.049

Table 5.43
Probability of *Poor* outcome after *Primary THR* with the tool: deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.168	Beta	22	109
Males, 60-70 years of age	0.165	Beta	38	192
Males, 70-80 years of age	0.133	Beta	23	150
Males, 80+ years of age	0.237	Beta	9	29
Females, 45-60 years of age	0.167	Beta	19	95
Females, 60-70 years of age	0.141	Beta	32	195
Females, 70-80 years of age	0.178	Beta	43	199
Females, 80+ years of age	0.234	Beta	15	49

5.7 Assumptions

During the process of identifying parameter values to populate the economic model, a number of assumptions were made, whether because of the simplification forced by the fact that we were modelling a complex reality or because of limitations in the data available. In this section we discuss the assumptions made on the various probabilities, costs and health utility estimates, their possible implications and general feasibility.

Although preoperative transition probabilities may vary between patient subgroups, the values extracted from the expert elicitation exercise were assumed to apply to all patients regardless of age or gender. The method of data collection posed an important limitation in this case. It would have been highly impractical to ask the same questions to all experts about each specific patient subgroup, and they may have not been able to provide different values for each group. Dividing the limited number of experts, to ask different surgeons about different subgroups, was not feasible either. A common estimate for the mean preoperative transition probabilities, therefore, may not capture the possible heterogeneity amongst groups. By including surgeons who specialise on a variety of patients, however, the uncertainty represented in their answers was transferred to the pooled probability distributions, incorporating this heterogeneity into the analysis, ultimately reflected in PSA results.

Transitions between *Good* and *Poor* outcomes after year two post operation were estimated based on follow up questionnaires only up to year five. Results from EPOS records point to diminishing improvements over the first five years and the extrapolation of estimated probabilities took the levels of each outcome category to a plateau. It is possible, however, that over time and especially after 10 years of a primary THR many of those *Good* outcomes worsen and the proportion of *Poor* outcomes increases. We do not have data to support this, yet it seems clinically plausible. We account for this with the distributions assigned to transition probabilities linked to the number of patients involved in EPOS from whom probabilities were estimated and which added some of this uncertainty into the results through PSA.

As patients transit in the model from either outcome category after a primary THR to a revision THR and then to an outcome category immediately after this, we are assuming that surgery outcome after the primary has no bearing on surgery outcome after revision. It may be possible that this is not the case but we have no data available to confirm either

hypothesis. However, our assumption is clinically plausible inasmuch as this assumption implies that patients requiring a revision would be in a similar situation concerning their prosthesis regardless of their state of origin when they transitioned into the *Revision THR* state. Such similarity would make them equally likely to perform well or poorly after the revision. Additionally, although *Good* and *Poor* outcome patients after the primary would not generate similar levels of HRQL or costs to the aggregate analysis, PSA did allow for variability such that these *Good* and *Poor* outcome patients after the primary would not be so different in these regards either.

An important assumption was made when we used the cut-off point derived for primary THRs to categorise outcomes after revisions. This was done because no similar cut-off point has been calculated for revision THR patients. The resulting probabilities of *Good* and *Poor* outcomes are, nonetheless, acceptable since they imply a slightly higher likelihood of performing poorly after a revision, which was consistently reported by surgeons in the various rounds of consultations. Transition probabilities between outcome categories were also assumed to be equivalent after primaries and after revisions when those calculated for the former were applied to the latter. This was done because there are no datasets available with long-term follow-up of revision THR patients. However, we considered it is very likely that these transition probabilities are indeed similar because they describe patients' response after one year following major surgery, which primaries and revisions both are. Assigning a distribution to these probabilities also reflects results accounting for the uncertainty around their true value.

Finally, we applied all-cause mortality rates from the general population to patients with OA or other conditions possibly leading to a THR assuming that such musculoskeletal problems do not affect their chance of dying. Also, mortality rates applied to the first year after surgery, whether primary or revision, were those reported by the NJR which only describe the risk of death without attempting to identify whether surgery itself had any effect [54]. We therefore assumed that those values were a true reflection of death rates of patients undergoing a THR regardless of the reason, which is what the model required. We further assumed that outcome at one year after surgery, again whether primary or revision, did not have any bearing on mortality rates during that period.

In regard to HRQL values, it is important to note that both pre and postoperative measures used for the economic model were taken at roughly the same time with respect

to the operation, but not necessarily at the same point in the progression of the disease. A recent study looking at the HES-PROMs data from 2009-10, a subset of which we used to inform HRQL parameters for our model, found that non-white and more deprived patients tend to have joint replacement operations at a point when their OHS are lower than their white and less deprived counterparts, suggesting they had reached a more advanced stage of the disease [127]. We did not explore these inequalities here, but as the outcome prediction tool uses preoperative OHS as the main predictor variable, it is likely that the tool already takes account of such different disease stages regardless of the patients' race or deprivation level.

Regarding the use of the outcome prediction tool, inequalities in access to health services in general, to appropriate referrals and to surgery itself may also have an impact on the tool's effects. These inequalities have already been identified in England based on gender, age, deprivation, and ethnicity [128], but their possible effects on the application of the outcome prediction tool are outside the scope of this research.

When estimating QALYs for the model's health states, we assumed that the pattern of progression by outcome category during the first year after the operation in EPOS is generalizable to the wider population. We also assumed that the connection between OHS in the first and second years is representative of the changes all or most patients would experience. Although this might not necessarily be strictly the case, these assumptions are highly plausible as EPOS is a multicentre study whose main limitation is that the prosthesis employed in the THR was of the Exeter brand. The most frequently used stem in cemented THRs in England and Wales with more than 60% of the interventions performed in 2011 is in fact the Exeter V40, the second most common accounting for less than 20% of arthroplasties [54], hence data from THRs performed exclusively on Exeter prostheses is likely to be generalizable. Also, although assumptions were made about the patterns of quality of life progression, these were applied to the HES-PROMs dataset, a highly representative source where data was ultimately extracted from.

The health states of *Good* and *Poor* outcome after primary or after revision are the states where most patients would remain for long periods of time, until death in many cases. We did not consider a utility decrement when assigning health utility estimates to these states, which resulted in patients dying whilst still at high HRQL levels, especially in the case of

Good outcomes. This is a potentially unfitting assumption, but it also becomes irrelevant in this analysis because results from the model employing the outcome prediction tool are compared to current practice, and if a decrement were to be applied on the grounds of ageing, it would affect *Good* and *Poor* outcomes equally and for both comparators, hence such effects would essentially cancel each other out. Therefore, final utility estimates from each separate model should not be considered an accurate estimation of health utilities obtained with or without the intervention, they should only be analysed with respect to one another.

Because of the lack of datasets with follow-up information on revision THR patients, we assumed that the progression of health utility estimated from primary THR patients in EPOS was also applicable to revision THR patients. Although primary and revision patients may evolve differently during the first few months after their operation, applying these patterns to observed pre and postoperative scores reported by the highly representative HES-PROMs meant that the estimation of the parameters would still be highly accurate. The reason for this is that the progression patterns applied only described how patients move from their preoperative to their postoperative scores and not the scores themselves. Health utility estimates for the model states describing the second and subsequent years after revision THR were also affected by our assumption that the connection between OHS at years one and two after primary is the same as that after a revision operation. Again, in the absence of data describing how revision patients evolve from years one to two after a revision, the best approximation available was what has been observed in primary patients, which is what we used to populate the model.

When estimating parameter distributions to characterise uncertainty around health utility estimates, we assumed the time-trade-off weights reported in the literature for the EQ-5D [62] without considering any uncertainty around such valuation. Although these values are commonly used when performing economic evaluations, it is important to acknowledge that other valuation methods exist which could ultimately produce different health utility estimates.

Regarding assumptions about cost parameters, the costs for the *Risk-factor modification* state considered only reported primary care consultations and prescriptions by patients before their THR. It did not include the cost of the risk-factor modification programme itself because these vary according to the type of problem needing to be addressed (e.g. weight

reduction, blood pressure) and to date we have no reliable data on the use of these programmes by THR patients before the operation. Moreover, the inclusion in the model of separate states for risk-factor modification and long-term medical management was primarily justified by the intention to include a non-surgical treatment alternative as well as the reality of delayed primaries due to risk-factor management. We did not expect costs of the risk-factor modification programmes to have any significant effect on overall results.

Surgery costs, on the other hand, were explicitly included because they are the most resource intensive state of the economic model and, furthermore, they were assumed to be the same regardless of outcome category one year after the operation. We had no reason to believe that there would be an association between the HRG assigned to the operation, whether primary or revision, and surgical outcome a year later.

Costs of complications were not explicitly included but in many cases they were already part of the cost estimations. Perioperative complications were considered in HRG reference costs and primary care resource use due to complications was also part of the CPRD data used to produce cost estimates. However, surgical complications such as deep vein thrombosis (DVT), pulmonary embolism (PE), fracture, and the more recently explored associations between THR and myocardial infarction [129] or stroke [130] were excluded from the analysis. As this economic evaluation assesses the implementation of an outcome prediction tool after THR, the effect of costs associated to complications would be relevant only inasmuch as the tool changes the proportion of patients going into surgery and these complications appear in statistically different rates between the outcome categories considered. Since we lack data on the differential incidence of complications between *Good* and *Poor* outcome patients as defined here, and the rate of complications such as DVT and PE reported in other economic evaluations of THR is as low as 1% [47], these were not incorporated into the analysis.

In using preliminary results from the COASt cohort for sections of the cost estimation exercise, we assumed that the cohort is representative of clinical practice and more generally of patients in the UK. More specifically, we assumed that the list of medications used after a THR as well as the pattern of resource use and its relationship with surgery outcome observed in COASt is similar to the overall pattern and connection in the country as a whole.

We assumed that estimating the surgery outcome predictive model at an OHS threshold of 33 on resource use data collected in COASt for the first year after a primary was a valid approximation of the coefficients and statistical significance that would have been obtained had the model been estimated from resource use collected during the second year. This was a necessary assumption, given the lack of data on resource use collected during the second postoperative year from THR patients with available OHS. It is also a feasible assumption considering that, if resource use is associated to the level of pain and limitations as measured by the OHS, then the timing of the measure should be irrelevant and the resulting coefficients would represent the number of consultations and prescriptions associated to the groups scoring above or below the new threshold.

Regarding the application of the outcome prediction tool, we indicated that it would have the effect of lowering the probabilities of being referred for a THR, whether directly or through risk-factor modification, and that the transition probability to long-term medical management would increase because patients not referred for a THR would be treated non-surgically. We assumed that the tool would not have any direct effect over the referral pattern of patients originally sent for long-term medical management because those patients had by definition not been considered for the operation either because their problem would not be solved by the THR, they were unfit for surgery, or they did not want to have it. None of these situations would feasibly be affected by the output of the outcome prediction tool.

Finally, we assumed that there was no correlation between model parameters within each model considering current practice or the application of the outcome prediction tool. The distinction is made because the difference between the two models is, in fact, that they are populated by a different set of certain probabilities and HRQL measures that are associated to whether the tool is used or not. Any correlation amongst parameters beyond the changing patterns due to implementing the prediction tool was not considered in the economic model.

5.8 Contributions

Section 5.2 of this chapter on the expert elicitation exercise was a direct result of the work by Rafael Pinedo (RP). An important contribution was received from Dr Laura Bojke and Marta Soares as they provided RP with a set of slides that had been previously designed and used to support the elicitation interview in a similar exercise at the University of York

and which RP adapted for this elicitation on THR preoperative transition probabilities. The input from all surgeons interviewed was of course greatly useful and appreciated. The decision to elicit experts' opinions directly in an interview, the design of the questions, format of the interview, recording of responses and analysis of results were all performed directly by RP.

To estimate all remaining parameters, a great deal of data from HES-PROMs was used. RP was responsible for completing and submitting the applications to the NHS Information Centre, obtaining approval, and ultimately uploading, accessing, cleaning, merging, linking and analysing the HES and PROMs data.

The CPRD extract was another of the main datasets used to estimate parameters, principally costs. For this, RP produced the set of criteria for the data to be extracted and Joe Maskell (JM), at the University of Southampton's School of Medicine at the time, accessed the dataset and produced large SAS files containing individual records. These files were then accessed, cleaned, merged, linked and analysed by RP to produce final parameter value estimations. As well as data from the CPRD, estimation of cost parameters was possible thanks to preliminary and unpublished data from the COASt cohort. The COASt and Biobank teams at the University of Oxford facilitated access to the follow-up forms of THR patients who had completed and returned them so that RP could enter the data, clean it and analyse it before estimating the model predicting surgery outcome based on resource use. Whilst organising and grouping the medication taken by THR patients, dsicussions with GP Dr Simon Frasier from the University of Southampton's School of Medicine Primary Care Department were particularly useful.

All other work presented in this chapter for the estimation of parameter values populating the economic model was performed entirely by RP.

6 Cost-effectiveness analysis of the THR outcome prediction tool

In Chapter 5 we estimated all the necessary parameter input values to populate the Markov model detailed in Chapter 4. In section 6.1 of this chapter we describe how the model was structured mathematically before presenting final results. Deterministic results are presented in section 6.2 for each subgroup entering the model. These are followed in section 6.3 by results from the sensitivity analyses on key model parameters, and crucially, on the cut-off point used with the outcome prediction tool to determine which patients should be referred for surgery. This section includes the graphical representation of probabilistic results in the cost-effectiveness plane as well as cost-effectiveness accetability curves (CEACs), also for each patient subgroup. The chapter closes with an indication of contributions received for the production of these results, discussed extensively in the final Chapter 7.

6.1 Mathematical structure of the model

The cohort Markov model was set up in Microsoft Excel with transition probabilities determining the distribution of the cohort amongst the various health states after each cycle. The model was run for each patient group, i.e. we produced results for females and males entering the model at 45, 60, 70 or 80 years of age. As it was a lifetime model, we recorded results for the number of yearly cycles necessary for all patients in the cohort to move into the absorbing state of death.

A two-tier system of dynamic probabilities was put into place. First, as the age of the patients entering the model was predefined, the cycle number was used to identify their age as they moved through the model and hence the appropriate all-cause age-related mortality was applied. All other transition probabilities were therefore adjusted at every cycle to reflect the lower proportion of patients still alive and moving into health states different from *death*. Second, because we obtained parameter values for four age groups (45-60, 60-70, 70-80, more than 80), as patients in each cohort reached the starting point of the next age-group, the corresponding set of probabilities was applied. As a result, patients entering the model at age 45, for example, would experience not only larger mortality rates and lower transition probabilities to other health states as they became older, but they also would transition amongst states at different relative probabilities when they entered the model as compared to when those still alive reach 60, 70, and finally 80 years of age.

In the particular case of mortality immediately following THR, for both primary and revision procedures, we observed that the probabilities estimated based on the most recent NJR report [54] and presented in Appendix 13 became, at a certain age, lower than the all-cause mortality rate reported by the ONS [106]. For females, this happened at age 63, whilst for males all-cause mortality rates became greater than the estimated mortality after THR at 56. We used the higher of the two as the mortality rate following primary or revision THR.

Health utility and cost estimates associated to each health state in the model were also updated as patients reached the following age bracket. Costs associated to model health states were the only ones considered since it was understood that implementation of the outcome prediction tool would not imply any relevant cost per patient borne by the health care system. As total undiscounted costs and health utility estimates were calculated for each yearly cycle, an annual discount rate of 3.5% was applied according to standard methods for technology appraisal [57] to both costs and health utilities to produce total present values and an estimate of the deterministic ICER for each patient subgroup.

For the probabilistic analysis, we followed the standard method of independent uniform draws from each parameter distribution for 1,000 repetitions. Specifically for the first set of transitions from the starting point of surgical assessment, after applying the corresponding all-cause mortality rate and counting THR followed by either outcome category as one transition, there were three possible destinations for patients: risk-factor modification, long-term medical management and THR. The standard approach would have been to apply a Dirichlet or a series of Beta distributions to this set of probabilities, which would have required a common parameter that was not available with our data. As reported in Section 5.2.8, no distribution could be adequately fitted to some of the surgeons' pooled judgements about pre-operative transition probabilities and we resorted to obtaining random draws from the observed empirical distribution. We therefore generated independent draws for the three probabilities above, whether from a fitted or empirical distribution, and subsequently adjusted their values maintaining their relative weights so that their sum was equal to 1 minus the corresponding death rate. For all other probabilities, alternative transitions were or could be grouped into pairs, hence only one random draw was necessary as the remaining probability would be the difference between the former and 1.

6.2 Cost-effectiveness results

Expected costs and QALYs over the lifetime of the cohorts entering the model were calculated for current practice and the hypothetical scenario of implementing the outcome prediction tool with predicted OHS of 38 as the threshold to direct patients to THR (above 38) or to long-term medical management (below 38). Results for each patient subgroup, including the corresponding incremental costs and QALYs (prediction tool minus current practice) and the respective ICERs are shown in Table 6.1.

Table 6.1
Deterministic cost-effectiveness results, by patient subgroup

Subgroup (gender, starting age)	Current practice		Prediction tool		Incremental		
	Costs	QALY	Costs	QALY	Costs	QALY	**ICER**
Females, 45	11,562	14.52	10,437	10.87	- 1,125	- 3.66	**308**
Females, 60	9,282	11.08	7,853	7.97	- 1,429	- 3.11	**460**
Females, 70	7,891	7.93	6,199	5.56	- 1,692	- 2.37	**714**
Females, 80	6,520	4.75	4,676	3.26	- 1,844	- 1.49	**1,240**
Males, 45	10,086	14.61	9,055	10.70	- 1,031	- 3.92	**263**
Males, 60	8,196	10.81	6,890	7.64	- 1,306	- 3.17	**412**
Males, 70	7,062	7.64	5,495	5.30	- 1,568	- 2.34	**669**
Males, 80	5,954	4.48	4,367	3.07	- 1,586	- 1.40	**1,130**

As Table 6.1 shows, implementation of the outcome prediction tool is associated with lower costs as well as lower QALY gain than current practice for all patient subgroups. Whilst current lifetime costs for the average patient assessed for a surgical intervention vary between £6,000 and £11,500 above that which patients without a hip condition cost, and this for a gain of 4.5 to 14.5 QALYs with values mainly depending on age, implementing the outcome prediction tool would reduce such costs by £1,000 to £1,500 but also reduce QALY gain by as much as four years in full health, or its equivalent.

As a result, ICERs were estimated to be around £250 to £300 per QALY forgone for men or women assessed at 45 years of age, up to £1,100 to £1,200 per QALY lost for 80-year-old men or women considered for a THR. If these results had originated from the NE quadrant of the cost-effectiveness plane, then such ICERs would have suggested implementing the outcome prediction tool as a highly cost-effective intervention costing significantly less than £30,000 per QALY gained. However, as the prediction tool would reduce costs at the expense of QALYs gained, thus placing the cost-effectiveness ratio in

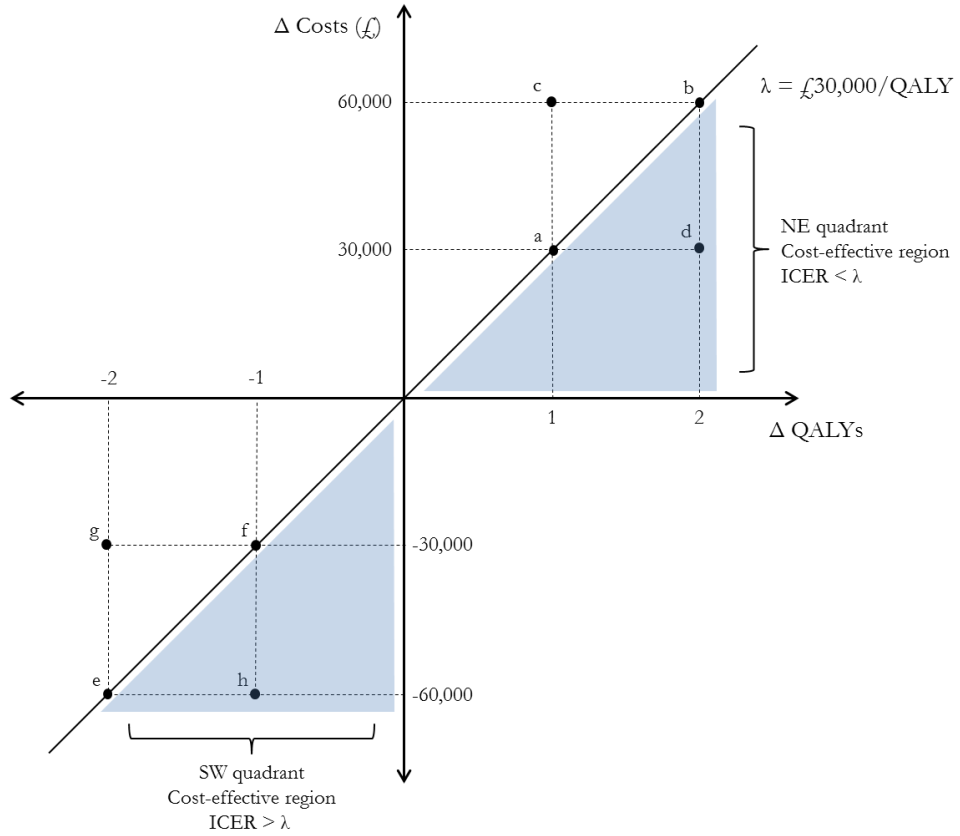
the SW quadrant, only ICERs above £30,000 per QALY forgone might be considered cost-effective under the assumption that the health care system would be willing to reduce costs at the expense of length and quality of life at the same rate that it is willing to adopt technologies which increase QALYs at a positive cost. Hence, the above deterministic results suggest that the outcome prediction tool would only be cost-effective if the health care system were willing to exchange reduction in costs for reduction in length and quality of life at a rate lower than the reported ICERs.

Figure 6.1 illustrates how, for typical assessments falling in the NE quadrant, cost-effectiveness is associated with ICERs lower than the threshold, whereas when results are in the SW quadrant this is the other way around. Saving costs at the expense of QALYs may be considered by many as unethical. Nevertheless, if the resources saved can be reallocated to technologies capable of producing more QALYs than those displaced at a cost per QALY lower than the threshold, then adoption of both interventions would lead to a net health benefit for society. This may justify extending the cost-effectiveness threshold applied in the NE to the SW quadrant as well. We show this in Figure 6.1 where point "h" in the SW quadrant could be deemed potentially cost-effective because it would save £60,000 whilst reducing one QALY, as long as those resources are reinvested in a cost-effective technology in the NE quadrant such as point "d", which with half of the money saved would produce two QALYs. Adopting both technologies associated with points "h" and "d" in Figure 6.1 would jointly save £30,000 and produce one additional QALY, a result that if provided by a single technology would have been considered dominant. Although it is not an objective of this thesis to explore in detail any disinvestment models, this analysis helps to clarify how unusual results in the SW quadrants, such as the ones obtained in our assessment, can be considered potentially cost-effective, among other conditions, if their ICER is higher than the cost-effectiveness threshold.

Another helpful way to analyse results where incremental costs and QALYs are both negative is the net benefit approach. We calculated the net monetary benefit (NMB) of each intervention by combining both the expected cost and QALY outputs of each alternative into one unique measure in monetary terms. To do this, QALYs were converted into the same unit as costs by multiplying them by a set willingness-to-pay ratio and subtracting costs from the result. The intervention producing the highest NMB could then be considered the more cost-effective alternative. As Table 6.2 shows, at a

willingness to pay threshold of £500 per QALY, using the outcome prediction tool produces a higher NMB than current practice for both men and women entering the model at 70 or 80 years of age. At this threshold level, the result would be a net monetary loss for those and all other subgroups, meaning that costs would exceed the value assigned to the QALYs gained. The higher NMB for patients above 70 years of age under the scenario of using the tool is consistent with ICERs reported in Table 6.1, which are above the willingness to pay threshold of £500 per QALY for these patient subgroups. At a threshold of £1,000 per QALY, only the patient subgroups surgically assessed at 80 years of age would get a higher NMB with the prediction tool than under current practice, again in line with ICERs reported above considering that results are in the SW quadrant. As the willingness to pay threshold increases, approaching the current acceptable range of £20,000 to £30,000 per QALY, current practice produces much higher (and positive) NMB for all patient subgroups.

Figure 6.1
Cost-effective regions in the NE and SW quadrants of the CE plane



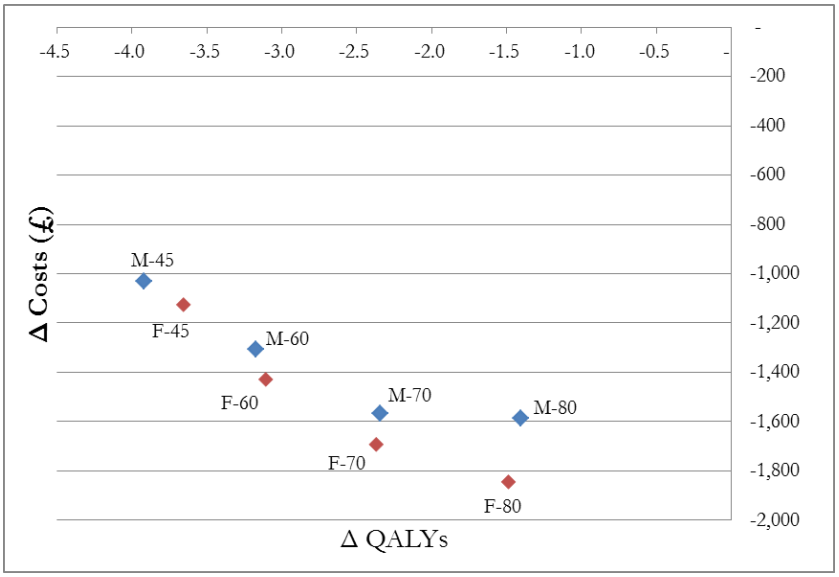
Both Tables 6.1 and 6.2 show how results differ only slightly by gender and that the majority of variation is driven by the age at which patients are assessed for a surgical

intervention. Figure 6.2 illustrates this, with incremental results for all patient subgroups found in the SW quadrant of the cost-effectiveness plane and results moving towards greater cost reduction and lower QALY loss as age increases. It is noticeable how results by gender also diverge as patients' age entering the model increases, particularly in terms of cost reduction.

Table 6.2
Net monetary benefit (£) by patient subgroup
at selected levels of willingness to pay

	£500/QALY		£1,000/QALY		£10,000/QALY		£20,000/QALY	
	Current	Tool	Current	Tool	Current	Tool	Current	Tool
Females, 45	- 4,301	- 5,003	2,960	430	133,664	98,235	278,890	206,908
Females, 60	- 3,743	- 3,867	1,796	119	101,496	71,862	212,273	151,577
Females, 70	- 3,927	- 3,419	37	- 640	71,394	49,390	150,679	104,979
Females, 80	- 4,144	- 3,044	- 1,768	- 1,412	41,001	27,970	88,522	60,616
Males, 45	- 2,778	- 3,705	4,529	1,644	136,057	97,937	282,199	204,928
Males, 60	- 2,792	- 3,072	2,612	746	99,884	69,463	207,965	145,815
Males, 70	- 3,241	- 2,845	580	- 196	69,362	47,494	145,787	100,484
Males, 80	- 3,716	- 2,831	- 1,478	- 1,295	38,807	26,354	83,568	57,076

Figure 6.2
Incremental cost-effectiveness results by patient subgroups

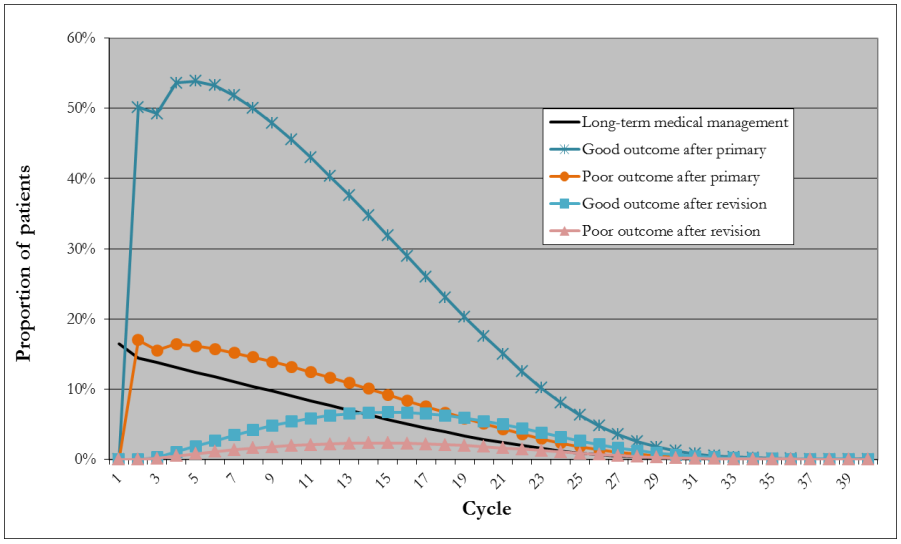


Results in Table 6.2, indicating that the introduction of the outcome prediction tool would reduce costs but also produce less QALYs than current practice of THR referral, are mainly driven by the tool's diverting patients from THR to long-term medical management in order to minimise *Poor* outcomes after surgery. This is confirmed by

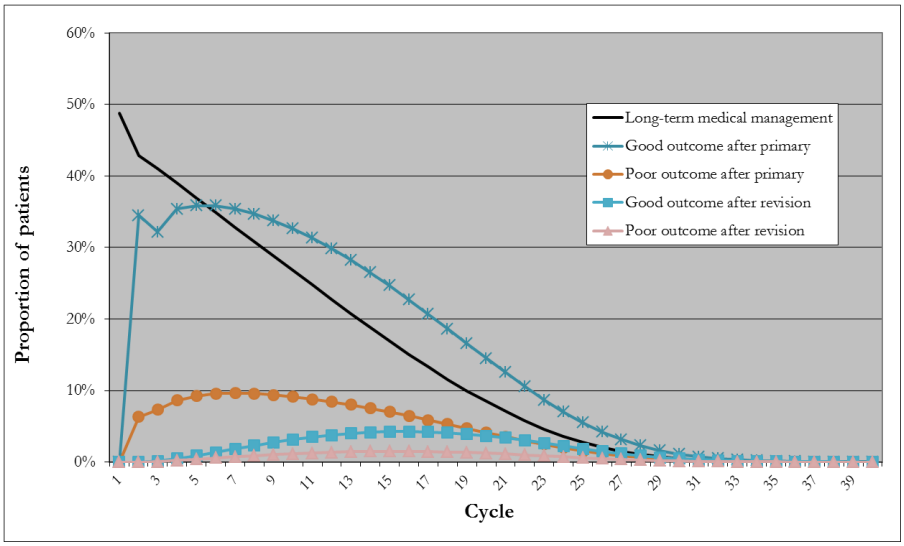
Figure 6.3 (A) which shows how, for women of 70 years of age (Females 70-80 being the subgroup with most THRs in the UK) under current practice, the proportion of patients in the non-surgical alternative starts under 20% and decreases progressively, with a high proportion of *Good* outcomes after primary THR exceeding 50% early in the simulation and *Poor* outcomes topping around 18% during the first few cycles of the model. If the outcome prediction tool were implemented, the model indicates that we should in fact expect *Poor* outcomes to stay under 10%, but by doing so many potential *Good* outcomes would be sacrificed. This health state would hold less than 40% of the cohort at any time, with roughly as many patients in it throughout all cycles as in long-term medical management, as shown in Figure 6.3 (B).

Figure 6.3
Relative survival in key model states, Females 70 years of age

(A) Current practice



(B) Outcome prediction tool



The different distribution of patients produced by the outcome prediction tool combined with the high QALY values associated to *Good* outcome (around 0.7) and low values whilst patients are in long-term medical management (roughly 0.05, derived from the disutilities reported in Table 5.42) explain the lower expected QALYs produced if the tool were applied as compared to current practice. This is also the case with costs. Although the annual primary care cost of a *Good* outcome after primary THR is no more than £34, being on long-term medical management costs the NHS only an estimated £90 to £120 a year. Keeping many patients from receiving a THR might hence increase slightly overall primary care costs, but it would also save or delay expenses of £6,000 to £6,600 on the primary THR, and therefore in some cases also a revision procedure costing £7,700 to £8,500. These sizeable savings explain why introducing the outcome prediction tool would be cost-saving with respect to current practice.

The reduction of QALY difference between the two alternatives as age increases, which was shown in Figure 6.2, can hence be understood based on the reduced survival of older patients, which also reduced the number of years adding less QALYs via the long-term medical management state compared to *Good* outcome after THR. In the case of the growing difference in costs, the outcome prediction tool would produce greater savings with older patients because a much smaller proportion of the total cohort would ultimately be referred for surgery compared with current practice and the slightly higher costs of being held at the long-term medical management state would not amount to any relevant magnitude in the opposite direction.

6.3 Sensitivity analysis

The results presented above indicate that implementing the outcome prediction tool would be cost-saving but it would also produce less QALY gain than current practice, on average. These results were the product of bringing together average probabilities linking 14 health states as well as average costs and QALYs for each in a mathematical simulation. Although most of the data analysed came from very large and representative patient-level datasets such as HES, PROMs, NJR and the CPRD, thus providing a high level of confidence around parameter value estimates, it is nonetheless important to explore the effect that uncertainty in key parameters may have on final results. We therefore conducted one-way sensitivity analyses on the discount rate and on the health utility estimate applied to long-term medical management, as this is the parameter value for which there was no highly representative data and therefore the one subject to the

largest uncertainty. We also performed a sensitivity analysis on the cut-off point assumed by the outcome prediction tool to direct patients to THR or to long-term medical management as the non-surgical alternative. The above analyses were performed only on females entering the model at 70 years of age to illustrate effects, because of the eight subgroups considered this is the largest one receiving THRs in the UK. Finally, we performed a fully probabilistic sensitivity analysis to reflect the uncertainty in all parameter values in decision uncertainty via CEACs.

As discussed in the previous section, the average health utility estimate for long-term medical management when the prediction tool is assumed to be in place played an important part in the tool producing less QALYs than current practice. Although its low value (around 0.05) was justified by the tool's discriminatory *raison d'etre* based primarily on preoperative OHS, as explained in Chapter 5, we performed sensitivity analysis on the QALY estimate associated to this state in order to ascertain whether it affected results in any significant manner. As Table 6.3 shows, varying the mean value of health utility assigned to this health state in five equal steps from the low 0.05 to the same value applied to the simulation for current practice (0.334) increases the QALY gain when using the tool but not enough to reach the levels attained by current practice, *ceteris paribus*. The difference in QALY gain is driven by the higher proportion of *Good* outcomes in current practice and which, if the prediction tool were implemented, would have been kept from surgery in long-term medical management because of the tool's imperfect specificity.

As is customary in economic evaluations and suggested in NICE's Guide to the methods of technology appraisal [57], we performed sensitivity analysis by dropping the discount rate for both costs and benefits from 3.5% to 1.5%. The results obtained when applying this lower rate did not affect results in any significant way. Both incremental costs and QALYs were again in the negative range, with costs savings slightly larger (£1,712 instead of £1,692) and QALYs lost increasing marginally, from 2.37 to 2.78. In neither of the above two cases would the expected effects of changing original mean values be enough to change the decision of not adopting the tool, unless the health service were willing to forgo QALYs for savings at a rate of only £1,600 to £1,700 saved per QALY lost.

Arguably the largest uncertainty of all parameters is that associated to the application of the outcome prediction tool itself, as all results modelled here are, although based on the patient-level data used to estimate the statistical tool, hypothetical. We therefore conducted a sensitivity analysis on the cut-off point at which the outcome prediction tool

would be used to direct patients into surgery, or not, to explore potential effects on final results. By changing this cut-off point, five probabilities and a number of health utility estimates would all change: the probabilities of being referred for a THR, for risk-factor modification, for long-term medical management, the probabilities of *Good* and *Poor* outcome, and the QALY estimate for all preoperative states with the exception of surgical assessment. Since the base case analysis used 38 as the reference cut-off point for the prediction tool, we adjusted the above model input parameters accordingly for scenarios where the tool would direct patients based on cut-off points of 32, 34, 36, 40 or 42. Table 6.4 shows the resulting total costs and QALYs of each alternative as well as the corresponding incremental differences and ICERs.

Table 6.3
Sensitivity analysis on health utility estimate for long term medical management, females 70 years of age

Mean health utility estimate	Current practice		Prediction tool		Incremental	
	Costs	QALY	Costs	QALY	Costs	QALY
0.057 (base case)	7,891	7.93	6,199	5.56	- 1,692	- 2.37
0.112	7,891	7.93	6,199	5.73	- 1,692	- 2.20
0.167	7,891	7.93	6,199	5.91	- 1,692	- 2.02
0.223	7,891	7.93	6,199	6.08	- 1,692	- 1.85
0.278	7,891	7.93	6,199	6.26	- 1,692	- 1.67
0.334 (current practice)	7,891	7.93	6,199	6.43	- 1,692	- 1.50

Table 6.4
Sensitivity analysis on the cut-off point applied by the outcome prediction tool, females 70 years of age

	Current practice		Prediction tool		Incremental		
Cut-off point	Costs	QALY	Costs	QALY	Costs	QALY	**ICER**
32	7,891	7.93	7,630	7.02	-261	-0.91	**288.72**
34	7,891	7.93	7,394	6.67	-497	-1.26	**394.74**
36	7,891	7.93	6,896	6.18	-995	-1.75	**569.83**
38 (base case)	7,891	7.93	6,199	5.56	-1,692	-2.37	**714.19**
40	7,891	7.93	5,337	4.95	-2,554	-2.97	**858.80**
42	7,891	7.93	4,648	4.58	-3,243	-3.35	**967.34**

As expected, with changes to the parameters of the simulation under application of the prediction tool, total costs and QALYs for current practice considering the cohort of 70-

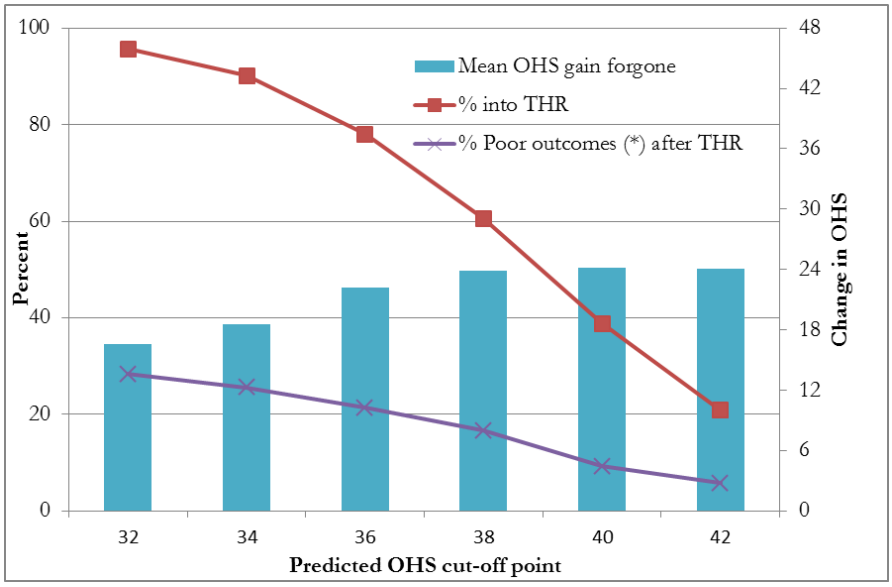
year-old women did not change, but those for the tool did. As the tool becomes more lenient and directs patients with predicted postoperative OHS under 38 for a THR, savings with respect to current practice are reduced because more patients ultimately have their hips replaced. Application of the tool thus calibrated would also mean that the QALYs generated would approach those attained by current practice because more potential *Good* outcomes wrongly held back before in long-term medical management would now be put forward for a THR, hence attaining the higher QALYs that the operation achieves on most patients. The opposite effect was obtained when parameters were adjusted to a prediction tool that applied cut-off points higher than 38 to decide which patients should receive a THR or not: more money would be saved but more QALYs would be forgone.

Figure 6.4 visibly illustrates this. As the cut-off point applied by the outcome prediction tool increases, the percentage of patients referred for THR decreases and so does the proportion of *Poor* outcomes, which is the main purpose of the tool. However, as most patients kept from receiving the potential benefits of surgery would have benefited from the THR, the average OHS gain due to surgery but forgone because those patients were kept back in long-term medical management, increases. A prediction tool operating with a cut-off point below 32 would not serve any purpose because at 32 it would be expected to produce only slightly above 30% of *Poor* outcomes, which some patient subgroups already report under current practice, as reported in Chapter 5. In summary, the more stringent the prediction tool, the deeper into the SW quadrant of the cost-effectiveness plane it would move with respect to current practice. It must be stressed that this sensitivity analysis did not affect the costs or postoperative health utilities for *Good* and *Poor* outcomes throughout the model. These were calculated assuming the 38 and 33 OHS cut-off points described in the corresponding section hence implying that if the tool used a different threshold to direct patients, transitions to either outcome category would change but the costs and utilities patients accrue whilst in those states would not.

Finally, in order to reflect the uncertainty of all model input parameters not only individually but also simultaneously as uncertainty around the decision to adopt the prediction tool, we conducted a full probabilistic sensitivity analysis by allowing all parameter values to change stochastically and independently based on their respective distribution (as identified in Chapter 5). Figure 6.5 shows the results of running 1,000 Monte Carlo simulations and placing the corresponding incremental costs and QALYs on

the cost-effectiveness plane for women and men entering the model at 70 years of age. As the Figure shows, the great majority of the simulations placed incremental cost-effectiveness results on the SW quadrant. More specifically, in 87% of cases for women and 88% of cases for men 70 years of age, implementing the tool was expected to cost less but also produce less QALYs than current practice. The mean incremental costs and QALYs of all simulations and respective ICERs are presented in Table 6.5, showing, as expected, nearly the exact same results as the deterministic analysis reported in Table 6.1.

Figure 6.4
Effects of applying the prediction tool at selected cut-off points



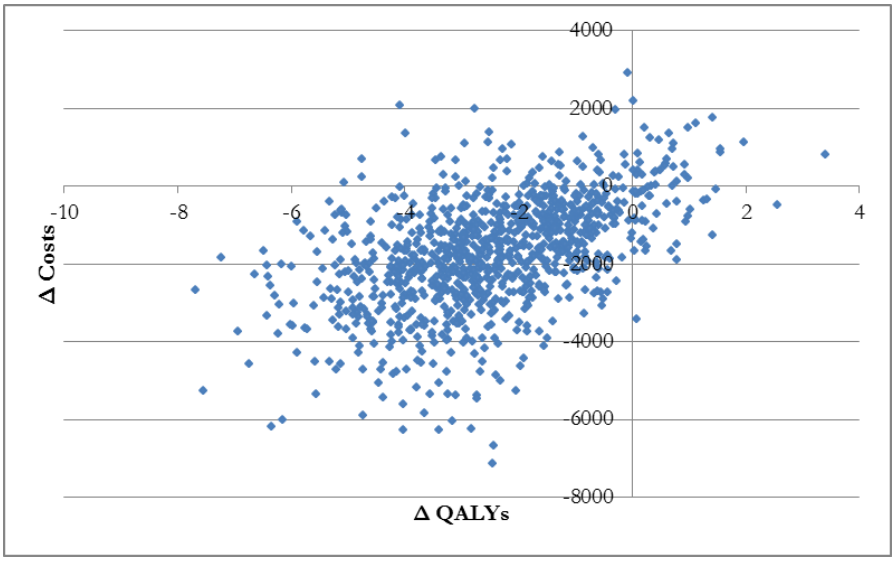
The pattern of high concentration of probabilistic results in the SW quadrant observed in Figure 6.5 for 70-year-old women and men changed only slightly with age, with the younger cohorts reporting the lowest percentage of simulations in the SW quadrant (78% for both men and women) and the oldest ones performing essentially the same as the 70 year olds. Equivalent plots for all remaining patient subgroups are shown in Appendix 55.

Running the series of Monte Carlo simulations produced a CEAC, which is shown in Figure 6.6 only for women entering the model at 70 years of age (as Figure 6.2 confirmed, there is no significant difference in results by gender). The curves representing the probability that either current practice or the tool would be deemed cost-effective at the various thresholds represented in the x-axis crossed at a point between £700 and £800 per QALY. This is consistent with the above-reported ICERs of £714 (deterministic) or £695 (mean probabilistic). As shown in the previous section, NMB was higher for the prediction tool (thus making it cost-effective) only when the ICER was higher than the

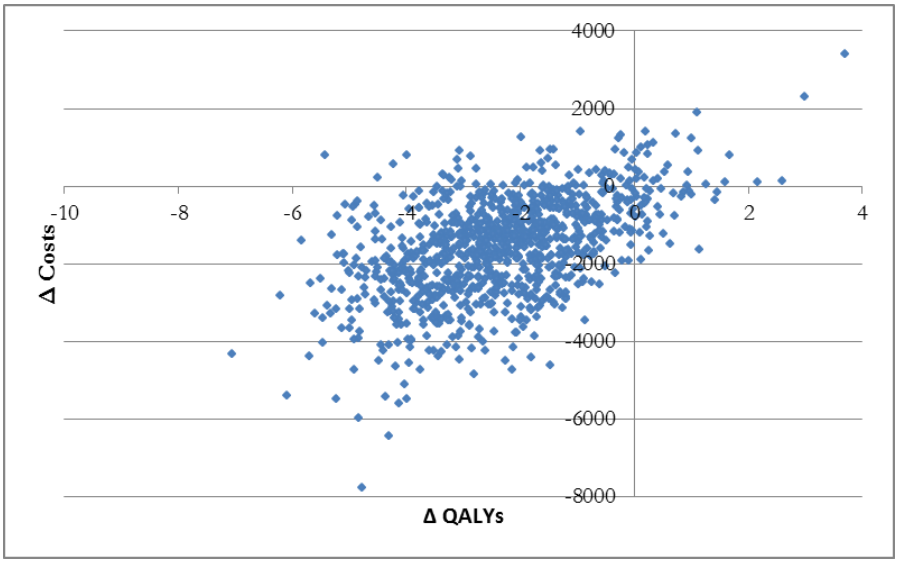
willingness to pay thresholds; once the threshold exceeded the ICER, current practice became the cost-effective intervention. This is shown in Figure 6.6 by the decreasing probability of the outcome prediction tool being cost-effective as the threshold increases, with this probability falling under that for current practice at a point exactly or near the ICER for women entering the model at 70 years of age.

Figure 6.5
Results of Monte Carlo simulations on the cost-effectiveness plane

(A) Females, 70 years of age



(B) Males, 70 years of age



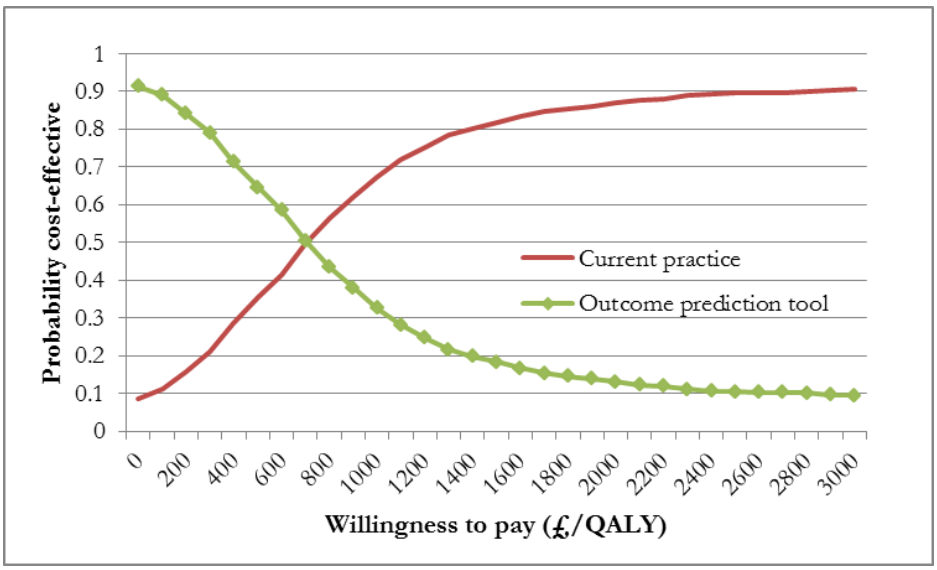
It is important to stress, however, that the range of willingness to pay thresholds within which implementing the outcome prediction tool would be cost-effective refers actually to scenarios of cost reduction and less QALYs generated with respect to current practice. This is effectively the range of willingness to pay thresholds at which both alternatives

produce net monetary losses, with the outcome prediction tool generating lower net losses than current practice. It is, ultimately, a range of 'willingness to save' resources at the expense of QALYs forgone. CEACs reported in Appendix 56 show this for all patient subgroups, for which we found no significant differences between genders and a slight trend depending on age. The range of willingness to pay thresholds at which the outcome prediction tool remains cost-effective increases as the age of patients gets higher. Figure 6.7 shows this for female patient subgroups. In other words, if the willingness to save resources at the expense of QALYs forgone progressively drops from £30,000 per QALY, for example, implementing the outcome prediction tool would become cost-effective for older patients first, and then gradually for younger cohorts.

Table 6.5
Mean probabilistic cost-effectiveness results

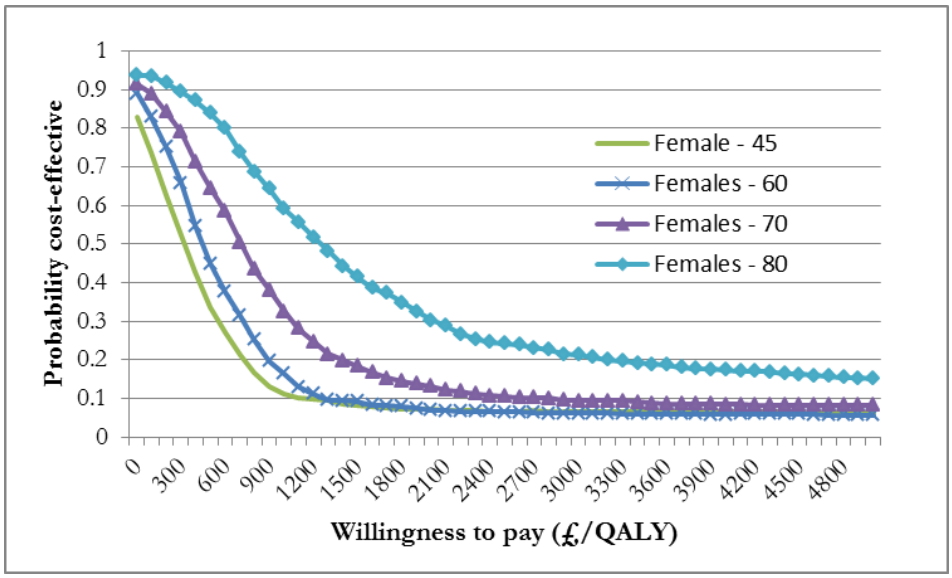
Subgroup (gender, starting age)	Incremental		
	Costs	QALY	**ICER**
Females, 45	- 1,236	- 3.91	**316**
Females, 60	- 1,533	- 3.29	**465**
Females, 70	- 1,720	- 2.47	**695**
Females, 80	- 1,990	- 1.54	**1,295**
Males, 45	- 1,055	- 3.98	**265**
Males, 60	- 1,403	- 3.33	**420**
Males, 70	- 1,581	- 2.34	**674**
Males, 80	- 1,755	- 1.47	**1,195**

Figure 6.6
Cost-effectiveness acceptability curve, Females 70 years of age



Finally, the probabilistic sensitivity analysis allowed for the calculation of the expected value of perfect information (EVPI), a measure of the loss in monetary units of making a decision under uncertainty [82]. Using the NMB approach, we calculated the mean net benefit of maintaining current practice and not implementing the prediction tool, which was the best decision based on the above results, and subtracted that value from the mean NMB of the best alternative, regardless of which it was, for all 1,000 Monte Carlo simulations for females entering the model at 70 years of age. We obtained an EVPI of £1,856, which is to be interpreted as the monetary equivalent of the loss per patient due to the uncertainty surrounding decision making. Combining this value with the expected lifetime of the technology being assessed and the number of patients who can benefit from it, an expected population EVPI could be calculated against which potential research funding could be compared in order to determine whether research aiming to address the uncertainties affecting the model would be a worthwhile investment.

Figure 6.7
Probability of cost-effectiveness of the outcome prediction tool



The sensitivity analyses reported in this chapter addressed only the uncertainty around parameter values, which was considered to be the most relevant for this assessment. Other sources of uncertainty such as methodological or structural were not addressed, but further research could explore whether those other sources of uncertainty could potentially be driving results as obtained with the model and parameter values presented here.

6.4 Contributions

This chapter was entirely developed by Rafael Pinedo.

7 General discussion and the road ahead

Based on the results presented in Chapter 6 as well as on the analyses described in the previous chapters, we discuss next some of the main considerations that we hope will contribute to a better understanding of the economics of THRs in the UK and the potential use of an outcome prediction tool to minimise less than satisfactory outcomes after such a successful operation. We begin by discussing the results from the economic evaluation of the outcome prediction tool for THRs and continue with a reflection on the findings identified during the estimation of model parameter values. We continue with a general discussion about data sources and then explore some of the implications of our model, linking resource use with surgery outcomes. Finally, we consider the potential meaning of our work in light of prediction tools in general as well as their economic evaluations, before closing the chapter with the strengths and limitations of our work and possible lines of future research not detailed in the previous sections of the chapter.

7.1 On the cost-effectiveness of the outcome prediction tool for THR

7.1.1 First known assessment

This assessment is the first recorded economic evaluation of an outcome prediction tool for THR that we know of. The literature review reported in Chapter 2 found economic evaluations of THRs exploring prophylactic interventions, blood donation, transfusions, an important number of studies comparing THR against other surgical and non-surgical interventions, as well as assessments of cost-effectiveness of different prostheses. Evaluations of THR against other surgical methods are the most common amongst the recent assessments, which is consistent with the belief that THRs are unarguably a highly effective and cost-effective intervention. As shown in the literature review, however, little or no attention has been paid to those patients who do not perform well after surgery, those who still feel pain and experience difficulty moving about and carrying out regular activities, and who are therefore understandably dissatisfied. This is the first known assessment exploring the expected costs and benefits, in terms of HRQL, of a tool aimed at reducing the number of such outcomes after THR not only in the UK but the world over.

This economic model joins only two previous ones which explicitly recognise that THR patients are not all similar after the operation and that, as a result, the model structure should acknowledge this with separate model states for the different surgery outcome

groups. As reported in Chapter 4, Chang and colleagues had used ACR functional classification to split outcomes into successful (Class I) or fair (Class 2), the remaining possible transitions being revision within a year and death [19]. This was in 1996, and the following year Saleh and colleagues built three categories into their model based on the HHS: good to excellent, fair, or poor [43]. Although eight years later Kalairajah and colleagues compared the HHS to the OHS and proposed a classification with four categories for the latter [61], no economic evaluation with separate health states reflecting these categories has been published to date. The present assessment provides substantial grounds to support the separation of outcome categories after THR based on significantly different PROMs and primary care costs for the two outcome categories we used in our model.

We found that, based on data from over 68,000 patients with records collected via the PROMs initiative, those scoring above 38 in their OHS one year after surgery were associated to a mean EQ-5D summary score of 0.86 to 0.90 depending on the patient subgroup also a year after the operation, whilst for those patients scoring under 38 the mean value was 0.49 to 0.59. This wide difference was also found when estimating primary care costs after surgery based on data from over 13,000 THR patients as compared to many more controls in the CPRD datasets. By applying a model estimating outcome category based on resource use, we identified mean primary care costs during the first year after surgery for those we termed *Good* outcomes that were lower than similar expenses by controls, whereas patients estimated to be *Poor* outcomes were associated with average costs ranging from £263 to £312 over that spent by the NHS on their respective controls. Data from large and representative datasets confirm that not all THR patients are the same and therefore they must not be pooled into a single model state when performing economic evaluations of THR for which this distinction is relevant.

As reported above and detailed throughout Chapter 5, the economic evaluation presented here used estimations for health utility and cost obtained from large patient-level data sets as parameters values for model states representing outcome categories. This had never been done before, as Chang and colleagues elicited utility values and did not include primary care costs after surgery hence assuming equal costs between their *successful* and *fair* categories [19], whilst Saleh and colleagues did not use health utilities but only probabilities and costs to produce their results, the latter varying by outcome categories

only by assuming that one of them would experience complications but the others would not [43]. Data sets such as HES-PROMs and the combination of CPRD with the COASt cohort allowed us to produce, for the first time, specific and accurate patient-level estimations of costs and health utilities of outcome categories after THR. The benefits of the model structure and data analyses applied for this economic evaluation can furthermore be used for similar assessments of TKRs, for example, for which PROMs are also routinely collected from NHS patients.

7.1.2 No grounds for rationing

The outcome prediction tool for THR developed under COASt would, as intended, reduce the number and proportion of unsatisfactory and *Poor* outcomes after the operation, saving NHS resources in the process. However, the tool would do so at the cost of keeping a number of patients from surgery who would have otherwise improved significantly in their OHS and HRQL, meaning that the tool would also produce less QALYs than current practice. The highest savings per QALY forgone were reported from the oldest patient subgroups (men and women 80 and above) with a reported ICER around £1,200 per QALY, a likely uninteresting alternative for a health care system normally willing to adopt interventions that would generate and not lose QALYs at a cost not exceeding £30,000 per QALY.

The tool described in Chapter 4 would not be a cost-effective alternative for the NHS in the UK. The net benefit analysis performed indicates that the application of the tool is consistently associated to lower net benefits when these are positive for both interventions, and it would only be preferred to current practice when both accept net benefit losses. This is mostly due to the QALYs potentially added by the THR but forgone by patients kept from surgery if the tool were implemented. Keeping patients from surgery, therefore, appears unlikely to be cost-effective for any tool applied to such a highly successful operation, unless the tool is extremely sensitive and specific, to a level that the one assessed here appears not to reach.

In this context, it seems highly unlikely that simple preoperative OHS could direct patients more efficiently than current practice, or even compared to the prediction tool assessed here. Nonetheless, documents such as the 2010/11 South West London Effective Commissioning Initiative [131] suggest that a primary THR should be provided to patients as long as they have a preoperative OHS lower than 26, or if other criteria

involving pain and functional limitation are met. Justification for this specific threshold is not provided, other than a reference about patients with preoperative OHS below 20 achieving the greatest benefit from THR [60], which does not appear clearly stated in the publication nor does it address cost-effectiveness considerations behind the definition of a cut-off point to consider THR. The same criterion was applied by the former Cheshire and Merseyside PCT [132], whilst Derby City and Derbyshire specified a cut-off of 30 or above to fund a primary THR [133], again with no indication of evidence to justify the specific OHS threshold and furthermore pointing in the opposite direction compared to the South West London document, i.e. that THRs should be performed on patients who are not at their worst in pain and mobility. The outcome prediction tool assessed in this evaluation considered not only preoperative OHS but also age, BMI, and a number of environmental and surgical variables to predict OHS at one year after surgery. This prediction model is more comprehensive and appropriate than using merely preoperative OHS to guide the decision about performing a THR or not, and it did not prove cost-effective. Using only preoperative OHS would most likely be associated to even higher net benefit losses than those found for the outcome prediction tool, which suggests that the rationing policy based on OHS should be stopped.

7.1.3 Synergy

We produced cost-effectiveness results for the use of an outcome prediction tool serving as a strict decision maker when it comes to referring patients for a THR or not. This is not necessarily the limit of the tool's cost-effectiveness because instead of stringently deciding whether patients should be referred for a THR or not, the prediction tool could potentially complement the surgeon's assessment and proposed course of action. In fact, a multi-criteria process is a more likely scenario since it is not reasonable to imagine any statistical tool as a substitute for the complex evaluation that orthopaedic surgeons conduct. The significant improvement experienced by THR patients is a testament to the marvels of the procedure, the expertise of surgeons, and their ability to identify patients who would benefit a great deal. The prediction tool, instead, was specifically designed to predict outcome by looking at the evidence from thousands of patients. The proper combination of the two could produce a joint THR surgical assessment capable of producing even more net benefits than current practice.

This research, as many previous ones, found solid ground to continue supporting THRs widely. It remains a fact, however, that many patients are unsatisfied and experiencing

pain and limitations after surgery, in some cases even worse than they had before the operation. Not all such cases are amenable to being prevented given the natural complexities of the operation and the human body, but the number and proportion of those worse outcomes could be reduced. Efforts to develop a more sensitive and specific tool must therefore continue. The outcome prediction tool developed under COASt moved us a great deal closer to understanding and being able to identify when a patient is most likely to perform poorly after a THR. It is possible that a modified assessment by surgeons which considers results from the outcome prediction tool would achieve not only the desired reduction of poor outcomes but also a higher net benefit than current practice.

7.1.4 New tool or new definition of outcome categories?

Figure 6.4 in the previous chapter summarised the main reason why implementing the tool is expected to produce less QALYs than current practice, showing the high and increasing levels of improvement in OHS forgone as the outcome prediction tool's cut-off point rises. The prediction tool simply isn't sensitive and specific enough; or in other words, a THR is just a remarkably effective intervention producing notable increases in the disease-specific outcome measure as well as in a generic HRQL one, even for patients labelled as *Poor* outcomes based on a combination of satisfaction and OHS.

One way forward is to work on improving the statistical tool. Other potential predictors of outcome such as the volume of operations performed in the hospital or the experience of the particular surgeon performing the operation, have previously been found to be associated with outcome not only for hip procedures [60, 134] but also for arthroplasties of the knee[135], and could be included. More complete data not requiring as much imputation of missing values could also be employed in the estimating sample to produce a more accurate tool.

Improving the predictive power of the tool seems necessary for it to achieve better QALY results by keeping from surgery only the small proportion of patients who would not improve, or would do so only slightly, whilst sending all others achieving significant QALY gains through to surgery. The sensitivity analysis conducted around the OHS cut-off point at which the tool would direct patients to THR or to long-term medical management showed that regardless of the cut-off, the prediction tool, as developed, would not be able to achieve better QALY results than current practice (see Table 6.4 in

Chapter 6). It is, therefore, not a matter of calibrating the current prediction tool. A second approach to improve performance of the tool could involve the adjustment of all model input parameters associated to what we termed *Good* and *Poor* outcomes based on OHS threshold of 38 to reflect the various thresholds identified by Arden and colleagues [79] for specific patient subgroups based on gender, age, baseline OHS, BMI and expectations.

Given the significant effectiveness and cost-effectiveness of THRs as they are performed in the UK, we believe that a new description of the outcome group intended to be prevented is the optimal way forward. We furthermore believe that this outcome group should be limited to those patients who do not improve in their OHS or EQ-5D, or who do so only very slightly. Using the postoperative OHS threshold of 38 to distinguish between two outcome categories and employing an outcome prediction tool to prevent patients from falling into the lower scoring group is a waste of potential significant improvements in HRQL. The basis for this OHS threshold was that it was found to be the level that best distinguishes between satisfied and unsatisfied THR patients. Satisfaction does not, however, seem to be a valid proxy for HRQL gain. Figure 5.31 in Chapter 5 showing the notable improvement in EQ-5D summary score of those labelled as *Poor,* and hence likely unsatisfied outcomes, confirms this. A new definition of the outcome category, grouping patients who do not or only hardly improve after the operation, needs to be established and if a prediction tool capable of accurately identifying them can be developed, then we may have found a way to make THRs in the UK even more cost-effective than they already are.

7.2 On findings based on parameter inputs

One of the important strengths of this economic evaluation is that the Markov model employed for the assessment was populated, with the exception of preoperative probabilities, with parameter inputs estimated from large patient-level datasets. The analyses performed on those data sets and the values obtained are informative in their own right.

7.2.1 Demonstrating improvement

We have referenced numerous publications indicating that THRs are highly effective and cost-effective interventions [8, 14, 16-18]. Since the early 1990's researchers have been pointing at this surgical procedure as one producing great benefits at relatively low costs.

But it has not been until now, with the PROMs initiative in place for several years, that we can confidently explore the improvements brought about by this intervention, based on a very large sample progressively approaching the entire population of THRs in the UK.

Based on the PROMs data that we analysed for this assessment, mean OHS improved 19 to 21 points in the period covering 6 months to one year between pre and postoperative measures for all patient subgroups. The lowest mean preoperative score was reported by women 80 years of age or older, who went from 16 before the operation to 36 after their hip had been replaced. The highest starting point was that of men aged between 60 and 70 years, and they went from 20 to 40. In both cases the change was 20 points in the OHS scale, hence the increase for the older women subgroup amounted to 62% of its maximum potential for improvement, whilst for the men in their 60's it was 72%. These were in fact the boundaries of the range of percentage of possible improvement reported for all patient subgroups, hence an impressive progress.

But one of the greatest benefits for economic evaluations is that PROMs also collects EQ-5D data, which reflects the improvement experienced by THR patients in their HRQL. The changes between pre and postoperative measures of the EQ-5D summary score ranged from 0.39 (men aged 45 to 60) to 0.44 (women 45 to 60 and 80 or older), and changes were between 61% and 68% of what was possible, on average, because a large group reported scores of perfect health. This is certainly a remarkable improvement in health utility and achieved in a very short period of time. Coupled with an intervention that is generally very safe, with very low mortality often similar to the age-related all-cause mortality rates, and followed by a rapid rehabilitation, THRs would have to be extremely expensive not to be cost-effective. Better yet, these significant improvements are not exclusive to the entire sample in average terms, and naturally to the *Good* outcomes, but it is also the case for those patients we have labelled as *Poor* outcomes.

7.2.2 The sooner, the better

The model presented here incorporated a long-term medical management arm that essentially worked as a surgery delay mechanism, which in a certain proportion of patients meant that they would not get a THR before they died. This was particularly important because if the outcome prediction tool were to be implemented, it would identify patients likely to perform poorly and those patients would be kept from surgery precisely by placing them in this medical management state. The PROMs data analysed in Chapter 5

and described in the previous paragraphs showed, however, that waiting until the disease affects patients more severely would reduce their improvement.

Figure 5.31 in Section 5.6 illustrated how for both *Good* and *Poor* outcomes the mean EQ-5D summary score increased after surgery, and it also showed that *Poor* outcomes started at a lower EQ-5D score than *Good* outcomes (0.18 vs 0.35 for the lowest scoring patient subgroups) and achieved a smaller improvement (0.25 vs 0.44). Assuming that the disease progresses with time and therefore that the longer patients remain without a replacement the lower their OHS and EQ-5D scores would be, a delay mechanism such as the one implicitly put into place by the outcome prediction tool would reduce the ability of patients to improve. Field and colleagues [60] have already suggested that delaying surgery could make it more difficult for patients to achieve the best possible improvement. At least one economic evaluation comparing THR against watchful waiting was structured assuming the exact opposite, i.e. that patients were to remain in watchful waiting until their quality of life dropped to very low levels [46]. Based on the above evidence, it would be important to perform similar assessments using as comparator a watchful waiting alternative where patients in need for a THR do not wait so long, perhaps until their pain, mobility and quality of life begins to decrease in a sustained manner but not beyond that point.

These findings must be handled with care since they may be viewed as an indication for THR for all OA patients early in their disease stage when it is also a fact that an important number of patients do perform poorly after surgery. The complex prediction tool assessed in this study included a measure of disease progression by incorporating preoperative OHS as one of the predictors, and yet it lacked the necessary accuracy to identify *Poor* outcomes with sufficient sensitivity and specificity to make it a cost-effective intervention. Having a THR when patients are not at their worst may increase the average improvement obtained, but that does not guarantee that *Poor* outcomes will be reduced.

The improvement reported above and shared, albeit in different magnitudes, by the outcome category groups we have called *Good* and *Poor* outcomes, suggest that the term *Poor* lacks accuracy. A more appropriate label for these groups would be *Better* and *Worse* outcomes.

7.2.3 Differentiated rates of revision

We indicated in Section 7.1.1 that our findings in this study support distinguishing between outcome categories when performing economic evaluations for which the clearly different outcome groups are relevant. This is supported even further by the revision rates reported by Rothwell and colleagues [108] for the four different outcome groups suggested by Kalairajah and colleagues [61] where patients with a postoperative score lower than 27 reported a two-year revision rate of 7.6%, compared to 0.5% for those patients who scored above 41. Although equivalent revision rates have not been calculated for the UK, it is sensible to expect a similar pattern whereby worse outcomes have their replacements revised at a significantly higher rate than better outcomes. Given the high cost of revision surgeries, this is yet another good reason to continue working on the development of an outcome prediction tool because by accurately preventing worse outcomes after a primary, it would not only be preventing the higher primary costs during the lifetime of the primary prosthesis, but it would also be preventing the much higher costs of a revision THR.

7.3 Not only good data, but the right data

Reflecting upon the assessment performed, and particularly about the data that was analysed, we found a number of key data issues that would help future research in the area answer questions much more precisely. This research undoubtedly benefited from privileged data. We had access to records on tens of thousands of patients that are representative of the UK population such as those from the CPRD, data very recently collected as is the case of the HES-PROMs, and a short but comprehensive follow up of patients of the COASt cohort. Nonetheless, we were able to identify important data gaps as well as room for improvement in the collection of these data sets that would improve the quality of further research in the field.

7.3.1 Data not being collected

Notwithstanding the valuable insights obtained from the expert elicitation exercise, the justification to resort to expert opinion was the lack of data about what happens to patients before they receive an arthroplasty. We found an abundant literature on joint replacements, both primary and revision procedures, their effects, limitations and costs. However, there is little if anything published about the actual care pathway of patients as their joint pain evolves, about their options, their choices, their opportunity costs. We

were able to learn via the expert elicitation exercise that over the last few years there has been much improvement in the referral of potential THR patients to an orthopaedic surgeon, as most referred patients are thought to be accurately diagnosed as candidates for a THR. But we also noticed, based on the variability of experts' responses, that these improvements are far from uniform across regions and even amongst hospitals of the same region. It is possible that the significant success of THR interventions might have drawn most of the attention to the procedure and its nearly miraculous effects, but in order to take the most advantage from such a successful operation the care pathway of patients before they reach the operation theatre must not be ignored.

7.3.2 No THR, no follow-up

A closer follow-up of patients with growing joint pain and reduced mobility but who did not have a THR is needed. This will help tackle the problem of the lack of counterfactual when assessing THRs against not performing any other intervention. As the cost-effectiveness of hip replacement has been repeatedly proven to be worth the investment, assessing the cost-effectiveness of arthroplasty has involved comparing the costs and effects of having the joint replaced against often unreasonable assumptions about what is expected to happen to patients if they do not get a replacement. The most recent publication reporting a cost-effectiveness analysis of THR against no surgery automatically assumed the hypothetical latter group to have health utility levels maintained over time and equal to the preoperative level of those who did have the operation performed, whilst costs were simply assumed to be zero [136]. As our research shows, primary care costs of patients both before and after a THR are non-zero and certainly not insignificant, especially for those with worse outcomes. Health utility levels, nonetheless, were also assumed to remain the same and at the preoperative level for our model because of a lack of follow-up of those patients.

As there are patients who never get a replacement, and the use of the prediction tool might increase that number, it would be important to move away from assumptions and populate cost-effectiveness models with actual data on the costs and QALYs associated to patients who were treated without the surgical intervention. This will require incorporating those patients who do not get joint replacements into studies where they are followed-up over time, their use of health care resources recorded and estimates of their HRQL collected.

7.3.3 A large prosthesis market

Over 100 different brands of acetabular cups and more than 140 brands of femoral components were used in the UK during 2011 [54]. These components of a THR can furthermore be fixed with cement, without cement, or with a combination thereof (hybrid), with an additional classification by head size (varying between 22 and 60 mm) and bearing surface (with different combinations of metal, ceramic and polyethylene). As a result, to speak of a THR in general terms as we have done for this assessment means that we did not make any distinction between the significant number of combinations of components and types of THR, all of which are associated to different prosthesis survival rates [54]. We intended to incorporate specific revision rates by fixation type which are reported to the NJR, but regrettably our request for the data was denied.

Nevertheless, having access to these data is essential not only to refine economic evaluations such as this one, but also to explore the effects that they may have on outcome after surgery. In their 8th annual report, the NJR reported that 935 different combinations of acetabular cups and femoral components had been used in the seven-year period during which the Registry had been collecting data. Of those, at least 20 had been used on 2,500 patients or more, reporting five-year revision rates as far apart as 0.58% (CI 0.42%-0.79%) for the Exeter V40 with Elite Plus Ogee (13,000 patients) compared to 3.6% (CI 2.72%-4.76%) for the SL-Plus Cementless Stem with Exceed (3,500 patients) [83]. Figures for the following year were not reported in the 9th annual report. As the NICE technology appraisal issued in 2000 allows the use of prostheses for primary THR that have at least three years of evidence indicating they can last for 10 years or more, research on comparative performance of prosthesis brands is of paramount importance. The above evidence on differential survival of the prostheses and the significant difference in prosthesis costs [53] support further research specific by prosthesis type, something Pennington and colleagues have recently started to address with a cost-effectiveness analysis of THR by fixation type in 2012 [15].

We understand that requesting brand details would have made our application a sensitive one given the particular interests of patients, surgeons, manufacturers, the NJR and the NHS in general. However, if the Registry is to contribute to the production of research that can potentially inform and shape policy for joint replacements in the UK, it should allow (subject to all corresponding quality checks) access to data such as fixation type and even prosthesis brand that they collect from hospitals, so that research can move forward.

In their 9th annual report, the NJR indicated the details of a series of research requests using data from the Registry among which there is one approved for the comparative cost-effectiveness of the most commonly used THR brands and types.

7.3.4 Health-related quality of life

We have also found room for improvement in the estimation of health utilities for hip patients. The commonly used EQ-5D measure has been criticised for its ceiling effects (low sensitivity at values close to one) [137, 138] as well as for having response levels too broad to make it sensitive to small changes in health [138, 139]. These problems largely affect the responses provided by hip patients and their subsequent health utility estimate as many patients may be experiencing small, but clear, changes in their HRQL that are not recorded when only three possible levels of severity are provided. Likewise, a significant proportion of postoperative patients report an EQ-5D summary index of one (perfect health) likely because their experience in pain or mobility limitations, even if not perfect, are closer to "no problem" than to the next response level, "some problems". In fact, a number of the EQ-5D forms completed by patients in the COASt cohort study in Oxford could not be used because patients were unable to choose a single response in one or more dimensions. Some patients marked two different levels, probably indicating that their health state would be best described by an intermediate level between the two selected, whilst others did not mark any of the options and instead handwrote their own response, for example, "minor problems", because none of the options available was close enough to it.

These occasional missing data as well as a the low sensitivity of the EQ-5D measure in some cases could potentially be minimised when the new five-level EQ-5D questionnaire and valuation weights are introduced [140]. The new measure will offer respondents five different levels of severity to describe their health in the usual five dimensions. Although the new EQ-5D-5L utility weights have not been presented yet and further research is still needed to explore its higher sensitivity and diminished ceiling effects, it appears to be a promising development for the estimation of health utility in general, and for THR patients in particular.

7.3.5 Missing data on HES-PROMs

Through the HES-PROMs we had access to invaluable patient-level HRQL data for a number of patients, unusually large compared to what most economic evaluations are able

to use. In fact, given this important number of observations, we decided to base our analysis on complete cases, leaving aside a non-negligible number of records with missing data. We are therefore confident about overall mean values estimated based on such a large sample being unbiased, but we also believe that the majority of missing values are likely not missing at random. Although missing values were a small proportion of the total number of observations, given that our analysis made a distinction between better and worse outcomes and that values were estimated for eight different patient subgroups, it would be important to explore the effects on results of addressing missing values by multiple imputation, for example.

The largest number of observations discarded because of missing data was that relating to a missing value on the link variable between the HES and PROMs separate data sets. This involved over 43,000 PROMs records. Since the missing variable is not one completed by the patient, it is more likely that it is not missing for a particular patient subgroup or outcome category and possibly affecting a particular hospital, region or range of dates. Nonetheless, all efforts directed at reducing the number of missing data will improve the quality of analyses performed with it. The PROMs initiative is in its initial stages and it would benefit greatly by exploring ways to better identify the sort of patients that are not returning forms, possible reasons for it, and especially the data linkage process so that once collected data can be cleaned, processed, and made available for analysis.

7.3.6 Long-term follow-up

Finally, as joint replacements are interventions that impact patients for a long time and revision surgeries have been found to be important drivers of cost-effectiveness [45], access to long term follow up data on THR patients is essential. The Swedish Hip Arthroplasty Register has been following up patients since the late 1970's and they have also been collecting HRQL data since 2002 [141]. This is a good example for the UK to follow, where the main commitment must be maintaining the collection of data over time; not only regarding the failure of prosthesis but also patient-reported outcomes, prosthesis types and brands, details about the hospital where the procedure was performed as well as the surgeons involved, and crucially socio-demographic information about the patients.

Important additions to the information collected would be all likely determinants of outcome such as stage of disease progression, diagnoses, coexistent conditions, and previous treatment received. In terms of health care use, it would be important to achieve

high degrees of effective linkage between the clinical follow-up and hospital as well as primary care records before and after the operation. In the case of the UK, this would mean building and maintaining links between an extended version of the PROMs initiative with the records being collected by the NJR, the NHS hospital episode records, the NHS outpatient data, and the new CPRD. Given that between 20% to 25% of THRs are performed privately, of which about half are privately funded [83], links from the NJR data to the corresponding records in the private sector will contribute to building a most complete database of relevant information about joint replacements in the country.

Efforts made by the UK in the direction of improving the data collected to evaluate THRs and joint replacements in general are noteworthy. The establishment of the NJR in 2002 was a major first step, followed by including hip and knee replacements as two of the four interventions for which PROMs are systematically collected as a measure of treatment outcome and quality of care in the NHS. Although the national initiative only involves one preoperative and one postoperative measure six months after surgery, the NJR has begun a project extending the follow up period for hip and knee replacement patients by sending PROM forms to 35,000 patients in England at one, three and five years after surgery [142]. These initiatives, combined with the measures described above, will make an important contribution to building a solid body of data which, available to researchers, will help shape policy on THR and joint replacement surgery for the benefit of patients and the efficiency of the health care system.

7.4 On the connection between resource use and outcome

The CPRD was an extraordinary source of resource use data, with records on patients and controls stretching for more than 10 years before and after the operation, and with monumental details on consultations and medication prescription. However, we needed estimates of resource use by outcome category and the lack of outcome measures in the CPRD posed a major challenge for the research. One of the main contributions of this work has been the development of a model linking resource use to THR outcome. This was done by quantifying resource use by participants in the COASt cohort who had completed a one-year follow-up form, estimating a logistic model of surgery outcome based on patient's resource use, and fitting that association to the much larger and representative CPRD. The link between resource use and surgery outcome is strong and clear: if patients are still feeling pain, having difficulties performing their regular activities, are not well in general as a result of the operation, they will seek out their GP and other

health care professionals at the primary care level and will likely be prescribed medication to alleviate their condition. The demand for health care resource reveals the outcome of surgery and we have no reason to believe that when such demands are legitimate, they will not lead to the provision of services that are subsequently reported in the GP practice records we analysed.

By developing this model, we were effectively able to analyse and estimate the resources used not only by all patients receiving a THR, but also by those patients likely to have had a better outcome and those likely to have had a worse outcome, all based on a data set lacking data on traditional outcome measures. There is room for improvement of this model as more patients in the COASt cohort complete their follow-up and as this follow-up goes beyond one year, because the trend of resource use over more than 12 months may prove a better predictor of outcome category than only that of the first year. More importantly, if the general state of the hip patient after surgery can be estimated based on their resource use, then it is likely that the progression of the disease can also be mapped out by looking at the use of resource over time even before the patient requires surgery. Therefore, by expanding this model, resource use could potentially contribute to monitoring patients' disease progression and making more informed decisions along the care pathway.

7.5 On prediction tools and their economic evaluations

Economic evaluations of health care interventions answer a question of cost-effectiveness. When results by subgroups are explored then specific allocative efficiency within the same intervention is also analysed. When results by subgroups are presented to the decision maker with significant differences in results by subgroups, it is often the case that the intervention would be recommended and funded for those patient subgroups reporting a favourable ICER, leaving others with weaker evidence or results surpassing the accepted threshold outside the recommendation for funded treatment. The case of treatment for Alzheimer in the UK is an example of an intervention funded for some subgroups and not others at first (NICE's Technology Appraisals 19 and 111), later revised to modify the inclusion criteria [143]. Although the decision of implementation can be selective, there will be a number of patients within those subgroups for which the treatment is recommended for whom the treatment would not be cost-effective. Likewise, for many patients in the groups not reaching the cost-effectiveness threshold the intervention would have produced results at a cost that would have made it cost-effective

if considered individually. This is due to the heterogeneity necessarily found within patient subgroups for which cost-effectiveness analyses are performed because we are not able to conduct these assessments at patient-level. Outcome prediction tools, however, may bring economic evaluations to a level considerably close to the individual. The cost-effectiveness analysis of implementing a prediction tool can, in theory, produce results that allow identifying the very specific groups for whom the intervention would be cost-effective based on the predictors of outcome included in the model. Though not strictly individualised care, outcome prediction tools can be an important step in the direction of attaining the maximum possible health outcome by identifying more precisely the patients for whom the intervention would produce the greatest benefit by unit of cost.

Because outcome prediction tools such as the one assessed here are based on patients' relevant baseline characteristics, it seems likely that patient-level simulations instead of cohort models would be a more appropriate approach to assess their cost-effectiveness. Employing a patient-level simulation is likely to provide more accurate results for individual patients about the cost-effectiveness of implementing the prediction tool, which would help define the subgroups for which implementation would provide the greatest net benefits. The usual difficulty performing patient-level simulation is access to appropriate and sufficient patient-level data, but as discussed above this is becoming less of a problem for joint replacements in the UK.

Whether through patient-level or cohort models, the economic evaluation of a prediction tool would also benefit greatly from an alternative approach to establishing the optimal cut-off point of the tool. In this study the definition of outcome categories was based on the cut-off point that maximised sensitivity and specificity as well as the area under the ROC curve of OHS predicting satisfaction [79]. But the cut-off point needs not be defined in this way. Lakin and colleagues [144] explored the concept of selecting cut-off points for diagnostic tests from the area of the ROC plane that would make the test cost-effective instead of the usual criteria of maximising sensitivity and specificity. By doing this, the calibration of the outcome prediction tool, which produces a similar ROC curve when the outcome variable is binary, can be done considering cost-effectiveness at the earlier and, we would argue, more appropriate stage of finding the tool's cut-off point. As opposed to conducting an economic evaluation after the cut-off point has been established, possibly leading to a suboptimal result, selecting the cut-off point whilst considering sensitivity, specificity and cost-effectiveness simultaneously could allow for

the identification of a trade-off in sensitivity against specificity falling within the range of cost-effectiveness which otherwise would not be known but which would be cost-effective. With the tool assessed here, results made it clear that the ROC curve does not cross the cost-effectiveness area; however, for a modified prediction tool, it may.

7.6 Strengths and limitations

The main strengths and limitations of this work have been reported in each chapter. In general, this research benefited from a full review of the published literature on the economic evaluations of THRs and from the best available sources of data to populate a cost-effectiveness model. First, the only source of data not based on patient-level records was an expert elicitation exercise reported in a fully comprehensive manner in Chapter 5. When expert opinion has been used in similar previous assessments, the details about how the elicitation was conducted were not reported [19, 43]. For our economic evaluation, every step of the process of collecting and synthesising experts' judgement was copiously described.

Apart from the expert judgements, all other sources of data consisted of patient-level data sets with the most appropriate, representative and up-to-date information on the probabilities, health utility and resource use associated to THRs in the UK, both before and after the operation. HES-PROMs, CPRD, EPOS and the COASt cohort provide the best data on hip arthroplasty in the UK, and the only model parameter estimated based on data from elsewhere was revision rates by outcome category, which were published on data from New Zealand but ultimately adjusted to match the UK's overall revision rate and relative sizes of the outcome groups.

Finally, the level of detail provided by the above data sources allowed estimating model parameter values for patient subgroups by age and gender. This made it possible not only to present results separately by these subgroups, but critically it also allowed adjusting all parameter values in the model so that not only death rates but almost all other parameters changed in the simulation as patients became older. If results are only as good as the data feeding the model, then those produced by this research are results in which we can have great confidence.

But no research is exempt from shortcomings. First of all, the intervention that was assessed with our economic evaluation has not been implemented yet. Although the final

work package of the COASt research programme involves a validation of the prediction tool on the cohort of patients recruited from Oxford and Southampton hospitals, our study was performed assuming that the results of the tool would be those of its internal validation. Although an internal validation would generally be associated with better results than external ones, the prediction tool was estimated after merging large datasets and performing a substantial imputation of values that were missing or simply not collected. As a result, a proper validation of the tool is fundamental to obtaining more robust results around the expected cost-effectiveness of its implementation.

As reported in the section detailing the assumptions made by the model's structure and parameterisation, there are a number of limitations that, although clinically feasible, constrain results. Although the expert elicitation was conducted with a sound methodology and the frequency of convergent results speak of understanding of the process by and agreement amongst surgeons, a validation of those values at a national level would improve the model's robustness. Revision rates by outcome groups were adjusted from those reported on patients from New Zealand whilst equivalent values can now be produced for the UK thanks to the PROMs initiative. The lack of long-term follow up of patients who do not receive a THR and even of those who do have their hips replaced for both primary and revision operations forced us to make a number of assumptions that, if replaced by evidence, would improve reliability of results. Further research can focus on these limitations.

7.7 A final research recommendation

Some of the former PCTs in England were using BMI thresholds for THR referrals, up until PCTs ceased to exist with the introduction of the new structure of the health care system in England in April of 2013 [145]. BMI thresholds of 25 [87, 88], 30 [88] and 35 kg/m^2 [89, 90] were defined as a basis to encourage weight reduction before referral for THR. It is not clear whether the newly formed clinical commissioning groups will continue applying these criteria to ration THRs but, as with OHS thresholds, they lack appropriate economic evaluations. We originally intended to include BMI as one of the defining criteria for the patient subgroups in our analysis, but were not able to do so because height and weight were available in only about 40% of CPRD records. BMI was also unavailable in the HES-PROMs dataset, a limitation that disappears if records are linked to the NJR as they do collect height and weight measures from hospitals performing THRs. Although our economic evaluation focused on the application of an

outcome prediction tool and did not include BMI groups in the analysis, it did show that current practice of THRs in the UK is remarkably cost-effective and therefore suggests that any rationing such as the one possibly still in place based on BMI must be carefully reviewed as it may be denying a significant improvement in health to patients and an opportunity to invest health care resources in a very cost-effective manner.

7.8 Contributions

This chapter was entirely developed by Rafael Pinedo.

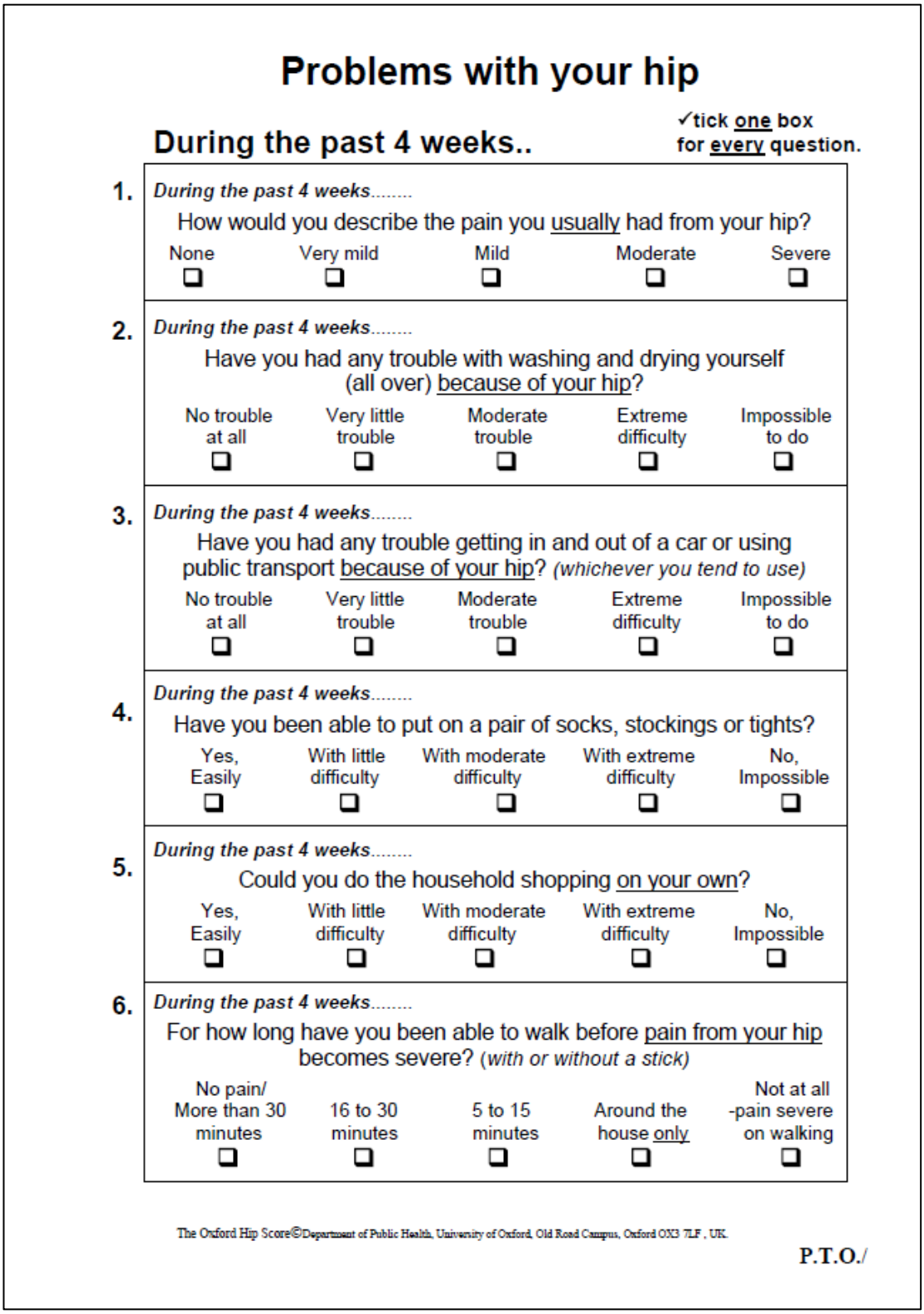
Appendices

Appendix 1 – Publications not meeting eligibility criteria

Authors	Year	Title	Reason for ineligibility
Daellenbach, et al	1990	Economic appraisal of new technology in the absence of survival data--the case of total hip replacement.	An appraisal without any empirical data
Birkmeyer, et al	1993	The cost-effectiveness of preoperative autologous blood donation for total hip and knee replacement.	Assessment of blood donation
Healy, et al	1994	Preoperative autologous blood donation in total-hip arthroplasty: A cost- effectiveness analysis	Assessment of blood donation
O'Brien, et al	1994	Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement.	Assessment of deep vein thrombosis (DVT) prophylaxis
Menzin, et al	1995	Cost-effectiveness of enoxaparin vs low-dose warfarin in the prevention of deep-vein thrombosis after total hip replacement surgery.	Assessment of DVT prophylaxis
Gillespie, et al	1995	Evaluation of new technologies for total hip replacement. Economic modelling and clinical trials	Not a full evaluation: present value
Sarasin, et al	1996	Antithrombotic strategy after total hip replacement: A cost-effectiveness analysis comparing prolonged oral anticoagulants with screening for deep vein thrombosis.	Assessment of DVT prophylaxis
Pynsent, et al	1996	The total cost of hip-joint replacement; a model for purchasers.	Not a full evaluation: costs only
bdool-Carrim, et al	1997	The cost and benefit of prophylaxis against deep-vein thrombosis in elective hip replacement.	Assessment of DVT prophylaxis
Detournay, et al	1998	Cost effectiveness of a low-molecular-weight heparin in prolonged prophylaxis against deep vein thrombosis after total hip replacement.	Assessment of DVT prophylaxis
Levin, et al	1998	Economic evaluation of desirudin vs heparin in deep vein thrombosis prevention after hip replacement surgery.	Assessment of DVT prophylaxis
Hawkins, et al	1998	A pharmacoeconomic assessment of enoxaparin and warfarin as prophylaxis for deep vein thrombosis in patients undergoing knee replacement surgery.	Assessment of surgical prophylaxis
Francis, et al	1999	A pharmacoeconomic evaluation of low-molecular-weight heparin in patients after total hip-replacement surgery.	Assessment of DVT prophylaxis
Marchetti, et al	1999	Long-term cost-effectiveness of low molecular weight heparin versus unfractionated heparin for the prophylaxis of venous thromboembolism in elective hip replacement.	Assessment of DVT prophylaxis
Nicolaides, et al	1999	Cost-effectiveness of desirudin in the prevention of the thromboembolic complications of surgery.	Assessment of DVT prophylaxis
Baxter, et al	1999	An economic model to estimate the relative costs over 20 years of different hip prostheses.	Not a full evaluation: costs only
Sonnenberg, et al	1999	The cost-effectiveness of autologous transfusion revisited: Implications of an increased risk of bacterial infection with allogeneic transfusion.	Assessment of autologous transfusion
Davies, et al	2000	Economic evaluation of enoxaparin as postdischarge prophylaxis for deep vein thrombosis (DVT) in elective hip surgery.	Assessment of DVT prophylaxis
Jackson, et al	2000	The cost-effectiveness of postoperative recovery of RBCs in preventing transfusion-associated virus transmission after joint arthroplasty.	Assessment of blood conservation technique
Wade, et al	2000	Cost effectiveness of outpatient anticoagulant prophylaxis after total hip arthroplasty.	Assessment of anticoagulant prophylaxis
Friedman, et al	2000	Cost analyses of extended prophylaxis with enoxaparin after hip arthroplasty.	Assessment of DVT prophylaxis
Levin, et al	2001	Cost effectiveness of desirudin compared with a low molecular weight heparin in the prevention of deep vein thrombosis after total hip replacement.	Assessment of DVT prophylaxis

Authors	Year	Title	Reason for ineligibility
Botteman, et al	2002	Results of an economic model to assess the cost-effectiveness of enoxaparin, a low-molecular-weight heparin, versus warfarin for the prophylaxis of deep vein thrombosis and associated long-term complications in total hip replacement surgery in the United States.	Assessment of DVT prophylaxis
Sarasin, et al	2002	Out of hospital antithrombotic prophylaxis after total hip replacement: low-molecular-weight heparin, warfarin, aspirin or nothing? A cost-effectiveness analysis.	Assessment of DVT prophylaxis
Sonnenberg, et al	2002	A health economic analysis of autologous transfusion.	Assessment of autologous transfusion
Gordois, et al	2003	The cost-effectiveness of fondaparinux compared with enoxaparin as prophylaxis against thromboembolism following major orthopedic surgery.	Assessment of DVT prophylaxis
Lundkvist, et al	2003	Cost-effectiveness of fondaparinux vs. enoxaparin as venous thromboembolism prophylaxis in Sweden.	Assessment of DVT prophylaxis
Caprini, et al	2003	Cohen, A. T. (2003). Economic burden of long-term complications of deep vein thrombosis after total hip replacement surgery in the United States.	Assessment of DVT prophylaxis
Spiegelhalter, et al	2003	Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling.	Used THR as example in application of statistical methods
Bozic, et al	2004	Economic evaluation in total hip arthroplasty: analysis and review of the literature.	A literature review not including an economic evaluation
Haentjens, et al	2004	Prolonged enoxaparin therapy to prevent venous thromboembolism after primary hip or knee replacement. A cost-utility analysis.	Assessment of DVT prophylaxis
Bjorvatn, et al	2005	Fondaparinux sodium compared with enoxaparin sodium: A cost-effectiveness analysis.	Assessment of DVT prophylaxis
Bischof, et al	2006	Cost-effectiveness of extended venous thromboembolism prophylaxis with fondaparinux in hip surgery patients	Assessment of DVT prophylaxis
Straumann, et al	2006	Cost-benefit analysis of MIS THA: Model-based analysis of the consequences for Switzerland	Not a full evaluation: costs only
Kurtz, et al	2007	Future clinical and economic impact of revision total hip and knee arthroplasty	Broadly explores economic impact of THRs
Graves, et al	2007	Economics and preventing hospital-acquired infection: Broadening the perspective.	Prophylaxis: Prevention of infection
Skedgel, et al	2007	The cost-effectiveness of extended-duration antithrombotic prophylaxis after total hip arthroplasty.	Assessment of DVT prophylaxis
Cranny, et al	2008	A systematic review and economic model of switching from non-glycopeptide to glycopeptide antibiotic prophylaxis for surgery.	Assessment of surgical infection prophylaxis
Wolowacz, et al	2008	Cost-effectiveness of venous thromboembolism prophylaxis in total hip and knee replacement surgery: the evolving application of health economic modelling over 20 years.	Assessment of DVT prophylaxis

Appendix 2 – Oxford Hip Score questionnaire



Problems with your hip

During the past 4 weeks..

✓tick one box for every question.

1. *During the past 4 weeks.....*

How would you describe the pain you usually 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) because of your hip?

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 hip? *(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 on your own?

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 pain from your hip 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 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 one box for every 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 because of your hip?

Not at all painful	Slightly painful	Moderately painful	Very painful	Unbearable
❑	❑	❑	❑	❑

9 *During the past 4 weeks.....*

Have you been limping when walking, because of your hip?

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, severe pain - 'shooting', 'stabbing' or 'spasms' - from the affected hip?

No days	Only 1 or 2 days	Some days	Most days	Every day
❑	❑	❑	❑	❑

11 *During the past 4 weeks.....*

How much has pain from your hip 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 pain from your hip in bed at night?

No nights	Only 1 or 2 nights	Some nights	Most nights	Every night
❑	❑	❑	❑	❑

Appendix 3 – EQ-5D questionnaire

EQ-5D – English version for the UK

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems in walking about ❑

I have some problems in walking about ❑

I am confined to bed ❑

Self-Care

I have no problems with self-care ❑

I have some problems washing or dressing myself ❑

I am unable to wash or dress myself ❑

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 ❑

Pain/Discomfort

I have no pain or discomfort ❑

I have moderate pain or discomfort ❑

I have extreme pain or discomfort ❑

Anxiety/Depression

I am not anxious or depressed ❑

I am moderately anxious or depressed ❑

I am extremely anxious or depressed ❑

Appendix 4 – Detailed explanation of EQ-5D summary scores

The EQ-5D questionnaire asks respondents to describe their health in five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) by marking one of three possible levels (1=no problems, 2=some problems, 3=extreme problems). Responses to these five questions constitute a descriptive health state, such that someone reporting no problems in any of the first four dimensions but some problems with anxiety or depression, for example, would be associated to the descriptive health state 11112. The combination of five dimensions at three possible levels means that there are 243 possible descriptive health states in total. Each of these health states can be translated into a unique summary index anchored in 1 being "perfect health" and 0 being "death" by applying a preference-based valuation set. We used the set derived from a sample of the UK general population.

This valuation set allows assigning an index that reflects the preferences of the general population between all states. For instance, state 11112 under the UK valuation set would be assigned a summary score of 0.848, whilst health state 11121 (some problems in pain/discomfort but no problems in the other dimensions) would be assigned 0.796. This suggests that, though there is only a slight difference between those generic health state summary scores, people in the UK would generally prefer having some problems with anxiety or depression (higher generic health score) than some problems with pain or discomfort. The valuation set also allows for negative summary scores, indicating health states that would be considered worse than death. This is the case, for example, of descriptive state 22233, which would produce a score of -0.181, or state 33333, the worst possible state with severe problems in all five dimensions, associated to a summary score of -0.594. Conversely, the maximum score attainable is 1, which can only be obtained when the respondent reports having no problem in all five dimensions, hence the common term "perfect health".

Appendix 5 – Multicollinearity check

Variance Inflation Factor (VIF) and R^2 among OHS questions

Mean VIF	Description of pain	Night pain	Sudden pain	Limping	Walk duration	Climb stairs	Socks	Rise chair	Car	Washing up	House shopping	Pain and work
Description of pain	-	2.24	1.96	2.20	1.64	1.67	1.44	2.53	2.01	1.47	1.36	2.92
Night pain	-	-	1.87	1.65	1.42	1.52	1.33	1.91	1.65	1.35	1.26	2.02
Sudden pain	-	-	-	1.57	1.42	1.46	1.27	1.78	1.50	1.31	1.22	1.91
Limping	-	-	-	-	1.51	1.59	1.42	1.86	1.75	1.41	1.35	2.28
Walk duration	-	-	-	-	-	1.64	1.34	1.57	1.52	1.34	1.55	1.80
Climb stairs	-	-	-	-	-	-	1.59	1.81	2.05	1.60	1.82	1.94
Socks	-	-	-	-	-	-	-	1.54	1.72	1.82	1.50	1.52
Rise chair	-	-	-	-	-	-	-	-	2.19	1.56	1.41	2.45
Car	-	-	-	-	-	-	-	-	-	1.78	1.64	2.18
Washing up	-	-	-	-	-	-	-	-	-	-	1.46	1.61
House shopping	-	-	-	-	-	-	-	-	-	-	-	1.62
Pain and work	-	-	-	-	-	-	-	-	-	-	-	-

Max:	2.92

R^2	Description of pain	Night pain	Sudden pain	Limping	Walk duration	Climb stairs	Socks	Rise chair	Car	Washing up	House shopping	Pain and work
Description of pain	-	0.55	0.49	0.55	0.39	0.40	0.30	0.61	0.50	0.32	0.27	0.66
Night pain	-	-	0.47	0.39	0.30	0.34	0.25	0.48	0.39	0.26	0.21	0.50
Sudden pain	-	-	-	0.36	0.30	0.32	0.21	0.44	0.34	0.24	0.18	0.48
Limping	-	-	-	-	0.34	0.37	0.29	0.46	0.43	0.29	0.26	0.56
Walk duration	-	-	-	-	-	0.39	0.25	0.36	0.34	0.26	0.35	0.44
Climb stairs	-	-	-	-	-	-	0.37	0.45	0.51	0.38	0.45	0.48
Socks	-	-	-	-	-	-	-	0.35	0.42	0.45	0.33	0.34
Rise chair	-	-	-	-	-	-	-	-	0.54	0.36	0.29	0.59
Car	-	-	-	-	-	-	-	-	-	0.44	0.39	0.54
Washing up	-	-	-	-	-	-	-	-	-	-	0.32	0.38
House shopping	-	-	-	-	-	-	-	-	-	-	-	0.38
Pain and work	-	-	-	-	-	-	-	-	-	-	-	-

Max:	0.66

Appendix 6 – Coefficients for Response Mapping model

[Part I]

Coefficients for Response Mapping model by EQ-5D dimension and response level†

OHS question: response level	Mobility		Self-care		Usual Activities	
	1	3	2	3	1	3
Description of pain: 0	Base case		Base case		Base case	
Description of pain: 1	0.270	-12.538 *	0.069	0.684	0.228	0.022
Description of pain: 2	0.239	17.937 *	0.495 *	0.505	0.009	0.028
Description of pain: 3	0.506	6.363	1.047 *	-15.188 *	-0.186	-0.387
Description of pain: 4	0.406	10.411	1.086 *	1.337	0.034	-0.236
Night pain: 0	Base case		Base case		Base case	
Night pain: 1	0.258	0.327	-0.145	-0.382	0.244	0.198
Night pain: 2	-0.206	-26.334 *	0.108	-1.002	0.185	-0.149
Night pain: 3	0.034	-13.243 *	0.305	-15.313 *	0.684 *	0.520
Night pain: 4	-0.120	-13.476 *	0.216	0.674	0.560	0.560
Sudden pain: 0	Base case		Base case		Base case	
Sudden pain: 1	-0.957 *	-0.789	0.488 *	0.914	-0.003	-0.097
Sudden pain: 2	-0.539	-14.095 *	0.182	-0.360	-0.232	-0.009
Sudden pain: 3	-0.742 *	-4.900	0.325	0.875	-0.673	-1.205 *
Sudden pain: 4	-0.716 *	-29.959 *	0.158	-0.366	-0.374	-0.038
Limping: 0	Base case		Base case		Base case	
Limping: 1	0.516 *	2.308	-0.075	-0.149	0.198	-0.212
Limping: 2	0.651 *	16.388 *	0.013	0.072	0.104	-0.544
Limping: 3	1.026 *	10.119 *	-0.099	-2.124	0.534 *	-0.336
Limping: 4	1.538 *	32.265 *	-0.004	-0.460	0.710 *	0.129
Walking duration: 0	Base case		Base case		Base case	
Walking duration: 1	-0.750	0.442	-0.126	0.931	-0.150	0.299
Walking duration: 2	-0.662 *	-13.361 *	0.013	-0.285	-0.232	0.219
Walking duration: 3	-0.136	-1.323	-0.340	-0.066	-0.051	-0.468
Walking duration: 4	0.365	17.524 *	-0.095	-0.786	0.042	-0.199
Climbing stairs: 0	Base case		Base case		Base case	
Climbing stairs: 1	-0.672	-0.926	0.233	0.403	-0.762	-0.390
Climbing stairs: 2	-0.373	-17.256 *	0.229	-0.964	-0.488	-0.690 *
Climbing stairs: 3	0.380	-11.117	-0.058	-1.480	-0.003	-0.505
Climbing stairs: 4	0.819 *	-10.422 *	-0.248	-1.547	0.398	-0.677
Socks and stockings: 0	Base case		Base case		Base case	
Socks and stockings: 1	0.393	-1.460	-0.350	-1.374 *	0.082	-0.452 *
Socks and stockings: 2	0.035	-28.254 *	-0.928 *	-1.029	-0.009	-0.376
Socks and stockings: 3	0.261	-2.077	-1.628 *	-2.364 *	0.047	-0.305
Socks and stockings: 4	0.735 *	7.397	-2.580 *	-1.118	0.256	-0.395
Pain from standing up from chair: 0	Base case		Base case		Base case	
Pain from standing up from chair: 1	0.718	-2.117	-0.462	-1.770 *	0.780	-0.400
Pain from standing up from chair: 2	0.445	-14.710 *	-0.341	0.553	0.788	-0.367
Pain from standing up from chair: 3	0.494	-16.799 *	-0.409	0.931	0.496	-0.616
Pain from standing up from chair: 4	0.400	-32.046 *	-0.329	1.235	0.305	-0.443
Car and public transport: 0	Base case		Base case		Base case	
Car and public transport: 1	0.152	1.356	-0.135	-0.339	-0.768	0.095
Car and public transport: 2	-0.209	-13.786 *	-0.204	-1.925 *	-1.223	0.113
Car and public transport: 3	0.101	-23.704 *	-0.454	-3.295	-0.657	-0.754
Car and public transport: 4	0.327	-18.237 *	-0.475	-2.200	-0.146	-0.018
Washing and drying: 0	Base case		Base case		Base case	
Washing and drying: 1	-1.340 *	0.904	1.034 *	-1.894 *	-0.543	0.504
Washing and drying: 2	-0.988	-28.771 *	-0.150	-3.159 *	-0.174	0.098
Washing and drying: 3	-1.000	-12.552 *	-1.161 *	-4.903 *	0.012	0.157
Washing and drying: 4	-1.079	-13.849 *	-2.183 *	-2.858 *	0.299	0.028
House shopping: 0	Base case		Base case		Base case	
House shopping: 1	0.255	-32.816 *	-0.090	-0.433	0.005	-0.634 *
House shopping: 2	0.443	-12.410 *	-0.304 *	-1.662	-0.199	-1.030 *
House shopping: 3	0.840 *	1.661	-0.652 *	-0.897	0.026	-0.913 *
House shopping: 4	1.659 *	-16.399 *	-1.031 *	-3.012 *	0.888 *	-1.109 *
Pain interfering work: 0	Base case		Base case		Base case	
Pain interfering work: 1	-0.369	0.363	-0.280	-1.262	-0.157	-1.401 *
Pain interfering work: 2	0.450	11.253 *	-0.206	-1.594	-0.282	-2.270 *
Pain interfering work: 3	0.807	-4.422	-0.295	0.956	0.865	-2.383 *

OHS question: response level	Mobility		Self-care		Usual Activities	
	1	3	2	3	1	3
Pain interfering work: 4	1.312 *	-0.642	-0.717 *	-0.148	2.174 *	-1.717 *
Constant	-3.041 *	-2.543	2.470 *	2.995 *	-2.602 *	1.637 *

[Part II]

Coefficients for Response mapping model by EQ-5D dimension and response level†

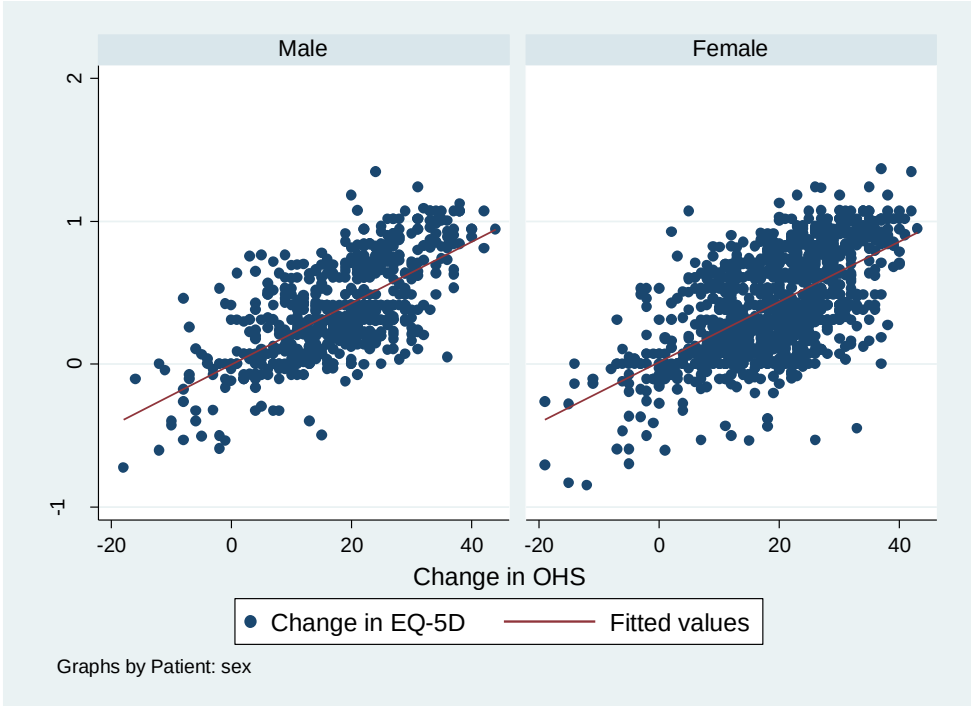
OHS question: response level	Pain / Discomfort		Anxiety / Depression	
	1	3	2	3
Description of pain: 0	Base case		Base case	
Description of pain: 1	-1.345 *	-2.340 *	-0.085	-0.229
Description of pain: 2	-0.793 *	-3.260 *	0.089	0.680
Description of pain: 3	0.197	-3.440 *	0.178	-0.629
Description of pain: 4	1.561 *	-1.837 *	-0.071	0.223
Night pain: 0	Base case			
Night pain: 1	-0.051	-0.320	0.029	-0.311
Night pain: 2	-0.182	-0.627 *	0.224	0.009
Night pain: 3	-0.161	-1.508 *	0.170	0.464
Night pain: 4	0.533	-0.645 *	-0.242	-0.235
Sudden pain: 0	Base case			
Sudden pain: 1	-0.234	-0.259	-0.124	-0.143
Sudden pain: 2	-0.458	-0.622 *	0.129	-0.061
Sudden pain: 3	0.124	-0.327	-0.030	-1.749 *
Sudden pain: 4	0.422	-0.597 *	-0.180	-0.728
Limping: 0	Base case			
Limping: 1	0.304	-0.533 *	-0.055	0.214
Limping: 2	0.364	-0.610	-0.145	-0.218
Limping: 3	0.567 *	-0.373	0.132	-0.156
Limping: 4	0.637 *	-0.009	-0.415	-0.162
Walking duration: 0	Base case			
Walking duration: 1	0.090	-0.527 *	0.344	0.560
Walking duration: 2	0.092	-0.416 *	0.419 *	0.467
Walking duration: 3	0.018	-0.548 *	0.123	0.276
Walking duration: 4	0.339	-0.822 *	0.061	0.339
Climbing stairs: 0	Base case			
Climbing stairs: 1	-0.198	0.175	0.204	-0.123
Climbing stairs: 2	-0.616	-0.161	0.051	-0.792 *
Climbing stairs: 3	-0.824	-0.688	-0.111	-1.233 *
Climbing stairs: 4	-0.488	-0.588	-0.328	-1.123
Socks and stockings: 0	Base case			
Socks and stockings: 1	0.133	-0.327	-0.365 *	-0.130
Socks and stockings: 2	-0.171	-0.331	-0.169	-0.032
Socks and stockings: 3	-0.042	-0.184	-0.033	0.188
Socks and stockings: 4	0.302	-0.776	-0.418	0.365
Pain from standing up from chair: 0	Base case			
Pain from standing up from chair: 1	-0.721	-0.562	-0.407	-0.856 *
Pain from standing up from chair: 2	-0.817	-1.111 *	-0.497	-0.597
Pain from standing up from chair: 3	-0.754	-1.445 *	-0.537	-1.324 *
Pain from standing up from chair: 4	-0.269	-0.887	-0.494	-1.629 *
Car and public transport: 0	Base case			
Car and public transport: 1	-1.893	0.104	0.096	-0.289
Car and public transport: 2	-1.579	-0.100	-0.021	-0.683
Car and public transport: 3	-1.963 *	0.280	-0.255	-0.293
Car and public transport: 4	-1.783	0.725	-0.170	0.137
Washing and drying: 0	Base case			
Washing and drying: 1	-0.121	0.353	-0.040	0.154
Washing and drying: 2	-0.269	0.421	-0.039	0.065
Washing and drying: 3	-0.019	0.498	-0.232	-0.598
Washing and drying: 4	0.020	0.350	-0.216	0.180
House shopping: 0	Base case			
House shopping: 1	0.497	0.558 *	-0.221	-0.303
House shopping: 2	0.137	0.139	-0.262	-0.441
House shopping: 3	0.419	0.005	-0.752 *	-0.774
House shopping: 4	0.594 *	-0.153	-1.025 *	-1.123 *
Pain interfering work: 0	Base case			
Pain interfering work: 1	0.455	-0.476 *	0.172	-0.361
Pain interfering work: 2	-0.070	-1.153 *	-0.206	-0.796
Pain interfering work: 3	0.617	-1.392 *	-0.326	-1.192
Pain interfering work: 4	1.365 *	-0.455	-0.631 *	-2.172 *
Constant	-0.213	3.175 *	0.769	0.485

† Of the three response levels for each EQ-5D dimension (one, two or three), one was automatically removed by Stata to become the base level; coefficients are given for the remaining two with respect to the base case.

* Statistically significant at 0.05 level

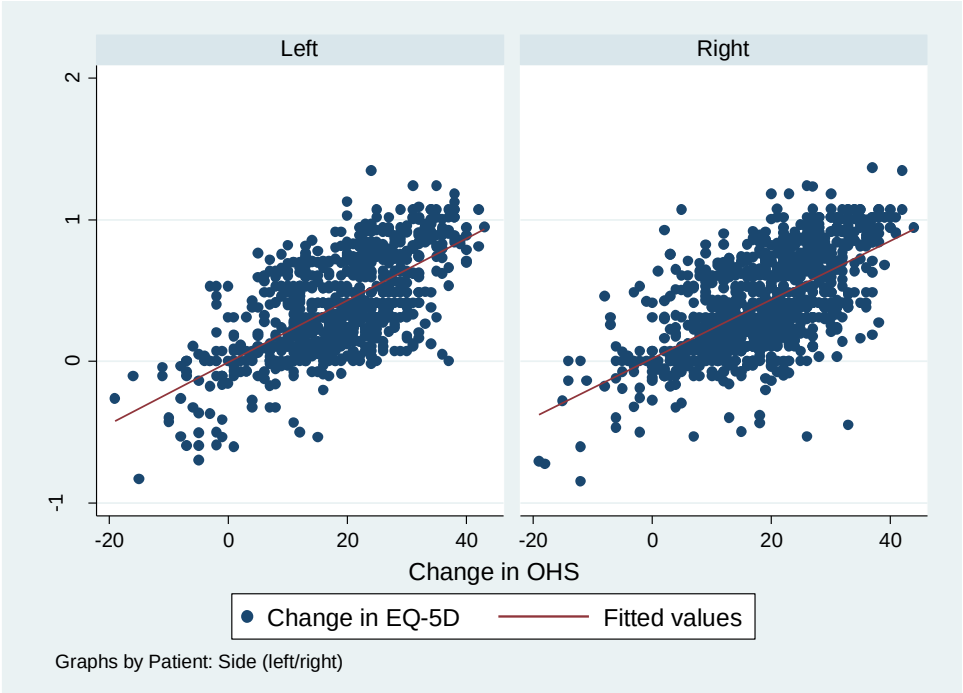
Appendix 7 – Scatterplots of change in OHS by change in EQ-5D

Change in OHS by change in EQ-5D
By gender *



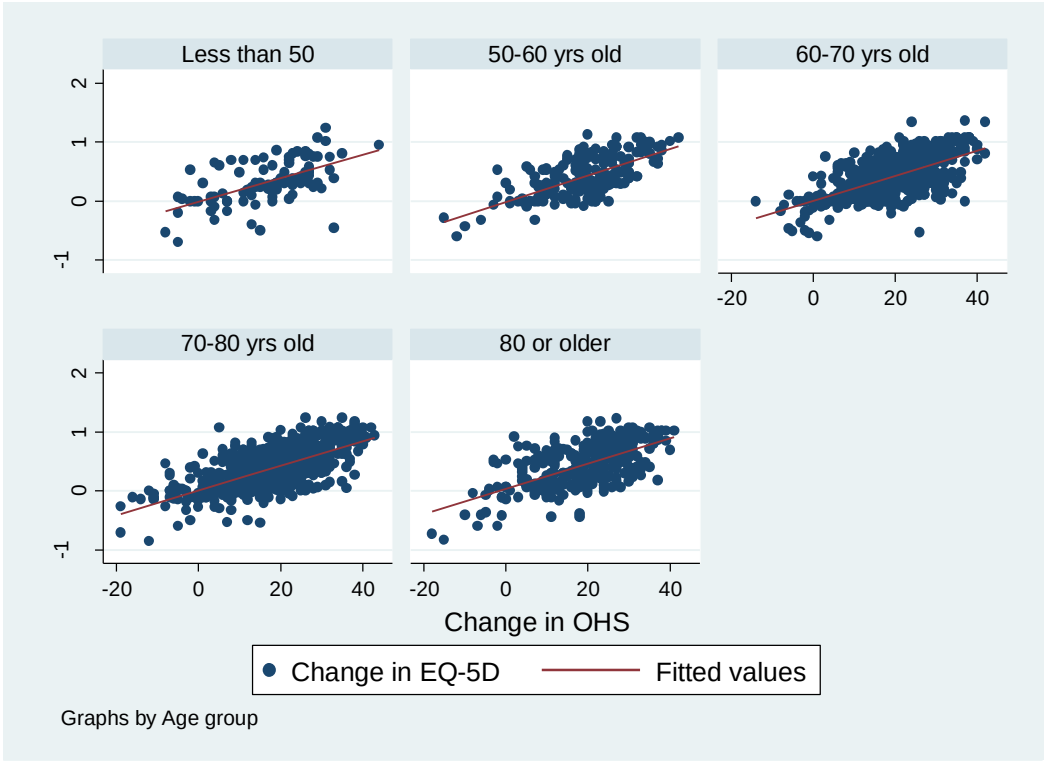
*: Number of observations = 1,759 (100% of sample)

Change in OHS by change in EQ-5D
By side of the hip *



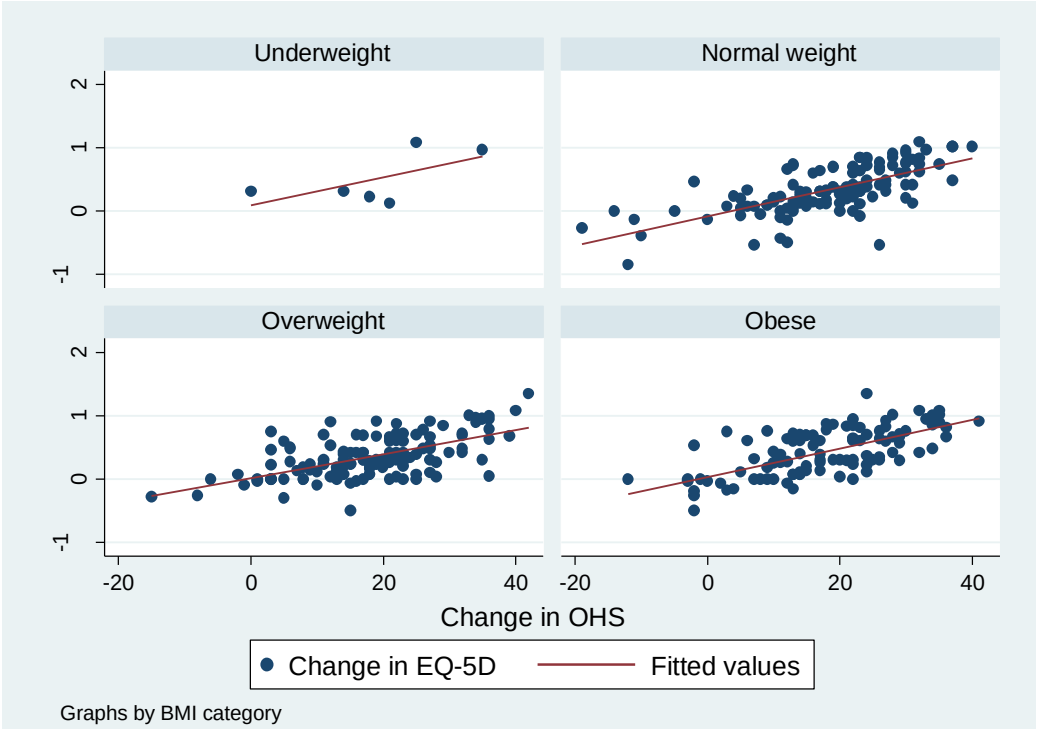
*: Number of observations = 1,759 (100% of sample)

Change in OHS by change in EQ-5D
By age group *



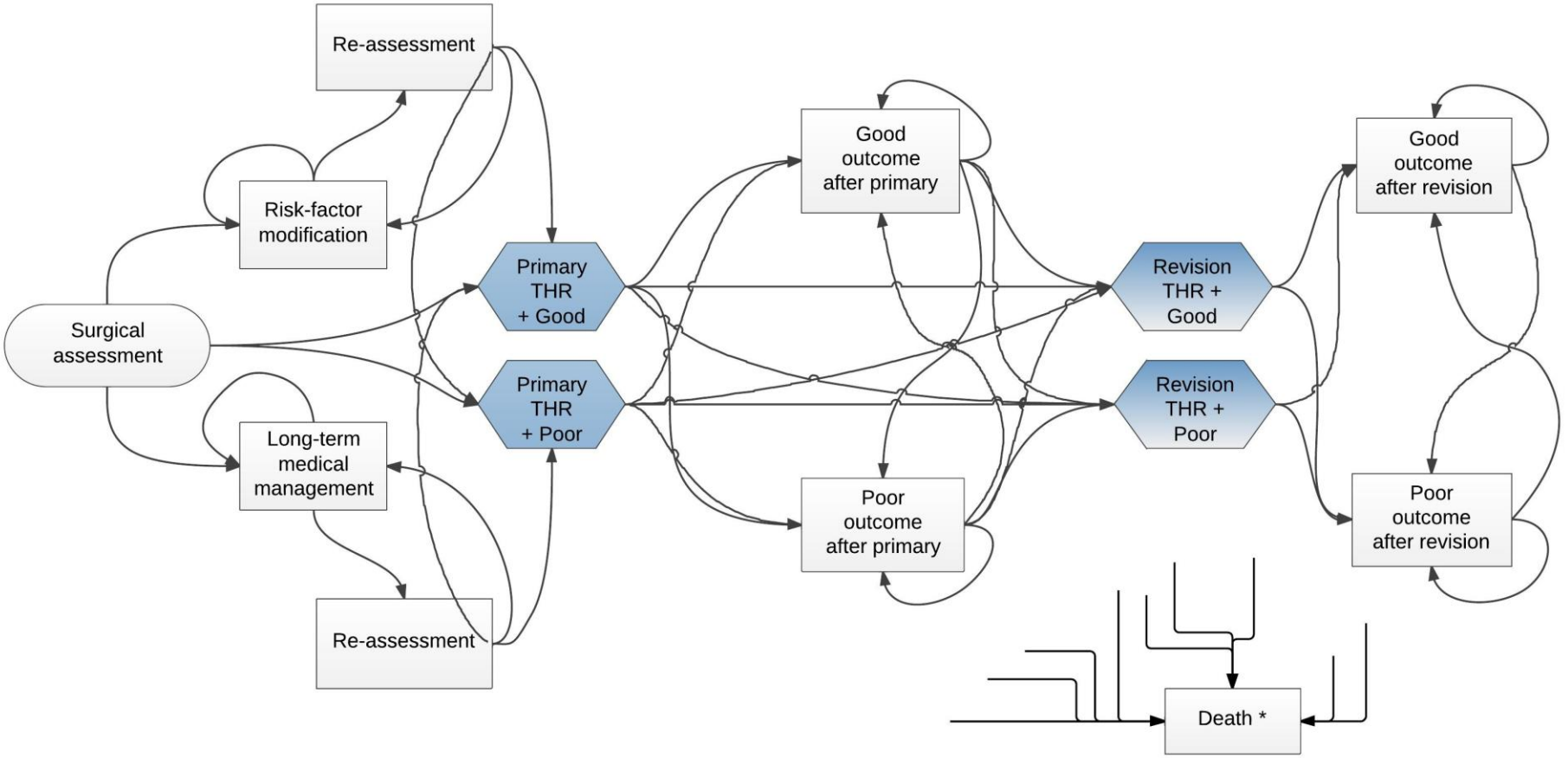
*: Number of observations = 1,759 (100% of sample)

Change in OHS by change in EQ-5D
By BMI group *



*: Number of observations = 361 (21% of sample)

Appendix 8 – Economic model with all transitions



* Represents transitions from each health state to the absorbing state of *Death*

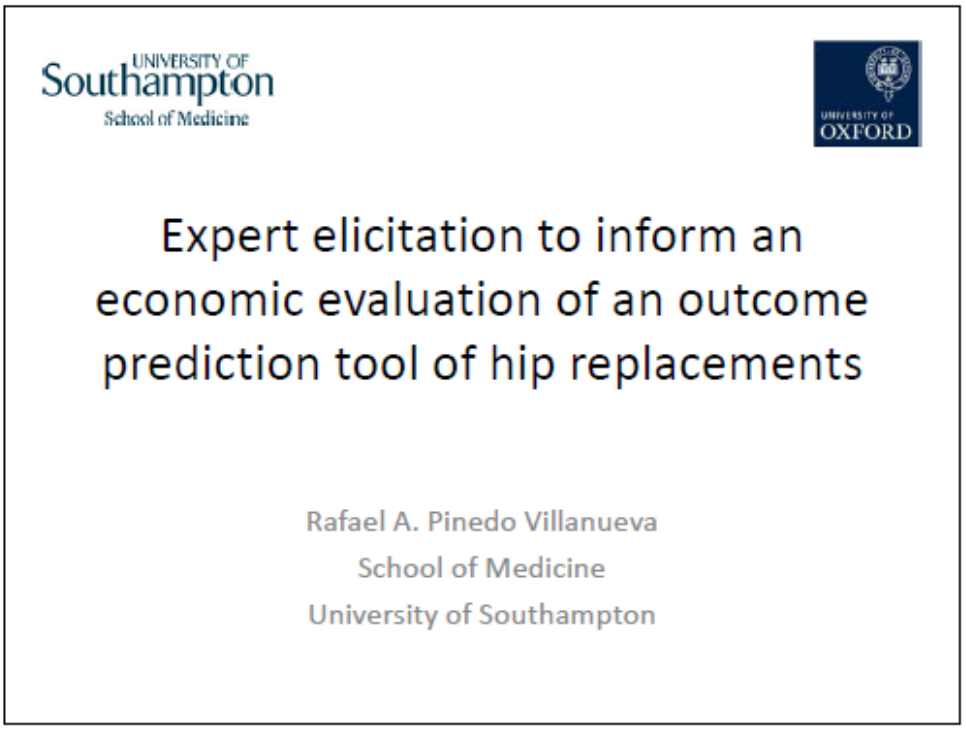
Appendix 9 – Expert elicitation grid, question 1

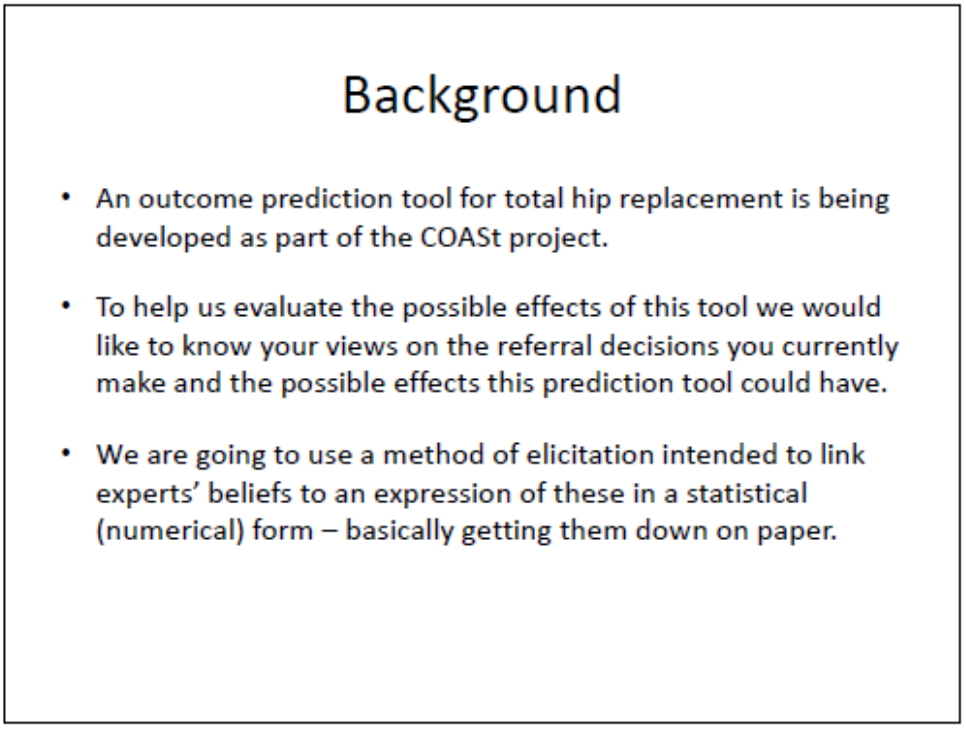
Section 1 - Question 1

What proportion of patients fall in group (a), i.e. those that are referred for a hip replacement?

0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%

Appendix 10 – Expert elicitation slides





Expert elicitation

- Your experience as an orthopaedic surgeon makes you the expert here!
- This doesn't mean that you are expected to know the answer to all questions.
- The questions may be about things that you already have opinions on or are quite knowledgeable about, but others may require some deep thinking.
- There are no right or wrong answers to these questions – we just want to know your opinions.

Expert elicitation (2)

- If you are unsure about (or don't know the answer to) a question you should still answer it. Just express how uncertain you are about it in your response (we will show you how to do this later on).
- The reason why we are using expert elicitation for our model is that some of the inputs for which we need data have not been explored systematically before.
- We have made every effort to make questions clear and fully self-explained, but if you have any doubt please feel free to ask before providing an answer.

Expert elicitation

- Your experience as an orthopaedic surgeon makes you the expert here!
- This doesn't mean that you are expected to know the answer to all questions.
- The questions may be about things that you already have opinions on or are quite knowledgeable about, but others may require some deep thinking.
- There are no right or wrong answers to these questions – we just want to know your opinions.

Expert elicitation (2)

- If you are unsure about (or don't know the answer to) a question you should still answer it. Just express how uncertain you are about it in your response (we will show you how to do this later on).
- The reason why we are using expert elicitation for our model is that some of the inputs for which we need data have not been explored systematically before.
- We have made every effort to make questions clear and fully self-explained, but if you have any doubt please feel free to ask before providing an answer.

Expert elicitation (3)

- We expect this exercise to last approximately 35 minutes.
- During that time, we will:
  - ✓ Present you with this **introduction**
  - – Gather basic information **about you**
  - – Provide you with **instructions** for the elicitation exercise
  - – Show you an **example** and **practice** question
  - – Remind you of **key aspects** before you provide your opinion
  - – Elicit your expert opinion on **section 1**: four questions
  - – Elicit your expert opinion on **section 2**: four questions
  - – Share with you what happens **next**

About you

- We would like to ask you a few questions that may help us to understand your responses and why we might get difference answers across the group.
  - – Job title
  - – Place of work
  - – Years performing hip replacements
  - – Approximate number of hip replacements performed last year
  - – Specialised in any patient subgroup or surgical technique?
  - – Percentage of referrals with diagnosis of *degenerative hip disease* (as opposed to *undiagnosed hip problem*)

Instructions

- For many of the questions we want to know how uncertain (or certain) you are about your beliefs.
- As we said in a previous slide there may be some questions which you feel you don't know the answer to – this is perfectly acceptable and reasonable, but we need to make sure you express your uncertainty.
- We are using grids to allow you to express any uncertainty in your answers. These will be shown in a bit.

Instructions (2)

- What do we mean by uncertainty?
  - Imagine explaining the chance of not healing to any particular patient. It could be 20% but it could be as low as 2% or as high as 50%. This **is** uncertainty!
  - This is different to thinking about differences between patients e.g. for one patient the risk of not healing may be 2%, for another it may be 20% and for another it may be 50%. All of these patients would have a different set of characteristics. This **isn't** uncertainty!

Instructions (3)

- All but one question require you to fill in a grid. We will present you with one sheet of paper with the grid for each question.
- Possible values are given along the bottom of the grid, ranging from 0% to 100% as all questions enquire about proportions.
- We will provide you with 21 chips which you are to place in the grid representing your belief and uncertainty about the question.

Instructions (4)

- Please begin by placing two of the chips at the bottom of the grid. One must represent the lowest value you believe is likely, the other the highest.
- Place the remaining 19 chips between these two extremes. The more likely you believe a response is, the more chips you should place in that column.
- You can place all 21 chips in one column if you feel that there is no uncertainty around the answer.

Example

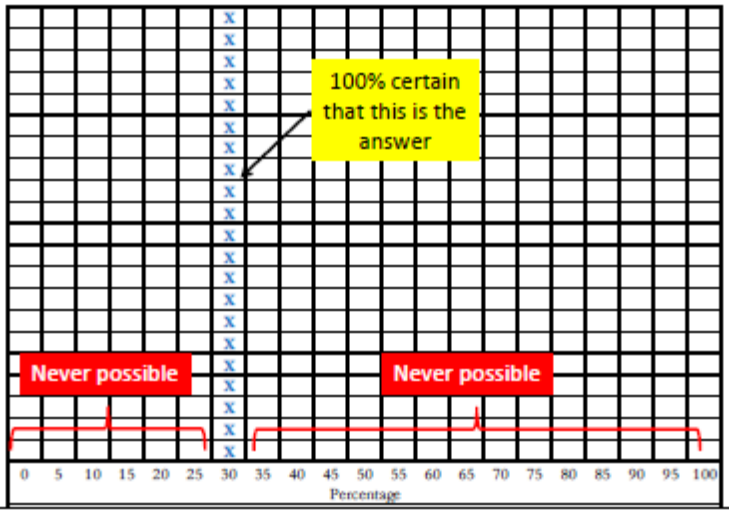
- The question is:

What proportion of pregnant women do you think are 50 years or older?

(Please be mindful about what your response says about the degree of certainty in your belief.)

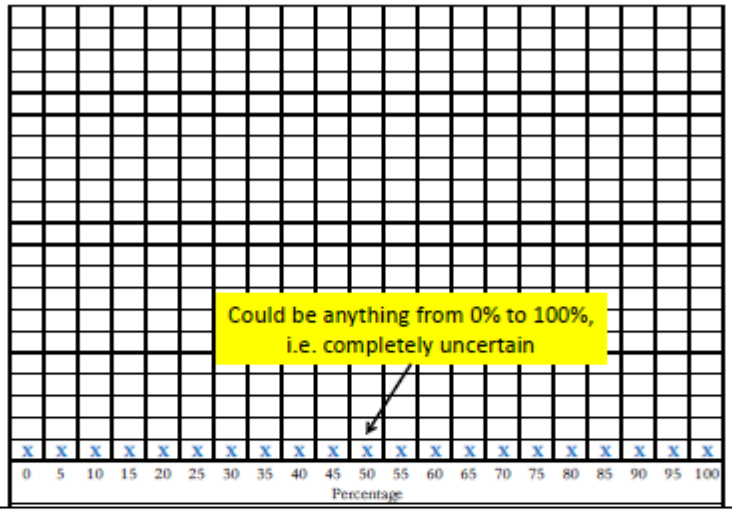
Example

What proportion of pregnant women do you think are 50 years or older?



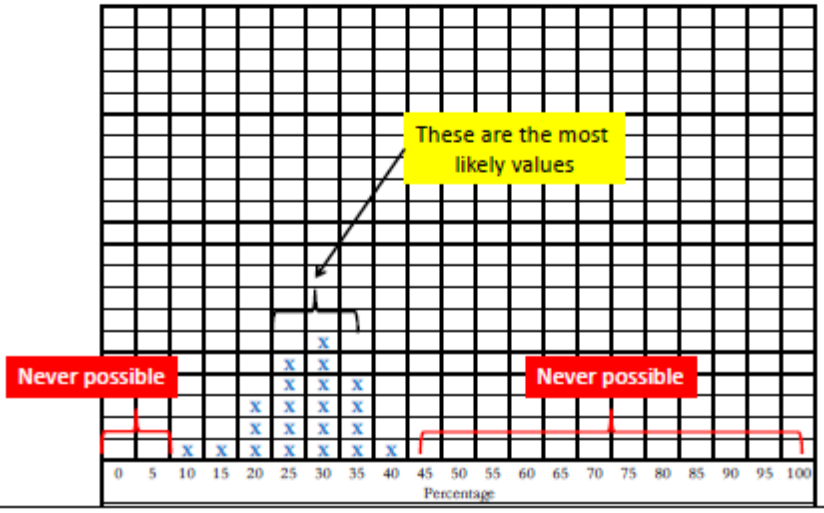
Example

What proportion of pregnant women do you think are 50 years or older?



Example

What proportion of pregnant women do you think are 50 years or older?



Practice

What proportion of pregnant women do you think are 50 years or older?

Key aspects

- It is your INDIVIDUAL beliefs we are interested in.
- To use these in our model we are going to combine them with the answers from other experts.
- Take time to carefully read each question before you answer.
- Make sure you know what your response means (previous slides – possible values, likely values).
- Try to be as objective as possible: put aside issues of which treatment you currently favour or would like to use in the future.

Key aspects (2)

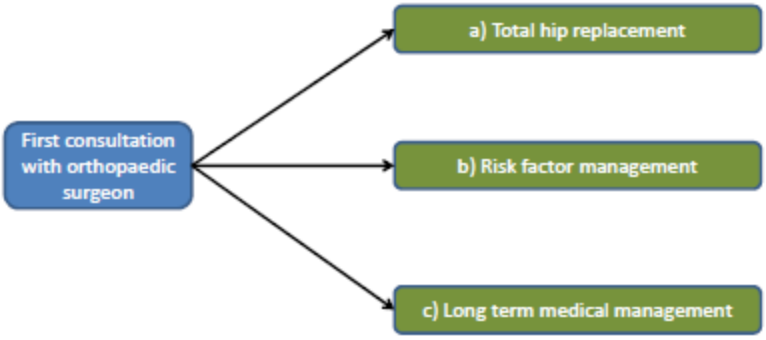
- Think about the patients you see, the relevant literature you have read, etc. when trying to answer the questions.
- With each new question, start afresh – try not to repeat the same grid for each question.
- You can go back and make any changes if you want to.
- If in doubt, please do ask questions.
- Let's begin!

Section 1

- We understand that the patients referred to you because of their hip problems (rightly diagnosed or not) can be:
  a) Referred for a hip replacement operation. This category includes patients for whom additional diagnostic tests may be needed to confirm their eligibility.
  b) Found to be candidates for hip replacement but not suitable to receive the operation until certain risk factors are properly dealt with. Examples of modifiable risk factors include excessive weight, high blood pressure, new onset diabetes or chronic varicose veins.
  c) Referred for long term medical management either because they (a) had a non-orthopaedic or non-hip problem, (b) were found unfit to receive a hip replacement, or (c) were unwilling to go through the operation.

Section 1

- These three options are summarised in the figure below. Keep them in mind as you answer the first three questions of this section.



Section 1
Question 1

What proportion of patients fall in group (a), i.e. those that are referred for a hip replacement?

Section 1
Question 2

What proportion of patients fall in group (b), i.e. those that are referred for risk factor management?

Section 1
Question 3

What proportion of patients fall in group (c), i.e. those that are referred for long term medical management?

Section 1
Question 4

- Suppose that a prediction tool that used patient characteristics (gender, age, BMI, Oxford Hip Score, etc.) to predict good/poor outcomes after a hip replacement operation were available to you,

4a) What do you think would happen to the proportion of patients referred for THR (group (a))?

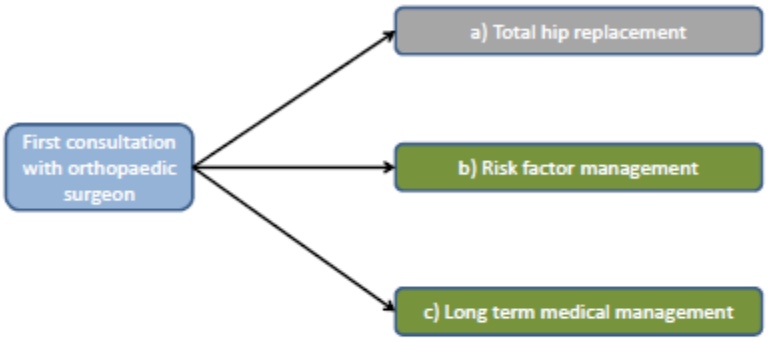
Section 1
Question 4

- The impact of the outcome prediction tool on the proportion of patients referred for THR would be approximately....

4b) How large?

Section 2

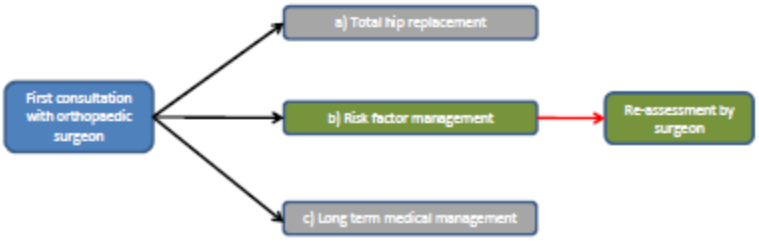
- For the four questions in Section 2, we are going to focus on groups (b) and (c) as defined previously.



Section 2
Question 5

- Thinking specifically of those patients on risk factor management,

which proportion would be back for re-assessment within a year?



Section 2
Question 6

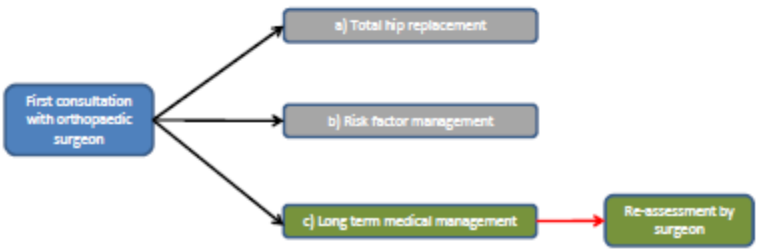
- And of that subgroup of patients re-assessed,

which proportion would be referred for a hip replacement?

Section 2
Question 7

- Now, let's focus on those patients on long-term medical management (remember, those were patients with a non-hip or non-orthopaedic problem, unfit for or unwilling to get a THR),

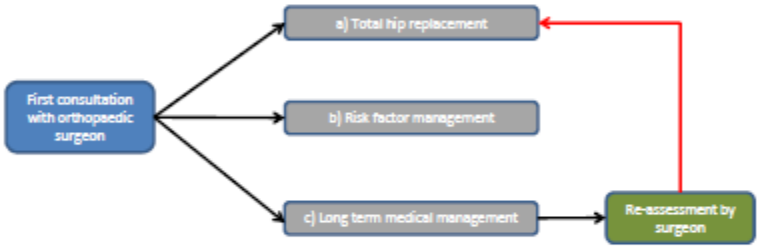
which proportion would be back for re-assessment within a year?



Section 2
Question 8

- And finally, of that subgroup of patients re-assessed,

which proportion would be referred for a hip replacement?



Thank you!

- Those are the questions we wanted your expert opinion on.
  We are very grateful for your support.

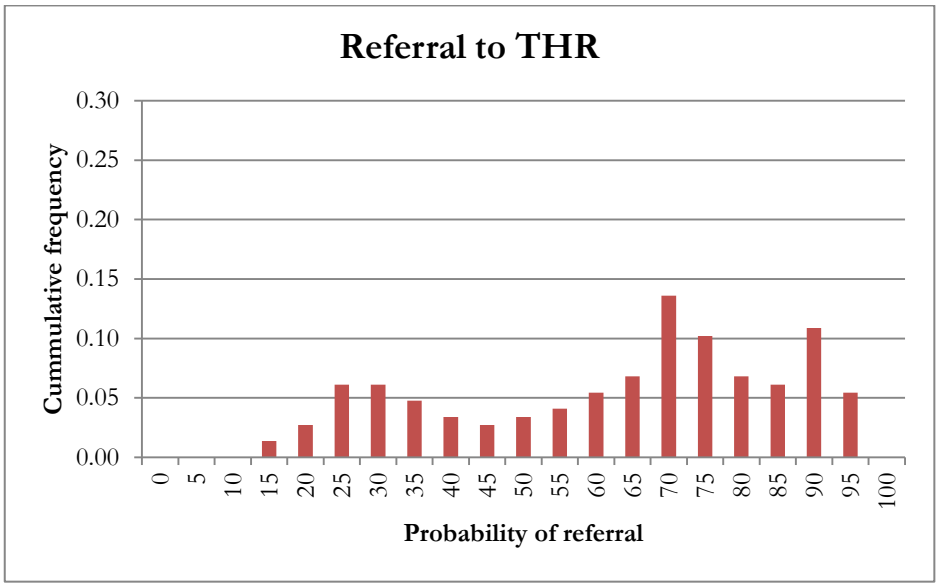
What happens next?

- We will combine your responses with those of other experts and feed an important part of our model with those values.
- Once the model has all the data it needs, we would have an idea of how cost-effective an outcome prediction tool for THR in the UK might be.
- We will also be doing this for Knee replacements and would like to know if you would be willing to participate in that expert elicitation exercise as well?

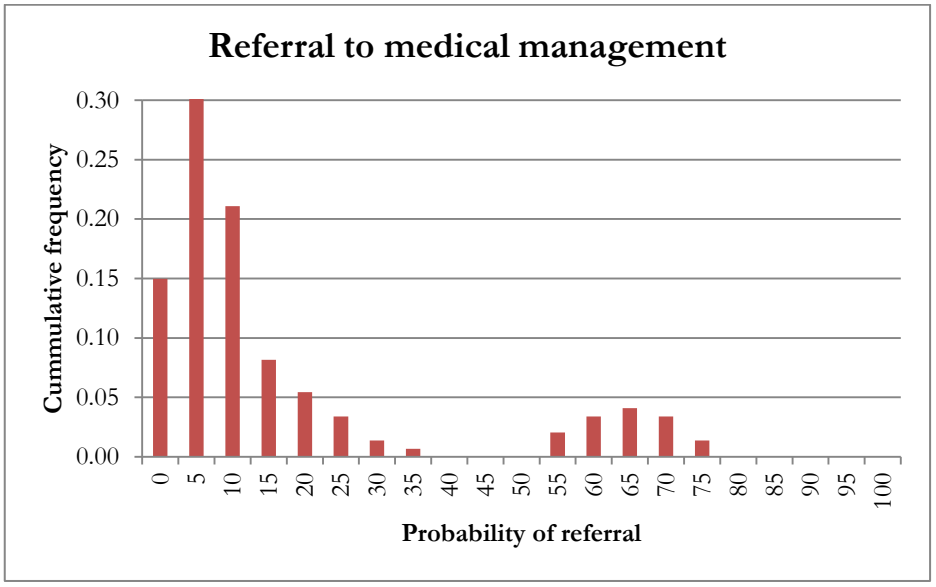
Once again, many thanks indeed!

Appendix 11 – Pooled distribution of transition probabilities from Surgical assessment health state, derived from expert elicitation

Pooled probability distribution of transition probability from Surgical assessment to Primary THR

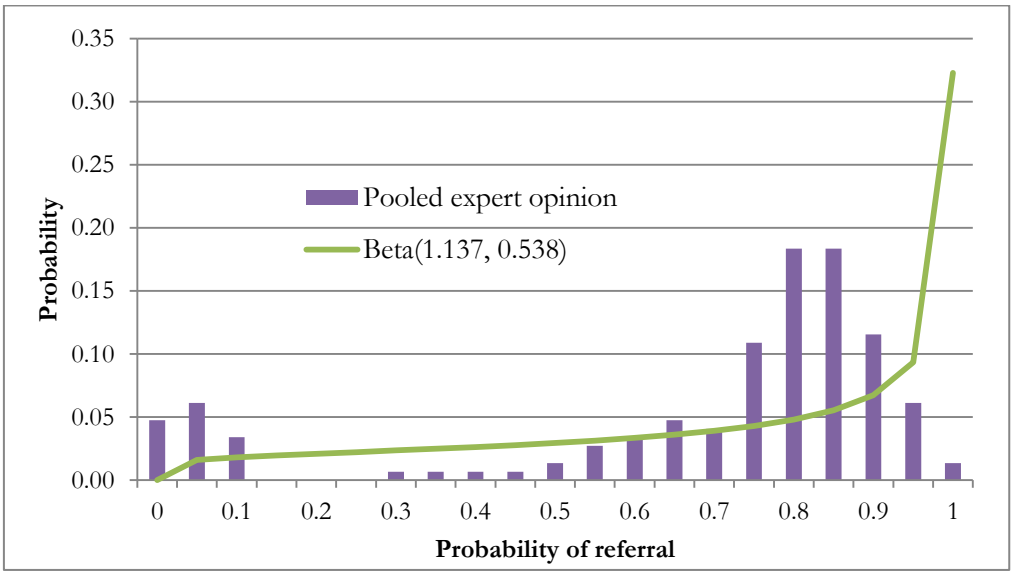


Pooled probability distribution of transition probability from Surgical assessment to Long-term medical management



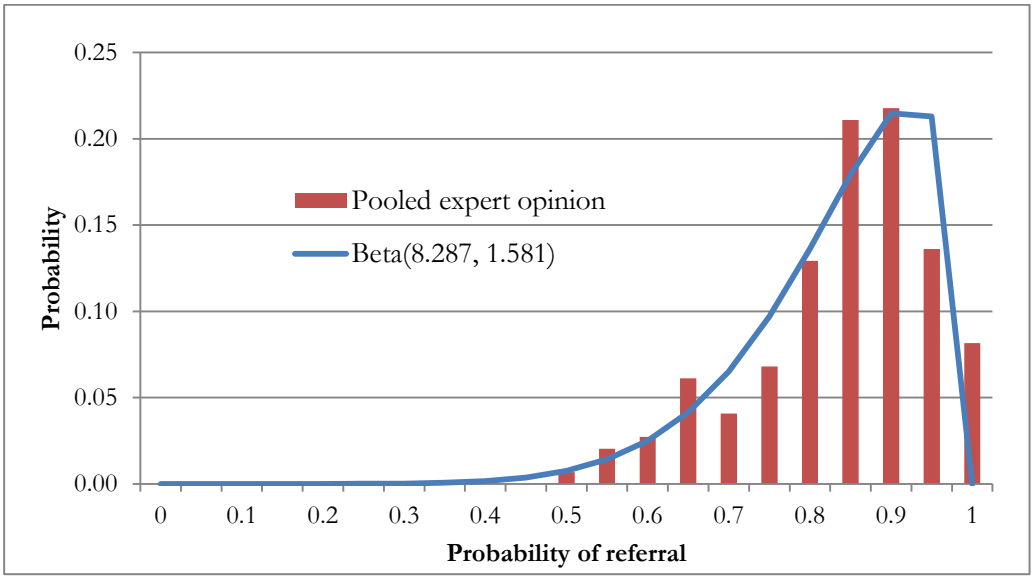
Appendix 12 – Aggregated probability distribution of experts' responses and respective fitted probability density function

Pooled probability distribution of experts' responses and respective fitted probability density function about referral to Re-assessment after Risk-factor modification



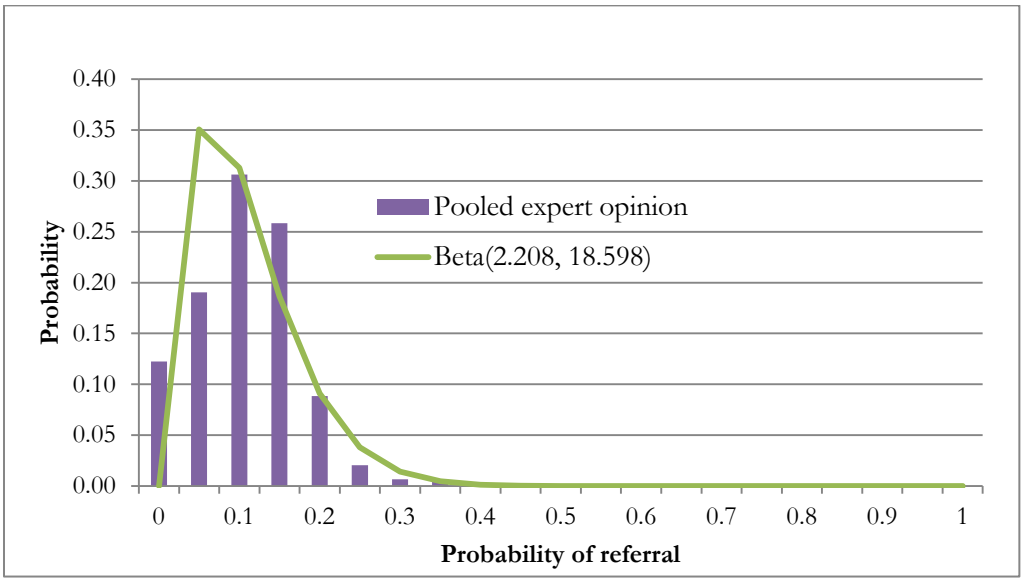
Observed mean: 0.679
Mean from fitted distribution: 0.758

Pooled probability distribution of experts' responses and respective fitted probability density function about referral to THR after Re-assessment post Risk-factor modification



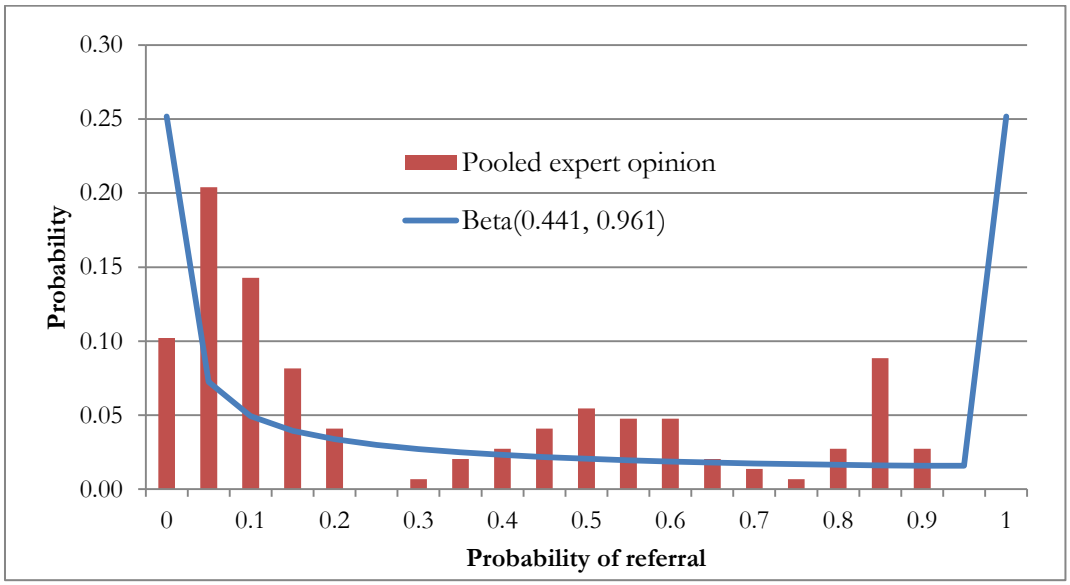
Observed mean: 0.840
Mean from fitted distribution: 0.863

Pooled probability distribution of experts' responses and respective fitted probability density function about referral to Re-assessment after Long-term medical management



Observed mean: 0.106
Mean from fitted distribution: 0.094

Pooled probability distribution of experts' responses and respective fitted probability density function about referral to THR after Re-assessment post Long-term medical management



Observed mean: 0.315
Mean from fitted distribution: 0.217

Appendix 13 – Mortality rates one year after THR

	Males	Females
45-60 years old	0.57%	0.61%
60-70 years old	1.00%	0.67%
70-80 years old	2.07%	1.32%
80+ years old	5.59%	3.62%

Notes:

1. Data obtained from the NJR's 9th Annual Report [54]
2. Since rates were reported by patient age groups in five year intervals starting at 55, rates indicated in the table above for 45-60 year olds correspond to that of 55 through 59, those of 60-70 and 70-80 are the weighted means of the two component groups reported by the NJR, and rates corresponding to patients older than 80 years of age were reported also in this way by the NJR hence no adjustment was applied.

Appendix 14 – EPOS patient counts by outcome category in years 1 and 2 after Primary THR

	Male		Female	
	Poor outcome	**Good outcome**	**Poor outcome**	**Good outcome**
45-70 years old	54	141	90	203
70+ years old	36	158	111	250
Total	**90**	**299**	**201**	**453**

Of these, *Good* outcomes transitioned into year 2 as follows:

	Male		Female	
	Poor outcome	**Good outcome**	**Poor outcome**	**Good outcome**
45-70 years old	10	131	7	196
70+ years old	2	156	20	230
Total	**12**	**287**	**27**	**426**

And *Poor* outcomes transitioned into year 2 as follows:

	Male		Female	
	Poor outcome	**Good outcome**	**Poor outcome**	**Good outcome**
45-70 years old	24	30	52	38
70+ years old	17	19	56	55
Total	**41**	**49**	**108**	**93**

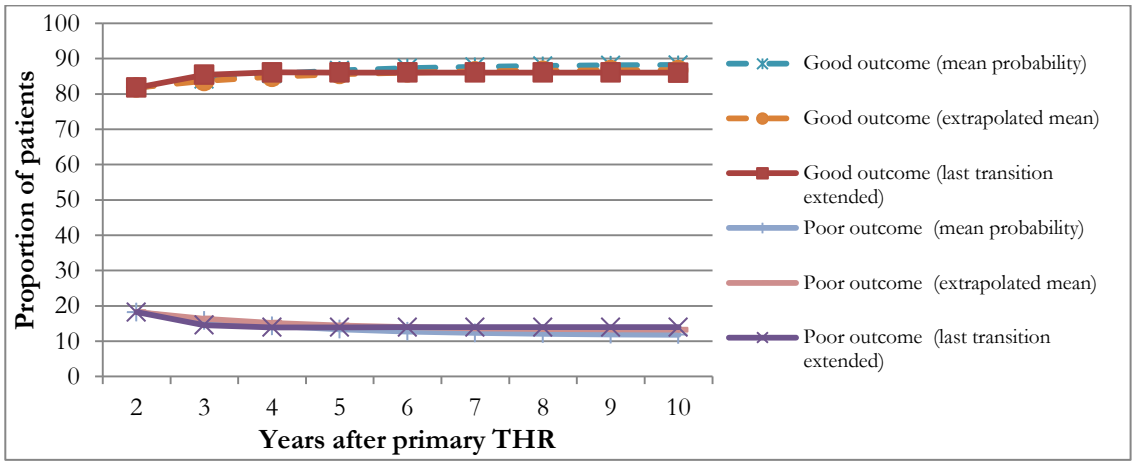
Appendix 15 – Observed proportion of EPOS patients remaining in *Good* outcome category between years two and five after Primary THR

	Years 2-3		Years 3-4		Years 4-5	
	n*	%**	n*	%**	n*	%**
Males, 45-70 years of age	156	96.8	165	95.8	172	94.8
Males, 70+ years of age	172	93.0	153	90.8	145	91.7
Females, 45-70 years of age	241	92.9	230	97.4	239	93.3
Females, 70+ years of age	284	90.1	270	88.9	246	90.7

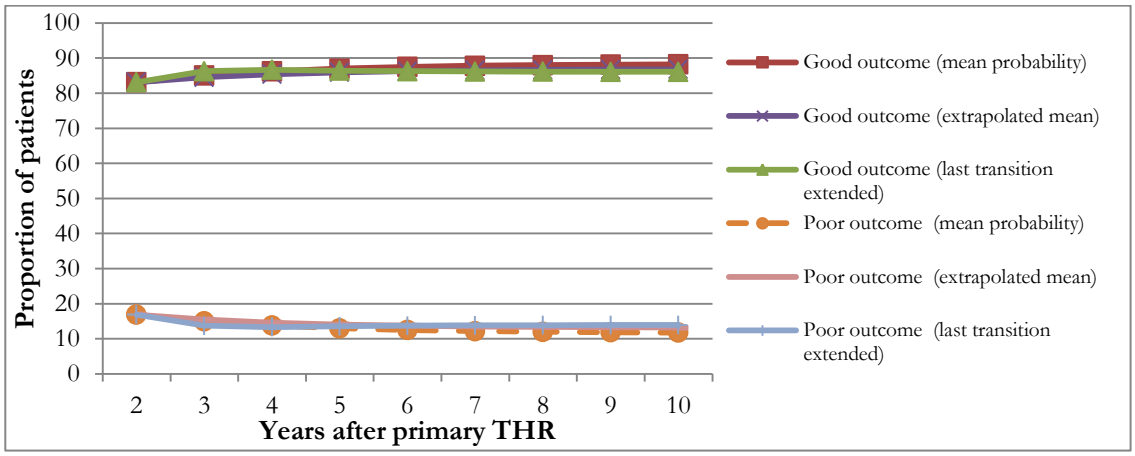
*: number of THR patients classified as *Good* outcomes in year pre transition
**: percentage of *n* patients who remained as *Good* outcome in the following year

Appendix 16 – Proportion estimates of *Good* and *Poor* outcomes from three scenarios extrapolating transition probabilities

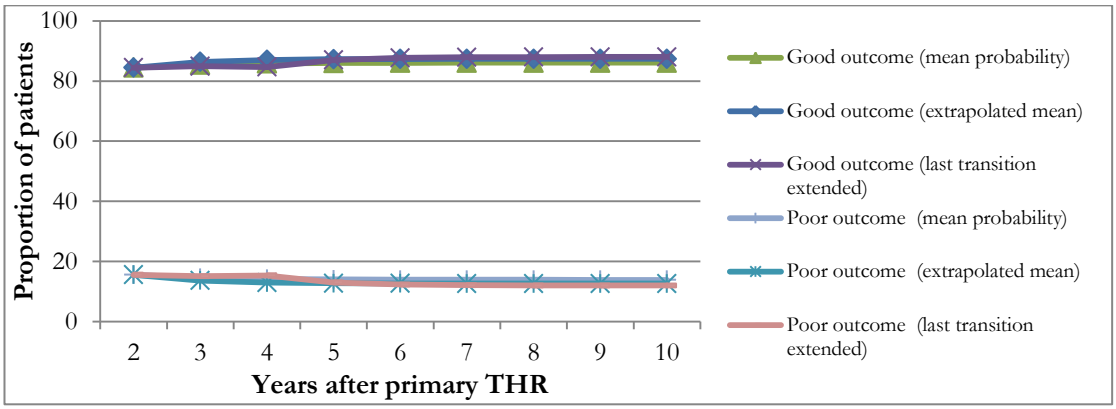
Males between 45 and 60 years of age



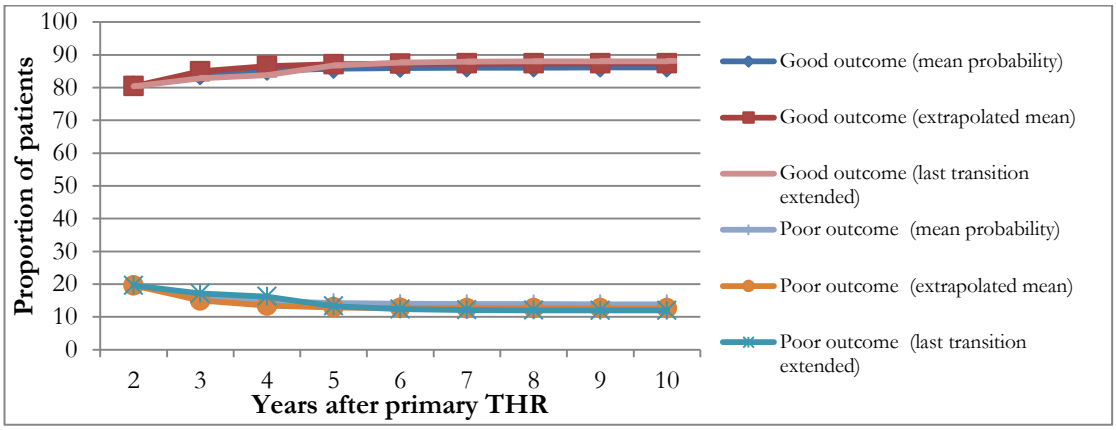
Males between 60 and 70 years of age



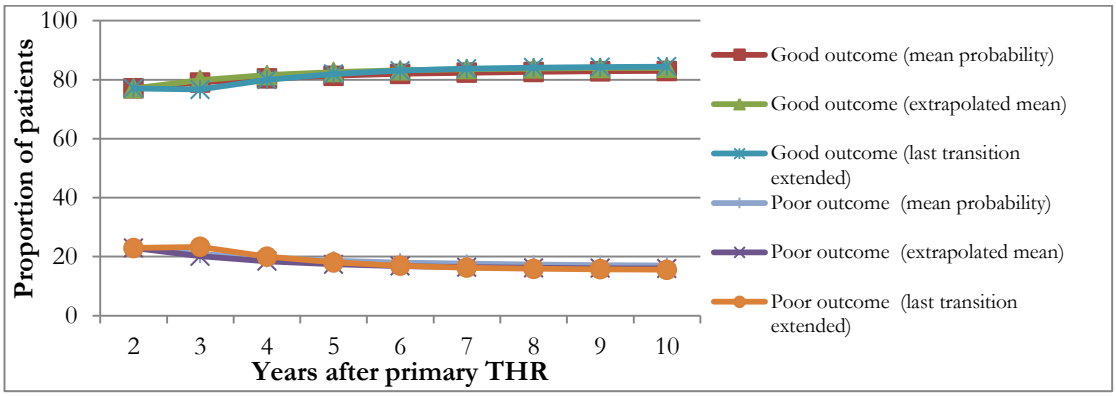
Males between 70 and 80 years of age



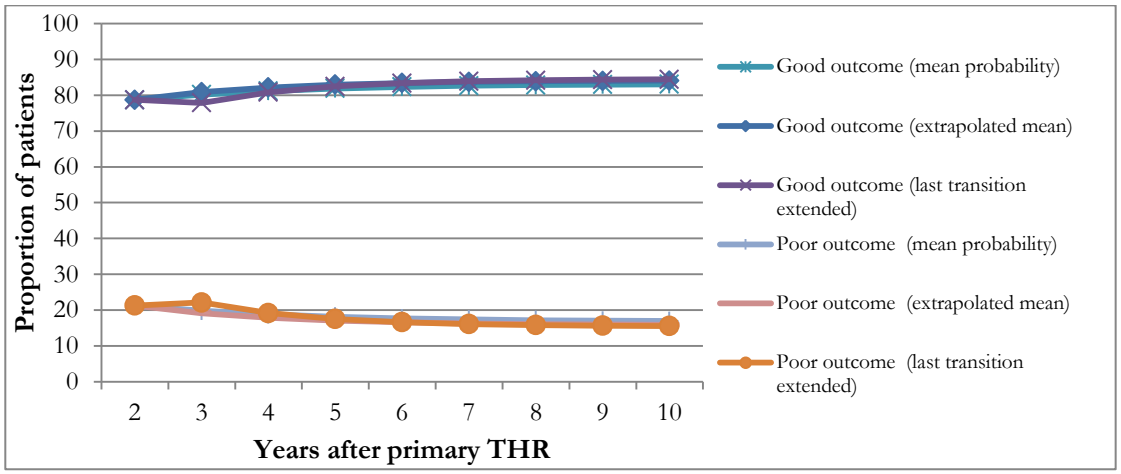
Males over 80 years of age



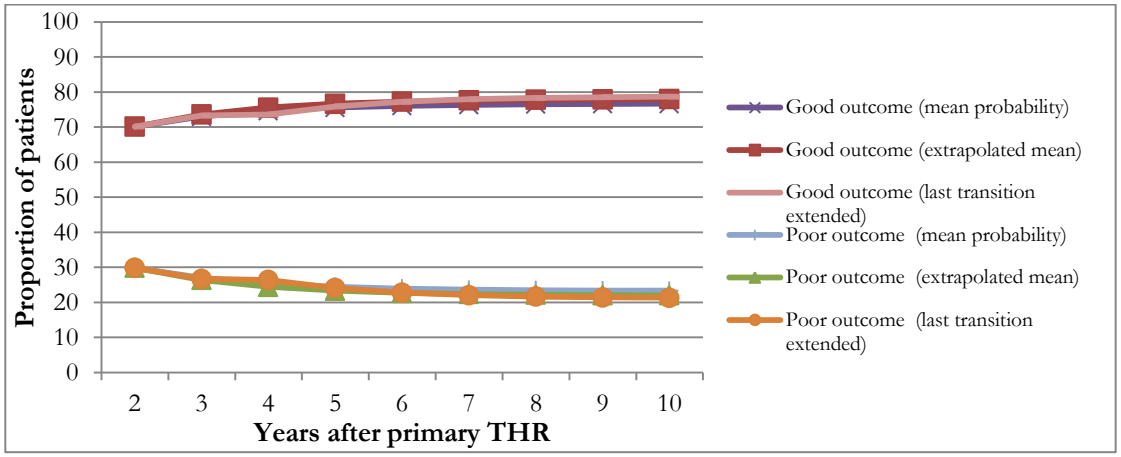
Females between 45 and 60 years of age



Females between 60 and 70 years of age

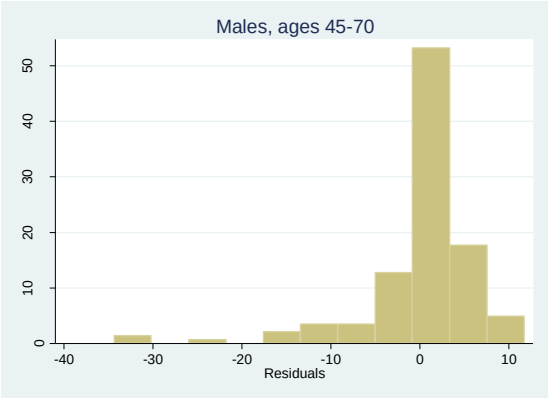


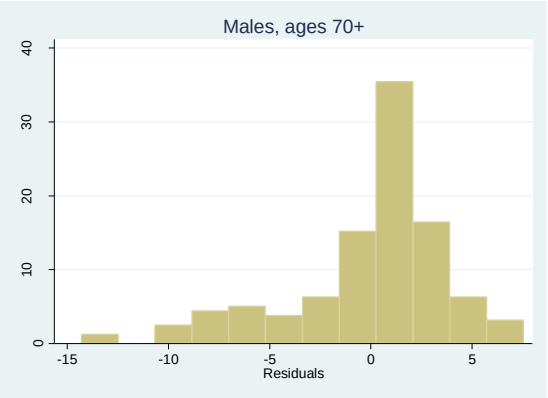
Females over 80 years of age

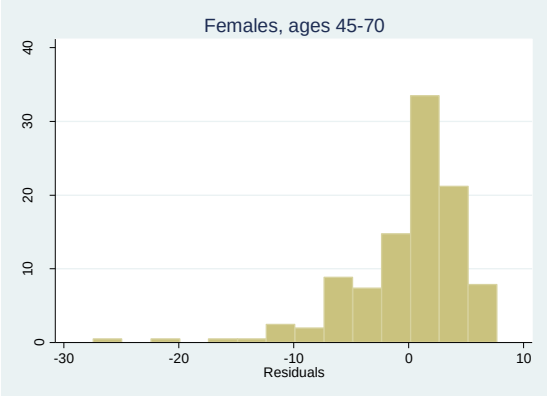


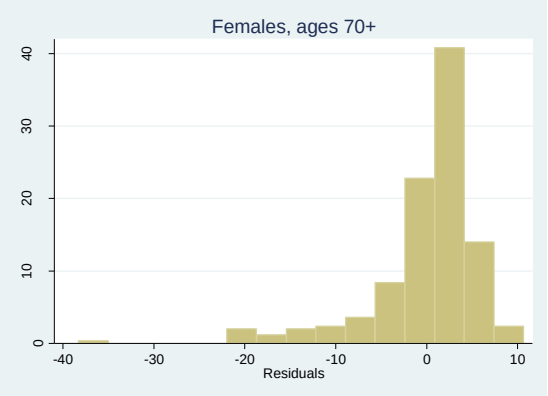
Appendix 17 – Distribution of residuals from OLS models predicting OHS two years after Primary THR

Good outcomes

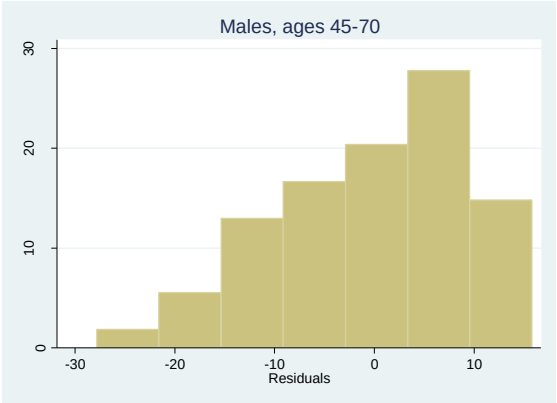


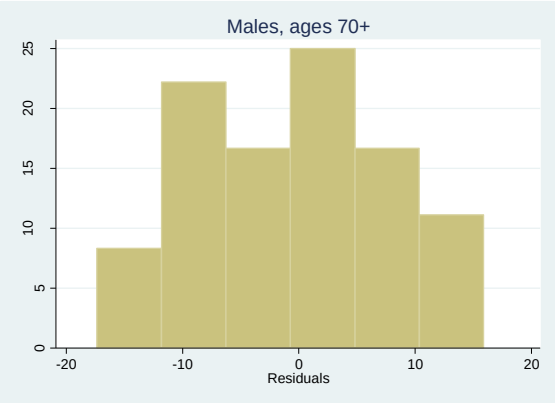


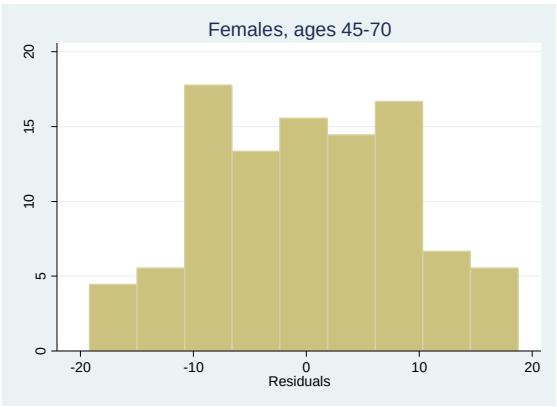


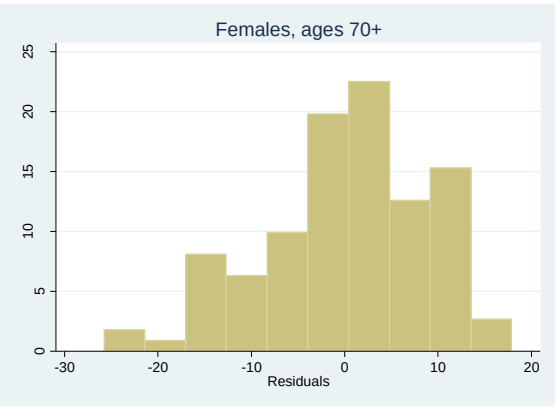


Poor outcomes









Appendix 18 – Mean QALY values associated to each model state

Health state	Males, 45-60 years of age		Males, 60-70 years of age		Males, 70-80 years of age		Males, 80+ years of age	
	Current practice	**Prediction tool**	**Current practice**	**Prediction tool**	**Current practice**	**Prediction tool**	**Current practice**	**Prediction tool**
Surgical assessment	0.385	0.385	0.408	0.408	0.403	0.403	0.344	0.344
Risk-factor modification (RFM)	0.385	0.503	0.408	0.516	0.403	0.510	0.344	0.482
Re-assessment after RFM	0.385	0.503	0.408	0.516	0.403	0.510	0.344	0.482
Long-term medical management (LTMM)	0.385	0.046	0.408	0.058	0.403	0.067	0.344	0.052
Re-assessment after LTMM	0.385	0.046	0.408	0.058	0.403	0.067	0.344	0.052
Primary THR + Good outcome first year	0.783	0.783	0.788	0.788	0.780	0.780	0.753	0.753
Primary THR + Poor outcome first year	0.460	0.460	0.515	0.515	0.550	0.550	0.543	0.543
Good outcome after primary THR	0.866	0.866	0.872	0.872	0.861	0.861	0.840	0.840
Poor outcome after primary THR	0.573	0.573	0.602	0.602	0.527	0.527	0.529	0.529
Revision THR + Good outcome first year	0.720	0.720	0.741	0.741	0.738	0.738	0.703	0.703
Revision THR + Poor outcome first year	0.345	0.345	0.373	0.373	0.435	0.435	0.465	0.465
Good outcome after revision THR	0.783	0.783	0.791	0.791	0.81	0.810	0.786	0.786
Poor outcome after revision THR	0.469	0.469	0.483	0.483	0.459	0.459	0.471	0.471
Death	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Health state	Females, 45-60 years		Females, 60-70 years		Females, 70-80 years		Females, 80+ years	
	Current practice	**Prediction tool**	**Current practice**	**Prediction tool**	**Current practice**	**Prediction tool**	**Current practice**	**Prediction tool**
Surgical assessment	0.385	0.385	0.347	0.347	0.334	0.334	0.276	0.276
Risk-factor modification (RFM)	0.385	0.464	0.347	0.489	0.334	0.478	0.276	0.461
Re-assessment after RFM	0.385	0.464	0.347	0.489	0.334	0.478	0.276	0.461
Long-term medical management (LTMM)	0.385	0.049	0.347	0.054	0.334	0.057	0.276	0.043
Re-assessment after LTMM	0.385	0.049	0.347	0.054	0.334	0.057	0.276	0.043
Primary THR + Good outcome first year	0.752	0.752	0.761	0.761	0.755	0.755	0.730	0.730
Primary THR + Poor outcome first year	0.464	0.464	0.522	0.522	0.546	0.546	0.538	0.538
Good outcome after primary THR	0.861	0.861	0.865	0.865	0.815	0.815	0.785	0.785
Poor outcome after primary THR	0.511	0.511	0.541	0.541	0.595	0.595	0.586	0.586
Revision THR + Good outcome first year	0.706	0.706	0.728	0.728	0.712	0.712	0.687	0.687
Revision THR + Poor outcome first year	0.384	0.384	0.408	0.408	0.451	0.451	0.457	0.457
Good outcome after revision THR	0.791	0.791	0.804	0.804	0.762	0.762	0.745	0.745
Poor outcome after revision THR	0.449	0.449	0.471	0.471	0.538	0.538	0.546	0.546
Death	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Appendix 19 – Estimating measures of resource use from CPRD cases and controls

In the work by Violato *et al* [123], the mean cost attributable to the condition of interest was obtained by subtracting the mean cost by controls from the mean cost by cases. The authors reported a confidence interval around this difference but the method for obtaining this interval was not specified, justified, or a reference provided. After a meeting with the author it was learnt that, given the lack of any known previous work on a similar analysis, a t-test was performed to compare the means of cases and controls, and the confidence interval reported in the output used as the confidence interval for the difference.

Obtaining a mean difference directly from the means of cases and controls is reasonable; however, using a t-test, expected to be employed on data that are normally distributed, to obtain a confidence interval from highly skewed and not-normally distributed data such as that of resource use does not seem appropriate. Estimating parameters that reflect variability of these data is a challenge, and in order to produce estimates as close as possible to observed variation we chose to conduct a patient-, or pseudo-patient-level data throughout the analysis. To this end, we estimated individual differences from the resource use reported by cases minus the average resource use reported by controls.

This approach generates two potential complications: first, it allows for negative differences when the mean of controls is higher than the values reported by cases; and second, it affects the weight applied to the values reported by controls, reducing them by a factor equal to the number of controls in each set per case. However, negative measures of resource use are understood in this study, as it is explained in the text, as an indication of one or a group of non-hip pain individuals who, in average, demand more healthcare resources than a comparable patient with hip pain. This is clinically plausible and hence there is no reason to avoid these data from being fed into the model, as neither are the cases in which cases spend significantly more than the average of their controls. Secondly, the reduced weight applied to controls when calculating overall mean resource use attributable to hip pain is not a problem either as the rest of the weight to make controls equivalent to the respective case is provided by other controls. In other words, averaging controls is a means to synthetizing a group of controls into only one so that individual differences can be estimated.

The above features mean that our estimates of mean costs attributable to hip pain will be different than if we used the method employed by Violato *et al*, but also that the uncertainty parameters will be more appropriately estimated being based on the actual distribution of observed data as opposed to a test whose parametric requirement does not correspond to the actual distribution of the data. Figure A19-1 below shows how the distribution of total preoperative consultation costs by cases and controls during the year prior to a THR is, as expected, right-skewed and far from normal. Following the method used by Violato *et al*, mean costs by cases and controls can be calculated and the difference between the means obtained (shown in Table A19-1); however, the distribution of the difference cannot be plotted as there is no patient-level data for it. The method employed in our analysis is able to produce (pseudo)patient-level records for these differences, which can then be used to estimate the mean preoperative consultation costs attributable to hip pain and its distribution, which is shown in Figure A19-2. Finally, confidence intervals estimated using a t-test as well as calculated from the actual distribution of differences are shown in Table A19-2.

Figure A19-1
Distribution of preoperative consultation costs (in 2011 £) during the year prior to THR, by cases and controls

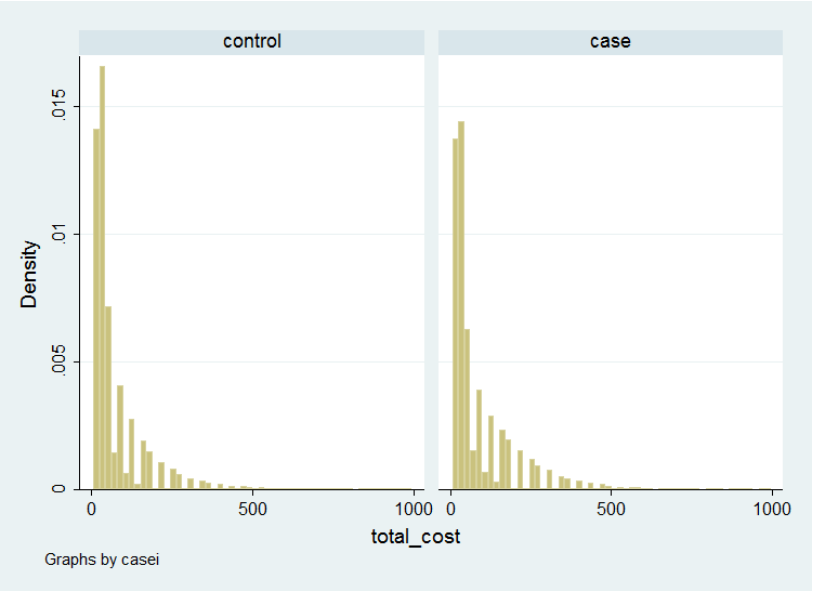


Table A19-1
Mean preoperative consultation costs (in 2011 £) during the year prior to THR, by cases and controls

	Cases	Controls	Difference
Females 45-60	90.8	70.8	**20.0**
Males 45-60	85.9	65.5	**20.4**
Females 60-70	91.4	74.3	**17.1**
Males 60-70	89.9	72.2	**17.6**
Females 70-80	92.0	78.6	**13.4**
Males 70-80	89.2	75.8	**13.3**
Females 80+	87.1	76.5	**10.6**
Males 80+	92.7	77.0	**15.8**
TOTAL	90.1	75.2	**14.9**

As shown in Table A19-2, there is a significant difference between the mean costs of consultations attributable to hip pain if calculated using the method followed by Violato *et al* (£15 for the entire sample), compared to those obtained by applying the method used in this study (£66). It is difficult to say whether the latter overestimates the mean difference or if the direct comparison of the means underestimates it. Whilst the approach we followed affects the weights of controls with respect to the overall mean difference, it has a critical strength over the simple operation of means because it considers who is a control for whom in the dataset. The mean values obtained from the approach used by Violato *et al*, nonetheless, suggest that results for consultation costs attributable to hip pain would be much lower than expected. Considering that the analysis is performed during the 12 months prior to a THR, a difference of only £15 would indicate that patients visited their GP, in average, only 0.5 times more than similar controls. That does not fit the experience of most patients during the year prior to going into major surgery. Results of £60 obtained from our approach would instead be equivalent to an average of

two visits to the GP by THR patients in excess of what their controls did, a more sensible figure about the demand of healthcare services by these patients.
Regarding confidence intervals, the method used by Violato *et al* violates the parametric requirement of normality of the t-test, whereas extracting parameters of uncertainty directly from the observed distribution of differences allows populating the PSA with values that actually correspond to the distribution of resource use we are after, even if it is normal and accepts negative values.

Figure A19-2
Distribution of differences between preoperative consultation costs (in 2011 £) by cases minus mean of controls, during the year prior to THR

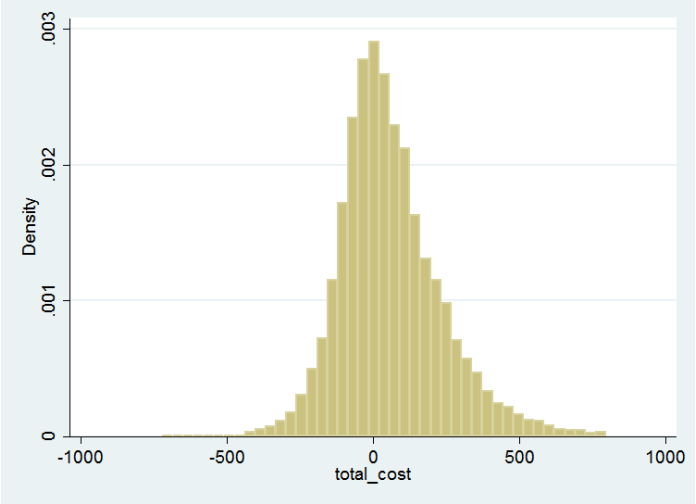


Table A19-1
Mean preoperative consultation costs (in 2011 £) during the year prior to THR, by cases and controls

	Diff of means and t-test			Mean of differences		
	Mean	**95% CI**		**Mean**	**95% CI**	
Females 45-60	20.00	17.11	22.90	81.74	73.57	89.92
Males 45-60	20.39	17.11	23.67	68.43	60.33	76.54
Females 60-70	17.10	15.03	19.18	74.65	68.49	80.80
Males 60-70	17.64	15.16	20.11	63.61	57.21	70.01
Females 70-80	13.37	11.48	15.26	61.56	56.01	67.11
Males 70-80	13.34	10.90	15.77	55.31	48.33	62.28
Females 80+	10.60	8.29	12.90	62.74	55.15	70.33
Males 80+	15.77	11.77	19.76	71.58	59.55	83.62
TOTAL	**14.91**			**66.15**	**63.60**	**68.69**

In conclusion, it appears that using administrative data from cases and sets of controls poses important challenges to appropriately estimating parameter values for resource use attributable to a condition, and particularly to describe uncertainty necessary for probabilistic sensitivity analysis in economic evaluations. The method used by Violato *et al* is reasonable for the estimation of mean differences but does not seem appropriate to obtain confidence intervals. The approach employed in this work produces patient-level

differences by averaging resource use or costs associated to controls, thus allowing for negative values and a normal distribution which is not commonly used to populate costs in an economic model. However, the method keeps track of which controls are matched to which cases, produces mean values that appear to be more clinically plausible, and allows extracting uncertainty parameters from an observed distribution which can more confidently be used in PSA.

This particular example shows that further research is warranted on the methods to estimate resource use parameter values when based on records from administrative datasets aimed at populating economic models. For future assessments, it would be advisable, at a minimum, to run sensitivity analysis on the methods used to extract these values.

Appendix 20 – CPRD records: staff roles inclusion criteria

Staff roles included in the analysis

Group	Staff role
GPs	Assistant
	Associate
	Commercial Deputising service
	Community Medical Officer
	GP Registrar
	Locum
	Non-commercial local rota of less than 10 GPs
	Partner
	Senior Partner
Nurses	Community Nurse
	Contact Tracing Nurse
	Health Visitor
	Hospital Nurse
	Practice Nurse
Other health professionals	Acupuncturist
	Chiropodist
	Consultant
	Dietician
	Other Health Care Professional
	Physiotherapist

Staff roles excluded from the analysis

Administrator	Midwife
Business Manager	No Data Entered
Chiropractor	Non-qualified Dispenser
Community Psychiatric Nurse	Osteopath
Computer Manager	Pharmacist
Counsellor	Practice Manager
Dispenser	Receptionist
Fund Manager	School Nurse
Health Education Officer	Secretary
Interpreter/Link Worker	Social Worker
Maintenance staff	Sole Practitioner

Appendix 21 – CPRD records: events inclusion criteria

Events included in the analysis

Category	Event description
Day visit	Acute visit
	Casualty Attendance
	Clinic
	Emergency Consultation
	Follow-up/routine visit
	Surgery consultation
Night visit	Night visit, Practice
	Night visit, Deputising service
	Night visit, Local rota
Telephone call	Telephone call from a patient
	Telephone call to a patient

Events excluded from the analysis

Event description
Administration
Discharge details
Letter from Outpatients
Mail from patient
Mail to patient
Other
Out of hours, Non Practice
Out of hours, Practice
Repeat Issue
Results recording
Third Party Consultation
(blank)

Appendix 22 – Medication included in the resource use analysis

Category	Drug name
Antidepressants	Amitriptyline
NSAIDs	Aspirin (300 mg)
	Celecoxib
	Diclofenac
	Ibuprofen
	Meloxicam
	Nabumetone
	Naproxen
	Piroxicam
Opioid analgesics	Codeine
	Co-codamol
	Co-dydramol
	Dihydrocodeine
	Morphine
	Oxycodone
	Tramadol
Non-opioid analgesics	Paracetamol
Laxatives	Lactulose
	Macrogol
	Movicol
	Senna
Ulcer prevention	Lansoprazole
	Omeprazole
	Rabeprazole
	Ranitidine

Appendix 23 – Unit costs for consultation events, by staff role and event category

Staff role	Event category	Cost (£)[1]	Note
All included in GP group	Day visit	31.00	Unit cost per surgery consultation lasting 11.7 minutes[2]
	Night visit	104.00	Unit cost per home visit lasting 23.4 minutes[2]
	Phone call	19.00	Unit cost per telephone consultation lasting 7.1 minutes[2]
Acupuncturist	Day visit	25.00[3]	Unit cost per session as charged by South Warwickshire NHS Foundation Trust
	Phone call	6.25	Equivalent to 15 minutes of the unit cost per hour[4]
Chiropodist	Day visit	31.00	Unit cost per hour provided by a Community chiropodist/podiatrist
	Phone call	7.75	Equivalent to 15 minutes of the unit cost per hour[4]
Community nurse	Day visit	36.50	Equivalent to 30 minutes of the unit cost per hour of home visiting (including travel) [4]
	Night visit	73.00	Unit cost per hour of home visiting (including travel)[4]
	Phone call	9.75	Equivalent to 15 minutes of the unit cost per hour of Practice nurse[4]
Consultant	Day visit	81.00	Equivalent to 30 minutes of the unit cost per contract hour [4]
	Night visit	104.00	Unit cost per home visit lasting 23.4 minutes by a GP[2]
	Phone call	19.00	Unit cost per telephone consultation lasting 7.1 minutes by a GP [2]
Contact Tracing Nurse	Day visit	36.50	Equivalent to 30 minutes of the unit cost per hour of home visiting by Community nurse (including travel) [4]
Dietician	Day visit	35.00	Unit cost per hour [4]
	Night visit	34.00	Lowest cost reported for a night visit (Physiotherapist)
	Phone call	8.75	Equivalent to 15 minutes of the unit cost per hour[4]
Health visitor	Day visit	36.50	Generally a nurse, hence same costs as Community nurse
	Night visit	73.00	
	Phone call	9.75	
Hospital Nurse	Day visit	20.00	Equivalent to 30 minutes of the unit cost per hour provided by a Nurse, day ward (includes staff nurse, registered nurse, registered practitioner) [4]
	Phone call	10.00	Equivalent to 15 minutes of the unit cost per hour[4]
Other Health Care Professional	Day visit	12.75	Lowest cost reported for a consultation (Practice Nurse)
	Night visit	34.00	Lowest cost reported for a night visit (Physiotherapist)
	Phone call	8.75	Lowest cost reported for a phone call (Dietician)
Physiotherapist	Day visit	35.00	Unit cost per hour provided by Hospital physiotherapist[4]
	Night visit	34.00	Unit cost per hour provided by Community physiotherapist[4]
	Phone call	8.75	Equivalent to 15 minutes of the unit cost per hour by Hospital physiotherapist[4]
Practice Nurse	Day visit	12.75	Equivalent to 15 minutes of the unit cost per hour of face-to-face contact[4]
	Night visit	73.00	Unit cost per hour of home visiting by a Community Nurse (including travel)[4]
	Phone call	9.75	Equivalent to 15 minutes of the unit cost per hour[4]

[1] All figures in 2011 pound sterling and, unless otherwise noted, taken from the Unit Costs of Health and Social Care 2011 published by the Personal Social Services Research Unit [116].

[2] Excluding direct care staff costs but including qualification costs.

[3] Since the Unit Costs of Health and Social Care 2011 does not report NHS reference costs for services provided by an acupuncturist, the figure was extracted from the South Warwickshire NHS Foundation Trust's website (http://www.swft.nhs.uk/our-services/acupuncture.aspx), where it is reported that they do not commission acupuncture services any more, but offer private sessions at a cost of £25.

[4] Including qualification costs.

Appendix 24 – Unit costs for medication

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
Antidepressant							
Amitriptyline	Oral	M07191001	Oral solution	50mg/5ml	150	18.19	0.1213
	Tablets	4005611	Tabs	25 mg	28	0.83	0.0296
		4005610	Tabs	10 mg	28	0.84	0.0300
		4013478	Tabs	50 mg	28	0.92	0.0329
Non-steroidal anti-inflammatory drugs (NSAIDs)							
Aspirin	Tablets	M05745001	Enteric coated tabs	300 mg	100	6.47	0.0647
(300 mg)		M01580002	Dispersible tabs	300 mg	100	2.88	0.0288
		M01580001	Tabs	300 mg	32	0.31	0.0097
		4000499	NU-seals EC tabs	300 mg	100	4.15	0.0415
		M05290001	Soluble tabs	500 mg	100	2.88	0.0288
		M05745002	Enteric coated tabs	600 mg	100	6.47	0.0647
		4000500	NU-seals EC tabs	600 mg	100	4.15	0.0415
		M05783002	Mod rel tabs	300 mg	100	6.47	0.0647
		M08158001	Orodispersible tabs	300 mg	100	2.88	0.0288
		4004577	Alphar/Cox disp tabs	300 mg	100	2.88	0.0288
	Suppository	M07114001	Suppository	300 mg	12	59.28	4.9400
Celebrex	Capsules	4085926	Capsules	200 mg	30	21.55	0.7183
(Celecoxib)		4085925	Capsules	100 mg	60	21.55	0.3592
Celecoxib	Capsules	M08418002	Capsules	200 mg	30	21.55	0.7183
		M08418001	Capsules	100 mg	60	21.55	0.3592
Diclofenac	Capsules	M01029001	Sodium, modified release	75 mg	56	11.40	0.2036
		M07347001	Sodium, dual release	75 mg	56	11.40	0.2036
	Suppository	M03600001	(suppository)	100 mg	10	3.23	0.3230
		M07395002	(suppository)	50 mg	10	3.23	0.3230
		M07395001	(suppository)	25 mg	10	3.23	0.3230
		M03600002	(suppository)	12.5 mg	10	3.23	0.3230
	Tablets	M10655002	Sodium	50 mg	84	1.42	0.0169
		M10631001	Sodium, modified release	75 mg	56	11.40	0.2036
		M03599002	Sodium	50 mg	84	1.42	0.0169
		M10631002	Sodium, modified release	100 mg	28	8.20	0.2929
		M06914001	Sodium + Misoprostol	50 mg + 200 mcg	60	11.98	0.1997
		M10655001	Sodium	25 mg	84	1.07	0.0127
		M06783003	Sodium, modified release	100 mg	28	8.20	0.2929
		M06914002	Sodium + Misoprostol	75 mg + 200 mcg	60	15.83	0.2638
		M03599001	Sodium	25 mg	84	1.07	0.0127
		4076703	Voltarol	25 mg	30	3.46	0.1153
		M06461001	Sodium	50 mg	84	1.42	0.0169

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
		M08579002	Potassium	50 mg	28	6.18	0.2207
		M08579001	Potassium	25 mg	28	3.23	0.1154
	Tablets	4074156	(Flexotard)	100 mg	28	8.20	0.2929
		4089401	Sodium, enteric coated tablets	50 mg	84	1.42	0.0169
		4013225	Sodium	50 mg	84	1.42	0.0169
		4070157	Sodium, enteric coated tablets	50 mg	84	1.42	0.0169
Ibuprofen	Capsules	M08918002	Capsules	200 mg	84	1.44	0.0171
	Granules	M00698002	Granules	600 mg	20	6.53	0.3265
		M08918001	Granules	400 mg	20	6.53	0.3265
	Tablets	M02873002	Tabs	400 mg	84	1.73	0.0206
		M02873003	Tabs	600 mg	84	3.96	0.0471
		M02873001	Tabs	200 mg	84	1.44	0.0171
		M00698001	Mod rel tabs	800 mg	56	6.48	0.1157
		M05124001	Tabs	800 mg	56	6.48	0.1157
		M01235001	Mod rel tabs	200 mg	84	1.44	0.0171
		M08918003	Orodispersible tabs	200 mg	84	1.44	0.0171
		M03859001	Mod rel tabs	300 mg	120	9.64	0.0803
		M06133001	Soluble tabs	200 mg	84	1.44	0.0171
		4006462	Hillcross tabs	400 mg	84	1.73	0.0206
		M13370001	Dexibuprofen	400 mg	60	9.47	0.1578
		M13364001	Dexibuprofen	300 mg	60	9.47	0.1578
		M11766001	Tabs	200 mg	84	1.44	0.0171
Meloxicam	Tablets	M09638001	Tabs	7.5 mg	30	1.12	0.0373
		M09638002	Tabs	15 mg	30	1.32	0.0440
Nabumetone	Tablets	M05092001	Tabs	500 mg	56	4.75	0.0848
	Oral	M05092002	Suspension	500 mg / 5 ml	300	24.08	0.0803
Naproxen	Tablets	M02771002	Tabs	500 mg	28	1.65	0.0589
		M02771001	Tabs	250 mg	28	1.25	0.0446
		M06830002	Enteric coated tabs	500 mg	56	4.98	0.0889
		M06830001	Enteric coated tabs	250 mg	56	3.01	0.0538
		M06830003	Enteric coated tabs	375 mg	56	26.82	0.4789
		M04687001	+ Misoprostol: Tabs	500 mg + 200 mcg	56	23.76	0.4243
		M04246001	Tabs	275 mg	60	7.1	0.1183
		M06864003	Mod rel tabs	500 mg	28	1.65	0.0589
		M02771003	Tabs	375 mg	56	26.82	0.4789
		M11866001	+ Misoprostol: Tabs	500 mg + 200 mcg	56	23.76	0.4243
		4012916	Timpron	500 mg	28	1.65	0.0589
		4012915	Timpron	250 mg	28	1.25	0.0446
	Granules	M06864001	Granules	500 mg	28	1.65	0.0589
		M06864002	Granules	375 mg	56	26.82	0.4789
Piroxicam	Capsules	M04501001	Capsules	10 mg	56	13.32	0.2379
		M04501002	Capsules	20 mg	28	19.04	0.6800
		4002893	Feldene caps	10 mg	30	3.86	0.1287

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
		4002894	Feldene caps	20 mg	30	7.71	0.2570
	Tablets	M00517001	Betadex tabs	20 mg	30	13.82	0.4607
		M04501003	Orodispersible tabs	20 mg	28	32.41	1.1575
		M05190002	Dispersible tabs	20 mg	28	32.41	1.1575
		M05190001	Dispersible tabs	10 mg	56	9.96	0.1779
		4012576	Feldene melt tabs	20 mg	30	10.53	0.3510
		4002915	Feldene dispersible tabs	20 mg	28	32.41	1.1575
		4002914	Feldene dispersible tabs	10 mg	56	9.96	0.1779
Opioid analgesics							
Codeine	Oral	M01078001	Codeine Phosphate: syrup	25 mg / 5 ml	100	0.98	0.0098
	Tablets	M01077002	Codeine Phosphate	30 mg	28	1.18	0.0421
		M01077001	Codeine Phosphate	15 mg	28	0.97	0.0346
		M01077003	Codeine Phosphate	60 mg	28	1.67	0.0596
		M01906002	Codeine Phos + Aspirin: Disp tabs	8 mg + 400 mg	100	42.03	0.4203
		M01906001	Codeine Phos + Aspirin: Tabs	8 mg + 400 mg	100	42.03	0.4203
		M02916001	Ibuprofen + Codeine: mod rel tab	300 mg + 20 mg	28	1.18	0.0421
		M07878002	Ibuprofen + Codeine: Tab	200 mg + 12.8 mg	28	1.18	0.0421
		M05643002	Aspirin + codeine + Cafeine: Tabs	300mg + 8mg + 105mg	100	42.03	0.4203
		M03782003	Codeine + Aspirin: Soluble tabs	8 mg + 500 mg	100	42.03	0.4203
		M07878001	Ibuprofen + Codeine: Tab	200 mg + 12.5 mg	28	1.18	0.0421
Co-codamol	Capsules	M08876001	Capsules	30 mg + 500 mg	100	5.38	0.0538
		M08876003	Capsules	8 mg + 500 mg	20	1.71	0.0855
	Tablets	M03246001	Tabs	8 mg + 500 mg	100	1.47	0.0147
		M03246002	Disp. Tabs	8 mg + 500 mg	100	4.2	0.0420
		M03246003	Tabs	30 mg + 500 mg	100	3.54	0.0354
		M08876002	Eff Tabs	8 mg + 500 mg	100	4.2	0.0420
		M08193001	Tabs	15 mg + 500 mg	100	8.25	0.0825
		4068499	Eff Tabs	30 mg + 500 mg	100	7.56	0.0756
		4067890	Tabs	30 mg + 500 mg	100	3.54	0.0354
		M06035001	Paracet + Codei + Buclizine: Tabs	500mg + 8mg + 6.25mg	48	4.53	0.0944
		M07951003	Paracet + Codei + Cafeine: Tabs	500mg + 8 mg + 30mg	100	1.47	0.0147
Co-dydramol	Oral	M13079001	Sugar-free suspension	500 mg + 10 mg / 5 ml	30	1.06	0.0353
	Tablets	M03623001	Tabs	500 mg + 10 mg	30	1.06	0.0353
		4056883	Tabs	500 mg + 10 mg	30	1.06	0.0353

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
	Tablets	M04541003	Tabs	500 mg + 20 mg	112	10.58	0.0945
		M07533001	Tabs	500 mg + 30 mg	56	6.54	0.1168
		M00561002	Effervescent tabs	500 mg + 20 mg	112	10.58	0.0945
		M00561003	W/Paracetamol forte: Effervescent tabs	500 mg + 30 mg	56	6.54	0.1168
		M04541002	Tabs	500 mg + 7.46 mg	30	1.06	0.0353
		4085966	Tabs	500 mg + 30mg	56	6.54	0.1168
Dihydrocodeine	Oral	M03620002	Elixir	10 mg / 5 ml	150	3.5	0.0233
	Tablets	M03620001	Tabs	30 mg	28	1.39	0.0496
		M03622001	Mod rel tabs	60 mg	56	5.18	0.0925
		M03622003	Mod rel tabs	120 mg	56	10.91	0.1948
		M03622002	Mod rel tabs	90 mg	56	8.66	0.1546
		M03620003	Tabs	40 mg	100	11.51	0.1151
Morphine	Capsules	M08798001	Mod rel caps	10 mg	60	3.3	0.0550
		M08798002	Mod rel caps	30 mg	60	9.24	0.1540
		M08798003	Mod rel caps	60 mg	60	15.39	0.2565
		4066862	Mod rel caps	100 mg	60	28.54	0.4757
		M10321001	Mod rel caps	20 mg	60	9.61	0.1602
		M10321003	Mod rel caps	100 mg	60	24.37	0.4062
		4068218	Mod rel caps	200 mg	60	48.74	0.8123
		M09645002	Mod rel caps	60 mg	28	14.95	0.5339
		M09645001	Mod rel caps	30 mg	28	10.91	0.3896
		M09645003	Mod rel caps	90 mg	28	22.04	0.7871
		M10321002	Mod rel caps	50 mg	60	15.39	0.2565
		M09646001	Mod rel caps	120 mg	28	29.15	1.0411
	Oral	M05227001	Oral solution	10 mg / 5 ml	500	7.47	0.0149
		M05227002	Unit dose vial	10 mg / 5 ml	500	7.47	0.0149
		M06390001	Concentrated oral sol	20 mg / ml	120	18.59	0.1549
		M05227003	Unit dose vial	30 mg / 5 ml	500	7.47	0.0149
		M12943001	Solution for infusion	50 mg / 10 ml	500	7.47	0.0149
		M04207001	Tincture	10 mg / ml	500	7.47	0.0149
	Sachets	4060027	Granules for suspension (sachets)	30 mg	30	25.54	0.8513
		4085443	Granules for suspension (sachets)	20 mg	30	24.58	0.8193
		M02206002	Mod rel gran for suspens (sachets)	20 mg	30	24.58	0.8193
	Tablets	4085451	MST continus	30 mg	60	12.41	0.2068
		4085452	MST continus	60 mg	60	24.22	0.4037
		4085453	MST continus	100 mg	60	38.34	0.6390
		M04209001	Mod rel tabs	10 mg	60	3.3	0.0550
		M04209002	Mod rel tabs	30 mg	60	7.89	0.1315

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
		4085450	MST continus	15 mg	60	9.61	0.1602
	Tablets	M06816001	Tabs	10 mg	56	5.29	0.0945
		M04209003	Mod rel tabs	60 mg	60	15.39	0.2565
		4085454	MST continus	200 mg	60	81.34	1.3557
		M04210001	Mod rel tabs	100 mg	60	24.37	0.4062
		M06816002	Tabs	20 mg	56	10.57	0.1888
		M02206001	Mod rel tabs	15 mg	60	9.61	0.1602
		M04210003	Mod rel tabs	5 mg	60	3.29	0.0548
		4063284	Diamorphine (Heroin)	10 mg	100	16.97	0.1697
		M06816003	Tabs	50 mg	56	28.02	0.5004
		M04210002	Mod rel tabs	200 mg	60	48.74	0.8123
		4060706	Tabs	5 mg	60	3.29	0.0548
		M01916001	Tabs	10 mg	56	5.29	0.0945
Oxycodone	Capsules	M09042002	Capsules	10 mg	56	22.76	0.4064
		M09042001	Capsules	5 mg	56	11.38	0.2032
		M09042003	Capsules	20 mg	56	45.51	0.8127
	Oral	M07693002	Oral liquid	5 mg / 5 ml	250	8.7	0.0348
		M07693003	Oral liquid concentrate	10 mg / ml	120	41.8	0.3483
	Tablets	M08575001	Mod rel tabs	10 mg	56	24.91	0.4448
		M08575002	Mod rel tabs	20 mg	56	49.82	0.8896
		M08575003	Mod rel tabs	40 mg	56	99.66	1.7796
		M07693001	Mod rel tabs	80 mg	56	199.33	3.5595
		M06549001	Mod rel tabs	5 mg	28	12.46	0.4450
Tramadol	Capsules	M07322001	Capsules	50 mg	100	1.99	0.0199
		M08846002	Mod rel caps	100 mg	60	12.14	0.2023
		M08841001	12 hr mod rel caps	100 mg	60	12.14	0.2023
		M08849001	Mod rel caps	200 mg	30	14.98	0.4993
		M08846001	Mod rel caps	50 mg	60	4.55	0.0758
		M08846003	Mod rel caps	150 mg	28	10.7	0.3821
		M08841003	12 hr mod rel caps	200 mg	60	24.28	0.4047
		M08841002	12 hr mod rel caps	150 mg	60	18.21	0.3035
		M09913002	24 hr mod rel caps	200 mg	30	14.98	0.4993
		M09913001	24 hr mod rel caps	150 mg	28	10.7	0.3821
		M09913003	24 hr mod rel caps	300 mg	30	22.47	0.7490
		M09914001	24 hr mod rel caps	400 mg	28	28.51	1.0182
		M08849002	12 hr mod rel caps	75 mg	60	12.14	0.2023
		4085920	Ivax capsules	50 mg	100	1.99	0.0199
	Tablets	M07322003	Soluble tabs	50 mg	100	13.33	0.1333
		M12135001	Orodispersible tabs	50 mg	60	7.12	0.1187
Non-opioid analgesics							
Paracetamol	Capsules	M04531003	Capsules	500 mg	100	3.13	0.0313
		M05802001	+ Isometheptene Mucate caps	325 mg + 65 mg	30	5.5	0.1833
		4021531	Sterwin caplets	500 mg	100	1.61	0.0161

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
		4108268	Ivax caplets	500 mg	100	1.61	0.0161
	Capsules	M09879001	Capsules	120 mg	100	3.13	0.0313
	Granules	4070130	Eli	125 mg	16	0.89	0.0556
	Oral	M05478001	Sugar-free suspension	120 mg / 5 ml	500	2.25	0.0045
		M02255002	Oral suspension	120 mg / 5 ml	500	2.25	0.0045
		M04531002	Suspension	250 mg / 5 ml	500	3.3	0.0066
		M05478002	Sugar-free suspension	250 mg / 5 ml	500	3.3	0.0066
		4074102	RP suspension	120 mg / 5 ml	500	2.25	0.0045
		M05478003	Sugar-free suspension	500 mg / 5 ml	500	3.3	0.0066
		M09879002	Syrup	125 mg / 5 ml	500	2.25	0.0045
		M08377001	+Diphenydramine hydrochloride sugar-free syrup	120 mg + 12.5 mg / 5 ml	500	2.25	0.0045
		4057888	AAH(Vant) suspension 6 plus	250 mg / 5 ml	500	3.3	0.0066
	Sachets	M05075002	+ Metoclopramide sachets	500 mg + 5 mg	42	12.52	0.2981
		M09879003	Shachets	1 g	42	12.52	0.2981
	Suppository	M06526001	Suppository	120 mg	10	11.5	1.1500
		M06526002	Suppository	125 mg	10	11.5	1.1500
		M06526003	Suppository	500 mg	10	37.74	3.7740
		M09876001	Suppository	240 mg	10	23	2.3000
		M09876003	Suppository	250 mg	10	23	2.3000
		M09876002	Suppository	60 mg	10	9.96	0.9960
		4070159	Aurumpharm suppository	500 mg	10	37.74	3.7740
		4091814	Suppository	60 mg	10	9.96	0.9960
	Tablets	M02764001	+ Dextropropoxyphene	325 mg + 32.5 mg	60	9.68	0.1613
		M02773001	Tabs	500 mg	100	1.61	0.0161
		M04531001	Soluble tabs	500 mg	60	4.18	0.0697
		4003216	Pandol soluble tabs	500 mg	100	1.61	0.0161
		M05075001	+ Metoclopramide tabs	500 mg + 5 mg	42	9.64	0.2295
		M12383001	+ Tramadol	325 mg + 37.5 mg	60	9.68	0.1613
		M11221001	+ Domperidone tabs	500 mg + 10 mg	42	9.64	0.2295
		M02255001	Soluble tabs	120 mg	16	0.89	0.0556
		M11800001	Dissolving tabs	250 mg	60	4.18	0.0697
		M09978002	+ Methionine tabs	500 mg + 100 mg	100	1.61	0.0161
		M04532002	+ Aspirin dispersible tabs	200 mg + 300 mg	100	1.61	0.0161
Laxatives							
Lactulose	Oral	M03969001	Solution	3.35 gr / 5 ml	500	2.25	0.0045
		M03969003	Solution	3.1-3.7 gr / 5 ml	500	2.25	0.0045
		M03969002	Solution (flavoured)	3.35 gr /	500	2.25	0.0045

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
				5 ml			
Macrogol	Sachets	M10872001	Polyethylene glycol w/electrolytes oral powder	(13.125)	30	6.68	0.2227
		M00398001	NPF Oral powder	10 gr	30	6.68	0.2227
		M10769001	Compound NPF Oral powder	6.9 gr	30	6.68	0.2227
		M11928001	4000 - Powder	10 gr	30	6.68	0.2227
		M06378001	+ Sodium sulphate + electrolytes powder		30	6.68	0.2227
Movicol	Sachets	4072795	Sachets	(13.125)	50	11.13	0.2226
		4103590	Half oral powder	(6.563)	30	4.01	0.1337
Senna	Granules	M05600001	Ispaghula husk w/ Senna fruits granules	54.2% + 12.4%	400	7.45	0.0186
		M04701003	Granules		400	7.45	0.0186
	Oral	M04701002	Syrup	7.5 mg / 5 ml	500	2.69	0.0054
		M13487001	Oral solution	7.5 mg / 5 ml	500	2.69	0.0054
	Tablets	M04701001	Tabs	7.5 mg	60	1.44	0.0240
		M10618001	Tabs	15 mg	60	1.44	0.0240
		4003130	Senokot	7.5 mg	60	1.44	0.0240
		M09140001	Tabs	12 mg	60	1.44	0.0240
		M10618002	Chewable tabs	15 mg	60	1.44	0.0240
		4001041	APS tabs	7.5 mg	60	1.44	0.0240
Ulcer prevention							
Lansoprazole	Capsules	M07262002	Caps of e/c granules	15 mg	28	1.2	0.0429
		M07262001	Caps of e/c granules	30 mg	28	1.86	0.0664
	Tablets	M10810001	Orodispersible gastro-resistant tab	15 mg	28	2.99	0.1068
		M10843001	Orodispersible gastro-resistant tab	30 mg	28	5.5	0.1964
		M07926001	Gastro-resistant granules for oral susp	30 mg	28	5.5	0.1964
Omeprazole	Capsules	M05588001	Caps	20 mg	28	1.62	0.0579
		M05588003	Caps	10 mg	28	1.62	0.0579
		M05588002	Gastro-resistant caps	40 mg	7	1.65	0.2357
		4085830	Ivax Gastro-resistant caps	20 mg	28	1.62	0.0579
	Tablets	M11034001	Esomeprazole tabs	20 mg	28	13.88	0.4957
		M11034002	Esomeprazole tabs	40 mg	28	18.89	0.6746
		M07553002	Dispersible tabs	20 mg	28	1.62	0.0579
		M07657001	Gastro-resistant tabs	20 mg	28	4.89	0.1746
		M07553001	Dispersible tabs	10 mg	28	1.62	0.0579
		M07656001	Gastro-resistant tabs	10 mg	28	5.72	0.2043
		M07553003	Dispersible tabs	40 mg	7	1.65	0.2357
		M10574001	Gastro-resistant tabs	40 mg	7	5.72	0.8171
Rabeprazole	Tablets	M10617002	Gastro-resistant tabs	20 mg	28	19.55	0.6982
		M10617001	Gastro-resistant tabs	10 mg	28	11.56	0.4129
Ranitidine	Oral	M06886001	Syrup	150 mg / 10 ml	300	19.31	0.0644

Drug	Presentation	CPRD			BNF		
		Product code	Description	Strength	Quantity	Price (£)	Unit price
	Tablets	M04666001	Tabs	150 mg	60	1.48	0.0247
		M04666002	Tabs	300 mg	30	1.57	0.0523
	Tablets	M04667002	Effervescent tabs	150 mg	60	16	0.2667
		M08514001	Tabs	75 mg	60	1.48	0.0247
		M04667003	Effervescent tabs	300 mg	30	15.05	0.5017
		M08830001	Bismuth citrate tabs	400 mg	30	15.05	0.5017
		4073460	Genus tabs	150 mg	60	1.48	0.0247
		M11500001	Effervescent tabs	75 mg	60	16	0.2667

Notes:

1. CPRD product codes were identified by searching the "product name" field of the CPRD product dictionary. It is hence possible that a number of CPRD product codes associated to the drugs we searched for had not been identified if the name of the drug was not part of the "product name" as recorded in the CPRD.
2. Of all CPRD product codes identified, any reporting a total count of 100 or more was included in the analysis.
3. Some presentations such as gel, creams, powder, injections, and ointments were excluded given the difficulty to ascertain the doses prescribed and corresponding cost.
4. Prices were obtained from the online version of the British National Formulary [117] during the second semester of 2012.

Appendix 25 – Number of consultation events attributable to hip pain on the year immediately prior to THR, data from CPRD

		FEMALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	2.19	4.78	1.96	5.12	1.70	5.28	1.66	5.44
	Night visit	0.00	0.10	-0.00	0.09	0.00	0.16	-0.00	0.13
	Phone call	0.32	1.66	0.32	1.60	0.29	1.51	0.37	2.11
Acupuncturist	Day visit	0.00	0.07	0.00	0.02	-0.00	0.04	-0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	-0.00	0.02	0.00	0.16	0.00	0.22	0.01	0.28
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Community nurse	Day visit	0.02	0.54	0.03	0.73	0.01	1.11	0.02	1.18
	Night visit	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.01
	Phone call	-0.00	0.04	-0.00	0.08	-0.00	0.07	0.00	0.16
Consultant	Day visit	-0.00	0.03	-0.00	0.03	-0.00	0.04	-0.00	0.01
	Night visit	0.00	0.00	-0.00	0.01	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.27	0.01	0.43	0.00	0.27	0.00	0.27
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
	Phone call	0.00	0.09	-0.00	0.03	0.00	0.06	0.00	0.07
Health visitor	Day visit	0.02	0.44	0.02	0.56	0.01	0.48	0.00	0.40
	Night visit	-0.00	0.01	0.00	0.00	-0.00	0.00	0.00	0.00
	Phone call	0.00	0.07	0.01	0.20	0.00	0.12	-0.00	0.09
Hospital nurse	Day visit	-0.00	0.02	0.00	0.07	0.00	0.08	-0.00	0.06
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	-0.00	0.00	-0.00	0.01
Other HCP	Day visit	0.07	0.85	0.06	0.95	0.00	0.92	0.03	1.02
	Night visit	0.00	0.02	-0.00	0.02	-0.00	0.02	-0.00	0.02
	Phone call	0.00	0.08	0.00	0.10	0.00	0.13	0.00	0.16
Physiotherapist	Day visit	0.02	0.58	0.03	0.73	0.03	0.63	0.02	0.50
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.03
Practice nurse	Day visit	0.20	2.62	0.19	2.76	0.04	3.66	0.04	3.10
	Night visit	0.00	0.00	0.00	0.02	-0.00	0.01	0.00	0.03
	Phone call	0.01	0.22	0.02	0.37	0.01	0.31	0.02	0.38

		MALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	1.87	4.43	1.78	4.67	1.63	4.99	2.02	5.37
	Night visit	0.00	0.08	-0.00	0.11	-0.00	0.15	0.01	0.16
	Phone call	0.22	1.24	0.20	1.08	0.19	1.22	0.29	1.82
Acupuncturist	Day visit	-0.00	0.01	0.00	0.05	0.00	0.07	0.00	0.07
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	-0.00	0.06	0.01	0.18	0.00	0.14	-0.00	0.23
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.02	0.49	-0.01	0.57	-0.02	0.83	0.01	1.28
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.08	0.00	0.06	0.00	0.04	0.00	0.05
Consultant	Day visit	-0.00	0.07	-0.00	0.02	-0.00	0.03	-0.00	0.04
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	0.00	0.22	0.01	0.37	0.01	0.33	0.01	0.46
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.03	-0.00	0.03	0.00	0.08	0.01	0.19
Health visitor	Day visit	-0.00	0.18	0.00	0.35	0.02	0.35	0.00	0.23
	Night visit	0.00	0.00	-0.00	0.01	-0.00	0.00	0.00	0.00
	Phone call	-0.00	0.03	0.00	0.06	0.00	0.07	0.00	0.05
Hospital nurse	Day visit	0.00	0.04	0.00	0.11	-0.00	0.05	-0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.00	0.00	0.02	0.00	0.00
Other HCP	Day visit	0.02	0.56	0.05	0.90	0.02	1.02	0.00	1.02
	Night visit	0.00	0.04	-0.00	0.02	-0.00	0.03	-0.00	0.04
	Phone call	-0.00	0.10	0.00	0.11	-0.00	0.10	0.00	0.10
Physiotherapist	Day visit	0.01	0.33	0.01	0.33	0.02	0.46	0.01	0.35
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Practice nurse	Day visit	0.26	2.74	0.17	2.84	-0.06	3.79	0.04	3.85
	Night visit	0.00	0.00	-0.00	0.01	-0.00	0.01	0.00	0.04
	Phone call	0.01	0.32	0.01	0.31	0.01	0.38	-0.00	0.26

Appendix 26 – Consultation costs attributable to hip pain on the year immediately prior to THR

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	74.35	159.71	66.80	169.09	58.20	173.36	58.31	181.41
Acupuncturist	0.03	1.78	0.01	0.57	-0.03	0.89	-0.00	0.79
Chiropractor	-0.04	0.52	0.00	5.11	0.06	6.81	0.30	8.63
Community nurse	0.79	19.85	1.02	26.49	0.22	40.53	0.72	43.18
Consultant	-0.02	2.22	-0.05	2.87	-0.04	3.18	-0.02	0.80
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.11	9.66	0.17	15.03	0.03	9.32	0.03	9.45
Health visitor	0.74	16.31	0.92	21.29	0.37	17.86	0.03	14.73
Hospital nurse	-0.01	0.37	0.02	1.40	0.01	1.67	-0.00	1.15
Other HCP	0.99	10.99	0.73	12.30	0.04	11.86	0.42	13.22
Physiotherapist	0.85	20.39	1.16	25.58	1.04	22.17	0.75	17.53
Practice nurse	2.60	33.57	2.57	35.88	0.62	47.02	0.71	40.17

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	62.56	144.28	58.70	150.69	53.87	162.96	68.74	176.79
Acupuncturist	-0.01	0.30	0.00	1.36	0.01	1.65	0.04	1.84
Chiropractor	-0.05	1.85	0.16	5.57	0.03	4.27	-0.10	7.19
Community nurse	0.93	18.10	-0.34	20.78	-0.70	30.21	0.44	46.88
Consultant	-0.19	6.06	-0.02	1.71	-0.05	2.48	-0.08	3.39
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	0.12	7.53	0.44	12.92	0.25	11.75	0.56	16.36
Health visitor	-0.04	6.61	0.17	13.08	0.66	13.06	0.18	8.18
Hospital nurse	0.03	0.76	0.08	2.11	-0.00	0.86	-0.01	0.55
Other HCP	0.30	7.74	0.66	11.60	0.22	13.25	-0.01	13.15
Physiotherapist	0.27	11.63	0.34	11.52	0.70	16.08	0.52	12.17
Practice nurse	3.41	35.62	2.21	36.57	-0.61	48.77	0.48	49.39

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from PPSRU [116].

Appendix 27 – Prescription costs attributable to hip pain on the year immediately prior to THR, data from CPRD

	FEMALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.20
Aspirin	0.18	5.20	-0.08	2.48	-0.10	2.08	-0.17	2.07
Celecoxib	3.11	32.07	6.92	46.23	5.75	42.36	5.47	38.84
Cocodamol	4.04	21.74	4.32	22.11	3.96	21.49	2.78	19.70
Codeine	2.12	12.20	2.15	14.19	1.57	12.29	1.71	13.83
Codydramol	0.63	14.42	0.61	24.52	0.56	15.17	0.69	24.91
Diclofenac	10.37	35.61	7.84	31.51	6.14	26.81	3.42	20.55
Dihydrocodeine	2.11	14.92	1.47	12.44	0.81	11.45	0.84	9.28
Ibuprofen	0.52	7.86	0.98	7.28	0.95	6.66	0.70	5.67
Lactulose	-0.07	2.21	-0.09	3.65	-0.24	5.05	-0.70	6.77
Lansoprazole	-0.13	4.32	-0.28	5.19	-0.19	5.63	-0.35	4.91
Macrogol	-0.01	0.36	0.00	0.56	-0.01	0.71	-0.00	1.97
Meloxicam	0.30	4.12	0.24	2.14	0.25	2.28	0.20	1.87
Morphine	0.60	11.31	0.72	18.83	-0.13	15.76	0.25	12.98
Movicol	0.03	5.14	0.16	6.65	-0.12	5.07	-0.14	7.15
Nabumetone	0.09	2.70	0.13	4.09	0.10	3.96	0.05	1.96
Naproxen	0.98	9.03	1.02	10.84	0.84	11.70	0.13	4.96
Omeprazole	-0.44	13.74	-0.44	14.81	-0.29	17.39	0.06	13.61
Oxycodone	0.10	15.43	0.50	20.83	0.20	14.96	1.06	47.12
Paracetamol	12.37	69.79	15.42	80.15	14.71	82.25	12.75	104.95
Piroxicam	1.01	12.63	0.62	15.95	0.79	17.28	1.08	17.77
Rabeprazole	-0.26	13.36	-0.17	17.57	-0.53	20.81	-0.21	20.86
Ranitidine	-0.15	8.59	-0.00	5.05	-0.13	5.93	-0.26	6.33
Senna	-0.07	2.35	0.02	1.56	-0.09	2.56	-0.28	3.17
Tramadol	4.10	29.22	3.50	26.57	2.04	20.06	1.64	18.01

	MALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	-0.08	2.67
Aspirin	-0.06	1.60	-0.24	2.70	-0.23	2.92	0.01	3.00
Celecoxib	2.93	31.47	3.82	41.26	5.34	38.87	4.33	33.23
Cocodamol	3.85	21.31	4.00	22.15	2.69	19.95	2.69	15.63
Codeine	1.39	8.87	1.92	12.50	1.99	11.58	1.20	10.56
Codydramol	0.29	5.75	0.39	12.69	0.29	4.97	-0.04	8.84
Diclofenac	8.45	32.10	8.62	32.10	7.41	29.27	6.40	25.76
Dihydrocodeine	1.35	12.72	1.11	11.68	0.97	9.26	0.66	7.09
Ibuprofen	0.49	5.42	0.87	6.95	1.17	7.39	1.13	7.09
Lactulose	-0.03	2.16	0.07	3.19	-0.15	4.84	-0.48	6.79
Lansoprazole	-0.26	4.44	-0.40	5.14	-0.29	5.71	-0.54	5.42
Macrogol	0.00	0.00	0.01	0.77	0.03	1.80	-0.05	0.84
Meloxicam	0.29	2.17	0.23	2.14	0.13	1.61	0.13	1.44
Morphine	0.87	19.22	0.32	20.64	-0.07	15.61	0.40	11.69
Movicol	0.02	1.81	-0.09	3.61	-0.18	9.67	-0.81	6.83
Nabumetone	0.08	1.98	0.06	3.20	0.03	2.77	0.02	0.57
Naproxen	0.19	17.22	1.19	15.11	0.93	9.21	0.84	11.14
Omeprazole	-0.25	10.13	-0.49	12.15	-0.92	11.12	-0.77	12.54
Oxycodone	0.06	1.54	0.59	30.37	0.34	14.91	0.00	0.00
Paracetamol	8.28	53.69	11.55	65.93	12.20	71.03	15.06	66.36
Piroxicam	0.93	15.08	0.80	13.70	1.07	18.69	0.20	9.64
Rabeprazole	-0.67	18.04	-0.96	15.19	-1.24	18.99	-0.43	18.49
Ranitidine	-0.10	2.60	-0.14	5.02	-0.14	5.19	-0.60	4.69
Senna	0.00	1.00	-0.06	1.49	-0.04	2.09	-0.13	3.20
Tramadol	2.62	26.79	2.67	24.25	1.91	21.69	1.26	11.81

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from BNF [117].

Appendix 28 – National average NHS reference costs by HRG

Currency Code	Currency Description	National Average Unit Cost (£)
HA11A	Major Hip Procedures Category 2 for Trauma with Major CC	13,600
HA11B	Major Hip Procedures Category 2 for Trauma with Intermediate CC	8,297
HA11C	Major Hip Procedures Category 2 for Trauma without CC	7,477
HA12B	Major Hip Procedures Category 1 for Trauma with CC	8,087
HA12C	Major Hip Procedures Category 1 for Trauma without CC	6,317
HA13A	Intermediate Hip Procedures for Trauma with Major CC	8,233
HA13B	Intermediate Hip Procedures for Trauma with Intermediate CC	6,101
HA13C	Intermediate Hip Procedures for Trauma without CC	5,603
HA14A	Minor Hip Procedures for Trauma with Major CC	7,973
HA14B	Minor Hip Procedures for Trauma with Intermediate CC	4,012
HA14C	Minor Hip Procedures for Trauma without CC	3,696
HA91Z	Hip Trauma Diagnosis without Procedure	3,014
HA96Z	Multiple Trauma Diagnoses without Procedure	2,770
HA99Z	Other Procedures for Trauma	2,415
HB11A	Major Hip Procedures for non Trauma Category 2 with Major CC	11,736
HB11B	Major Hip Procedures for non Trauma Category 2 with CC	6,643
HB11C	Major Hip Procedures for non Trauma Category 2 without CC	6,412
HB12A	Major Hip Procedures for non Trauma Category 1 with Major CC	8,830
HB12B	Major Hip Procedures for non Trauma Category 1 with CC	6,583
HB12C	Major Hip Procedures for non Trauma Category 1 without CC	5,958
HB13Z	Intermediate Hip Procedures for non Trauma Category 2	4,492
HB14B	Intermediate Hip Procedures for non Trauma Category 1 with CC	4,834
HB14C	Intermediate Hip Procedures for non Trauma Category 1 without CC	2,453
HB15D	Minor Hip Procedures for non Trauma Category 2 19 years and over with CC	4,902
HB15E	Minor Hip Procedures for non Trauma Category 2 19 years and over without CC	1,574
HB15F	Minor Hip Procedures for non Trauma Category 2 18 years and under with CC	4,700
HB15G	Minor Hip Procedures for non Trauma Category 2 18 years and under without CC	2,072
HB16B	Minor Hip Procedures for non Trauma Category 1 with CC	5,239
HB16C	Minor Hip Procedures for non Trauma Category 1 without CC	1,429
HB91Z	Other non Trauma Diagnosis without Procedure	2,262
HB99Z	Other Procedures for non Trauma	714
HR01B	Reconstruction Procedures Category 6 with CC	20,400
HR01C	Reconstruction Procedures Category 6 without CC	16,130
HR02Z	Reconstruction Procedures Category 5	9,254
HR03Z	Reconstruction Procedures Category 4	11,062
HR04B	Reconstruction Procedures Category 3 with CC	12,080
HR04C	Reconstruction Procedures Category 3 without CC	8,492
HR05Z	Reconstruction Procedures Category 2	7,340
HR06A	Reconstruction Procedures Category 1 19 years and over	4,587
HR06B	Reconstruction Procedures Category 1 18 years and under	3,900

Note: Extracted from the National Schedule of Reference Costs Year : 2011-12 - All NHS trusts and NHS foundation trusts - HRG Data [126]

Appendix 29 – Relative frequencies of HRGs by patient subgroup for HES primary THR records, 2011-2012

Females, 45-60 years

HRG	Relative frequency
HA11A	0.04%
HA11B	0.04%
HA11C	0.12%
HA12B	0.08%
HA12C	0.33%
HA13B	0.02%
HA13C	0.33%
HA14A	0.02%
HA99Z	0.04%
HB11A	0.16%
HB11B	0.06%
HB11C	2.56%
HB12A	2.36%
HB12B	2.05%
HB12C	90.96%
HB13Z	0.33%
HB14B	0.02%
HB14C	0.06%
HB15D	0.02%
HR01B	0.02%
HR05Z	0.02%
Invalid	0.38%
TOTAL	**100%**

Females, 60-70 years

HRG	Relative frequency
HA11A	0.01%
HA11B	0.01%
HA11C	0.05%
HA12B	0.08%
HA12C	0.13%
HA13A	0.01%
HA13C	0.18%
HA14B	0.01%
HA99Z	0.05%
HB11A	0.17%
HB11B	0.11%
HB11C	1.89%
HB12A	3.01%
HB12B	2.62%
HB12C	91.23%
HB13Z	0.14%
HR01B	0.01%
HR01C	0.02%
HR06A	0.02%
Invalid	0.27%
TOTAL	**100%**

Females, 70-80 years

HRG	Relative frequency
HA11A	0.02%
HA11C	0.04%
HA12B	0.05%
HA12C	0.08%
HA13A	0.02%
HA13B	0.02%
HA13C	0.08%
HA99Z	0.04%
HB11A	0.24%
HB11B	0.10%
HB11C	1.81%
HB12A	4.36%
HB12B	5.24%
HB12C	87.46%
HB13Z	0.17%
HB14C	0.01%
HB16C	0.01%
HR06A	0.01%
Invalid	0.24%
TOTAL	**100%**

Females, 80+ years

HRG	Relative frequency
HA11C	0.07%
HA12B	0.03%
HA12C	0.18%
HA13A	0.10%
HA13B	0.03%
HA13C	0.45%
HA99Z	0.10%
HB11A	0.73%
HB11B	0.23%
HB11C	2.71%
HB12A	8.82%
HB12B	6.74%
HB12C	79.46%
HB13Z	0.17%
HR06A	0.02%
Invalid	0.17%
TOTAL	**100%**

Males, 45-60 years

HRG	Relative frequency
HA11C	0.09%
HA12B	0.09%
HA12C	0.13%
HA13A	0.02%
HA13C	0.24%
HA99Z	0.02%
HB11A	0.15%
HB11B	0.22%
HB11C	2.45%
HB12A	2.62%
HB12B	2.88%
HB12C	89.27%
HB13Z	0.58%
HB14C	0.06%
HR01B	0.02%
Invalid	1.16%
TOTAL	**100%**

Males, 60-70 years

HRG	Relative frequency
HA11B	0.01%
HA11C	0.04%
HA12B	0.04%
HA12C	0.08%
HA13A	0.03%
HA13B	0.01%
HA13C	0.26%
HA99Z	0.01%
HB11A	0.13%
HB11B	0.08%
HB11C	1.66%
HB12A	3.42%
HB12B	4.40%
HB12C	88.98%
HB13Z	0.17%
HB14C	0.03%
HB16C	0.01%
HR01B	0.01%
HR01C	0.03%
HR06A	0.03%
Invalid	0.56%
TOTAL	**100%**

Males, 70-80 years

HRG	Relative frequency
HA11A	0.04%
HA11B	0.01%
HA11C	0.04%
HA12B	0.01%
HA12C	0.13%
HA13A	0.03%
HA13B	0.01%
HA13C	0.23%
HA99Z	0.03%
HB11A	0.26%
HB11B	0.15%
HB11C	1.41%
HB12A	5.32%
HB12B	8.03%
HB12C	83.58%
HB13Z	0.20%
HB14B	0.01%
HB14C	0.01%
HB16C	0.01%
HR01C	0.01%
Invalid	0.47%
TOTAL	**100%**

Males, 80+ years

HRG	Relative frequency
HA11C	0.04%
HA12B	0.16%
HA12C	0.35%
HA13A	0.16%
HA13B	0.08%
HA13C	0.35%
HB11A	0.47%
HB11B	0.27%
HB11C	1.78%
HB12A	9.97%
HB12B	10.47%
HB12C	75.29%
HB13Z	0.16%
HB15D	0.04%
HR06A	0.08%
Invalid	0.35%
TOTAL	**100%**

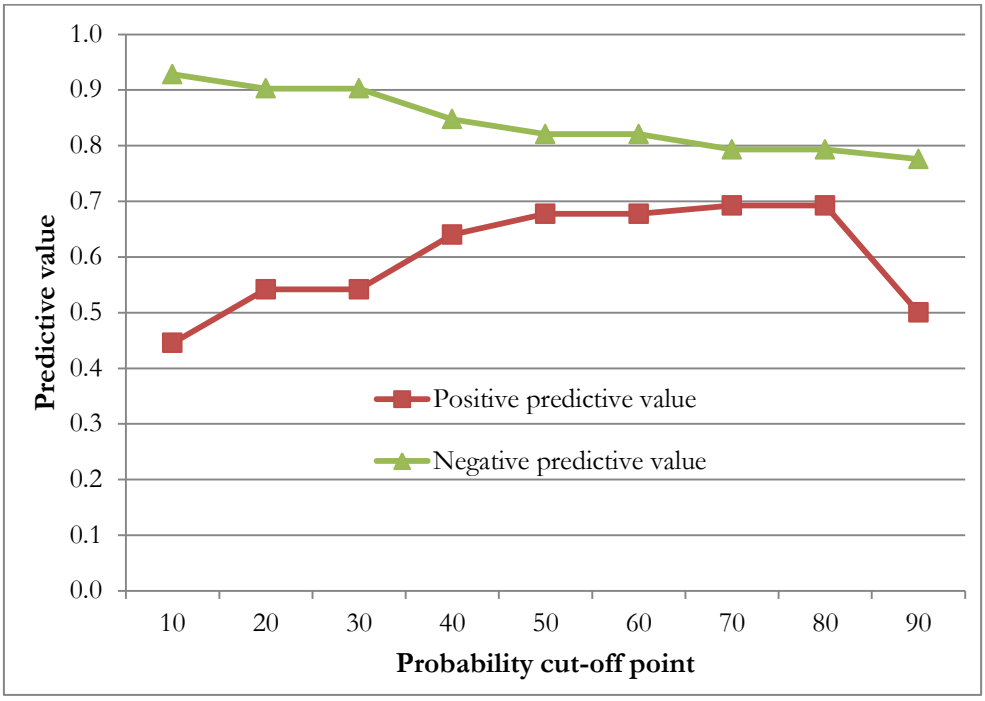
Appendix 30 – Classifying CPRD patients as Good or Poor outcomes based on their reported resource use

After fitting the logistic model predicting *Poor* surgery outcome to CPRD data, the default choice would have been to label as *Poor* all those patients with probability equal to or greater than 0.5, and similarly for *Good* outcomes. Nevertheless, the probability cut-off point affected greatly the proportion of *Good* and *Poor* outcomes and therefore the relative estimates of primary care postoperative costs associated with each.

Applying the 0.5 cut-off point would have assigned only 683 of the 13,756 cases in the CPRD extract for the first year after a primary THR to the *Poor* outcome group. This represented 5% of the cases, when the PROMs data we held, largely representative of the UK's population, reported a 35% proportion of *Poor* outcomes. Keeping the criterion for *Poor* outcome at 0.5 would have forced an unrealistic low proportion of such outcomes which in turn would have produced an unrealistically high estimate of costs for *Poor* outcomes as the model associated higher resource use with higher probability of being in that category. We avoided this because we were aware of the circular connection between resource use and costs through the model for surgical outcome, as well as the fact that some patients labelled as *Poor* outcomes do in fact use very little NHS healthcare resources.

A cut-off point of 0.4 meant, instead, that 1,138 or 30% of the CPRD cases in the first postoperative year were labelled as *Poor* outcomes, very close to the 35% found in PROMs. The proximity in the relative proportion of surgery outcome categories suggested the use of 0.4 as a cut-off point, but we were still concerned about the make-up of each group in terms of accurately predicted *Good* and *Poor* outcomes. We concentrated therefore in observing not only the model's sensitivity and specificity but also and more importantly its predictive values. More than anything, we were after the highest possible positive predictive value (percentage of likely *Poor* outcomes that were actually *Poor* outcomes) whilst keeping the negative predictive value as high as possible. Predictive values were most relevant in this particular case because we were using the records from predicted outcomes to produce estimates of resource use. Hence, it was of outmost importance to have a high proportion of true *Poor outcomes* in the predicted *Poor* outcomes group than to have a high proportion of correctly classified poor outcomes (sensitivity).

Given the circular nature of the logit model (by using GP visits to identify cases with high resource use and associate those to *Poor* outcome), we monitored predictive values and resulting mean use of resources by outcome group to identify an appropriate cut-off point to classify patients when fitting the model to the CPRD. A similar breakdown of *Good* and *Poor* outcomes was produced by the model when using a cut-off point of 0.4. Nonetheless, as shown in the figure below, it was also the case that the positive predictive rate of the model fitted to the COASt data increased with as the probability cut-off point also increased, up to 0.8, and it did so more rapidly than the negative predictive value decreased. A cut-off point of 0.4 therefore produced not only a breakdown of *Good* and *Poor* outcomes close to the proportions observed in the PROMs data, but more importantly it did so by sacrificing only a small portion of correctly classified *Poor* outcome patients whilst predicting accurately a higher proportion of *Good* outcomes than if the default 0.5 cut-off point had been used.



Appendix 31 – Number of consultation events attributable to hip pain on the year immediately after THR for *likely Poor* outcomes

		FEMALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	6.50	4.84	6.02	5.30	6.12	5.08	5.78	5.08
	Night visit	0.04	0.23	0.02	0.19	0.03	0.22	-0.00	0.12
	Phone call	0.83	1.99	0.85	2.03	0.81	2.05	0.89	2.50
Acupuncturist	Day visit	0.00	0.06	-0.00	0.05	0.00	0.05	-0.01	0.16
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.00	0.04	0.02	0.22	0.02	0.29	0.01	0.33
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.32	1.54	0.21	0.93	0.30	1.69	0.48	2.38
	Night visit	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.02
	Phone call	0.02	0.19	-0.00	0.03	0.00	0.08	0.02	0.20
Consultant	Day visit	0.01	0.17	0.00	0.00	0.00	0.06	-0.00	0.01
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.11	0.00	0.12	-0.01	0.24	-0.01	0.14
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.04	-0.00	0.01	0.00	0.10	0.00	0.00
Health visitor	Day visit	0.00	0.28	-0.00	0.21	-0.00	0.30	0.01	0.30
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.01	0.10	-0.00	0.01	-0.00	0.04
Hospital nurse	Day visit	-0.00	0.06	-0.00	0.06	0.00	0.11	-0.00	0.10
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.18	1.08	0.15	1.16	0.13	1.17	0.11	1.11
	Night visit	-0.00	0.01	-0.00	0.01	0.00	0.06	-0.00	0.05
	Phone call	0.00	0.06	0.01	0.12	0.01	0.13	0.00	0.11
Physiotherapist	Day visit	0.14	0.99	0.09	0.78	0.08	0.84	0.04	0.48
	Night visit	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	-0.00	0.01	-0.00	0.01
Practice nurse	Day visit	1.26	7.53	0.62	3.04	0.69	4.99	0.44	4.27
	Night visit	0.00	0.00	-0.00	0.01	0.00	0.03	0.00	0.00
	Phone call	0.10	0.70	0.05	0.41	0.04	0.46	0.03	0.43

		MALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	6.61	4.21	6.01	4.41	6.34	4.47	7.09	5.65
	Night visit	0.00	0.15	0.02	0.21	0.00	0.10	0.01	0.14
	Phone call	0.57	1.33	0.46	1.49	0.62	1.90	0.65	1.71
Acupuncturist	Day visit	0.00	0.00	0.00	0.05	0.00	0.08	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.00	0.00	0.03	0.49	0.00	0.25	0.02	0.27
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.37	2.45	0.17	0.94	0.46	3.73	0.37	1.82
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
	Phone call	0.00	0.00	0.00	0.07	0.00	0.12	0.01	0.14
Consultant	Day visit	0.00	0.00	0.00	0.05	0.00	0.07	-0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.02	0.00	0.32	-0.02	0.17	-0.03	0.33
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	-0.00	0.01	-0.00	0.02
Health visitor	Day visit	0.04	0.93	0.01	0.25	0.02	0.33	0.00	0.24
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	0.00	0.09	0.00	0.00
Hospital nurse	Day visit	0.00	0.00	0.05	0.76	-0.00	0.08	0.00	0.06
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	-0.00	0.03	0.00	0.00
Other HCP	Day visit	0.06	0.43	0.18	1.05	0.35	2.81	0.07	0.95
	Night visit	0.00	0.00	0.00	0.07	0.00	0.04	-0.00	0.02
	Phone call	0.00	0.06	0.01	0.15	0.00	0.08	-0.01	0.09
Physiotherapist	Day visit	0.09	0.70	0.05	0.56	0.04	0.54	0.02	0.35
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Practice nurse	Day visit	0.95	4.02	0.93	3.67	1.03	5.28	0.94	4.10
	Night visit	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.06
	Phone call	0.07	0.45	0.05	0.32	0.03	0.34	-0.01	0.18

Note: Resource use data obtained from CPRD. Classification as *likely Poor* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 32 – Number of consultation events attributable to hip pain on the year immediately after THR for *likely Good* outcomes

		FEMALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	-0.74	2.56	-1.23	2.70	-1.42	2.94	-1.58	3.21
	Night visit	-0.00	0.07	-0.00	0.08	-0.00	0.10	-0.01	0.16
	Phone call	0.26	1.72	0.13	1.14	0.13	1.23	0.18	1.87
Acupuncturist	Day visit	-0.00	0.05	-0.00	0.04	0.00	0.11	0.00	0.16
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.01	0.18	0.00	0.18	-0.00	0.21	-0.01	0.31
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.01	0.38	0.01	0.45	-0.03	0.74	-0.13	0.98
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	-0.00	0.02
	Phone call	-0.00	0.08	0.00	0.06	0.01	0.13	-0.00	0.09
Consultant	Day visit	-0.00	0.01	-0.00	0.03	-0.00	0.02	0.00	0.03
	Night visit	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.02	0.14	-0.01	0.18	-0.00	0.26	-0.01	0.12
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.11	-0.00	0.01	0.00	0.07	0.00	0.08
Health visitor	Day visit	-0.01	0.29	0.01	0.43	0.00	0.41	-0.02	0.25
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.05	-0.00	0.03	0.00	0.16	-0.00	0.07
Hospital nurse	Day visit	0.00	0.03	-0.00	0.05	-0.00	0.05	-0.00	0.02
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	-0.00	0.01	0.00	0.00
Other HCP	Day visit	0.04	0.97	-0.01	0.88	-0.04	0.82	0.01	0.94
	Night visit	-0.00	0.01	-0.00	0.01	-0.00	0.02	-0.00	0.03
	Phone call	-0.00	0.05	-0.00	0.09	0.00	0.13	-0.00	0.10
Physiotherapist	Day visit	0.02	0.50	0.02	0.58	0.01	0.46	0.01	0.42
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Practice nurse	Day visit	-0.08	1.87	-0.18	2.16	-0.28	2.78	-0.29	3.09
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.01
	Phone call	0.01	0.22	-0.00	0.25	0.00	0.24	-0.00	0.38

		MALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	-0.39	2.73	-0.83	2.67	-1.31	3.05	-1.21	3.09
	Night visit	-0.01	0.07	0.00	0.12	-0.01	0.12	-0.01	0.17
	Phone call	0.24	1.28	0.12	0.94	0.08	1.12	0.23	1.41
Acupuncturist	Day visit	-0.00	0.03	-0.00	0.01	-0.00	0.07	0.00	0.04
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	-0.00	0.05	-0.00	0.20	0.00	0.19	-0.01	0.27
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.04	0.44	0.01	0.43	-0.01	0.48	-0.12	1.85
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.01	0.07	0.00	0.07	0.00	0.11	-0.00	0.08
Consultant	Day visit	0.00	0.04	-0.00	0.02	-0.00	0.01	0.00	0.05
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	0.00	0.24	0.03	0.55	0.02	0.44	0.03	0.67
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
	Phone call	0.00	0.00	0.00	0.10	0.00	0.12	0.00	0.02
Health visitor	Day visit	0.01	0.34	0.01	0.32	0.02	0.44	0.03	0.73
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.08	0.00	0.09	-0.00	0.04	-0.00	0.03
Hospital nurse	Day visit	-0.00	0.01	-0.00	0.04	0.00	0.06	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
Other HCP	Day visit	0.01	0.64	0.03	0.87	-0.00	1.00	0.02	1.19
	Night visit	-0.00	0.01	-0.00	0.01	-0.00	0.04	-0.00	0.01
	Phone call	-0.00	0.05	-0.00	0.06	0.00	0.11	-0.00	0.13
Physiotherapist	Day visit	0.01	0.36	0.02	0.41	0.03	0.45	0.01	0.22
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.04	-0.00	0.01	0.00	0.00	-0.00	0.01
Practice nurse	Day visit	0.16	2.41	-0.06	2.52	-0.33	3.31	-0.11	3.32
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	-0.00	0.02
	Phone call	-0.00	0.27	-0.00	0.20	0.00	0.21	-0.00	0.32

Note: Resource use data obtained from CPRD. Classification as *likely Good* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 33 – Consultations costs attributable to hip pain on the year immediately after THR for *likely Poor* outcome patients

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	221.28	168.48	204.82	180.50	207.74	169.77	196.25	172.20
Acupuncturist	0.08	1.44	-0.06	1.28	0.04	1.16	-0.16	3.88
Chiropractor	0.08	1.34	0.57	6.94	0.66	8.88	0.33	10.21
Community nurse	11.72	56.31	7.53	34.00	10.88	61.56	17.66	86.81
Consultant	0.74	14.06	-0.01	0.15	0.06	4.63	-0.06	1.11
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.04	3.73	0.01	4.30	-0.37	8.34	-0.18	4.99
Health visitor	0.08	10.12	-0.04	8.00	-0.03	11.11	0.22	10.89
Hospital nurse	-0.07	1.15	-0.07	1.12	0.10	2.29	-0.08	1.94
Other HCP	2.29	13.84	1.96	14.94	1.85	15.02	1.43	14.18
Physiotherapist	4.77	34.65	3.06	27.13	2.88	29.54	1.42	16.66
Practice nurse	16.97	96.36	8.44	39.30	9.24	63.87	5.88	55.09

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	216.09	140.84	196.62	144.20	208.60	150.27	232.94	183.04
Acupuncturist	0.00	0.00	0.04	1.21	0.09	2.04	0.00	0.00
Chiropractor	0.00	0.00	1.02	15.11	0.10	7.81	0.48	8.27
Community nurse	13.38	89.41	6.12	34.31	16.87	136.06	13.41	66.25
Consultant	0.00	0.00	0.14	3.89	0.21	5.54	-0.13	2.07
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.09	0.87	0.08	11.25	-0.58	6.12	-0.92	11.68
Health visitor	1.28	34.03	0.19	9.14	0.83	12.47	0.16	8.65
Hospital nurse	0.00	0.00	1.02	15.13	-0.05	1.49	0.08	1.28
Other HCP	0.82	5.56	2.43	13.63	4.54	35.88	0.80	12.43
Physiotherapist	3.25	24.56	1.69	19.77	1.27	18.90	0.53	12.26
Practice nurse	12.78	51.99	12.38	47.21	13.54	67.32	12.19	52.37

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 34 – Consultations costs attributable to hip pain on the year immediately after THR for *likely Good* outcome patients

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	-18.21	90.78	-36.23	89.81	-41.86	98.14	-46.28	111.70
Acupuncturist	-0.02	1.13	-0.04	0.99	0.02	2.85	0.02	3.93
Chiropractor	0.25	5.72	0.03	5.52	-0.01	6.58	-0.35	9.73
Community nurse	0.43	13.70	0.54	16.39	-1.30	26.90	-4.65	35.83
Consultant	-0.02	0.60	0.05	3.53	-0.09	1.48	0.01	2.63
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.53	5.11	-0.32	6.17	-0.11	9.26	-0.31	4.46
Health visitor	-0.26	10.79	0.35	15.70	0.03	15.94	-0.57	9.44
Hospital nurse	0.03	0.66	-0.03	1.05	-0.02	1.12	-0.02	0.41
Other HCP	0.47	12.45	-0.15	11.31	-0.48	10.52	0.13	12.11
Physiotherapist	0.59	17.36	0.66	20.18	0.35	16.27	0.52	14.72
Practice nurse	-1.00	24.17	-2.37	27.79	-3.52	35.69	-3.74	39.98

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	-8.05	91.94	-22.91	88.28	-39.69	101.63	-34.11	104.52
Acupuncturist	-0.05	0.66	-0.01	0.24	-0.08	1.77	0.03	0.91
Chiropractor	-0.01	1.58	-0.13	6.07	0.09	5.87	-0.31	8.37
Community nurse	1.59	16.14	0.54	15.71	-0.49	17.57	-4.21	67.44
Consultant	0.15	3.82	-0.11	1.67	-0.04	1.21	0.14	3.80
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	0.03	8.27	1.18	19.72	0.74	15.88	1.15	23.54
Health visitor	0.44	12.27	0.36	11.53	0.66	16.19	1.23	26.81
Hospital nurse	-0.01	0.20	-0.04	0.88	0.01	1.13	0.00	0.00
Other HCP	0.08	8.16	0.35	11.16	-0.05	12.82	0.19	15.32
Physiotherapist	0.47	12.85	0.84	14.40	1.03	15.73	0.35	7.71
Practice nurse	2.06	31.02	-0.76	32.40	-4.17	42.37	-1.45	42.89

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Good* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 35 – Prescription costs attributable to hip pain on the year immediately after THR for *likely Poor* outcome patients

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	-0.11	2.82
Aspirin	0.00	0.18	-0.13	2.44	-0.11	2.10	-0.20	1.56
Celecoxib	2.33	45.84	2.36	41.29	5.33	42.39	3.63	31.60
Cocodamol	6.18	25.08	5.02	23.46	4.53	25.81	3.37	24.31
Codeine	3.13	12.55	3.37	17.62	2.43	14.96	2.39	15.78
Codydramol	0.94	8.69	1.31	11.84	0.61	5.59	0.87	8.15
Diclofenac	6.43	26.06	3.43	23.33	3.38	23.49	1.62	15.55
Dihydrocodeine	3.21	20.20	2.09	13.18	1.74	18.13	2.29	13.68
Ibuprofen	-0.19	4.10	0.06	4.78	-0.09	4.58	0.04	3.97
Lactulose	0.13	2.00	0.33	3.91	0.35	6.10	0.28	7.84
Lansoprazole	0.66	5.38	0.02	6.32	0.02	7.35	0.08	6.69
Macrogol	0.01	0.26	0.03	1.07	0.09	2.70	-0.06	1.43
Meloxicam	0.29	2.18	0.20	2.02	0.10	1.77	0.27	2.04
Morphine	1.66	20.52	0.76	12.11	0.72	10.15	0.71	12.58
Movicol	-0.01	5.97	0.42	4.46	0.05	6.77	0.68	10.39
Nabumetone	0.19	4.26	0.08	2.31	0.09	3.11	0.05	1.37
Naproxen	0.54	5.47	0.74	6.74	0.38	5.33	0.10	3.95
Omeprazole	-0.23	26.14	1.25	21.90	1.16	23.95	1.66	17.68
Oxycodone	0.24	3.15	0.23	5.16	0.93	22.61	6.76	172.23
Paracetamol	20.12	77.23	25.32	88.80	22.63	83.71	13.95	88.34
Piroxicam	0.23	7.99	-0.55	11.20	0.71	14.58	-0.37	5.80
Rabeprazole	1.54	24.06	1.03	26.83	0.43	29.17	0.12	20.28
Ranitidine	0.05	2.81	0.23	3.79	0.10	4.18	-0.35	8.99
Senna	0.08	2.01	0.08	1.51	0.01	2.31	0.22	4.18
Tramadol	7.43	36.67	5.24	28.97	3.03	20.24	2.59	17.33

	MALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aspirin	-0.11	2.08	-0.37	3.14	-0.34	2.82	0.00	3.06
Celecoxib	-0.68	7.71	2.79	36.01	5.39	42.97	2.07	27.93
Cocodamol	4.94	23.43	6.03	23.05	2.96	20.58	4.06	19.78
Codeine	1.19	7.71	1.54	11.08	2.45	12.26	1.18	9.53
Codydramol	1.41	10.41	0.78	6.46	0.53	7.02	0.59	5.15
Diclofenac	6.22	27.75	5.15	24.87	3.14	21.11	3.25	18.64
Dihydrocodeine	2.02	12.07	1.88	12.43	2.00	12.31	0.50	4.09
Ibuprofen	0.72	8.86	0.16	4.78	0.13	6.46	0.28	4.69
Lactulose	-0.16	5.25	0.69	5.35	0.79	5.96	0.10	8.27
Lansoprazole	-0.03	4.01	0.30	6.47	0.41	7.77	0.00	6.59
Macrogol	0.00	0.00	-0.01	0.16	0.15	3.27	-0.05	1.06
Meloxicam	0.27	2.11	0.17	1.80	0.04	0.93	0.02	0.52
Morphine	-0.08	1.09	1.74	36.92	0.17	5.32	0.77	14.23
Movicol	0.14	2.20	-0.08	2.58	0.48	9.31	-0.65	11.08
Nabumetone	0.07	1.01	0.20	5.78	-0.07	1.29	-0.01	0.21
Naproxen	0.21	4.87	0.43	6.39	0.22	4.06	-0.05	3.50
Omeprazole	-0.17	5.85	1.89	25.31	0.01	23.32	0.66	16.65
Oxycodone	0.00	0.00	7.27	120.20	0.31	5.96	0.00	0.00
Paracetamol	13.33	67.69	19.58	72.66	16.57	62.57	7.54	50.87
Piroxicam	-0.40	7.22	1.29	15.47	0.43	13.14	0.12	2.52
Rabeprazole	0.52	29.53	-0.50	13.16	-1.11	26.27	1.05	34.97
Ranitidine	-0.07	2.37	0.12	3.48	0.24	7.65	-0.50	4.99
Senna	0.02	0.70	0.05	2.38	-0.06	3.04	0.35	4.59
Tramadol	5.71	34.54	4.18	24.59	2.71	25.15	3.81	26.15

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from BNF [117]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 36 – Prescription costs attributable to hip pain on the year immediately after THR for *likely Good* outcome patients

	FEMALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.10	2.67	0.00	0.00	0.00	0.00	0.00	0.00
Aspirin	0.21	6.62	-0.07	0.96	-0.08	2.69	-0.15	1.98
Celecoxib	0.08	16.89	1.21	26.11	0.88	21.54	1.44	26.75
Cocodamol	0.48	14.97	0.76	15.22	0.84	15.05	0.26	12.74
Codeine	0.03	5.03	0.15	9.47	0.19	8.75	0.07	9.59
Codydramol	-0.06	1.68	0.37	19.41	0.24	12.37	-0.11	2.36
Diclofenac	2.06	18.02	1.02	16.84	1.46	15.87	0.98	13.63
Dihydrocodeine	-0.07	8.16	0.19	6.08	0.21	6.76	-0.17	4.58
Ibuprofen	0.04	3.54	0.08	4.48	0.04	4.73	0.20	3.86
Lactulose	-0.14	2.13	-0.17	3.43	-0.30	4.60	-0.95	6.12
Lansoprazole	-0.20	3.00	-0.49	4.41	-0.17	4.70	-0.45	5.01
Macrogol	-0.00	0.12	-0.03	1.45	-0.06	2.08	-0.00	0.99
Meloxicam	0.12	1.80	0.10	1.63	0.07	1.50	0.04	1.16
Morphine	-0.06	0.80	0.26	15.46	-0.33	6.30	-0.06	6.54
Movicol	-0.07	2.04	-0.15	7.26	-0.23	5.35	-0.87	6.42
Nabumetone	0.21	3.80	-0.02	1.76	0.05	3.10	-0.05	1.84
Naproxen	0.25	11.80	0.21	4.83	0.36	7.16	-0.07	5.02
Omeprazole	-0.56	8.99	-1.48	12.25	-0.76	15.04	-0.97	11.45
Oxycodone	-0.76	19.70	1.88	61.23	-0.23	7.57	0.00	0.00
Paracetamol	2.36	53.21	3.67	52.66	2.80	64.39	-1.25	58.42
Piroxicam	0.36	8.61	0.92	20.74	-0.02	7.34	0.37	13.77
Rabeprazole	-0.22	13.92	-0.60	13.67	-1.01	18.11	0.06	19.96
Ranitidine	-0.34	6.03	-0.11	2.83	-0.24	6.38	-0.07	6.48
Senna	-0.08	1.13	-0.01	1.17	-0.13	2.35	-0.29	3.47
Tramadol	0.73	14.43	0.64	15.24	-0.11	7.32	0.22	9.86

	MALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aspirin	-0.11	1.00	-0.23	2.46	-0.22	2.40	-0.02	2.19
Celecoxib	0.41	19.04	0.59	27.50	0.75	19.61	0.03	19.23
Cocodamol	0.27	9.17	0.49	12.83	0.19	10.22	-0.23	9.93
Codeine	0.20	5.30	0.37	8.48	0.64	7.91	0.04	7.45
Codydramol	0.06	1.50	-0.39	14.20	-0.11	4.00	0.03	1.93
Diclofenac	1.82	15.77	1.32	19.35	2.08	19.72	1.81	17.61
Dihydrocodeine	0.23	6.28	-0.10	5.63	-0.32	5.33	0.09	1.71
Ibuprofen	-0.16	2.77	0.12	4.01	0.22	5.07	0.04	2.75
Lactulose	-0.03	0.79	-0.06	2.82	-0.39	4.51	-0.56	7.51
Lansoprazole	-0.21	3.90	-0.67	4.29	-0.54	5.72	-0.79	4.91
Macrogol	0.00	0.00	-0.01	0.31	-0.10	1.75	-0.04	0.70
Meloxicam	0.05	1.45	0.04	1.41	-0.01	1.19	0.07	0.89
Morphine	-0.24	7.22	-0.29	4.28	-0.30	5.12	-0.49	9.29
Movicol	-0.08	1.47	-0.00	5.21	-0.18	3.19	-0.74	6.36
Nabumetone	-0.10	2.42	-0.06	1.24	0.02	1.61	0.00	0.00
Naproxen	-0.34	8.73	0.18	8.96	0.08	3.63	0.06	2.87
Omeprazole	-0.47	5.89	-1.04	8.13	-0.87	7.78	-1.50	16.34
Oxycodone	0.00	0.00	-0.08	4.45	-0.58	11.58	-0.12	3.41
Paracetamol	3.01	42.00	0.78	39.46	-0.29	46.18	6.25	48.80
Piroxicam	0.41	6.78	0.14	8.19	0.55	7.96	-0.09	1.44
Rabeprazole	-0.28	16.29	-0.56	14.17	-1.47	14.42	1.58	23.58
Ranitidine	-0.17	1.75	-0.18	2.52	-0.13	4.69	-0.23	5.48
Senna	-0.03	0.80	-0.09	1.57	-0.03	2.05	-0.19	2.78
Tramadol	0.17	14.79	0.14	10.89	-0.36	7.86	-0.12	2.35

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from BNF [117]
4. Classification as *likely Good* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.4.

Appendix 37 – Surgery outcome predictive model assuming a PASS score of 33

Logit regression for *Poor* outcome, estimated on year one primary THR COASt cohort data

Predictor	Coefficient	p-value	95% confidence interval	
Number GP visists = 1 to 4	2.167	0.000	1.345	2.989
Number GP visists = 5 or more	1.261	0.181	-0.587	3.108
Paracetamol?	1.421	0.002	0.541	2.301
Number of opioid drugs	1.101	0.006	0.323	1.878
Constant	-3.493	0.000	-4.191	-2.795
n = 314				
Pseudo R2 = 0.266				

Appendix 38 – Number of consultation events attributable to hip pain for years two through 10 after THR for *likely Poor* outcomes

		FEMALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	4.09	4.40	4.19	4.28	4.28	5.25	4.51	6.31
	Night visit	0.00	0.12	0.01	0.15	0.02	0.25	0.00	0.14
	Phone call	0.70	2.44	0.54	2.72	0.59	2.50	0.73	2.64
Acupuncturist	Day visit	0.01	0.09	0.00	0.04	-0.00	0.10	-0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.01	0.19	0.05	0.47	0.02	0.39	0.00	0.19
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.34	1.69	0.08	0.94	0.31	2.79	0.38	2.19
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
	Phone call	0.00	0.09	0.00	0.35	0.01	0.16	0.01	0.18
Consultant	Day visit	0.00	0.06	0.00	0.04	-0.00	0.01	0.00	0.09
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.08	-0.01	0.24	-0.01	0.15	-0.01	0.10
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.04	0.00	0.00	-0.00	0.01
Health visitor	Day visit	-0.00	0.27	-0.02	0.22	0.00	0.42	0.06	1.84
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.02	0.00	0.04	0.00	0.11	-0.01	0.09
Hospital nurse	Day visit	0.00	0.00	-0.00	0.04	0.00	0.05	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.19	1.74	0.29	2.30	0.15	1.38	0.18	2.30
	Night visit	0.00	0.06	-0.00	0.05	-0.00	0.02	0.00	0.07
	Phone call	0.01	0.10	0.01	0.17	0.00	0.13	0.01	0.15
Physiotherapist	Day visit	0.07	1.05	0.06	0.73	0.07	0.92	0.02	0.60
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Practice nurse	Day visit	1.17	5.62	0.73	4.27	0.45	4.70	0.52	4.79
	Night visit	-0.00	0.03	-0.00	0.02	0.00	0.00	0.00	0.00
	Phone call	0.01	0.26	0.13	1.63	-0.00	0.37	0.01	0.33

		MALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	3.79	4.26	3.88	4.22	4.35	4.63	4.59	5.35
	Night visit	0.02	0.15	0.00	0.10	-0.00	0.09	-0.00	0.16
	Phone call	0.67	1.74	0.28	1.63	0.29	1.58	0.60	1.34
Acupuncturist	Day visit	0.00	0.00	0.00	0.05	0.01	0.09	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	-0.00	0.02	0.01	0.45	-0.01	0.14	0.03	0.39
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.15	1.24	0.07	0.67	0.12	1.11	1.06	6.43
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.03
	Phone call	0.01	0.09	-0.00	0.07	0.01	0.25	0.16	1.08
Consultant	Day visit	0.00	0.00	0.00	0.13	-0.01	0.09	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.06	-0.00	0.03	-0.01	0.19	-0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
Health visitor	Day visit	0.01	0.32	0.01	0.22	0.01	0.21	0.06	0.46
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.04	0.00	0.00	-0.00	0.01	-0.00	0.04
Hospital nurse	Day visit	0.00	0.00	-0.00	0.07	0.00	0.02	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.22	1.18	0.08	1.21	0.24	1.79	0.04	1.14
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.03
	Phone call	-0.00	0.06	-0.00	0.07	0.00	0.08	0.01	0.08
Physiotherapist	Day visit	-0.01	0.05	0.02	0.33	0.01	0.25	0.02	0.25
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Practice nurse	Day visit	1.61	7.18	0.88	4.11	0.98	6.39	0.73	4.46
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.02	0.20	0.06	0.44	0.02	0.45	0.01	0.22

Note: Resource use data obtained from CPRD pooling together records from years two through 10. Classification as *likely Poor* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 39 – Number of consultation events attributable to hip pain for years two through 10 after THR for *likely Good* outcomes

		FEMALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	0.49	4.91	0.15	5.22	0.01	5.27	-0.18	5.38
	Night visit	-0.00	0.09	-0.00	0.10	-0.00	0.10	-0.00	0.19
	Phone call	0.06	1.25	0.06	1.40	0.09	1.69	0.11	1.88
Acupuncturist	Day visit	0.00	0.08	0.00	0.08	0.00	0.04	-0.00	0.05
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.00	0.19	0.00	0.23	0.00	0.29	-0.01	0.26
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	0.01
Community nurse	Day visit	0.04	1.05	0.06	1.95	0.01	1.56	-0.01	1.99
	Night visit	0.00	0.00	0.00	0.00	0.00	0.01	-0.00	0.02
	Phone call	-0.00	0.13	-0.00	0.10	0.00	0.20	0.00	0.14
Consultant	Day visit	0.00	0.05	0.00	0.07	-0.00	0.04	0.00	0.05
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.01	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.00	0.00	0.01	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.00	0.23	-0.01	0.23	-0.00	0.23	-0.01	0.16
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.02	-0.00	0.03	-0.00	0.10	0.00	0.03
Health visitor	Day visit	0.01	0.35	-0.01	0.35	-0.01	0.41	-0.01	0.34
	Night visit	0.00	0.00	0.00	0.01	-0.00	0.01	0.00	0.00
	Phone call	-0.00	0.03	-0.00	0.07	-0.00	0.09	-0.00	0.06
Hospital nurse	Day visit	-0.00	0.04	0.00	0.09	0.00	0.05	0.00	0.05
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.01	-0.00	0.03	-0.00	0.01	0.00	0.00
Other HCP	Day visit	0.05	0.93	-0.02	0.98	-0.02	1.47	0.02	1.38
	Night visit	-0.00	0.01	0.00	0.03	-0.00	0.03	0.00	0.06
	Phone call	0.00	0.07	0.00	0.09	-0.00	0.11	-0.01	0.19
Physiotherapist	Day visit	0.01	0.41	0.01	0.48	0.01	0.57	0.01	0.43
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	0.02
Practice nurse	Day visit	0.22	2.81	0.02	3.24	0.04	4.09	0.11	3.57
	Night visit	-0.00	0.01	0.00	0.01	0.00	0.03	-0.00	0.03
	Phone call	-0.00	0.36	-0.00	0.48	0.00	0.36	-0.00	0.35

		MALES							
		45-60 years		60-70 years old		70-80 years old		80+ years old	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	0.20	4.50	0.32	4.56	0.05	5.22	0.61	6.03
	Night visit	0.00	0.11	-0.00	0.09	-0.00	0.12	0.00	0.18
	Phone call	0.01	0.93	0.04	1.13	0.00	1.38	-0.02	1.40
Acupuncturist	Day visit	-0.00	0.11	0.00	0.07	-0.00	0.04	0.00	0.03
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.00	-0.00	0.01	-0.00	0.01
Chiropractor	Day visit	-0.00	0.08	0.00	0.30	-0.00	0.22	0.00	0.31
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.02	1.08	0.00	1.08	0.01	1.10	0.07	1.68
	Night visit	0.00	0.04	0.00	0.00	0.00	0.00	-0.00	0.01
	Phone call	-0.00	0.03	0.00	0.06	-0.00	0.10	0.00	0.14
Consultant	Day visit	-0.00	0.03	0.00	0.07	0.00	0.04	0.00	0.15
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.01	0.10	0.01	0.30	0.01	0.33	-0.00	0.40
	Night visit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Phone call	-0.00	0.01	-0.00	0.03	0.00	0.07	0.00	0.06
Health visitor	Day visit	0.01	0.42	0.01	0.45	0.01	0.46	0.01	0.49
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00
	Phone call	0.00	0.11	-0.00	0.05	-0.00	0.02	-0.00	0.05
Hospital nurse	Day visit	0.00	0.08	0.01	0.15	0.00	0.08	0.00	0.04
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00
	Phone call	-0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.06	1.18	0.04	1.13	0.04	1.39	0.03	1.37
	Night visit	-0.00	0.01	0.00	0.05	0.00	0.03	0.00	0.08
	Phone call	0.00	0.14	-0.00	0.08	0.00	0.10	0.00	0.11
Physiotherapist	Day visit	0.02	0.55	0.01	0.39	-0.00	0.34	0.00	0.21
	Night visit	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.03
Practice nurse	Day visit	0.22	3.56	0.09	3.59	-0.13	4.56	0.32	4.86
	Night visit	0.00	0.00	0.00	0.02	0.00	0.03	-0.00	0.02
	Phone call	0.00	0.32	-0.00	0.28	-0.01	0.26	-0.01	0.28

Note: Resource use data obtained from CPRD pooling together records from years two through 10. Classification as *likely Poor* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 40 – Consultations costs attributable to hip pain for years two through 10 after THR for *likely Poor* outcome patients

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	140.23	153.51	141.25	150.08	145.52	186.90	153.75	219.90
Acupuncturist	0.14	2.15	0.05	1.01	-0.09	2.53	-0.04	0.79
Chiropractor	0.46	6.04	1.46	14.53	0.67	11.94	0.12	5.85
Community nurse	12.58	61.55	2.87	34.17	11.32	101.75	13.96	80.07
Consultant	0.29	4.81	0.07	3.41	-0.01	0.49	0.26	7.55
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.08	2.66	-0.39	8.49	-0.18	5.19	-0.35	3.54
Health visitor	-0.02	10.01	-0.60	8.16	0.03	15.52	2.13	67.32
Hospital nurse	0.00	0.00	-0.03	0.77	0.03	0.90	0.00	0.00
Other HCP	2.59	22.43	3.61	29.44	1.86	17.69	2.45	29.44
Physiotherapist	2.43	36.74	2.11	25.60	2.43	32.03	0.78	21.20
Practice nurse	14.86	71.84	10.55	56.93	5.70	60.20	6.68	61.26

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	132.07	146.22	125.94	142.78	140.09	149.41	153.13	172.00
Acupuncturist	0.00	0.00	0.07	1.28	0.19	2.17	0.00	0.00
Chiropractor	-0.04	0.52	0.41	14.05	-0.18	4.33	1.04	12.04
Community nurse	5.49	45.16	2.69	24.62	4.50	40.63	38.44	234.72
Consultant	0.00	0.00	0.11	10.39	-0.48	7.03	0.00	0.00
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.16	1.96	-0.16	1.21	-0.38	6.78	-0.11	0.92
Health visitor	0.46	11.32	0.34	7.90	0.19	7.75	2.02	16.72
Hospital nurse	0.00	0.00	-0.01	1.42	0.02	0.33	0.00	0.00
Other HCP	2.78	15.06	1.03	15.36	3.05	22.90	0.47	14.62
Physiotherapist	-0.21	1.90	0.87	11.54	0.30	8.75	0.71	8.63
Practice nurse	20.32	91.56	11.81	53.65	12.74	82.07	9.35	57.39

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case and pooling together records from years two through 10
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 41 – Consultations costs attributable to hip pain for years two through 10 after THR for *likely Good* outcome patients

FEMALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	16.25	162.10	5.69	172.87	1.52	176.15	-3.81	182.98
Acupuncturist	0.02	1.88	0.01	1.92	0.00	1.11	-0.03	1.18
Chiropractor	0.12	5.90	0.00	7.05	0.10	9.00	-0.28	8.15
Community nurse	1.29	38.17	2.36	71.13	0.33	56.81	-0.50	72.65
Consultant	0.10	3.77	0.01	5.44	-0.02	3.00	0.09	3.72
Contact tracing nurse	0.00	0.00	-0.00	0.11	0.00	0.00	0.00	0.00
Dietician	-0.01	8.16	-0.34	8.13	-0.07	8.12	-0.24	5.52
Health visitor	0.45	12.90	-0.39	13.14	-0.22	15.35	-0.28	12.52
Hospital nurse	-0.00	0.86	0.01	1.89	0.01	1.04	0.03	1.03
Other HCP	0.60	11.92	-0.20	12.44	-0.21	18.85	0.15	17.87
Physiotherapist	0.18	14.36	0.29	16.67	0.40	19.88	0.51	15.02
Practice nurse	2.78	36.33	0.25	41.85	0.49	52.63	1.41	46.02

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
GPs	6.60	145.74	10.69	149.33	1.29	171.72	18.64	197.91
Acupuncturist	-0.11	2.79	0.02	1.84	-0.02	1.12	0.02	0.66
Chiropractor	-0.06	2.52	0.12	9.30	-0.14	6.70	0.15	9.66
Community nurse	0.64	39.46	0.15	39.32	0.34	40.02	2.65	61.49
Consultant	-0.08	2.06	0.08	5.33	0.02	3.61	0.19	12.17
Contact tracing nurse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dietician	-0.19	3.49	0.38	10.59	0.40	11.85	-0.12	13.98
Health visitor	0.29	15.39	0.20	16.45	0.31	16.55	0.37	17.92
Hospital nurse	0.03	1.53	0.13	2.90	0.01	1.55	0.02	0.77
Other HCP	0.75	15.17	0.46	14.50	0.49	17.86	0.41	17.78
Physiotherapist	0.82	19.36	0.23	13.77	-0.10	11.97	0.07	7.46
Practice nurse	2.83	45.76	1.11	46.24	-1.71	58.49	3.90	62.26

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case case and pooling together records from years two through 10
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Good* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 42 – Prescription costs attributable to hip pain for years two through 10 after THR for *likely Poor* outcome patients

	FEMALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	-0.10	1.79
Aspirin	0.00	0.00	-0.19	3.23	-0.09	1.96	-0.18	1.20
Celecoxib	5.45	45.19	2.48	42.59	4.98	38.57	5.41	34.23
Cocodamol	4.41	26.42	6.39	29.35	4.25	27.18	1.14	25.10
Codeine	7.43	22.95	3.65	19.83	4.60	17.94	4.95	20.33
Codydramol	0.67	5.59	1.12	8.06	0.58	23.99	1.16	7.93
Diclofenac	12.79	41.37	3.49	22.81	1.91	20.44	1.49	14.28
Dihydrocodeine	4.80	22.46	2.88	14.72	3.76	24.25	2.72	14.78
Ibuprofen	0.03	4.84	-0.09	4.53	-0.27	8.85	0.11	2.83
Lactulose	0.47	4.11	0.20	5.52	0.16	7.32	0.40	10.23
Lansoprazole	0.74	8.04	0.57	7.44	0.45	7.68	0.83	8.53
Macrogol	0.00	0.00	0.12	2.26	0.22	4.43	0.04	2.08
Meloxicam	0.39	2.55	0.26	2.46	0.21	2.33	0.05	1.17
Morphine	3.53	60.57	3.91	48.38	0.54	12.74	1.49	24.32
Movicol	1.61	14.13	0.25	7.92	0.93	10.46	0.86	11.47
Nabumetone	1.09	8.17	0.09	1.83	0.17	3.69	-0.13	2.54
Naproxen	0.78	6.84	0.91	12.92	0.65	6.92	0.07	3.47
Omeprazole	-1.65	20.47	3.48	31.13	0.20	20.53	1.50	19.08
Oxycodone	0.00	0.00	3.84	89.53	1.07	22.78	0.13	3.32
Paracetamol	83.93	140.85	80.83	132.66	57.13	108.65	45.21	105.52
Piroxicam	0.10	1.71	0.82	12.88	0.76	13.52	0.71	13.91
Rabeprazole	2.62	25.47	1.66	27.85	0.79	23.65	2.85	28.03
Ranitidine	0.67	3.93	0.50	4.01	0.60	6.51	-0.39	9.99
Senna	0.13	1.20	0.15	2.18	0.19	3.87	0.03	4.84
Tramadol	10.66	44.46	6.19	37.44	4.02	34.67	3.19	30.79

	MALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aspirin	-0.27	2.51	-0.36	3.77	-0.35	2.16	-0.34	1.55
Celecoxib	7.40	32.90	3.34	33.23	2.01	36.84	-2.13	52.21
Cocodamol	2.51	40.29	7.69	42.39	5.89	26.78	2.86	13.56
Codeine	5.07	14.74	3.27	14.47	2.61	14.17	4.33	12.72
Codydramol	1.57	8.21	2.96	14.54	0.11	6.74	0.18	3.98
Diclofenac	11.77	50.83	4.18	22.48	1.35	21.61	5.69	31.06
Dihydrocodeine	11.98	37.00	5.29	21.02	2.58	12.86	1.58	10.37
Ibuprofen	0.99	10.73	-0.13	5.77	0.03	3.60	1.64	10.53
Lactulose	0.71	6.04	0.95	5.52	0.80	6.52	0.55	8.44
Lansoprazole	1.77	10.22	0.29	7.74	0.42	10.68	0.47	8.89
Macrogol	0.00	0.00	-0.04	0.68	-0.01	1.01	-0.16	1.92
Meloxicam	0.31	2.14	0.20	2.00	0.19	2.03	0.32	1.83
Morphine	-7.55	57.30	0.15	7.54	0.97	16.71	5.41	46.36
Movicol	0.09	1.12	-0.80	9.76	-0.47	6.08	0.78	6.33
Nabumetone	0.37	4.38	-0.01	0.24	0.11	4.95	0.01	0.44
Naproxen	-1.38	7.95	0.61	5.41	0.01	2.49	-1.31	5.64
Omeprazole	5.40	29.49	1.85	23.56	-0.37	23.16	2.06	15.26
Oxycodone	0.43	5.11	5.82	112.33	-0.15	2.94	0.00	0.00
Paracetamol	88.11	154.46	49.78	105.90	60.73	110.97	72.93	113.91
Piroxicam	0.38	8.39	3.13	25.06	0.45	12.58	0.00	0.00
Rabeprazole	1.98	17.12	0.01	34.46	0.26	32.67	-5.03	39.90
Ranitidine	0.94	4.42	0.37	3.75	0.08	4.37	-0.38	2.74
Senna	0.59	3.21	0.24	2.69	0.11	6.12	0.55	3.36
Tramadol	22.49	68.81	5.25	23.73	3.47	21.30	3.40	18.91

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case and pooling together records from years two through 10
3. Unit costs obtained from BNF [117]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 43 – Prescription costs attributable to hip pain for years two through 10 after THR for *likely Good* outcome patients

	FEMALES							
	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.29	10.65	0.00	0.00	0.00	0.00	-0.00	0.19
Aspirin	0.66	11.63	-0.05	1.60	-0.04	2.14	-0.05	2.26
Celecoxib	0.92	31.47	1.77	28.04	1.25	24.10	1.80	26.63
Cocodamol	0.82	11.96	1.00	16.07	0.58	16.36	0.63	15.01
Codeine	0.29	11.99	0.78	10.33	0.82	10.54	0.74	10.22
Codydramol	-0.05	4.89	0.03	3.87	0.29	16.81	0.05	4.01
Diclofenac	2.54	20.90	1.84	19.54	0.65	14.93	0.75	13.48
Dihydrocodeine	0.51	11.52	0.29	7.88	0.66	10.33	0.30	6.91
Ibuprofen	0.27	4.70	0.03	4.37	-0.02	3.84	-0.01	3.08
Lactulose	0.07	2.56	-0.09	3.48	-0.13	5.60	-0.53	7.26
Lansoprazole	-0.07	4.56	-0.22	5.12	0.09	6.31	-0.12	5.75
Macrogol	-0.01	0.24	-0.01	0.45	0.05	2.13	0.02	2.05
Meloxicam	0.10	1.72	0.09	1.65	0.06	1.35	0.02	1.05
Morphine	0.21	20.03	-0.12	7.90	-0.22	6.66	-0.27	7.99
Movicol	-0.05	2.88	-0.02	4.20	-0.04	6.26	-0.49	10.24
Nabumetone	0.19	3.80	-0.04	1.37	0.06	2.54	0.03	1.42
Naproxen	-0.17	3.13	0.31	6.97	0.28	5.62	0.07	3.04
Omeprazole	0.40	14.31	-0.23	18.30	-0.42	16.87	-0.08	13.86
Oxycodone	-0.06	3.16	0.47	28.28	0.15	16.18	0.81	54.98
Paracetamol	2.96	47.82	1.98	50.25	1.38	59.29	-2.32	60.54
Piroxicam	0.18	7.19	0.35	13.15	0.16	8.31	-0.06	3.77
Rabeprazole	0.25	12.42	0.17	18.58	0.70	19.83	0.92	23.36
Ranitidine	0.29	6.46	0.06	2.56	-0.14	4.34	-0.30	8.22
Senna	-0.02	1.83	-0.02	1.61	-0.06	2.62	-0.13	3.65
Tramadol	0.64	15.78	0.55	15.79	0.40	15.27	0.24	10.25

MALES

	45-60 years		60-70 years old		70-80 years old		80+ years old	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aspirin	-0.02	1.94	-0.08	2.33	-0.12	2.56	0.00	2.55
Celecoxib	0.42	13.84	1.11	21.71	0.76	18.09	0.13	11.53
Cocodamol	-0.03	12.05	0.42	12.72	0.50	11.27	0.64	12.42
Codeine	0.33	6.01	0.38	6.84	0.39	7.84	0.22	5.34
Codydramol	0.09	3.04	0.03	7.61	-0.05	3.16	-0.06	4.03
Diclofenac	2.07	18.78	1.96	18.07	1.74	18.74	2.93	22.87
Dihydrocodeine	0.33	6.68	0.16	6.19	-0.00	5.60	0.03	1.88
Ibuprofen	-0.06	3.29	0.01	3.82	0.16	4.42	0.28	4.82
Lactulose	0.03	1.60	0.00	3.40	-0.10	4.43	0.07	7.64
Lansoprazole	0.26	5.37	-0.04	4.76	-0.11	5.93	0.09	6.00
Macrogol	0.03	1.67	-0.01	0.30	-0.00	2.51	-0.10	2.03
Meloxicam	0.16	1.87	0.05	1.47	0.06	1.31	0.03	0.95
Morphine	0.04	9.65	0.26	22.33	-0.29	7.47	-0.16	6.90
Movicol	-0.02	1.72	0.00	5.51	-0.02	6.37	-0.19	7.06
Nabumetone	0.06	1.87	0.05	2.07	0.00	1.98	-0.01	0.34
Naproxen	0.05	4.70	0.20	4.77	0.14	6.14	-0.06	3.66
Omeprazole	0.05	10.33	-0.51	11.83	-0.04	11.74	0.38	12.16
Oxycodone	0.30	14.19	0.05	15.76	0.08	14.35	-0.13	4.31
Paracetamol	3.01	40.15	1.38	35.19	-0.08	39.42	1.44	43.87
Piroxicam	0.21	7.64	0.24	6.15	0.67	12.44	0.03	2.91
Rabeprazole	1.25	18.53	-0.78	18.85	-0.15	18.77	2.16	25.18
Ranitidine	-0.01	2.16	-0.05	3.10	0.09	7.15	-0.33	4.07
Senna	-0.01	0.79	-0.08	1.33	-0.14	2.10	-0.16	2.39
Tramadol	0.77	14.24	0.28	10.08	0.36	9.77	0.13	6.98

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case and pooling together records from years two through 10
3. Unit costs obtained from BNF [117]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 44 – Relative frequencies of HRGs by patient subgroup for HES revision THR records, 2011-2012

Females, 45-60 years

HRG	Relative frequency
HB11A	1.89%
HB11C	3.93%
HB12A	1.42%
HB12B	0.31%
HB12C	8.33%
HB13Z	0.47%
HB14C	1.10%
HB99Z	0.47%
HR01B	0.31%
HR01C	0.31%
HR03Z	0.16%
HR04B	3.77%
HR04C	29.40%
HR05Z	44.65%
HR06A	2.04%
Invalid	1.42%
TOTAL	**100%**

Females, 60-70 years

HRG	Relative frequency
HA14C	0.09%
HA99Z	0.19%
HB11A	2.50%
HB11B	0.19%
HB11C	3.89%
HB12A	1.48%
HB12B	0.28%
HB12C	6.67%
HB13Z	0.37%
HB14B	0.09%
HB14C	0.65%
HB15E	0.37%
HB99Z	0.19%
HR01B	0.09%
HR01C	0.09%
HR03Z	0.74%
HR04B	5.93%
HR04C	33.43%
HR05Z	40.28%
HR06A	2.04%
Invalid	0.46%
TOTAL	**100%**

Females, 70-80 years

HRG	Relative frequency
HA13B	0.07%
HA99Z	0.13%
HB11A	1.89%
HB11B	0.26%
HB11C	3.06%
HB12A	0.98%
HB12B	0.20%
HB12C	4.62%
HB13Z	0.52%
HB14B	0.20%
HB14C	0.85%
HB15D	0.26%
HB15E	0.07%
HB99Z	0.13%
HR01C	0.07%
HR03Z	0.72%
HR04B	7.81%
HR04C	33.96%
HR05Z	41.44%
HR06A	2.08%
Invalid	0.72%
TOTAL	**100%**

Females, 80+ years

HRG	Relative frequency
HA11A	0.11%
HA12C	0.11%
HA13A	0.11%
HA13C	0.11%
HA99Z	0.34%
HB11A	1.81%
HB11C	1.92%
HB12A	1.24%
HB12B	0.23%
HB12C	3.05%
HB13Z	0.90%
HB14B	0.11%
HB14C	0.90%
HB15D	0.34%
HB15E	0.11%
HB99Z	0.34%
HR03Z	0.45%
HR04B	9.27%
HR04C	28.47%
HR05Z	47.68%
HR06A	1.58%
Invalid	0.79%
TOTAL	**100%**

Males, 45-60 years

HRG	Relative frequency
HB11A	4.21%
HB11B	0.40%
HB11C	4.41%
HB12A	3.01%
HB12B	0.80%
HB12C	7.62%
HB13Z	0.40%
HB14B	0.60%
HB14C	0.40%
HB99Z	0.40%
HR01B	0.20%
HR01C	0.40%
HR03Z	0.40%
HR04B	4.41%
HR04C	30.66%
HR05Z	38.48%
HR06A	3.01%
Invalid	0.20%
TOTAL	**100%**

Males, 60-70 years

HRG	Relative frequency
HA99Z	0.12%
HB11A	4.31%
HB11B	0.36%
HB11C	3.35%
HB12A	2.40%
HB12B	0.24%
HB12C	7.07%
HB13Z	0.96%
HB14B	0.24%
HB14C	1.08%
HB15E	0.48%
HB99Z	0.12%
HR01B	0.12%
HR03Z	0.60%
HR04B	9.10%
HR04C	31.38%
HR05Z	34.49%
HR06A	1.92%
Invalid	1.68%
TOTAL	**100%**

Males, 70-80 years

HRG	Relative frequency
HA12C	0.08%
HA99Z	0.16%
HB11A	3.53%
HB11B	0.40%
HB11C	3.69%
HB12A	1.68%
HB12B	0.64%
HB12C	4.41%
HB13Z	0.24%
HB14B	0.16%
HB14C	0.56%
HB15D	0.24%
HB99Z	0.16%
HR01C	0.08%
HR03Z	0.32%
HR04B	9.06%
HR04C	34.08%
HR05Z	38.17%
HR06A	1.92%
Invalid	0.40%
TOTAL	**100%**

Males, 80+ years

HRG	Relative frequency
HA99Z	0.41%
HB11A	2.87%
HB11C	2.25%
HB12A	1.23%
HB12B	0.82%
HB12C	3.69%
HB13Z	0.82%
HB14B	0.41%
HB14C	1.23%
HB15D	0.20%
HB99Z	1.02%
HR01B	0.20%
HR01C	0.20%
HR03Z	0.41%
HR04B	12.70%
HR04C	32.38%
HR05Z	36.27%
HR06A	2.46%
Invalid	0.41%
TOTAL	**100%**

Appendix 45 – Number of consultation events attributable to hip pain on the year immediately after revision THR, by outcome category

		Likely Poor outcomes		*Likely Good* outcomes	
		Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	6.03	5.84	0.71	5.38
	Night visit	0.01	0.14	-0.00	0.09
	Phone call	1.14	2.04	0.59	1.80
Acupuncturist	Day visit	0.00	0.00	0.01	0.14
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Chiropractor	Day visit	0.00	0.00	-0.00	0.14
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.15	0.72	0.01	1.30
	Night visit	0.00	0.00	0.00	0.04
	Phone call	0.00	0.00	0.01	0.16
Consultant	Day visit	0.00	0.00	0.00	0.05
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Dietician	Day visit	0.00	0.00	-0.01	0.13
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.02
Health visitor	Day visit	-0.01	0.07	-0.00	0.35
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.05	0.89
Hospital nurse	Day visit	0.00	0.00	0.00	0.05
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.15	0.70	-0.05	1.00
	Night visit	0.00	0.00	-0.00	0.01
	Phone call	0.00	0.00	0.00	0.10
Physiotherapist	Day visit	0.05	0.32	0.00	0.45
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Practice nurse	Day visit	0.79	4.71	-0.26	3.15
	Night visit	0.00	0.00	0.00	0.00
	Phone call	-0.07	0.25	0.01	0.24

Note: Resource use data obtained from CPRD. Classification as *likely Poor* or *Good* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 46 – Consultations costs attributable to hip pain on the year immediately after revision THR, by outcome category

	All patients subgroups			
	Likely Poor **outcomes**		***Likely Good*** **outcomes**	
	Mean	Std. Dev.	Mean	Std. Dev.
GPs	210.12	183.03	32.76	178.72
Acupuncturist	0.00	0.00	0.17	3.56
Chiropractor	0.00	0.00	-0.02	4.40
Community nurse	5.60	26.13	0.70	48.19
Consultant	0.00	0.00	0.14	3.96
Contact tracing nurse	0.00	0.00	0.00	0.00
Dietician	0.00	0.00	-0.48	4.44
Health visitor	-0.35	2.53	0.44	20.06
Hospital nurse	0.00	0.00	0.03	0.98
Other HCP	1.90	8.91	-0.66	12.81
Physiotherapist	1.62	11.23	0.06	15.91
Practice nurse	9.46	60.40	-3.28	40.33

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 47 – Prescription costs attributable to hip pain on the year immediately after revision THR, by outcome category

	All patients subgroups			
	Likely Poor outcomes		*Likely Good* outcomes	
	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00
Aspirin	0.18	1.60	-0.03	1.70
Celecoxib	10.20	74.27	0.07	26.22
Cocodamol	11.40	36.33	0.98	12.64
Codeine	1.92	7.88	0.07	3.18
Codydramol	0.66	6.34	0.61	9.59
Diclofenac	1.69	37.12	2.60	15.57
Dihydrocodeine	6.14	26.63	0.48	14.57
Ibuprofen	1.03	10.59	0.05	4.23
Lactulose	1.14	7.03	0.20	4.26
Lansoprazole	2.56	9.49	-0.00	4.66
Macrogol	-0.04	0.31	-0.04	0.65
Meloxicam	0.11	2.38	0.09	1.21
Morphine	18.24	141.02	0.04	4.45
Movicol	1.97	12.64	0.20	3.83
Nabumetone	0.05	0.33	-0.02	0.46
Naproxen	-2.04	24.33	-0.13	8.33
Omeprazole	-1.81	18.13	0.39	18.66
Oxycodone	0.00	0.00	-0.06	1.18
Paracetamol	67.16	94.17	-0.51	48.71
Piroxicam	5.85	32.83	0.70	8.86
Rabeprazole	0.74	31.62	-0.08	10.73
Ranitidine	0.05	2.85	-0.52	11.89
Senna	0.45	2.64	0.05	2.53
Tramadol	8.45	30.84	3.04	29.63

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from BNF [117]
4. Classification as *likely Good* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 48 – Number of consultation events attributable to hip pain for years two through eight after revision THR, by outcome category

		Likely Poor outcomes		*Likely Good* outcomes	
		Mean	Std. Dev.	Mean	Std. Dev.
GPs	Day visit	3.90	4.56	0.66	5.04
	Night visit	0.02	0.14	-0.00	0.09
	Phone call	0.30	1.42	0.34	1.85
Acupuncturist	Day visit	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01
Chiropractor	Day visit	0.03	0.20	-0.00	0.22
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Community nurse	Day visit	0.31	1.31	0.13	1.87
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.13
Consultant	Day visit	0.00	0.00	-0.01	0.11
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Contact tracing nurse	Day visit	0.00	0.00	0.00	0.00
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Dietician	Day visit	-0.01	0.04	-0.00	0.13
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	-0.00	0.01
Health visitor	Day visit	-0.03	0.31	0.00	0.36
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.10
Hospital nurse	Day visit	0.00	0.00	0.01	0.13
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Other HCP	Day visit	0.20	1.17	0.06	1.06
	Night visit	0.00	0.00	0.00	0.04
	Phone call	-0.01	0.08	0.00	0.10
Physiotherapist	Day visit	0.19	1.17	0.01	0.87
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.00	0.00	0.00	0.00
Practice nurse	Day visit	0.48	3.88	0.41	4.78
	Night visit	0.00	0.00	0.00	0.00
	Phone call	0.05	0.34	0.02	0.34

Note: Resource use data obtained from CPRD. Classification as *likely Poor* or *Good* outcomes derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 49 – Consultations costs attributable to hip pain for years two through eight after revision THR, by outcome category

	All patients subgroups			
	Likely Poor outcomes		*Likely Good* outcomes	
	Mean	Std. Dev.	Mean	Std. Dev.
GPs	128.37	142.12	26.90	170.53
Acupuncturist	0.00	0.00	0.00	0.00
Chiropractor	0.84	6.19	-0.06	6.71
Community nurse	11.38	47.84	4.81	68.29
Consultant	0.00	0.00	-0.52	8.68
Contact tracing nurse	0.00	0.00	0.00	0.00
Dietician	-0.26	1.49	-0.11	4.54
Health visitor	-1.02	11.34	0.14	13.61
Hospital nurse	0.00	0.00	0.14	2.53
Other HCP	2.43	14.90	0.75	13.64
Physiotherapist	6.68	41.11	0.37	30.37
Practice nurse	6.56	50.23	5.45	61.52

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from PPSRU [116]
4. Classification as *likely Good* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 50 – Prescription costs attributable to hip pain for years two through eight after revision THR, by outcome category

	All patients subgroups			
	Likely Poor outcomes		*Likely Good* outcomes	
	Mean	Std. Dev.	Mean	Std. Dev.
Amitriptyline	0.00	0.00	0.00	0.00
Aspirin	0.22	1.42	-0.06	1.49
Celecoxib	6.14	47.17	1.04	18.85
Cocodamol	13.21	35.04	1.15	10.51
Codeine	0.34	3.10	0.05	1.65
Codydramol	1.29	9.00	0.28	6.78
Diclofenac	-0.25	19.39	2.00	16.64
Dihydrocodeine	3.60	29.72	0.26	12.62
Ibuprofen	-0.55	5.45	-0.03	5.70
Lactulose	2.04	7.95	0.15	4.37
Lansoprazole	1.38	9.44	-0.21	4.61
Macrogol	0.00	0.00	-0.02	0.86
Meloxicam	0.26	1.78	0.07	1.26
Morphine	-1.74	18.41	-0.04	2.60
Movicol	2.40	13.79	0.29	5.81
Nabumetone	0.00	0.00	-0.00	0.02
Naproxen	-6.56	48.39	-0.25	7.49
Omeprazole	2.31	25.04	-0.02	10.43
Oxycodone	-0.68	5.16	-0.24	5.95
Paracetamol	49.88	79.01	1.54	39.84
Piroxicam	6.05	34.78	0.73	11.26
Rabeprazole	7.64	49.32	-2.38	21.28
Ranitidine	0.88	4.37	0.21	2.67
Senna	0.20	2.26	0.03	1.88
Tramadol	17.84	64.61	0.91	22.24

Note:

1. All figures in Pound sterling.
2. Number of events obtained from the CPRD by subtracting mean values of controls from value reported by case
3. Unit costs obtained from BNF [117]
4. Classification as *likely Poor* outcome patients derived from fitting a logistic model predicting surgery outcome, estimated on data from COASt, and using a probability cut-off point of 0.3.

Appendix 51 – Preoperative transition probabilities when outcome prediction tool is applied

Preoperative probabilities with the tool at threshold point = 32, deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.130	0.093	Beta	1.573	10.516
Surgical assessment to Long-term medical management	0.203	0.208	Empirical		

Preoperative probabilities with the tool at threshold point = 34, deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.123	0.093	Beta	1.401	10.031
Surgical assessment to Long-term medical management	0.249	0.208	Empirical		

Preoperative probabilities with the tool at threshold point = 36, deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.106	0.093	Beta	1.057	8.907
Surgical assessment to Long-term medical management	0.350	0.208	Beta	1.492	2.768

Preoperative probabilities with the tool at threshold point = 40, deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.053	0.093	Empirical		
Surgical assessment to Long-term medical management	0.677	0.208	Beta	2.745	1.311

Preoperative probabilities with the tool at threshold point = 42, deterministic and probabilistic parameters

Transition probability	Mean	SD	Distribution	α	β
Surgical assessment to Risk factor modification	0.028	0.093	Empirical		
Surgical assessment to Long-term medical management	0.827	0.208	Empirical		

Appendix 52 – Simplified outcome prediction tool model output

Linear regression for continuous OHS at one year after primary THR

Predictor	Coefficient	p-value	95% confidence interval	
Preoperative OHS	0.351	0.000	0.306	0.395
Constant	33.211	0.000	32.411	34.011
n = 2,092				
Pseudo R2 = 0.103				

Appendix 53 – Disutility associated to preoperative states with the tool: parameter values by prediction model threshold point

Preoperative disutilities with the tool at threshold point = 34, deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Risk-factor modification				
Reassessment after Risk-factor modification				
Males, 45-60 years of age	0.613	Gamma	3.82	0.160
Males, 60-70 years of age	0.590	Gamma	3.68	0.160
Males, 70-80 years of age	0.594	Gamma	3.75	0.159
Males, 80+ years of age	0.653	Gamma	4.33	0.151
Females, 45-60 years of age	0.692	Gamma	4.61	0.150
Females, 60-70 years of age	0.650	Gamma	4.16	0.156
Females, 70-80 years of age	0.662	Gamma	4.33	0.153
Females, 80+ years of age	0.715	Gamma	5.08	0.141
Long-term medical management				
Reassessment after Long-term medical management				
Males, 45-60 years of age	1.130	Gamma	15.92	0.071
Males, 60-70 years of age	1.145	Gamma	47.32	0.024
Males, 70-80 years of age	1.172	Gamma	60.78	0.019
Males, 80+ years of age	1.164	Gamma	149.54	0.008
Females, 45-60 years of age	1.155	Gamma	44.43	0.026
Females, 60-70 years of age	1.162	Gamma	46.78	0.025
Females, 70-80 years of age	1.182	Gamma	39.40	0.030
Females, 80+ years of age	1.211	Gamma	41.53	0.029

Note: When fitting the simplified model using 32 as the threshold point to direct patients to surgery (or Risk-factor modification) when predicted OHS is above this value, or to Long-term medical management when below, there were no predicted OHS scores under 32. We used therefore the above values instead when performing the sensitivity analysis at a threshold of 32 as well.

Preoperative disutilities with the tool at threshold point = 36, deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Risk-factor modification				
Reassessment after Risk-factor modification				
Males, 45-60 years of age	0.581	Gamma	3.82	0.152
Males, 60-70 years of age	0.564	Gamma	3.68	0.153
Males, 70-80 years of age	0.565	Gamma	3.75	0.151
Males, 80+ years of age	0.607	Gamma	4.19	0.145
Females, 45-60 years of age	0.648	Gamma	4.34	0.149
Females, 60-70 years of age	0.608	Gamma	4.01	0.152
Females, 70-80 years of age	0.619	Gamma	4.18	0.148
Females, 80+ years of age	0.656	Gamma	4.68	0.140
Long-term medical management				
Reassessment after Long-term medical management				
Males, 45-60 years of age	1.063	Gamma	46.98	0.023
Males, 60-70 years of age	1.048	Gamma	59.95	0.017
Males, 70-80 years of age	1.049	Gamma	52.51	0.020
Males, 80+ years of age	1.048	Gamma	63.51	0.017
Females, 45-60 years of age	1.053	Gamma	55.13	0.019
Females, 60-70 years of age	1.051	Gamma	58.63	0.018
Females, 70-80 years of age	1.059	Gamma	49.56	0.021
Females, 80+ years of age	1.060	Gamma	43.25	0.025

Preoperative disutilities with the tool at threshold point = 40, deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Risk-factor modification				
Reassessment after Risk-factor modification				
Males, 45-60 years of age	0.405	Gamma	4.64	0.087
Males, 60-70 years of age	0.398	Gamma	4.69	0.085
Males, 70-80 years of age	0.406	Gamma	4.59	0.088
Males, 80+ years of age	0.422	Gamma	4.62	0.091
Females, 45-60 years of age	0.422	Gamma	4.19	0.101
Females, 60-70 years of age	0.413	Gamma	4.19	0.098
Females, 70-80 years of age	0.426	Gamma	4.13	0.103
Females, 80+ years of age	0.435	Gamma	4.35	0.100
Long-term medical management				
Reassessment after Long-term medical management				
Males, 45-60 years of age	0.799	Gamma	7.74	0.103
Males, 60-70 years of age	0.792	Gamma	7.85	0.101
Males, 70-80 years of age	0.791	Gamma	7.72	0.102
Males, 80+ years of age	0.830	Gamma	9.36	0.089
Females, 45-60 years of age	0.831	Gamma	8.66	0.096
Females, 60-70 years of age	0.810	Gamma	8.13	0.100
Females, 70-80 years of age	0.821	Gamma	8.49	0.097
Females, 80+ years of age	0.854	Gamma	9.56	0.089

Preoperative disutilities with the tool at threshold point = 42, deterministic and probabilistic parameters

State / Patient subgroup	Mean	Distribution	α	β
Risk-factor modification				
Reassessment after Risk-factor modification				
Males, 45-60 years of age	0.344	Gamma	5.68	0.061
Males, 60-70 years of age	0.341	Gamma	5.46	0.062
Males, 70-80 years of age	0.349	Gamma	5.57	0.063
Males, 80+ years of age	0.355	Gamma	5.78	0.061
Females, 45-60 years of age	0.351	Gamma	4.77	0.074
Females, 60-70 years of age	0.354	Gamma	4.51	0.078
Females, 70-80 years of age	0.361	Gamma	4.56	0.079
Females, 80+ years of age	0.384	Gamma	4.31	0.089
Long-term medical management				
Reassessment after Long-term medical management				
Males, 45-60 years of age	0.692	Gamma	5.08	0.136
Males, 60-70 years of age	0.675	Gamma	4.95	0.136
Males, 70-80 years of age	0.682	Gamma	5.04	0.135
Males, 80+ years of age	0.734	Gamma	6.00	0.122
Females, 45-60 years of age	0.748	Gamma	5.86	0.128
Females, 60-70 years of age	0.717	Gamma	5.40	0.133
Females, 70-80 years of age	0.733	Gamma	5.71	0.128
Females, 80+ years of age	0.776	Gamma	6.42	0.121

Appendix 54 – Probability of *Poor* outcome after primary THR when outcome prediction tool is applied

Probability of *Poor* outcome after *Primary THR* with the tool at threshold point = 32, deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.296	Beta	55	131
Males, 60-70 years of age	0.240	Beta	74	234
Males, 70-80 years of age	0.199	Beta	47	189
Males, 80+ years of age	0.353	Beta	18	33
Females, 45-60 years of age	0.307	Beta	63	142
Females, 60-70 years of age	0.281	Beta	110	282
Females, 70-80 years of age	0.314	Beta	131	286
Females, 80+ years of age	0.388	Beta	52	82

Probability of *Poor* outcome after *Primary THR* with the tool at threshold point = 34, deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.286	Beta	52	130
Males, 60-70 years of age	0.217	Beta	63	288
Males, 70-80 years of age	0.182	Beta	42	189
Males, 80+ years of age	0.353	Beta	18	33
Females, 45-60 years of age	0.255	Beta	47	137
Females, 60-70 years of age	0.241	Beta	87	274
Females, 70-80 years of age	0.287	Beta	112	278
Females, 80+ years of age	0.365	Beta	46	80

Probability of *Poor* outcome after *Primary THR* with the tool at threshold point = 36, deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.248	Beta	40	121
Males, 60-70 years of age	0.190	Beta	50	213
Males, 70-80 years of age	0.164	Beta	35	179
Males, 80+ years of age	0.289	Beta	13	32
Females, 45-60 years of age	0.201	Beta	31	123
Females, 60-70 years of age	0.192	Beta	58	244
Females, 70-80 years of age	0.238	Beta	79	253
Females, 80+ years of age	0.298	Beta	31	73

Probability of *Poor* outcome after *Primary THR* with the tool at threshold point = 40, deterministic and probabilistic parameters

Transition probability / Patient subgroup	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-60 years of age	0.078	Beta	7	83
Males, 60-70 years of age	0.095	Beta	15	143
Males, 70-80 years of age	0.077	Beta	10	120
Males, 80+ years of age	0.167	Beta	5	25
Females, 45-60 years of age	0.109	Beta	7	57
Females, 60-70 years of age	0.085	Beta	12	130
Females, 70-80 years of age	0.096	Beta	12	113
Females, 80+ years of age	0.162	Beta	6	31

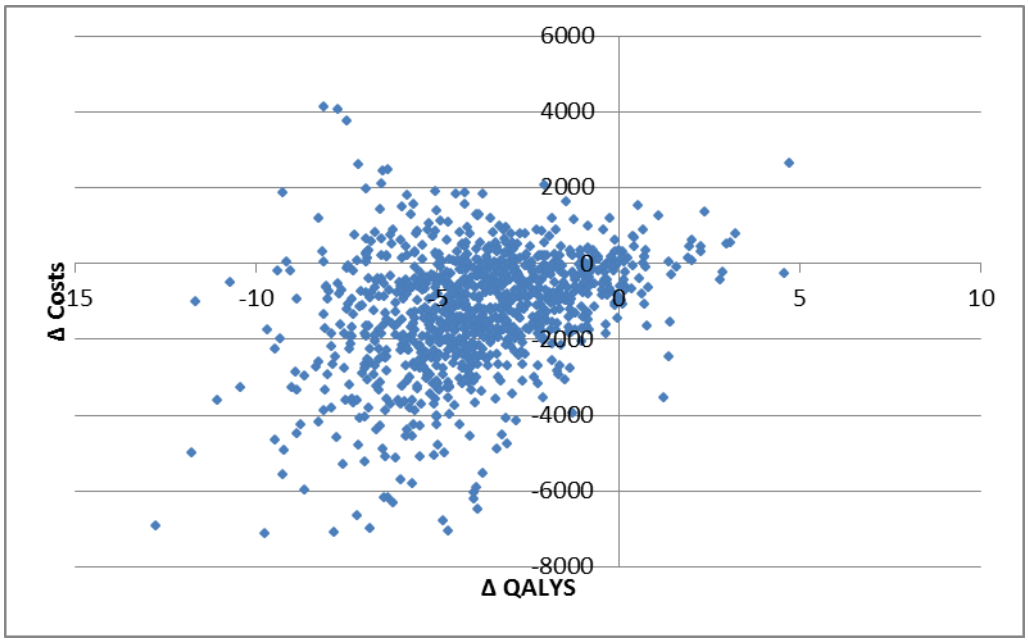
Probability of *Poor* outcome after *Primary THR* with the tool at threshold point = 42, deterministic and probabilistic parameters

Transition probability / Patient subgroup*	Mean	Distribution	α	β
Poor outcome first year after Primary THR				
Males, 45-70 years of age	0.074	Beta	11	138
Males,70+ years of age	0.047	Beta	4	81
Females, 45-70 years of age	0.039	Beta	4	98
Females, 70+ years of age	0.055	Beta	4	69

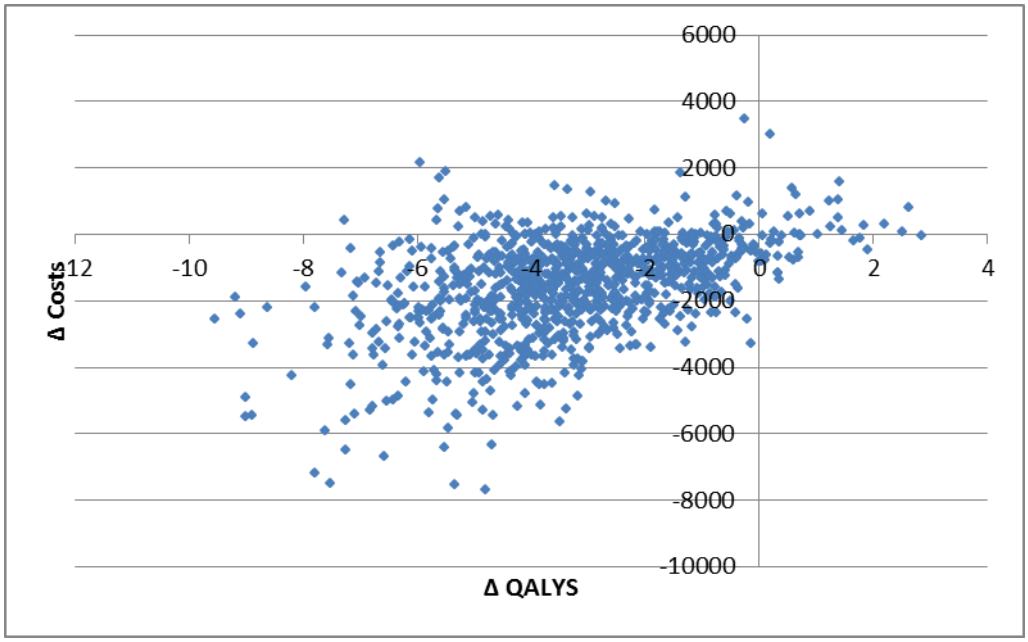
* Younger and older pairs of patient subgroups were merged given the low number of *Poor* outcomes reported

Appendix 55 – Results of Monte Carlo simulations on the cost-effectiveness plane

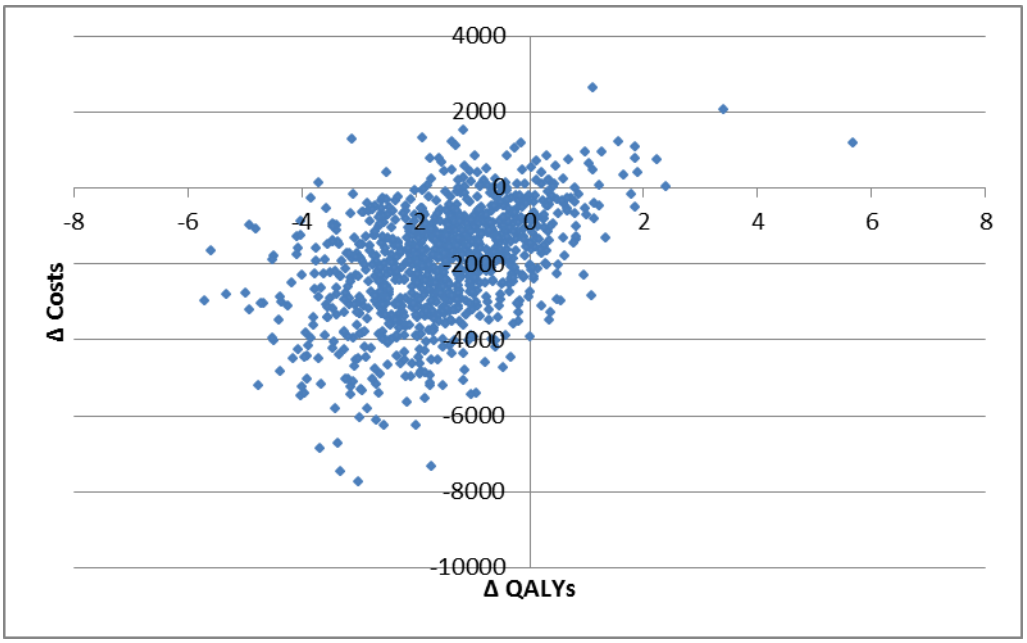
Females, 45 years of age



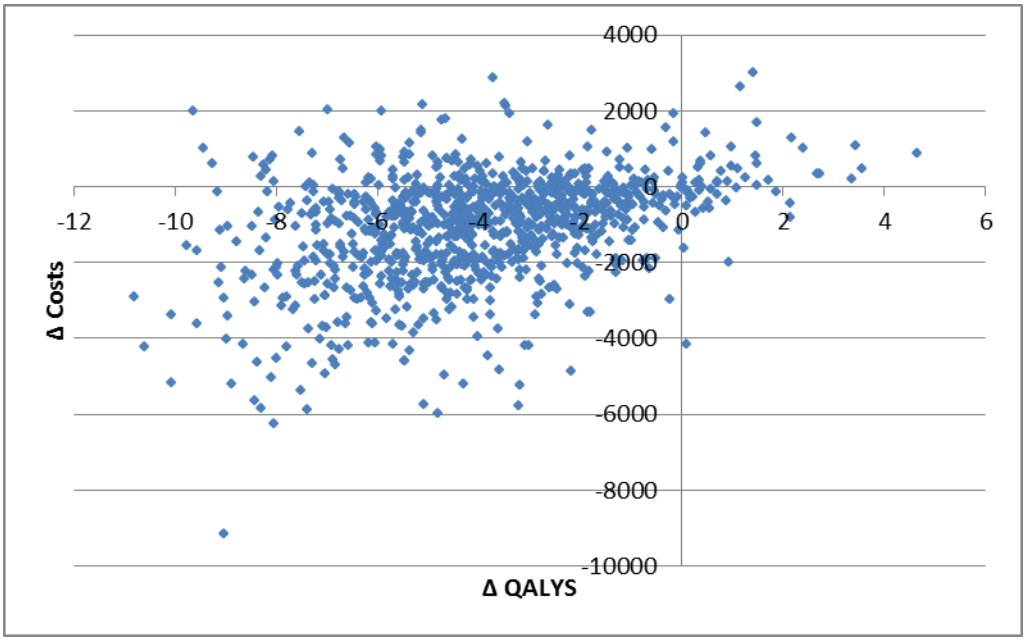
Females, 60 years of age



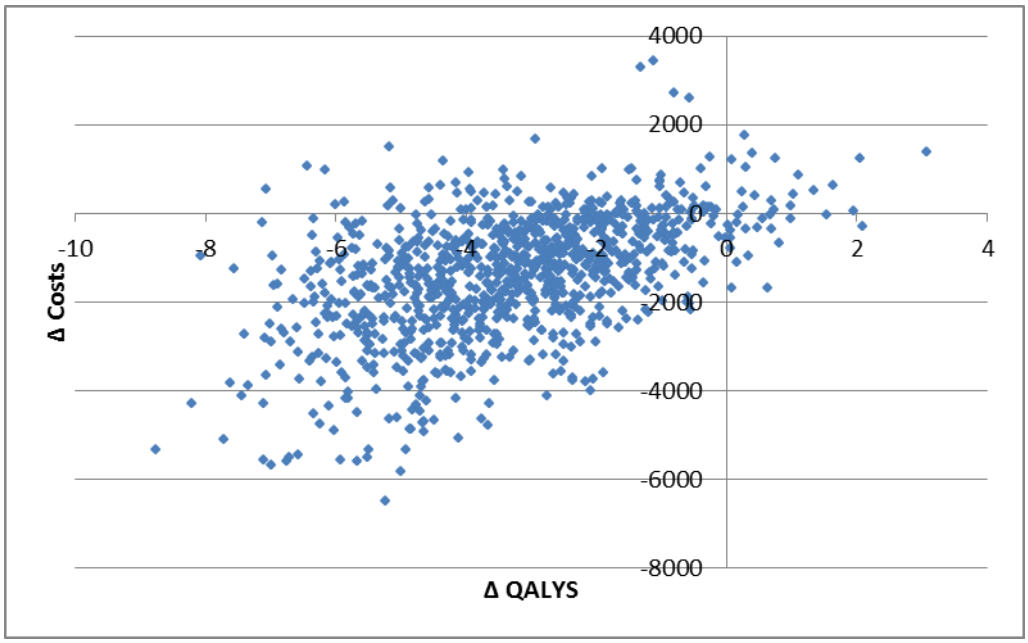
Females, 80 years of age



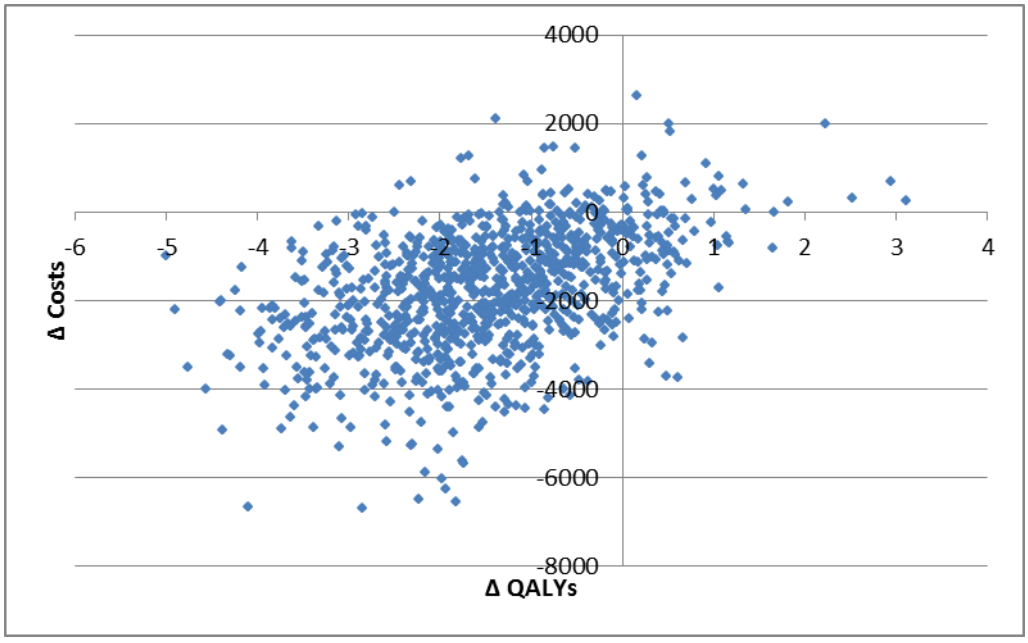
Males, 45 years of age



Males, 60 years of age

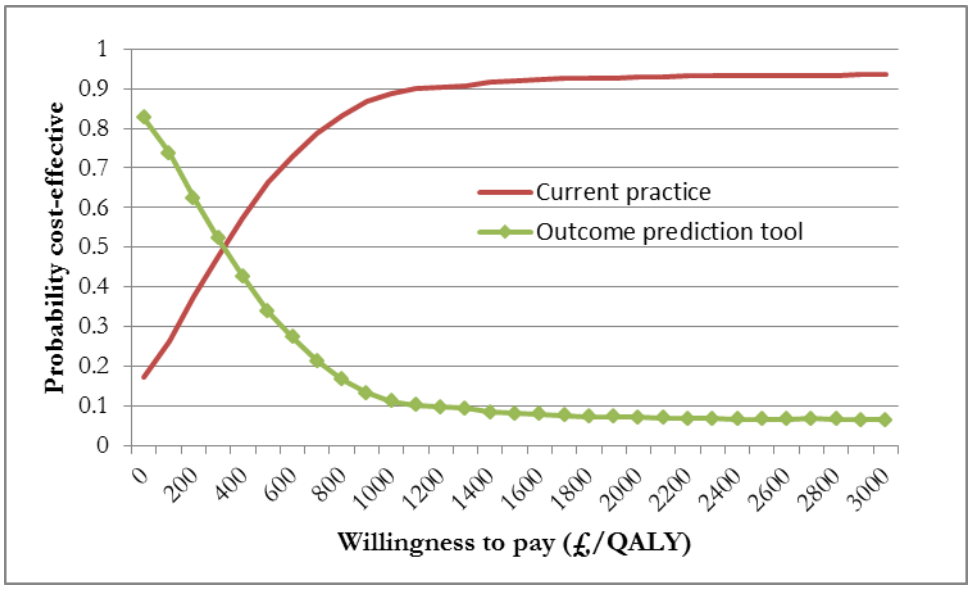


Males, 80 years of age

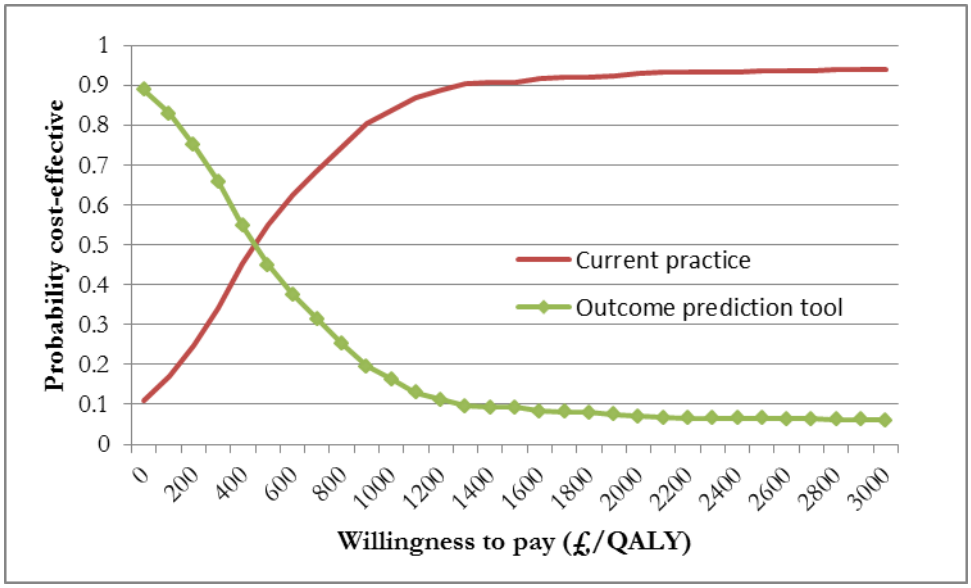


Appendix 56 – Cost-effectiveness acceptability curves

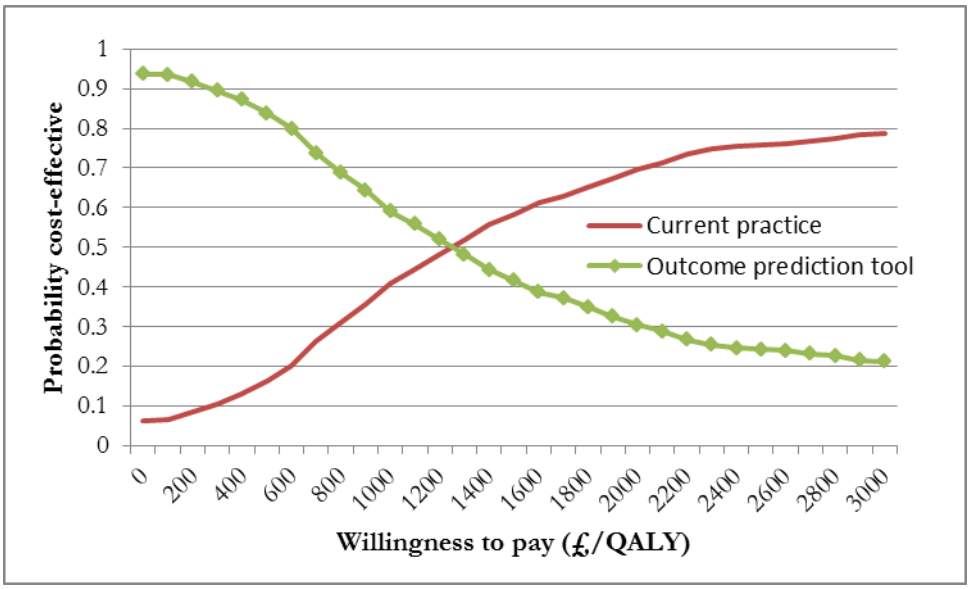
Females, 45 years of age



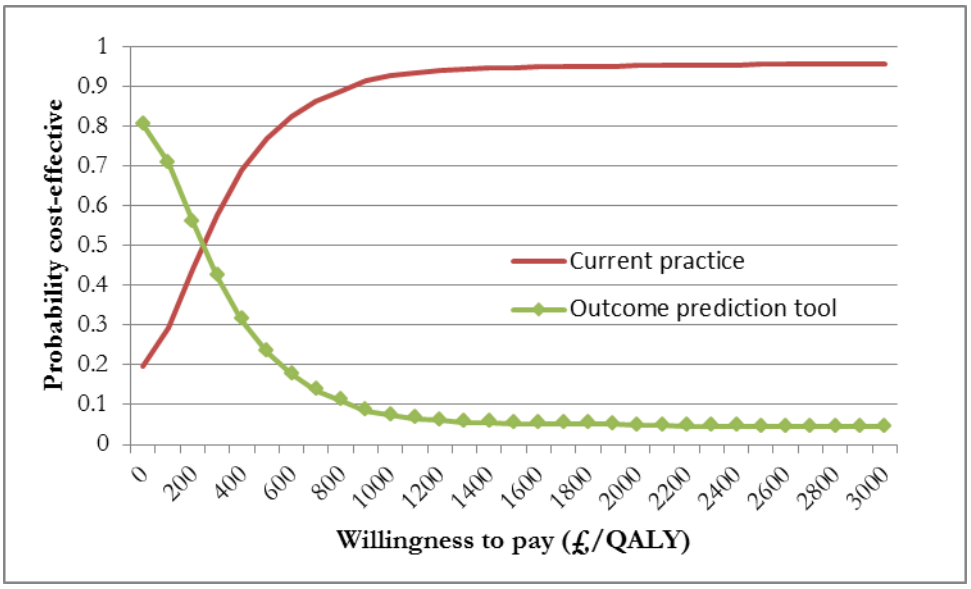
Females, 60 years of age



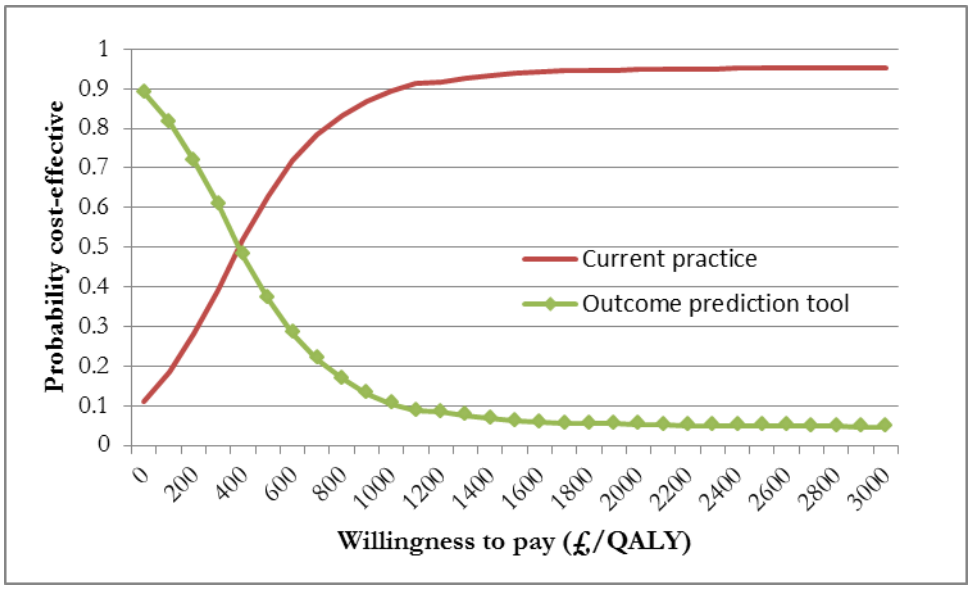
Females, 80 years of age



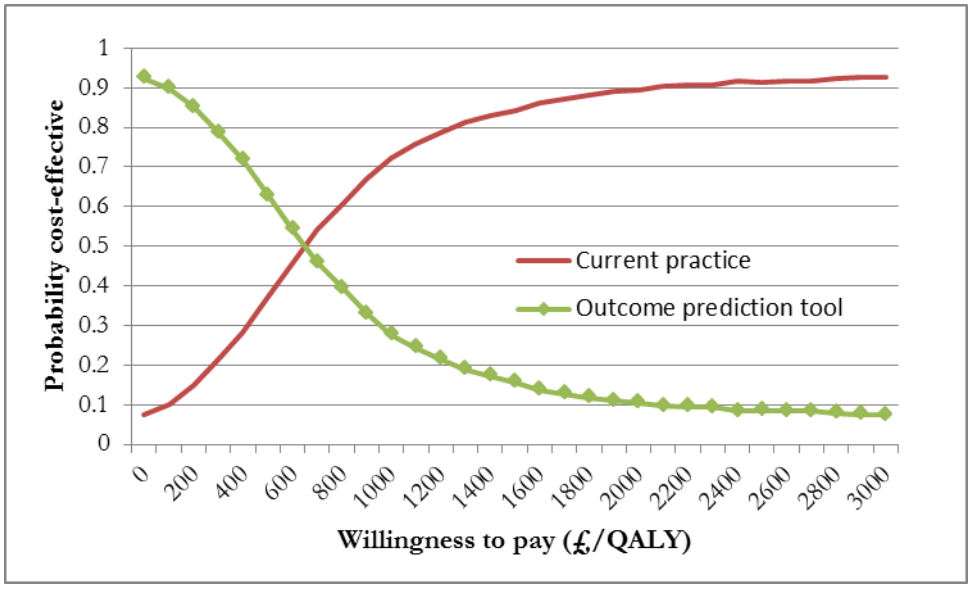
Males, 45 years of age



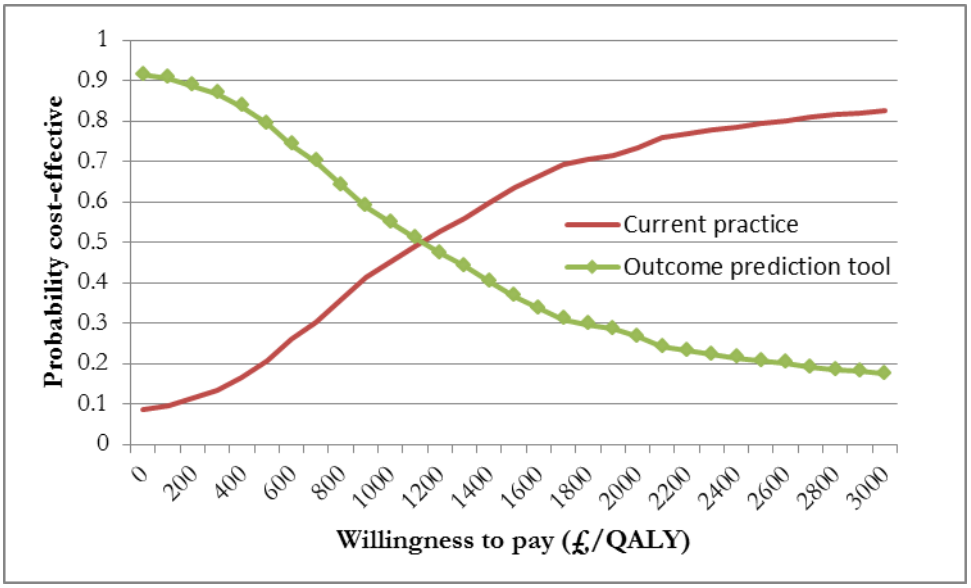
Males, 60 years of age



Males, 70 years of age



Males, 80 years of age



List of References

1. Park H, *An account of a new method of treating diseases of the joints of the knee and elbow: in a letter to Mr. Percival Pott*. 1782, London: Printed for J. Johnson.

2. Gomez PF and Morcuende JA, Early attempts at hip arthroplasty--1700s to 1950s. *Iowa Orthop J*. 2005; 25: p. 25-9.

3. Siopack JS and Jergesen HE, Total hip arthroplasty. *West J Med*. 1995; 162(3): p. 243-9.

4. Ethgen O and Reginster J-Y, Degenerative musculoskeletal disease. *Annals of the Rheumatic Diseases* 2004; 63(1): p. 1-3.

5. Bijlsma JWJ, Berenbaum F, and Lafeber FPJG, Osteoarthritis: an update with relevance for clinical practice. *The Lancet* 2011; 377(9783): p. 2115-26.

6. Dawson J, Fitzpatrick R, Fletcher J, and Wilson R, *Osteoarthritis Affecting the Hip and Knee*, in *Health Care Needs Assessment - The epidemiologically based needs assessment review*, Stevens A., Raftery J., Mant J., and Simpson S., Editors. 2004, Radcliffe Publishing Ltd. p. 549-634.

7. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, Birnbaum NS, Burmester GR, Bykerk VP, Cohen MD, Combe B, Costenbader KH, Dougados M, Emery P, Ferraccioli G, Hazes JM, Hobbs K, Huizinga TW, Kavanaugh A, Kay J, Kvien TK, Laing T, Mease P, Ménard HA, Moreland LW, Naden RL, Pincus T, Smolen JS, Stanislawska-Biernat E, Symmons D, Tak PP, Upchurch KS, Vencovský J, Wolfe F, and Hawker G, 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Annals of the Rheumatic Diseases* 2010; 69(9): p. 1580-8.

8. Frankel S, Eachus J, Pearson N, Greenwood R, Chan P, Peters TJ, Donovan J, Smith GD, and Dieppe P, Population requirement for primary hip-replacement surgery: a cross-sectional study. *The Lancet* 1999; 353(9161): p. 1304-9.

9. NIH Consensus Development Panel on Total Hip Replacement, Murray DG, Crown RS, Dickersin K, Duncan PW, Epps CH, Jr, Ettinger WH, Jr, Friedlaender GE, Lane JM, Lemons JE, Lewis JL, MacKenzie EJ, Pinals RS, Treuhaft PS, Bobyn JD, Callaghan JJ, Capello WN, Cauley JA, Davy DT, deAndrade JR, Engh CA, Fitzgerald RH, Jr, Galante JO, Gilbert JL, Goldberg VM, Goldring SR, Harris WH, Herberts P, Hungerford DS, Jacobs JJ, Jasty M, Johanson NA, Lewallen DG, Maloney WJ, Rubash HE, Salvati EA, Sledge CB, Sumner DR, Willert HG, Wright TM, Gordon S, Chwirut DJ, Dutta C, Elliott JM, Ferguson JH, Hall WH, Johnston RC, McGunagle DS, Quatrano LA, Wortman JS, Shulman LE, Hodes RJ, Alexander DF, and Pinn VW, Total Hip Replacement. *JAMA* 1995; 273(24): p. 1950-6.

10. National Institutes of Health, NIH Consensus Statement on Total Knee Replacement December 8-10, 2003. *J Bone Joint Surg Am* 2004; 86(6): p. 1328-35.

11. Hadorn DC and Holmes AC, The New Zealand priority criteria project. Part 1: Overview. *BMJ* 1997; 314(7074): p. 131.

12. The National Collaborating Centre for Chronic Conditions, *NICE Guideline 59 - Osteoarthritis: National clinical guideline for care and management in adults*, 2008, Royal College of Physicians: London.

13. NHS Hospital Episode Statistics. *Main procedures and interventions: 1999-2000 through 2011-2012*. Available from: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=205 (accessed 28 November 2012).

14. Bourne RB, Rorabeck CH, Laupacis A, Feeny D, Wong C, Tugwell P, Leslie K, and Bullas R, A randomized clinical trial comparing cemented to cementless total hip replacement in 250 osteoarthritic patients: the impact on health related quality of life and cost effectiveness. *Iowa Orthop J* 1994; 14: p. 108–14.

15. Pennington M, Grieve R, Sekhon JS, Gregg PJ, Black N, and van der Meulen J, Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis. *BMJ* 2013; 346.

16. Shields RK, Enloe LJ, and Leo KC, Health related quality of life in patients with total hip or knee replacement. *Archives of Physical Medicine and Rehabilitation* 1999; 80(5): p. 572-9.

17. Bozic KJ, Saleh KJ, Rosenberg AG, and Rubash HE, Economic evaluation in total hip arthroplasty: analysis and review of the literature. *The Journal of Arthroplasty* 2004; 19(2): p. 180-9.

18. Maynard A, Developing the Health Care Market. *The Economic Journal* 1991; 101(408): p. 1277-86.

19. Chang RW, Pellissier JM, and Hazen GB, A Cost-effectiveness Analysis of Total Hip Arthroplasty for Osteoarthritis of the Hip. *JAMA* 1996; 275(11): p. 858-65.

20. Learmonth ID, Young C, and Rorabeck C, The operation of the century: total hip replacement. *The Lancet* 2007; 370(9597): p. 1508-19.

21. Husted H, Holm G, and Jacobsen S, Predictors of length of stay and patient satisfaction after hip and knee replacement surgery: Fast-track experience in 712 patients. *Acta Orthopaedica* 2008; 79(2): p. 168 - 73.

22. Rolfson O, Dahlberg L, Nilsson J, Malchau H, and Garellick G, Variables determining outcome in total hip replacement surgery. *J Bone Joint Surg Br* 2009; 91(2): p. 157-61.

23. Schäfer T, Krummenauer F, Mettelsiefen J, Kirschner S, and Günther KP, Social, educational, and occupational predictors of total hip replacement outcome. *Osteoarthritis and Cartilage* 2010; 18(8): p. 1036-42.

24. Williams O, Fitzpatrick R, Hajat S, Reeves BC, Stimpson A, Morris RW, Murray DW, Rigge M, and Gregg PJ, Mortality, morbidity, and 1-year outcomes of primary elective total hip arthroplasty. *The Journal of Arthroplasty* 2002; 17(2): p. 165-71.

25. Department of Health, *Guidance on the routine collection of Patient Reported Outcome Measures (PROMs) - For the NHS in England 2009/10*, 2008, Department of Health.

26. Briggs A, Sculpher M, Britton A, Murray D, and Fitzpatrick R, The costs and benefits of primary total hip replacement: how likely are new prostheses to be cost-effective? *International Journal of Technology Assessment in Health Care* 1998; 14(4): p. 743 - 61.

27. Murray DW, Fitzpatrick R, Rogers K, Pandit H, Beard DJ, Carr AJ, and Dawson J, The use of the Oxford hip and knee scores. *J Bone Joint Surg Br* 2007; 89-B(8): p. 1010-4.

28. National Institute for Health and Clinical Excellence. *Hip disease - replacement prostheses: review decision - August 2011*. Available from: http://guidance.nice.org.uk/TA2/ReviewDecision (accessed 20 November 2012).

29. National Institute for Clinical Excellence. *Health Technology Appraisal - Total hip replacement and resurfacing arthroplasty for the treatment of pain or disability resulting from end stage arthritis of the hip (Review of technology appraisal guidance 2 and 44): Final scope*. Available from: http://www.nice.org.uk/nicemedia/live/13690/61348/61348.pdf (accessed 20 November 2012).

30. Claxton K, Sculpher M, and Drummond M, A rational framework for decision making by the National Institute For Clinical Excellence (NICE). *The Lancet* 2002; 360(9334): p. 711-5.

31. Centre for Reviews and Dissemination. *Systematic reviews: CRD's guidance for undertaking reviews in health care [Internet]*. Available from: http://www.york.ac.uk/inst/crd/systematic_reviews_book.htm (accessed 15 September 2010).

32. Moher D, Liberati A, Tetzlaff J, Altman DG, and for the PRISMA Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; 339(j): p. b2535.

33. Daellenbach HG, Gillespie WJ, Crosbie P, and Daellenbach US, Economic appraisal of new technology in the absence of survival data--the case of total hip replacement. *Soc Sci Med* 1990; 31(12): p. 1287-93.

34. Gillespie W, Pekarsky B, and O'Connell D, Evaluation of new technologies for total hip replacement. Economic modelling and clinical trials. *J Bone Joint Surg Br* 1995; 77-B(4): p. 528-33.

35. Baxter K and Bevan G, An economic model to estimate the relative costs over 20 years of different hip prostheses. *J Epidemiol Community Health* 1999; 53(9): p. 542-7.

36. McKenzie L, Vale L, Stearns S, and McCormack K, Metal on metal hip resurfacing arthroplasty. An economic analysis. *Eur J Health Econ* 2003; 4(2): p. 122-9.

37. Briggs A, Sculpher M, Dawson J, Fitzpatrick R, Murray D, and Malchau H, *Modelling the cost-effectiveness of primary hip replacement: how cost-effective is the Spectron compared to the Charnley prosthesis?*, in *CHE Technical Paper*2003, Centre for Health Economics, The University of York: York, UK.

38. Soohoo NF, Sharifi H, Kominski G, and Lieberman JR, Cost-Effectiveness Analysis of Unicompartmental Knee Arthroplasty as an Alternative to Total Knee Arthroplasty for Unicompartmental Osteoarthritis. *J Bone Joint Surg Am* 2006; 88(9): p. 1975-82.

39. Sharifi E, Sharifi H, Morshed S, Bozic K, and Diab M, Cost-Effectiveness Analysis of Periacetabular Osteotomy. *J Bone Joint Surg Am* 2008; 90(7): p. 1447-56.

40. Segal L, Day SE, Chapman AB, and Osborne RH, Can we reduce disease burden from osteoarthritis? *Med J Aust* 2004; 180(5 Suppl): p. S11-7.

41. Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, and Bevan G, Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. *Health Technol Assess* 1998; 2(6): p. 1-133.

42. Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, Dawson J, Carr A, Britton A, and Briggs A, Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses. *Health Technology Assessment* 1998; 2(20): p. 76.

43. Saleh KJ, Wood KC, Gafni A, and Gross AE, Immediate surgery versus waiting list policy in revision total hip arthroplasty : An economic evaluation. *The Journal of Arthroplasty* 1997; 12(1): p. 1-10.

44. Fisman DN, Reilly DT, Karchmer AW, and Goldie SJ, Clinical Effectiveness and Cost-Effectiveness of 2 Management Strategies for Infected Total Hip Arthroplasty in the Elderly. *Clinical Infectious Diseases* 2001; 32(3): p. 419-30.

45. Briggs A, Sculpher M, Dawson J, Fitzpatrick R, Murray D, and Malchau H, The Use of Probabilistic Decision Models in Technology Assessment: The Case of Total Hip Replacement. *Applied Health Economics and Health Policy* 2004; 3(2): p. 79-89.

46. Vale L, Wyness L, McCormack K, McKenzie L, Brazzelli M, and Stearns S, A systematic review of the effectiveness and cost-effectiveness of metal-on-metal

hip resurfacing arthroplasty for treatment of hip disease. *Health Technol Assess* 2002; 6(15): p. 1-109.

47. de Verteuil R, Imamura M, Zhu S, Glazener C, Fraser C, Munro N, Hutchison J, Grant A, Coyle D, Coyle K, and Vale L, A systematic review of the clinical effectiveness and cost-effectiveness and economic modelling of minimal incision total hip replacement approaches in the management of arthritic disease of the hip. *Health Technol Assess* 2008; 12(26): p. iii-iv, ix-223.

48. Slover J, Hoffman MV, Malchau H, Tosteson ANA, and Koval KJ, A Cost-Effectiveness Analysis of the Arthroplasty Options for Displaced Femoral Neck Fractures in the Active, Healthy, Elderly Population. *The Journal of Arthroplasty* 2009; 24(6): p. 854-60.

49. Bozic K, Pui C, Ludeman M, Vail T, and Silverstein M, Do the Potential Benefits of Metal-on-Metal Hip Resurfacing Justify the Increased Cost and Risk of Complications? *Clinical Orthopaedics and Related Research®* 2010; 468(9): p. 2301-12.

50. Bozic KJ, Morshed S, Silverstein MD, Rubash HE, and Kahn JG, Use of Cost-Effectiveness Analysis to Evaluate New Technologies in Orthopaedics. The Case of Alternative Bearing Surfaces in Total Hip Arthroplasty. *J Bone Joint Surg Am* 2006; 88(4): p. 706-14.

51. Marinelli M, Soccetti A, Panfoli N, and de Palma L, Cost-effectiveness of cemented versus cementless total hip arthroplasty. A Markov decision analysis based on implant cost. *J Orthop Traumatol* 2008; 9(1): p. 23-8.

52. Green WS, Toy P, and Bozic KJ, Cost Minimization Analysis of Preoperative Erythropoietin vs Autologous and Allogeneic Blood Donation in Total Joint Arthroplasty. *The Journal of Arthroplasty* 2010; 25(1): p. 93-6.

53. Murray D, Carr A, and Bulstrode C, Which primary total hip replacement? *J Bone Joint Surg Br* 1995; 77-B(4): p. 520-7.

54. National Joint Registry for England and Wales, *9th Annual Report*, 2012.

55. The Health and Social Care Information Centre, Hospital Episode Statistics for England. Inpatient statistics, 2011-12. 2012.

56. Caro JJ, Briggs AH, Siebert U, and Kuntz KM, Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2012; 15(6): p. 796-803.

57. National Institute for Health and Care Excellence. *Guide to the methods of technology appraisal 2013*. Available from: http://publications.nice.org.uk/pmg9 (accessed 9 July 2013).

58. Dawson J, Fitzpatrick R, Carr A, and Murray D, Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996; 78-B(2): p. 185-90.

59. Dawson J, Fitzpatrick R, Frost S, Gundle R, McLardy-Smith P, and Murray D, Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *J Bone Joint Surg Br* 2001; 83-B(8): p. 1125-9.

60. Field RE, Cronin MD, and Singh PJ, The Oxford hip scores for primary and revision hip replacement. *J Bone Joint Surg Br* 2005; 87-B(5): p. 618-22.

61. Kalairajah Y, Azurza K, Hulme C, Molloy S, and Drabu KJ, Health Outcome Measures in the Evaluation of Total Hip Arthroplasties—A Comparison Between the Harris Hip Score and the Oxford Hip Score. *The Journal of Arthroplasty* 2005; 20(8): p. 1037-41.

62. Dolan P, Gudex C, Kind P, and Williams A, The time trade-off method: Results from a general population study. *Health Economics* 1996; 5(2): p. 141-54.

63. EuroQuol Group. *EQ-5D Health Questionnaire - English version for the UK*. Available from: http://www.euroqol.org/eq-5d-products/how-to-obtain-eq-5d.html (accessed 5 February 2010).

64. Brazier J, Yang Y, Tsuchiya A, and Rowen D, A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European Journal of Health Economics* 2010; 11(2): p. 215-25.

65. Bansback N, Marra C, Tsuchiya A, Anis A, Guh D, Hammond T, and Brazier J, Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. *Arthritis Care & Research* 2007; 57(6): p. 963-71.

66. Grootendorst P, Marshall D, Pericak D, Bellamy N, Feeny D, and Torrance GW, A model to estimate health utilities index mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the knee. *The Journal of Rheumatology* 2007; 34(3): p. 534-42.

67. Browne J, Jamieson L, Lewsey J, Meulen Jvd, Black N, Cairns J, Lamping D, Smith S, Copley L, and Horrocks J. *Patient Reported Outcome Measures (PROMs) in Elective Surgery - Report to the Department of Health*. Available from: http://www.lshtm.ac.uk/php/hsrp/research/proms_report_12_dec_07.pdf (accessed 17 November 2010).

68. Mortimer D and Segal L, Comparing the Incomparable? A Systematic Review of Competing Techniques for Converting Descriptive Measures of Health Status into QALY-Weights. *Med Decis Making* 2008; 28(1): p. 66-89.

69. Gray AM, Rivero-Arias O, and Clarke PM, Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping. *Med Decis Making* 2006; 26(1): p. 18-29.

70. Franks P, Lubetkin EI, Gold MR, and Tancredi DJ, Mapping the SF-12 to preference-based instruments: convergent validity in a low-income, minority population. *Med Care.* 2003; 41(11): p. 1277-83.

71. World Health Organization. *BMI classification*. Available from: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html (accessed April 17th 2012).

72. Tsuchiya A, Brazier J, McColl E, and Parkin D, *Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices*, in *Sheffield Health Economics Group Discussion Paper Series*2002, School of Health and Related Research, The University of Sheffield.

73. Crott R and Briggs A, Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. *The European Journal of Health Economics* 2010; 11(4): p. 427-34.

74. Pinedo-Villanueva R, Turner D, Judge A, Raftery J, and Arden N, Mapping the Oxford hip score onto the EQ-5D utility index. *Quality of Life Research* 2013; 22(3): p. 665-75.

75. Judge A, Javaid MK, Arden NK, Cushnaghan J, Reading I, Croft P, Dieppe PA, and Cooper C, A clinical tool to identify patients who are most likely to receive long term improvement in physical function after total hip arthroplasty. *Arthritis Care & Research* 2012: p. n/a-n/a.

76. Ware JE, Jr., SF-36 health survey update. *Spine (Phila Pa 1976)* 2000; 25(24): p. 3130-9.

77. Steinbrocker O, Traeger CH, and Batterman RC, Therapeutic Criteria in Rheumatoid Arthritis. *Journal of the American Medical Association* 1949; 140(8): p. 659-62.

78. Harris WH, Traumatic Arthritis of the Hip after Dislocation and Acetabular Fractures: Treatment by Mold Arthroplasty: An end-result Study Using a New Method of Result Evaluation. *J Bone Joint Surg Am* 1969; 51(4): p. 737-55.

79. Arden NK, Kiran A, Judge A, Biant LC, Javaid MK, Murray DW, Carr AJ, Cooper C, and Field RE, What is a good patient reported outcome after total hip replacement? *Osteoarthritis and Cartilage* 2011; 19(2): p. 155-62.

80. Barton P, Bryan S, and Robinson S, Modelling in the economic evaluation of health care: selecting the appropriate approach. *J Health Serv Res Policy* 2004; 9(2): p. 110-8.

81. Cooper K, Brailsford SC, and Davies R, Choice of modelling technique for evaluating health care interventions. *J Oper Res Soc* 2006; 58(2): p. 168-76.

82. Briggs AH, Claxton K, and Sculpher MJ, *Decision modelling for health economic evaluation*. 2006: Oxford University Press.

83. National Joint Registry for England and Wales, *8th Annual Report*, 2011.

84. Siciliani L, Sivey P, and Street A, Differences in length of stay for hip replacement between public hospitals, specialised treatment centres and private providers: selection or efficiency? *Health Economics* 2012: p. n/a-n/a.

85. National Health Service, *The NHS Constitution: the NHS belongs to us all*, 2012.

86. Department of Health SC, Local Government & Care Partnerships,, *Implementing a ban on age discrimination in the NHS - making effective, appropriate decisions*, 2012.

87. NHS Hertfordshire, *Referral criteria for patients from primary care presenting with hip pain due to ostoarthritis, and clinical thresholds for elective primary hip replacement surgery*, 2012.

88. NHS Warwickshire, Referral and surgical threshold criteria for elective primary Hip Replacement Surgery. 2011.

89. Peninsula Commissioning Priorities Group, *Commissioning decision: Hip and knee replacement surgery in obese patients (those with a body mass index of 30 or greater)*, 2011.

90. Yorkshire and Humber Public Health Observatory, *Clinical Thresholds: Hip replacement for the treatment of joint symptoms and functional limitation*, 2010.

91. Garthwaite PH, Kadane JB, and O'Hagan A, Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* 2005; 100(470): p. 680-701.

92. O'Hagan A, *Uncertain judgements: eliciting experts' probabilities*. 2006: Wiley.

93. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, Woolacoot N, and Glanville J, Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004; 8(36): p. iii-iv, ix-xi, 1-158.

94. Linstone HA and Turoff M, *The Delphi method: techniques and applications*. 1975: Addison-Wesley Pub. Co., Advanced Book Program.

95. Clemen RT and Winkler RL, Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis* 1999; 19(2): p. 187-203.

96. Leal J, Wordsworth S, Legood R, and Blair E, Eliciting Expert Opinion for Economic Models: An Applied Example. *Value in Health* 2007; 10(3): p. 195-203.

97. McKenna C, McDaid C, Suekarran S, Hawkins N, Claxton K, Light K, Chester M, Cleland J, Woolacott N, and Sculpher M, Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis. *Health Technol Assess.* 2009; 13(24): p. iii-iv, ix-xi, 1-90.

98. Bojke L, Claxton KP, Bravo-Vergel Y, Sculpher M, Palmer SJ, and Abrams K, Eliciting Distributions to Populate Decision Analytic Models. *Value in Health* 2010.

99. van Noortwijk J, Dekker R, Cooke R, and Mazzuchi T, Expert judgement in maintenance optimization. *IEEE Transactions on Reliability* 1992; 41(3): p. 427-32.

100. Gigerenzer G, Hoffrage U, and Kleinbolting H, Probabilistic mental models: a Brunswikian theory of confidence. *Psychol Rev.* 1991; 98(4): p. 506-28.

101. Cooke R, *Experts in Uncertainty: Opinion and Subjective Probability in Science*. 1991: Oxford University Press.

102. Bojke L and Soares M, *Training slides for expert elicitation* Pinedo Villanueva R., Editor 2012.

103. Winkler RL and Clemen RT, Multiple Experts vs. Multiple Methods: Combining Correlation Assessments. *Decision Analysis* 2004; 1(3): p. 167-76.

104. Juslin P, Winman A, and Olsson H, Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes* 2003; 92(1–2): p. 34-51.

105. Stone M, The Opinion Pool. *The Annals of Mathematical Statistics* 1961; 32(4): p. 1339-42.

106. Office for National Statistics. *United Kingdom, Interim Life Tables, 1980-82 to 2007-09*. Available from: http://www.statistics.gov.uk/statbase/Product.asp?vlnk=14459 (accessed 8-February 2011).

107. The Health and Social Care Information Centre. *Provisional Monthly Patient Reported Outcome Measures (PROMs) in England - A guide to PROMs methodology*. Available from: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1488 (accessed 18 September 2012).

108. Rothwell AG, Hooper GJ, Hobbs A, and Frampton CM, An analysis of the Oxford hip and knee scores and their relationship to early joint revision in the New Zealand Joint Registry. *Journal of Bone & Joint Surgery, British Volume* 2010; 92-B(3): p. 413-8.

109. Brazier J, Roberts J, and Deverill M, The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* 2002; 21(2): p. 271-92.

110. Brazier J, Roberts J, Tsuchiya A, and Busschbach J, A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ.* 2004; 13(9): p. 873-84.

111. Whitehurst DGT and Bryan S, Another Study Showing that Two Preference-Based Measures of Health-Related Quality of Life (EQ-5D and SF-6D) are Not Interchangeable. But Why Should We Expect Them to be? *Value in Health* 2011; 14(4): p. 531-8.

112. Ara R and Wailoo A, Using Health State Utility Values in Models Exploring the Cost-Effectiveness of Health Technologies. *Value in Health* 2012; 15(6): p. 971-4.

113. Siani C, De Peretti C, Castelli C, Phung T, Daures J-P, and Duru G. *Uncertainty around the Mean Utility Assessment Accounting for Mapping Extrapolation: Application to Prostate Cancer*. Available from: http://www.ces-asso.org/sites/default/files/Siani_dePeretti_2010.pdf (accessed 13 March 2013).

114. Siani C, de Peretti C, Castelli C, Duru G, and Daures J-P. *Uncertainty around the Incremental Cost Utility Ratio Accounting for Mapping Prediction: Application to Hepatitis C*. Available from: http://www.ces-asso.org/sites/default/files/ArtCS2.pdf (accessed 13 March 2013).

115. Department of Health. *Publication of 2010-11 reference costs*. Available from: http://www.dh.gov.uk/health/2011/11/reference-costs/ (accessed July 3 2012).

116. Personal Social Services Research Unit, *Unit Costs of Health & Social Care 2011*, 2011, University of Kent: Canterbury.

117. Joint Formulary Committee. *British National Formulary (online)*. Available from: http://www.bnf.org (accessed June 20 2012).

118. Department of Health. *NHS reference costs: financial year 2011 to 2012*. Available from: https://www.gov.uk/government/publications/nhs-reference-costs-financial-year-2011-to-2012 (accessed 22 May 2013).

119. Medicines and Healthcare products Regulatory Agency. *The General Practice Research Database (GPRD) - Further Information for Patients*. Available from: http://www.erskinepractice.scot.nhs.uk/website/S11486/files/GPRD_PatientLeaflet.pdf (accessed 26 April 2013.

120. Lafuma A and Berdeaux G, Costs and effectiveness of travoprost versus a dorzolamide + timolol fixed combination in first-line treatment of glaucoma: analysis conducted on the United Kingdom General Practitioner Research Database*. *Current Medical Research and Opinion* 2007; 23(12): p. 3009-16.

121. Lafuma A and Berdeaux G, Costs and persistence of carbonic anhydrase inhibitor versus alpha-2 agonists, associated with beta-blockers, in glaucoma and ocular hypertension: an analysis of the UK-GPRD database*. *Current Medical Research and Opinion* 2008; 24(5): p. 1519-27.

122. Moore L, Remy V, Martin M, Beillat M, and McGuire A, A health economic model for evaluating a vaccine for the prevention of herpes zoster and post-herpetic neuralgia in the UK. *Cost Effectiveness and Resource Allocation* 2010; 8(1): p. 7.

123. Violato M, Gray A, Papanicolas I, and Ouellet M, Resource Use and Costs Associated with Coeliac Disease before and after Diagnosis in 3,646 Cases: Results of a UK Primary Care Database Analysis. *PLoS ONE* 2012; 7(7): p. e41308.

124. Administrative Data Liaison Service. *Administrative data introduction*. Available from: http://www.adls.ac.uk/adls-resources/guidance/introduction/ (accessed 04 Sept 2012.

125. Gutacker N, *[Electronic mail to Rafael Pinedo] Breakdown of total hip replacement HRGs by age and gender groups, primary and revision extracted from HES and NHS PbR*, 2013.

126. National Health Service. *NHS reference costs: financial year 2011 to 2012 - National Schedules of Reference Costs*. Available from: https://www.gov.uk/government/publications/nhs-reference-costs-financial-year-2011-to-2012 (accessed 24 May 2013).

127. Neuburger J, Hutchings A, Allwood D, Black N, and van der Meulen JH, Sociodemographic differences in the severity and duration of disease amongst patients undergoing hip or knee replacement surgery. *Journal of Public Health* 2012; 34(3): p. 421-9.

128. Judge A, Welton NJ, Sandhu J, and Ben-Shlomo Y, Equity in access to total joint replacement of the hip and knee in England: cross sectional study. *BMJ* 2010; 341.

129. Lalmohamed A VPKC and et al., Timing of acute myocardial infarction in patients undergoing total hip or knee replacement: A nationwide cohort study. *Archives of Internal Medicine* 2012; 172(16): p. 1229-35.

130. Lalmohamed A, Vestergaard P, Cooper C, de Boer A, Leufkens HGM, van Staa TP, and de Vries F, Timing of Stroke in Patients Undergoing Total Hip Replacement and Matched Controls. *Stroke* 2012.

131. South West London Public Health Network. *2010/11 South West London Effective Commissioning Initiative*. Available from: http://www.kingstonpct.nhs.uk/Downloads/Board%20Papers/24%20May%202010/Att%20D%20-ECI%20%20MASTER%20VERSION%202010-11%20DRAFT%207.pdf (accessed 13 June 2011).

132. Contract and Information Shared Services Unit. *Cheshire & Merseyside prior approval scheme: Incorporating Procedures of Lower Clinical Priority*. Available from: http://www.haltonandsthelenspct.nhs.uk/library/documents/HTSHcmfinalforcontractspapolicyfor2011122303114.pdf (accessed 17 April 2012).

133. NHS Derby City and NHS Derbyshire County. *Commissioning policy for procedures of limited clinical value (PLCV)*. Available from: http://www.derbycitypct.nhs.uk/UserFiles/Documents/1101%20Policy%20for%20PLCV%202011.pdf (accessed 13 June 2011).

134. Losina E, Walensky RP, Kessler CL, Emrani PS, Reichmann WM, Wright EA, Holt HL, Solomon DH, Yelin E, Paltiel AD, and Katz JN, Cost-effectiveness of Total Knee Arthroplasty in the United States: Patient Risk and Hospital Volume. *Arch Intern Med* 2009; 169(12): p. 1113-21.

135. Slover JD, Tosteson ANA, Bozic KJ, Rubash HE, and Malchau H, Impact of Hospital Volume on the Economic Value of Computer Navigation for Total Knee Replacement. *J Bone Joint Surg Am* 2008; 90(7): p. 1492-500.

136. Fordham R, Skinner J, Wang X, Nolan J, and Group tEPOS, The economic benefit of hip replacement: a 5-year follow-up of costs and outcomes in the Exeter Primary Outcomes Study. *BMJ Open* 2012; 2(3).

137. Brazier J, Jones N, and Kind P, Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Quality of Life Research* 1993; 2(3): p. 169-80.

138. Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, and Brown H, Validity of EuroQol -a generic health status instrument- in patients with rheumatoid arthritis. *Rheumatology* 1994; 33(7): p. 655-62.

139. van Agt H and Bonsel G, *EQ-5D concepts and methods: A developmental history*, Kind P., Brooks R., and Rabin R., Editors. 2005, Springer Netherlands. p. 29-33.

140. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, and Badia X, Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research* 2011; 20(10): p. 1727-36.

141. Department of Ortopaedics - Sahlgrenska University Hospital. *Swedish Hip Arthroplasty Register - Annual Report 2008, Shortened Version*. Available from: https://www.jru.orthop.gu.se/ (accessed 28 March 2011).

142. National Joint Registry for England and Wales. *NJR PROMs*. Available from: http://www.njrcentre.org.uk/njrcentre/Research/NJRPROMs/tabid/203/Default.aspx (accessed July 3 2012).

143. National Institute for Clinical Excellence. *Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease. Review of NICE technology appraisal guidance 111*. Available from:

http://www.nice.org.uk/nicemedia/live/13419/53619/53619.pdf (accessed 22 July 2013).

144. Laking G, Lord J, and Fischer A, The economics of diagnosis. *Health Economics* 2006; 15(10): p. 1109-20.

145. National Health Service. *The NHS structure explained*. Available from: http://www.nhs.uk/NHSEngland/thenhs/about/Pages/nhsstructure.aspx (accessed 22 July 2013).