Author:
Fowler, Tim J

Title:
The association between surgeon grade and implant survival following hip and knee replacement

Author:
Fowler, Tim J

Title:
The association between surgeon grade and implant survival following hip and knee replacement

THE ASSOCIATION BETWEEN SURGEON GRADE AND IMPLANT SURVIVAL FOLLOWING HIP AND KNEE REPLACEMENT

Timothy John Fowler

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Medicine (MD) in the Faculty of Health Sciences, School of Translational Health Sciences.

May 2022

Word Count: 49,445

Abstract

Over 2.5 million primary hip and knee replacements have been performed in England and Wales since the inception of the National Joint Registry (NJR) in 2003. These common procedures provide a cost-effective means of treating end-stage osteoarthritis. However, implant constructs are susceptible to failure, which often requires revision surgery. The revision-free survival of a joint replacement is a commonly used measure of surgical performance.

Hip and knee replacements are performed by surgeons of different grades with varying levels of supervision. It is not known if total hip replacements (THRs), total knee replacements (TKRs), or unicompartmental knee replacements (UKRs) performed by trainees last as long as those performed by consultants. The primary aim of this thesis was to investigate the association between surgeon grade (trainee or consultant) and implant survival following THR, TKR, and UKR.

A preliminary study using methods of evidence synthesis, found no strong evidence in the existing literature that trainees achieve worse outcomes compared to consultants, in terms of the net survival or crude revision rate of hip and knee replacements with 5 to 10 years follow-up. Three studies using NJR data were conducted to investigate the association between surgeon grade, the supervision of trainees, and the risk of all-cause revision following THR, TKR, and UKR. The indication for revision was explored as a secondary outcome measure. A combination of statistical methods including Kaplan-Meier, Cox regression, and flexible parametric survival analysis were employed. A final study using a local database explored the association between surgeon grade and a range of clinical, functional, and radiological outcome measures.

In general, the findings of this thesis support the conclusion that trainees in England and Wales achieve comparable all-cause implant survival to consultants following THR, TKR, and UKR. However, trainees should ideally be supervised by a scrubbed consultant when performing these procedures.

Dedication

I dedicate this thesis to my wife Kirsten and my son William, who was born during the course of this degree. Kirsten has made significant personal sacrifices to support me in my pursuit of completing this thesis and I am indebted to her for her patience, understanding and encouragement. This significant volume of work was completed alongside my ongoing surgical training and predominantly during the COVID-19 pandemic. It has been a challenging period in our lives, but a more enriching and enjoyable experience thanks to Kirsten. This was only possible thanks to her unwavering support and generosity.

Acknowledgements

The completion of this work has required the support of numerous individuals to whom I am most grateful. I would like to give particular thanks to my supervisors, Professor Michael Whitehouse and Dr Adrian Sayers, whose expert supervision has been of the highest quality. Their insight into this subject and the research methodologies employed has been invaluable. Prior to commencing this degree, my practical experience of epidemiological research and medical statistics was very limited. The completion of this work would not have been possible without their patience and guidance.

I would like to thank Professors Ashley Blom and Michael Whitehouse for giving me this opportunity and for supporting me in my application for a NIHR Academic Clinical Fellowship. This gave me the time, training, and resources to complete this work alongside my surgical training. It has been a privilege to work alongside other members of the Musculoskeletal Research Unit at the University of Bristol. While the pandemic has meant that much of this work has been completed remotely, I have always found my colleagues to be generous and professional in the advice they have given me. I am particularly grateful to Erik Lenguerrand and Setor Kunutsor for conducting my annual reviews. Their guidance has enhanced this work.

I would like to thank the patients and staff who have contributed data to the National Joint Registry and to our own local study. I am grateful to the Healthcare Quality Improvement Partnership, the National Joint Registry Steering Committee, and staff at the NJR Centre for facilitating this work. Finally, I am grateful to the trainees who performed many of the operations included in this thesis, and to the consultants who trained them. The findings of this thesis highlight the exceptional quality of their work and the excellent outcomes that trainees can achieve with consultant guidance. I hope that the outputs of this thesis will further improve surgical training practices in the United Kingdom (UK) and improve outcomes for patients.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ...  ... DATE: 11th October 2022

Statement of publications

The following publications have been included in this thesis according to the University's guidance on the integration of publications as chapters within a dissertation:

1) Fowler, T. J., Aquilina, A. L., Blom, A. W., Sayers, A. & Whitehouse, M. R. 2021. Association between surgeon grade and implant survival following hip and knee replacement: a systematic review and meta-analysis. *BMJ Open,* 11**,** e047882.
    a. This publication forms Chapter 2 of this thesis
    b. Published contribution statement: TF, AB, AS and MW conceived and designed the study; TF and AA independently screened the articles and performed data extraction in duplicate; TF and AS were responsible for data analysis; all authors were responsible for interpreting the data; TF drafted the manuscript; AB, AA, AS and MW revised the article critically for important intellectual content; all authors reviewed the final version of the manuscript and gave approval for submission for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.
    c. doi: 10.1136/bmjopen-2020-047882

2) Fowler, T. J., Aquilina, A. L., Reed, M. R., Blom, A. W., Sayers, A. & Whitehouse, M. R. 2022. The association between surgeon grade and risk of revision following total hip arthroplasty : an analysis of National Joint Registry data. *Bone Joint J,* 104-B**,** 341-351.
    a. This publication forms Chapter 3 of this thesis
    b. Unpublished contribition statement: TF was responsible for study concept, design, literature review, data analysis, interpretation of the results, and writing the manuscript. AA contributed to the literature review and was responsible for interpretation of the results and review of the manuscript. AB, MR and MW were responsible for study concept, design, interpretation of the results, and review of the manuscript. AS was responsible for study concept, design, data analysis, interpretation of the results, and review of the manuscript.
    c. doi: 10.1302/0301-620X.104B3.BJJ-2021-1389.R1.

3) Chapters 4, 5 and 6 have not yet been submitted for publication, but are presented in a publishable format. Author contribution statements for these chapters are included at the start of each chapter. TF was the first authour and principal researcher for every chapter.

SIGNED:

Mr Timothy Fowler
(MD candidate & first author)

Dr Adrian Sayers
(Supervisor & senior author)

Professor Michael Whitehouse
(Main supervisor & senior author)

Table of contents

List of figures

List of tables

"In a speciality as vast as that of orthopaedics, surgeons yearn for an easy hip operation, or, if a good operation is difficult, they hope that having mastered its performance through trial and tribulation it should be universally applicable. The only type of operation that could ever be universal would be an arthroplasty because this is the reconstruction of a normal joint. This type of surgery demands training in mechanical techniques which, though elementary in practical engineering, are as yet unknown in the training of a surgeon."

Sir John Charnley, 1959

Chapter 1 Introduction

1.1 Context of the thesis

Hip and knee replacements are common surgical interventions, primarily for the treatment of painful degenerative conditions affecting the hip and knee, such as osteoarthritis (OA). The burden of disease and the demand for these interventions is increasing (Ferguson et al., 2018, Price et al., 2018). This represents a significant challenge for healthcare systems, which rely on a highly-trained surgical workforce to meet this demand.

Surgical training programmes must train a sufficient number of surgeons to meet evolving workforce requirements (Centre for Workforce Intelligence, 2014). The UK Trauma and Orthopaedic Surgery curriculum mandates that trainees perform a minimum of 80 major joint replacement procedures before specialist certification (ISCP, 2021). In this context, hip and knee replacements are performed by surgeons of different grades, with varying levels of supervision. However, we do not know if this is a safe practice or whether trainees achieve comparable outcomes to fully-trained consultants.

There are challenges when quantifying the safety and efficacy of surgical interventions, but objective clinical endpoints can be useful measures of success. Hip and knee replacements are susceptible to failure and may require revision surgery. The revision-free survival of a joint replacement is a commonly used objective surgical outcome measure, calculated using statistical methods of survival analysis, such as Kaplan-Meier (KM). Joint replacement registries, such as the National Joint Registry (NJR), are a valuable source of implant survival data.

The body of work presented in this thesis concerns the outcomes of trainee-performed hip and knee replacements. The aim is to investigate the association between surgeon grade (consultant or trainee), the supervision of trainees, and implant survival following primary hip and knee replacement. This introductory chapter gives a comprehensive background of the key themes of this thesis, provides an overview of the existing literature on this subject, and establishes the aims and objectives.

1.2 Total hip replacement

Total hip replacement (THR) is a cost-effective surgical intervention for the treatment of painful degenerative conditions of the hip joint (Daigle et al., 2012). It is estimated that over one million THRs are performed worldwide per year (OECD, 2021), which reflects the effectiveness of THR to improve pain, function, and quality of life in the majority of patients (Ferguson et al., 2018).

1.2.1 Epidemiology

OA is a degenerative condition that causes progressive damage to articular cartilage and the surrounding anatomical structures, resulting in pain and disability. It affects over 250 million people worldwide and its prevalence is increasing, which represents a worsening health and socioeconomic burden (Hunter and Bierma-Zeinstra, 2019). The underlying aetiology of hip OA is multifactorial and represents an interaction between biological and environmental risk factors. Major risk factors include age, female sex, obesity, genetics, high impact activities, and conditions of joint incongruity (Hunter and Bierma-Zeinstra, 2019).

OA accounts for over 90% of primary THRs performed in England and Wales (The National Joint Registry, 2021a). Other indications include neck of femur fracture, avascular necrosis, and inflammatory arthritis. The median age of patients who undergo primary THR in England and Wales is 69 years (interquartile range [IQR] 61 to 76), and 60% of patients are female (The National Joint Registry, 2021a).

More than 98,000 primary THRs were performed in England and Wales in 2019, and over 1.25 million have been recorded in the NJR since its inception in 2003. The number of THRs performed per year in the UK rose by over 40% between 2009 and 2019 (The National Joint Registry, 2021a). A recent NJR study predicts that demand for THR in the UK will rise by a further 40% by 2060 (Matharu et al., 2021). Over 500,000 primary THRs were performed in the USA in 2018 (Agency for Healthcare Research and Quality, 2018), suggesting that the number of cases performed annually has been rising faster than previously predicted (Kurtz et al., 2014). This reflects the growing burden of

disease and the improved provision of arthroplasty services. However, the trajectory of recovery of these services in the wake of the COVID-19 pandemic is uncertain.

1.2.2 Historical context

The evolution of modern hip arthroplasty began in the 20th century with pioneering attempts attributed to Smith-Peterson, Wiles, and Charnley (Smith-Petersen, 2006, Wiles, 1958, Charnley, 1961). Sir John Charnley's low frictional torque arthroplasty, which was developed during the early 1960s, was a revolutionary innovation that set the standard for the modern era of THR. Among Charnley's significant contributions to implant design were the use of polymethylmethacrylate (PMMA) cement for component fixation, and the use of high-density polyethylene as a bearing material (Learmonth et al., 2007). Robin Ling and Clive Lee built on these principles in their design of the Exeter Hip (1970). Their collarless, polished, double-tapered stem relies on the viscoelastic properties of cement and established the taper slip theory that is central to the design of contemporary cemented femoral components (Fowler et al., 1988). The Exeter V40 stem (Stryker, Newbury, UK), an evolution of the original Exeter stem, is the most commonly used femoral component in several countries, including the UK, Australia, and New Zealand (The National Joint Registry, 2021a, AOANJRR, 2020, The New Zealand Joint Registry, 2020).

1.2.3 Implant components

Six decades of innovation has left surgeons with a significant catalogue of THR components to choose from. Recent evidence suggests that over 4,400 different implant construct combinations were used for primary THR in England and Wales between 2003 and 2016 (Deere et al., 2019).

The basic components of a contemporary THR include: 1) a metal femoral stem, which is fixed within the proximal femur and attaches to a modular head through a tapered trunnion; and 2) an acetabular cup (with or without modular liner), which is fixed into the prepared acetabulum and articulates with the prosthetic head. Component fixation is either cemented (cemented stem/cemented cup), uncemented (uncemented stem/uncemented cup), hybrid (cemented stem/uncemented cup), or reverse hybrid (uncemented stem/cemented cup). Cemented implants use PMMA for fixation, whereas

uncemented implants typically rely on initial press-fit fixation and subsequent osseointegration. A combination of head and cup/liner materials defines the bearing surface. Metal-on-polyethylene, ceramic-on-polyethylene, and ceramic-on-ceramic are the most commonly used bearing surfaces in the UK (The National Joint Registry, 2021a).

Incremental developments have refined the components that we use today in an attempt to improve longevity by mitigating the risk of failure due to loosening, lysis, wear, and instability. However, some innovations, such as the use of metal-on-metal bearing surfaces on stemmed prostheses, have had catastrophic consequences. These designs gained popularity due to perceived tribological advantages that allowed larger head sizes to be used, thereby reducing dislocation rates. However, registry analyses identified strong associations between the use of stemmed metal-on-metal implants and high failure rates (Smith et al., 2012b, Hunt et al., 2018).

A network meta-analysis has suggested that newer implant combinations are no better than a reference implant combination (small head (<36mm) cemented metal-on-polyethylene), in terms of risk of revision or Harris hip score (HHS) following primary THR (Lopez-Lopez et al., 2017). Fawsitt et al. conducted a cost-effectiveness analysis using NJR and Swedish Hip Arthroplasty Register (SHAR) data to compare the cost-effectiveness of various implants against small head (<36mm) cemented metal-on-polyethylene implants. They found that the cheapest implants (small head cemented metal-on-polyethylene) were the most cost-effective for patients over the age of 65. In contrast, small head cemented ceramic-on-polyethylene implants were the most cost-effective for patients under the age of 65 (Fawsitt et al., 2019).

1.3 Knee replacement

Over 100,000 knee replacements are performed per year in the UK alone, owing to the reduced pain and improved function experienced by the majority of patients (Price et al., 2018, The National Joint Registry, 2021a). The two main types of knee replacement include total knee replacement (TKR) and unicompartmental knee replacement (UKR); the use of each is typically dependent on the pattern of disease, surgeon expertise, and patient choice.

1.3.1 Epidemiology

It is estimated that knee OA accounts for approximately 85% of the burden of OA worldwide (Hunter and Bierma-Zeinstra, 2019). The underlying aetiology is multifactorial and represents an interaction between biological and environmental risk factors. Moderate to strong risk factors include increasing age, female sex, obesity, high-impact sports, and previous trauma (Hunter and Bierma-Zeinstra, 2019). OA is the indication for more than 95% of knee replacements performed in the UK (The National Joint Registry, 2021a). Estimates using data from the UK Clinical Research Practice Datalink (CRPD) suggest that the lifetime risk of TKR in men and women over the age of 50 is 7.7% and 10.6%, respectively (Culliford et al., 2012).

The number of primary TKRs performed in England and Wales increased from 61,000 in 2009 to 90,000 in 2019 (The National Joint Registry, 2021a). Conservative estimates suggest that this will increase to 118,000 by 2035 (Culliford et al., 2015), but these do not consider the recent impact of COVID-19 (The National Joint Registry, 2021a). Perhaps the most striking trend is the increased utilisation rates in younger patients, which is of particular concern given the increased risk of revision in patients who undergo primary TKR at a younger age (Pabinger et al., 2015, Bayliss et al., 2017).

Using CRPD data, Bayliss et al. demonstrated that patients aged ≤60 years at primary surgery have a significantly increased lifetime risk of revision. This effect is particularly pronounced for men who undergo TKR aged 50-55 years, who have an estimated lifetime risk of revision of 24%. This observation, combined with the increasing number of knee replacements performed per year, poses a significant and increasing burden for healthcare systems (Bayliss et al., 2017).

1.3.2 Historical context

Gunston is credited with establishing the concept of condylar TKR. His polycentric knee replacement established the following basic principles: 1) a metal femoral component that articulates with a polyethylene tibial component; 2) PMMA cemented fixation; 3) minimal bone resection; 4) stability primarily provided by native soft tissue balance; and 5) preservation of a physiological range of movement (Gunston, 1971). The evolution of TKR technology throughout the 1970s produced

implants with varying levels of constraint, different methods of fixation, and the option for patellar resurfacing. Continued incremental advances in design have led to the development of contemporary TKR constructs with excellent long-term survivorship (Evans et al., 2019b).

The concept of UKR was introduced in the 1950s with medial tibial plateau resurfacing designs (McKeever, 2005, MacIntosh and Hunter, 1972). Early UKR designs, including the Marmor and the St. Georg sled, were introduced in the 1970s and utilised fixed bearing tibial components (Marmor, 1979, Engelbrecht et al., 1976). Goodfellow and O'Connor subsequently developed the Oxford Partial Knee, which uses a mobile sliding bearing to provide a large contact area and facilitate efficient load transmission through a physiological range of movement (Goodfellow and O'Connor, 1978).

1.3.3 Implant components

1.3.3.1 Total knee replacement

The basic components of a contemporary primary TKR include: 1) a metal femoral component; 2) a tibial component, which is either modular with a metal baseplate and polyethylene insert, or monobloc polyethylene; and 3) an optional polyethylene patellar resurfacing. Variations in implant design allow surgeons to specify the method of fixation (cemented, uncemented, or hybrid), the level of constraint, bearing mobility (fixed or mobile), and whether or not the patella is resurfaced.

Constraint is a spectrum, which describes the level of stability and the freedom of movement conferred by the articulation between the tibial and femoral components. For example, the level of constraint varies depending on whether the posterior cruciate ligament is preserved (cruciate-retaining) or sacrificed (posterior-stabilised) during the operation. Cruciate-retaining and posterior-stabilised implants are the most commonly used TKR designs in England and Wales (The National Joint Registry, 2021a). Bearing mobility describes whether the polyethylene tibial insert is mobile or fixed within the baseplate. The patella may be selectively resurfaced during TKR. However, recent evidence suggests that performing TKR without patellar resurfacing is associated with an increased risk of revision, and secondary resurfacing is associated with a high risk of re-revision (Hunt et al., 2021b).

1.3.3.2 Unicompartmental knee replacement

The majority of UKRs performed in the UK involve the medial compartment, 5 to 10% involve the lateral compartment, and 1% involve the patellofemoral joint (The National Joint Registry, 2021a, Scott, 2005). In this thesis, the term UKR refers to arthroplasty of the medial or lateral compartment; alternative terms used in the literature include partial knee replacement and unicondylar knee replacement. The basic components of a UKR include: 1) a metal femoral component; and 2) a tibial component, which is typically modular with a metal baseplate and polyethylene insert. The main variations in implant design include the method of fixation (cemented, uncemented, or hybrid), bearing mobility (fixed or mobile), and laterality (medial or lateral). There is an ongoing debate on the relative merits of different UKR designs, particularly with regards to fixation and bearing mobility (Mohammad et al., 2021, Abu Al-Rub et al., 2020).

End-stage medial compartment OA can be treated with either TKR or UKR. While fewer than 10% of knee replacements in the UK are unicompartmental, it is estimated that up to 47% of patients have isolated medial compartment disease and may be eligible for either implant (Beard et al., 2019). Reported advantages of UKR over TKR include superior functional outcomes, reduced length of stay, fewer medical complications, greater cost-effectiveness, and lower mortality (Wilson et al., 2019, Beard et al., 2019). However, UKR revision rates are considerably higher than primary TKR revision rates and may have a higher risk of revision if performed by low-volume surgeons (Baker et al., 2013, Liddle et al., 2016).

1.4 Surgical training

The aim of surgical training is to develop competent surgeons who are capable of safe independent practice. Contemporary surgical training programmes aim to achieve this by ensuring an appropriate balance between service provision and training, by professionalising trainers, and by providing a learning environment that supports trainees to meet the requirements of a competency-based curriculum (The Royal College of Surgeons of England, 2015). Furthermore, they must ensure that a sufficient number of surgeons are trained in order to meet projected workforce requirements (The

British Orthopaedic Association, 2021, Centre for Workforce Intelligence, 2014). This section gives an overview of the current practice and regulatory structure of orthopaedic training in the UK. It aims to give insight into the learning environment in which surgeons in the UK are trained to perform hip and knee replacements.

1.4.1 Orthopaedic training in the United Kingdom

1.4.1.1 Recent historical context

Surgical training in the UK before the 1990s typically relied on an apprenticeship-based model. Progression through the traditional hierarchy of grades was generally based on local service provision requirements and the informal reports of supervising consultants (Hurreiz, 2019). Concerns about protracted training programmes, which consisted "of mainly unsupervised service; and unstructured, unreliable, and invalid feedback on the grapevine" (Hunter and McLaren, 1993), and reports of unacceptably low levels of supervision were commonplace (Collins, 1999, Lourie, 1999, Lourie, 1998, Wilson, 1997). Significant reforms to the structure of postgraduate surgical training in the UK were prompted by a growing concern for patient safety, the need to meet evolving workforce demands, and limitations to working hours introduced by the European Working Time Directive (EWTD) (Kelly and Canter, 2007, Purcell Jackson and Tarpley, 2009, Carr et al., 2002).

The 1993 Calman report recommended the introduction of a Certificate of Completion of Training (CCT) as a definitive educational endpoint, clearly defined curricula, and greater emphasis on assessment and feedback (Calman et al., 1999, Hurreiz, 2019). The Modernising Medical Careers (MMC) initiative (2005) saw the introduction of the Foundation Programme, Specialty Training, and new competency-based methods of assessment (Health et al., 2003, Kelly and Canter, 2007). The Shape of Training Review (2013) has continued the evolution of surgical training in the UK into its current form (Greenaway, 2013).

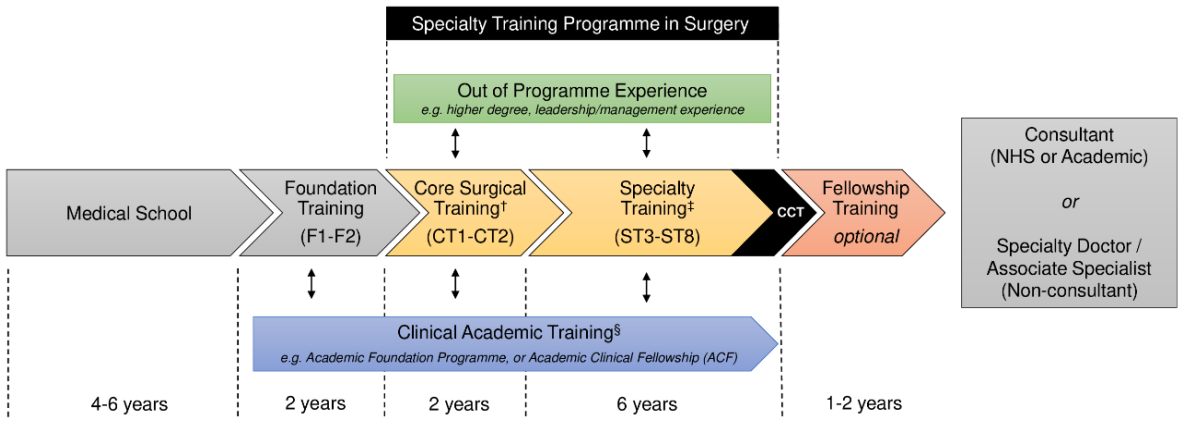
1.4.1.2 Current practice

On graduating from medical school, newly qualified doctors commence the 2-year Foundation Programme (F1-F2), in which they rotate through placements in a variety of medical and surgical

specialties. During this time, aspiring surgeons build a portfolio to support their application to Core Surgical Training (CST); a 2-year basic surgical training programme. Admittance to CST is by a competitive national selection process. Applicants are selected based on interview performance and achievement in the following areas: higher degrees; teaching qualifications; presentations and publications; involvement in quality improvement projects; and professional examinations (Health Education England, 2021).

Second-year CSTs who have completed the Membership of the Royal College of Surgeons (MRCS) examination are eligible to apply to Specialty Training, admittance to which is based on a further national selection process. Successful applicants are issued a National Training Number (NTN) and allocated to a regional training deanery based on preference and interview ranking. Specialty Training in Trauma and Orthopaedic Surgery (T&O) is typically a 6-year programme (ST3-ST8), during which trainees rotate through subspecialty placements on a 6-monthly basis. T&O trainees who have completed ST6 are eligible to sit the examination for Fellowship of the Royal College of Surgeons (FRCS), which is mandatory for subsequent CCT. A schematic summary of the stages of training in the UK is illustrated in Figure 1 (Fitzgerald et al., 2012).

Figure 1 - Schematic overview of surgical training in the UK. Adapted from: Fitzgerald JEF et al. (2012).



†General or themed posts; MRCS required for progression to ST3. ‡Specialty training in Trauma & Orthopaedic Surgery; FRCS required for progression to CCT. §Trainees have allotted time for academic training.

The UK curriculum is produced by the Specialty Advisory Committee (SAC) and implemented through the Intercollegiate Surgical Curriculum Programme (ISCP) (JCST, 2020). It provides a framework for training up to the expected level of a "day-one consultant". The current T&O curriculum assesses nine Generic Professional Capabilities (GPCs) and five Capabilities in Practice (CiPs), which are summarised in Appendix 1 (ISCP, 2021).

A trainee's competence is assessed continuously through the completion of workplace-based assessments (WBAs), which take three main forms: case-based discussion (CBD); clinical evaluation exercise (CEX); and procedure-based assessment (PBA). For each WBA, the trainee is awarded a score from level 1-4, based on the supervising consultant's judgement of the trainee's competence in the corresponding area of the curriculum.

PBAs were developed by the British Orthopaedic Association and the T&O SAC as part of the Orthopaedic Competence Assessment Project (Howells et al., 2008). They form an integral part of the T&O curriculum for the assessment of operative competence and have been adapted for use in all surgical specialties. An example of a PBA for primary THR is included in Appendix 2, along with a detailed summary of the rating system. The T&O curriculum defines an indicative number of 1,800 operations that trainees must perform before CCT. This includes 13 index procedures, including hip and knee replacements, for which trainees must achieve an indicative number and demonstrate level 4 competency (Appendix 3).

Training progress is formally assessed at the Annual Review of Competence Progression (ARCP), where trainees are required to demonstrate to a panel of experienced trainers that they have met their annual training requirements. On completion of Specialty Training, T&O trainees are recommended to the General Medical Council (GMC) for CCT based on minimum requirements for certification. Orthopaedic training in the UK is longer in duration and has the highest standards in terms of minimum operative experience compared to other countries, including the USA, Canada, Australia, and the European Union (Tahir et al., 2021).

1.4.1.3 Regulation of training

The regulatory structure of surgical training in the UK involves several key stakeholders. The GMC is the regulatory body for postgraduate medical training in the UK. It is responsible for setting the standards for postgraduate curricula and admitting clinicians to the Specialist Register; a register of doctors with a CCT who are eligible for consultant practice in the National Health Service (NHS) (General Medical Council, 2021a).

The Joint Committee on Surgical Training (JCST) acts on behalf of the four surgical Royal Colleges of the UK and Ireland. It is the parent organisation for the ten SACs, which work with the JCST to determine the content, structure, and implementation of training programmes. Through the SACs, the JCST develops curricula and is responsible for recommending appropriately qualified candidates for admission to the GMC Specialist Register (General Medical Council, 2021b).

The curriculum is implemented through the ISCP; an online portal through which trainees document their progress throughout the training programme (JCST, 2020). Implementation of the curriculum is delivered at a regional level by the Schools of Surgery, which are managed by Local Education Training Boards under the umbrella of NHS Health Education England (or equivalent in the devolved nations).

1.4.1.4 Surgical skills training

Purcell Jackson estimated that it takes 20,000 hours to train a surgeon (Purcell Jackson and Tarpley, 2009). The reduction of time in training due to MMC and the EWTD was met with concerns that trainees were completing training having performed significantly fewer operations, including hip and knee replacements, than their predecessors (Nzeako and Back, 2016, Sher et al., 2005). Lord Darzi has suggested that “It is no longer acceptable, or appropriate, for students at any level of training to practice new skills on patients, even if they have a patient's explicit consent (Aggarwal and Darzi, 2006).” Such factors have driven reforms to training programmes and advances in alternative methods used to train inexperienced surgeons (e.g. simulation and virtual reality technology).

The UK T&O curriculum outlines how surgical skills training within clinical practice should progress according to the following framework: 1) trainee observes trainer; 2) trainee assists trainer; 3) trainee is supervised by trainer; and 4) trainee operates independently with easy access to trainer. It states that trainees should have the opportunity for increased autonomy without compromising safe and effective patient care. There should be a gradual reduction in the level of supervision and an increase in case complexity until the level of competence for independent practice is acquired. It emphasises the fundamental role of supervision in the delivery of safe and effective training, and that the ultimate responsibility for the quality of patient care lies with the supervising consultant (ISCP, 2021, General Medical Council, 2021b). Trainees in the UK are required to record the level of supervision for operations they have performed using the following codes: supervised – trainer scrubbed (S-TS); supervised – trainer unscrubbed (S-TU); performed (P); and training more junior trainee (T). Full descriptions of these supervision codes are included in Appendix 4.

Simulation training is recommended as an adjunct to supervised clinical training in the UK (ISCP, 2021). This encompasses several formats, including Sawbone models, cadaveric workshops, arthroscopic simulators, and virtual reality training. Evidence from randomised controlled trials (RCTs) supports the efficacy of arthroscopic simulation, which has been shown to improve the technical performance of trainees in the operating theatre (Polce et al., 2020, Howells et al., 2008). Furthermore, recent studies have demonstrated that virtual reality simulation can improve trainee performance in THR (Logishetty et al., 2019, Logishetty et al., 2020).

1.4.2 Defining a surgeon's level of experience: clarification of terminology

Surgical experience is a broad term which encompasses several definitions. This thesis focuses on the specific exposure variable 'surgeon grade'. Therefore, it is important to distinguish between the alternative measures of surgical experience that have been used in the orthopaedic literature.

1.4.2.1 Surgeon grade

The term 'surgeon grade' is a commonly used measure of surgical experience, which is used to define a surgeon's designated level of training (e.g. trainee or consultant). In the UK, consultants have been

awarded a CCT and appointed to a senior clinical position in which they can be responsible for the supervision of trainees. The term ‘consultant’ is equivalent to ‘attending’, which is used in several healthcare settings, including North America. In the context of surgical training, the term ‘trainee’ refers to a postgraduate surgical trainee, who practices under the supervision of a consultant (or international equivalent).

In the existing orthopaedic literature, ‘surgeon grade’ has been used as a binary variable to describe whether a procedure has been performed by a trainee or a consultant (Bottomley et al., 2016, Palan et al., 2009, Faulkner et al., 2018, Jain et al., 2018, Reidy et al., 2016). It is also established practice for trainee-performed procedures to be subcategorised according to whether or not a trainee was supervised by a consultant (Inglis et al., 2013, Storey et al., 2018). Throughout this thesis, the term ‘surgeon grade’ is used as a binary variable to describe whether a procedure was performed by a trainee or a consultant.

1.4.2.2 Surgeon volume

The term ‘surgeon volume’, which is often used interchangeably with ‘surgeon caseload’, is typically used to describe the number of operations performed by an individual surgeon in a calendar year. In the arthroplasty literature, it has been used to describe a surgeon’s mean annual volume (Liddle et al., 2016, Mohammad et al., 2020), or a surgeon’s volume in the 365 days preceding an index procedure (Sayers et al., 2020a, Ravi et al., 2014, Jolbäck et al., 2019). Researchers have used several different methodological approaches to investigate the association between surgeon volume and the risk of revision following hip and knee replacement (Liddle et al., 2016, Sayers et al., 2020a, Ravi et al., 2014, Baker et al., 2013).

1.4.2.3 Duration of clinical practice

Jolbäck et al. used SHAR data to define surgical experience according to a surgeon’s duration of clinical practice. In an analysis of patient-reported outcome measures (PROMs), the authors compared THRs performed by trainees to three other surgeon groups, defined according to the number of years

between specialist certification and an index procedure (<8 years, 8-15 years, or >15 years) (Jolbäck et al., 2018).

1.4.2.4 Learning curve

The term 'learning curve' is generally used to describe a surgeon's rate of skills acquisition during an early period of practice, or the minimum number of cases required for a surgeon to become "competent" in a specific procedure. The term has been widely used in the orthopaedic literature, including in the context of training in hip and knee replacement (Nzeako and Back, 2016, McCulloch et al., 2021, de Steiger et al., 2015, Alvand et al., 2021).

1.5 Outcomes of hip and knee replacement

There are challenges when quantifying the safety and efficacy of a joint replacement, but implant survival and patient-reported outcomes can be useful metrics of success. These commonly used outcome measures can facilitate the comparative analysis of interventions, e.g. whether a joint replacement was performed by a trainee or a consultant. In this section, implant survival and PROMs are discussed, with the aim of introducing these common methods of measuring orthopaedic outcomes.

1.5.1 Implant survival

Implant survival is an objective outcome measure, defined according to the absence of revision surgery over time. This measure of implant longevity is the principal measure used in the current benchmarking strategies for implant components and is a commonly used measure of surgical performance. It is estimated using methods of survival analysis, which are discussed in Section 1.6 of this introduction.

The NJR defines revision as "any operation where one or more components are added to, removed from, or modified in a joint replacement (The National Joint Registry, 2021a)." The revision of hip and knee replacements poses an increasingly significant clinical and economic burden. It is estimated that by 2030 the number of revision THRs and TKRs performed in the USA will have increased by

137% and 601%, respectively, compared to 2005 (Kurtz et al., 2007). A similar analysis of NJR data has suggested that the volume of revision THRs and TKRs in England and Wales will have increased by 31% and 332%, respectively, between 2012 and 2030 (Patel et al., 2015). It is predicted that the annual cost of revision knee replacement in the USA alone will exceed $13 billion by 2030 (Bhandari et al., 2012). These estimates emphasise the importance of implant survival as a clinical outcome.

It has been suggested that all-cause revision of any component of the construct should be the primary endpoint of any analysis reporting joint replacement survival (Evans et al., 2019a). However, alternative endpoints, such as the selective revision of an individual component within a construct, are often used. Component-specific survival may be useful for manufacturers and medical device regulators, but it assumes that the failure of one component is independent of the construct as a whole. It can therefore give a misleading underestimate of the overall revision burden (Wylde and Blom, 2011).

There are limitations to the use of implant survival as an outcome, particularly when success is defined by survivorship alone. It does not account for variations in the threshold for revision between surgeons and in different healthcare settings. Furthermore, potentially eligible patients with poorly functioning joint replacements may not be offered or may choose not to undergo revision surgery. Thus, this outcome will not capture a proportion of patients with symptomatic joint replacements and/or radiographic evidence of failure (Wylde and Blom, 2011).

In a study using PROMs, Wylde et al. identified that 15% of TKR patients and 6% of THR patients reported severe-extreme persistent postsurgical pain (Wylde et al., 2011). Subsequent results of evidence synthesis suggest that 10-34% of TKR patients and 7-23% of THR patients experience an unfavourable long-term pain outcome (Beswick et al., 2012). Studies that use implant survival alone will fail to capture a proportion of this subgroup of patients who experience severe persistent pain and may feel that their joint replacement has been a failure (Murray and Frost, 1998, Bullens et al., 2001). Therefore, additional patient-centred outcome measures are valuable and should be reported alongside implant survival when available.

1.5.2 Patient-reported outcome measures

Validated patient-reported outcome questionnaires offer a cost-effective, minimally intrusive means of assessing patient-centred outcomes relating to various health-related quality of life (HRQoL) domains. These subjective instruments can be categorised according to whether they assess generic health, disease-specific, or joint-specific measures (Wylde and Blom, 2009). Generic measures are designed to give a global assessment of a patient's HRQoL. In contrast, joint-specific measures are typically designed to be sensitive to the outcomes of joint replacement and reduce the influence of other factors (Murray et al., 2007).

The EuroQol-5-dimension score (EQ-5D) is an example of a generic instrument, which facilitates the assessment of general health status using five domains (mobility, self-care, usual activities, pain and discomfort, and anxiety and depression). The EQ-5D is the most commonly used generic PROM instrument among joint replacement registries (Bohm et al., 2021). Generic instruments are particularly valuable in the economic evaluation of interventions, but their sensitivity can be limited as they can fail to capture HRQoL domains relating to specific interventions.

Disease-specific measures, such as the Western Ontario and McMaster University Osteoarthritis Index (WOMAC), are more responsive to change following hip and knee replacement compared to generic measures such as the SF-36 (Bachmeier et al., 2001). Joint-specific measures, such as the Hip Disability and Osteoarthritis Outcome Score (HOOS) and the Knee Injury and Osteoarthritis Outcome Score (KOOS), have additional advantages in terms of responsiveness compared to the WOMAC (Roos and Toksvig-Larsen, 2003, Nilsdotter et al., 2003). The Oxford Hip Score (OHS) and the Oxford Knee Score (OKS) are additional examples of extensively used joint-specific measures (Murray et al., 2007).

PROMs have been routinely collected by the NHS England PROM programme since 2009 (Devlin et al., 2010). In the context of hip and knee replacement, the NHS uses a combination of generic and joint-specific measures, including the EQ-5D, the OHS, and the OKS. PROMs are not routinely reported by the NJR (The National Joint Registry, 2018), but it is possible to link NHS PROM records

to the NJR. However, complete preoperative and postoperative PROMs are only linkable for a relatively small proportion of NJR records, which is a potential source of selection bias (Sayers et al., 2020b, Liddle et al., 2015). An important limitation of patient-reported outcome questionnaires in general, is their susceptibility to floor and ceiling effects, which relates to the possibility for patients to report minimum or maximum scores (Lim et al., 2015, Harris et al., 2017).

1.6 Methods of survival analysis

1.6.1 Overview

As discussed in Section 1.5, implant survival is an important and commonly used objective measure of joint replacement outcome. Estimates of survival are calculated using statistical methods, which consider the time taken for a well-defined failure event (e.g. revision) to occur and account for censored data that arise due to administrative censoring, incomplete follow-up, or death (Sayers et al., 2018). The basic statistical principles and commonly used methods of survival analysis are discussed in more detail, with particular attention to the methods used in this thesis.

Survival analyses are characterised by the following properties: 1) they measure the occurrence of a clearly defined binary event (e.g. revision), which may only occur in a small proportion of participants during the study period; 2) they measure the length of time for the event to occur among participants who may enter or exit the study at different time points; and 3) they account for censoring due to incomplete observations, such as patients who are lost to follow-up or exit the study due to an unrelated event (e.g. death). These methods of analysis are well suited to joint registry studies, in which patients undergo clearly defined interventions (e.g. primary and revision procedures) at recordable timepoints. The population of interest is predominantly elderly; therefore, in longer-term studies, a relatively high proportion of patients will die with an unrevised prosthesis.

The KM method is the methodological workhorse of the NJR and has been commonly used for the survival analysis of joint replacements since the late 1980s (Kaplan and Meier, 1958, Murray et al., 1993, The National Joint Registry, 2021a). It is a non-parametric method, which estimates the net survival function (or failure function in a one minus KM analysis) (Lambert et al., 2010). In other

words, it estimates the probability that the event of interest has not occurred by a specified timepoint. Confidence intervals for this estimate can be calculated using various methods (Murray et al., 1993, Greenwood, 1926, Peto et al., 1977). The use of KM in the context of joint replacement relies on the assumption that mortality is independent of whether or not implants are revised (i.e. the failure event and death are non-informative).

Competing risk (CR) models are an alternative to KM in the context of joint replacement survival analysis (Fennema and Lubsen, 2010). The use of KM has drawn criticism from proponents of CR models (Lampropoulou-Adamidou et al., 2018, Lacny et al., 2015). However, Sayers et al. use a simple simulation to illustrate the differences between KM and CR methods, which estimate net survival and crude survival, respectively. Sayers emphasises that each method estimates a different quantity, and the appropriate use of each method depends on the application (Sayers et al., 2018).

KM is useful for describing the failure rate of an implant, or comparing the failure rates between two groups, but does not facilitate adjustment for confounding variables. A commonly used method of adjusted survival analysis is the Cox proportional hazards (PH) model (Cox, 1972). PH models estimate the hazard function, which represents the instantaneous risk of an event (e.g. revision) at any time during follow-up. The exponentiated output of a PH model is the hazard ratio (HR), which is the ratio of the hazard functions between two intervention groups. While the hazard function for either intervention group may vary throughout follow-up, PH models assume that the ratio of hazard functions is constant (Sedgwick and Joekes, 2015). In the context of this thesis, the HR represents the risk of revision in one group compared to another group at any time during the study period.

If the PH assumption is not satisfied (i.e. the ratio of hazard functions varies with time) then a HR will either overestimate or underestimate the risk of revision. Royston and Lambert describe how this limitation of the Cox model relates to its semiparametric nature by which, “we make parametric assumptions about the effects of covariates on the hazard function, but not about the shape of the hazard function itself.” This limitation of the Cox model can be overcome by using alternative methods, such as flexible parametric survival modelling (FPM), which uses restricted cubic spline

functions to model the baseline hazard function and account for the time-dependent effects of specified covariates (Royston and Lambert, 2011, Lambert and Royston, 2009b). This method has been used in a number of previous NJR studies (Blom et al., 2020, Blom et al., 2021, Smith et al., 2012b, Hunt et al., 2018).

1.6.2 Rationale for the use of net estimates of failure

Throughout this thesis, unadjusted estimates of implant failure are calculated using the KM method, and the association between surgeon grade and the risk of revision is modelled using Cox regression and/or FPM. These methods estimate net failure and are based solely on the hazard profile of the event of interest (i.e. revision). An alternative approach would be to use competing risk (CR) methods (e.g. the cumulative incidence method, or Fine and Gray models), which produce estimates of crude failure and are based on both the hazard profile of the event of interest (i.e. revision) and the hazard of a competing event (i.e. death). Our use of KM, Cox regression and FPM is based on the understanding that net estimates of failure are favourable in the context of this thesis, in which we investigate the relative risk of revision of hip and knee replacements performed by different groups of surgeons. This notion is explored in detail in two articles on this subject, which are discussed throughout this section (Sayers et al., 2018, Ranstam and Robertsson, 2017).

Estimates of net failure represent the risk of revision in a hypothetical immortal cohort, in which all patients live until they experience revision. Ranstam et al. explain that "Competing risks are simply assumed to be eliminated. An estimate of the 10-year revision risk corresponds to the risk that a patient can expect if he or she lives that long (Ranstam and Robertsson, 2017)." Estimates of crude failure represent the likely number of revision events that will be observed in practice, where a proportion of patients will die during the follow-up period without undergoing revision. Sayers states that such estimates are "a composite of both the failure of implants and the mortality process (Sayers et al., 2018)." A key distinction between these two approaches depends on the assumptions that are made in the censoring process.

The use of KM in the context of joint replacement relies on the assumption that mortality is independent of whether or not implants are revised (i.e. the failure event and death are non-informative). The KM method was developed to account for incomplete observations due to non-informative right censoring, e.g. death (Kaplan and Meier, 1958). Sayers et al. state that "Individuals cease to be at risk of failure but have not failed where the reason that they cease to be at risk of failure is completely independent of the cause of failure… Our belief in this assumption is based on the observation that even when an implant or group of implants fail in a large number of patients, e.g. metal-on-metal, this is not associated with any increase in pathologies, in the short term, such as cancer that in turn may lead to an excess of mortality (Sayers et al., 2018, Smith et al., 2012a)." The Cox method relies on independent censoring, which is less restrictive compared to non-informative censoring, and assumes that censored patients are representative of those under observation (Ranstam and Robertsson, 2017).

Sayers et al. present a simple simulation of implant failure, which compares the estimates of KM and CR models in hypothetical cohorts of mortal and immortal patients. KM and CR produce statistically unbiased estimates of net and crude failure, respectively. However, the simulation demonstrates that the CR estimate is a biased underestimate of net failure (Sayers et al., 2018). In their article entitled 'The Cox model is better than the Fine and Gray model when estimating relative revision risks from arthroplasty register data', Ranstam et al. discuss the relative merits of these two approaches. In a simulated analysis of relative revision risk using Swedish Knee Arthroplasty Register data, they compare the estimates of Cox models with the estimates of a Fine and Gray model. The results of this simulation demonstrate that the Fine and Gray model underestimates the risk of revision in male patients; an effect which is explained by the way the model accounts for the competing risk of death. The authors conclude that the estimates of Fine and Gray models can be misleading when estimating relative revision risks using registry data, and thus recommend using the Cox model in this context (Ranstam and Robertsson, 2017).

Non-competing risk and competing risk methods estimate different quantities (net and crude failure, respectively), and the method used should depend on the application and study design. Crude

estimates of failure are useful for health system resource planning and economic projections, whereas net estimates of failure from non-competing risk methods are preferable when estimating relative revision risks (Sayers et al., 2018). Throughout this thesis, in which we aim to compare the risk of revision for joint replacements performed by trainees compared to joint replacements performed by consultants, estimates of net failure are preferable.

1.7 Joint replacement registries

Joint replacement registries, many of which operate at a national level, are the principal sources of implant survival data. They collect patient, operation, and unit-level data for patients undergoing joint replacement and act as an early warning system for issues relating to patient safety. NJR data are used extensively throughout this thesis. Therefore, an introductory overview of the NJR is given here, along with a brief discussion of the advantages and disadvantages of registry data.

1.7.1 The National Joint Registry

The NJR was established by the Department of Health in 2002 on the recommendation of the Royal College of Surgeons of England, in response to an investigation of the high early failure rate of the 3M Cemented Capital Hip (Muirhead-Allwood, 1998, Riordan et al., 1998). The report recommended that the primary objectives of a UK national joint registry should be to detect poorly performing implants at an early stage, improve the quality of practice and outcomes, and restrict the uptake of unevaluated prostheses (The Royal College of Surgeons of England, 2001).

The NJR began data collection in England and Wales in April 2003 with the following goals: to monitor outcomes at brand, hospital, and surgeon-levels; to support quality and cost-effectiveness; and to provide post-market implant surveillance. The NJR now collects data for England, Wales, Northern Ireland, The Isle of Man, and The States of Guernsey. It has been managed by the Healthcare Quality Improvement Partnership since 2008 and submission of records has been mandatory for eligible procedures since 2011.

The NJR links primary procedures to episodes of revision and the date of death. This process of linkage produces time-to-event data that facilitate survival analysis. Outputs of the NJR include an annual report of implant performance, surgeon and hospital-level reports, outlier analyses, and peer-reviewed publications based on external research.

1.7.2 Strengths and limitations of registry data

National registry studies have several advantages (Thygesen and Ersboll, 2014). Registry data already exist, which is economical in terms of time and financial cost. The most recent NJR annual report contains records of over 2.6 million linked primary joint replacements with up to 17.8 years of follow-up (The National Joint Registry, 2021a). Such large datasets give statistical power to identify rare outcomes over a long period of follow-up. The range of patient, operation, and unit-level data recorded by the NJR makes it possible to adjust for multiple confounding variables. Mandatory data submission reduces the proportion of missing records. The quality of NJR data has been automatically audited on a rolling monthly basis since 2019. In August 2021, over 97% of primary hip and knee replacements and 95% of revision procedures were recorded in the NJR (The National Joint Registry, 2021a). This high level of compliance means that data are largely complete for the target population, which limits selection bias and increases the generalisability of results.

Despite these strengths, registry studies have several limitations. Registry data are observational and can identify associations but not causation. Data quality is limited by compliance and the accuracy of input, which can result in missing records. Systematically missing data, where the absence of a record is “not at random” and related to its value (e.g. revision), are a potential source of bias that may lead to an underestimation of the risk of revision (Little and Rubin, 1987). However, missing NJR records are generally considered “missing completely at random” (i.e. there is no difference between records with missing data and records with complete data). The pattern of missingness of NJR data is discussed in more detail in Section 1.7.3. An additional limitation is that while the process of linking primary and revision procedures is audited, there is no robust system in place to audit the quality of confounding data within the NJR (Konan and Haddad, 2013). Registry studies offer a valuable

contribution to the literature and have had significant implications for clinical practice. However, caution must be taken to interpret findings in the context of these limitations.

1.7.3 Missing data

Missing data are commonplace in epidemiological studies and improper handling can impair the validity of results. Records may be completely missing, missing information on a specific confounding variable, or missing outcome data (e.g. due to a failure to link primary and revision procedures). The risk of bias due to missing data depends on the pattern of missingness, which is described using commonly used nomenclature: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Sterne et al., 2009). Data which are MNAR, where the missingness of a record is associated with its value, are a potential source of selection bias. For example, records that are systematically missing due to a high failure rate will lead to a biased underestimation of the risk of revision. Data that are MCAR, or MAR are less problematic as the missingness of records is not associated with observed values.

The pattern of missingness of NJR data is discussed in the introductory section of the NJR annual report, where the authors state the following: "Analysis of data which is missing in either a random fashion (MCAR) or random within known strata (MAR), e.g. method of fixation, is known to yield unbiased results. We believe that a coordinated systematic agreement of individuals across the registry to under-report the failure of a specific implant is exceedingly unlikely – We believe that missing data within the registry can be considered MCAR. We propose that this missing data mechanism will ensure that the quality assurance process of implants entered into the registry, consultant and units is statistically valid (The National Joint Registry, 2021a)."

Body mass index (BMI) is missing in a large proportion of NJR records. BMI was not recorded on the NJR Minimum Data Set (MDS) Form Version 1. It has been reported that approximately 40% of patients did not have a BMI recorded in the NJR in 2009, compared to approximately 18% in 2016 (Sayers et al., 2020b). Previous NJR studies have dealt with missing BMI data using different approaches, including complete-case primary analysis without BMI as a confounding variable (Sayers

et al., 2020a), complete-case sensitivity analysis with BMI as a confounding variable (Evans et al., 2020), and completion of missing values using multiple imputation (Liddle et al., 2014). Due to the large fraction of records with missing values, BMI was not included as a confounding variable in the analyses used in this thesis.

The preferred method of dealing with missing data throughout the studies included in this thesis was to use complete-case analysis with the exclusion of records with missing data in confounding variable fields used in the statistical models. This is based on the assumption that the pattern of missingness of NJR data is independent of the primary exposure and the outcome. Considering the large datasets, the small fraction of incomplete cases (at least 95% of all eligible records were included in the analyses), and the assumed pattern of missingness, any potential improvement in efficiency from using multiple imputation compared to complete-case analysis is likely to be negligible (White and Carlin, 2010, Sayers et al., 2020a).

1.8 Review of the existing literature

The following narrative review gives an overview of the existing literature on the outcomes of trainee-performed hip and knee replacements. The primary outcome of interest of this thesis is implant survival, which is systematically reviewed in Chapter 2. However, a range of outcomes are discussed here in order to give a broad overview of our current understanding of this subject.

1.8.1 Total hip replacement

1.8.1.1 Implant survival & revision rate

Palan and colleagues conducted a multicentre cohort study of 1,501 primary THRs performed in seven units between 1999 and 2002. Cases were categorised according to surgeon grade (trainee or consultant). They found no significant difference in the revision rate at 5 years for THRs performed by trainees (1.3%) and consultants (1.4%). Limitations of this study include the lack of data on supervision and the absence of an adjusted survival analysis (Palan et al., 2009).

In a smaller observational study, Reidy et al. included 879 patients who underwent primary THR in three centres between 2003 and 2004. Junior and senior trainees performed 138 (15.7%) and 148 (16.8%) procedures, respectively. Trainees were directly supervised by a consultant in 85% of cases, with the highest rates of supervision observed in the junior trainee group (93.5%). The net survival of THRs at 10 years was 95.9% for consultants, 96.7% for junior trainees, and 97.8% for senior trainees. Using an adjusted Cox analysis, the authors found no evidence of an association between surgeon grade or supervision and implant survival up to 10 years (Reidy et al., 2016).

Inglis et al. used New Zealand Joint Registry (NZJR) data for 35,415 primary THRs with up to 6 years of follow-up. Records were categorised according to whether the operation was performed by a consultant (n=30,344), a supervised trainee (n=2,982), or an unsupervised trainee (n=1,067). The authors found no significant difference between the revision rate per 100 component years for THRs performed by consultants (0.75, 95% CI 0.68 to 0.82) compared to supervised trainees (0.97, 95% CI 0.72 to 1.28), and unsupervised trainees (0.70, 95% CI 0.36 to 1.22). There was no significant difference in the indication for revision (including dislocation, infection, periprosthetic fracture, loosening and pain) between the three groups. Strengths of this study include the use of generalisable registry data and the inclusion of PROMs data, which is discussed in the following subsection. However, the author's description of their statistical methods is limited, and it is not clear if they used a formal adjusted survival analysis. Furthermore, the period of follow-up gives no insight into the survival of trainee-performed THRs beyond 6 years. Excluding this thesis, this is the only other registry study that has examined the association between trainee-performed THR and revision, which leaves scope for further investigation (Inglis et al., 2013).

1.8.1.2 Patient-reported outcome measures

Inglis et al. used NZJR data to analyse the 6-month postoperative OHS according to surgeon grade and supervision. They found significant differences in the OHS for consultants (OHS 40.70, 95% CI 40.51 to 40.88) compared to supervised trainees (OHS 38.95, 95% CI 38.29 to 39.62; $p<0.001$), and unsupervised trainees (OHS 38.27, 95% CI 37.10 to 39.44; $p=0.001$). The authors acknowledge that

their study is limited by the absence of preoperative scores and question the significance of these findings with reference to a minimal clinically important difference (MCID) in OHS of 3 to 5 points (Inglis et al., 2013, Murray et al., 2007). Preoperative scores and longer-term functional outcomes have been reported by other authors (Palan et al., 2009, Reidy et al., 2016, Moran et al., 2004).

A SHAR study of 6,713 cases examined the association between surgical experience and a range of PROMs at preoperative assessment and 1-year follow-up (Jolbäck et al., 2018). Surgeons were categorised into four groups: orthopaedic trainees, or specialists with <8 years, 8-15 years, or >15 years clinical practice after specialist certification. Linear regression models were adjusted for several patient-level factors. The authors found no association between surgeon experience and EQ-5D, EQ-Visual Analogue Scale (VAS), or Pain VAS. Patients operated on by trainees reported lower satisfaction on a VAS at 1-year compared to patients operated on by the most experienced specialists (those with >15 years practice).

The functional outcomes of trainee- compared to consultant-performed THRs are summarised in a recent systematic review and meta-analysis of observational studies. Singh et al. conducted separate meta-analyses according to the quality of included studies. In their overall analysis (three studies), there was no significant difference in HHS between consultants and trainees at 6 months follow-up. However, in their analysis of high-quality studies (two studies), consultant-performed THR was associated with a small improvement in in HHS compared to trainee-performed THR (Singh et al., 2019).

1.8.1.3 Complication rates

National joint registries, such as the NJR, do not typically record postoperative complications other than revision. Therefore, our understanding of the incidence of postoperative complications is based on cohort studies and case series. In their systematic review and meta-analysis on this subject, Singh et al. included six studies for dislocation and five studies for deep infection. They found no evidence of an association between trainee-performed THR and an increased risk of dislocation, or deep infection compared to consultant-performed THR (Singh et al., 2019).

1.8.1.4 Other outcomes

Additional outcomes explored in the literature include operative duration and radiological outcomes. Palan et al. found a significant difference in mean operation time for trainees compared to consultants (104 mins; range 40 to 240 vs. 85 mins; range 28 to 254; $p<0.001$) (Palan et al., 2009). Similar findings were observed by Weber et al. who found that the mean operative duration was 9 minutes longer for trainees compared to senior surgeons (78.1 mins; SD 25.4 vs. 69.3 mins; SD 2.8; $p<0.001$) (Weber et al., 2017). The finding that trainees take longer than consultants to perform a THR is consistent throughout the literature (Singh et al., 2019).

The radiological outcomes of trainee-performed THR are poorly understood. Moran et al. conducted a comparative analysis of 228 matched trainee and consultant radiographs. Trainees demonstrated inconsistent acetabular component alignment and positioned the cup with less anteversion compared to consultants (Moran et al., 2004). A more recent study of THRs performed in the context of hip fracture, reported that supervised trainees achieved equivalent radiological outcomes in terms of acetabular inclination, leg length, and cementation quality (MacDonald et al., 2020).

1.8.2 Total knee replacement

1.8.2.1 Implant survival & revision rate

In a multicentre observational study, Faulkner et al. included 686 patients who underwent primary TKR in three units between 2003 and 2004. Trainees performed 236 TKRs (34.4%) and 450 (65.6%) were performed by consultants. The level of supervision was not reported. The net survival of TKRs at 10 years was 94.6% for consultants and 96.3% for trainees (confidence intervals not available). Using an adjusted Cox regression analysis, the authors found no evidence of an association between surgeon grade and implant survival up to 10 years. This study also reported PROMs and complication rates but is limited by the low number of patients included (Faulkner et al., 2018).

Storey et al. conducted a registry study using NZJR data, including 79,671 primary TKRs with up to 9 years of follow-up. Cases were categorised according to whether the operation was performed by a consultant (n=71,519), a supervised senior trainee (n=5,073), a supervised junior trainee (n=1,770), or

an unsupervised senior trainee (n=1,309). The authors found no significant difference between the revision rate per 100 component years for TKRs performed by consultants (0.49, 95% CI 0.47 to 0.51) compared to supervised senior trainees (0.48, 95% CI 0.40 to 0.57), supervised junior trainees (0.22, 95% CI 0.14 to 0.32), and unsupervised senior trainees (0.46, 95% CI 0.31 to 0.66) (Storey et al., 2018).

Strengths of this study include the large number of patients, the use of generalisable national registry data, and the inclusion of PROMs data, which is discussed in the following subsection. However, the author's description of their statistical methodology is limited and the indication for revision was not reported. Other than this thesis, this is the only other registry study that has investigated the association between trainee-performed knee replacement and revision, which leaves scope for further investigation with robust statistical methodology.

1.8.2.2 Patient-reported outcome measures

In their NZJR study, Storey et al. analysed the 6-month postoperative OKS for 22,155 patients who had undergone TKR. They found a significant difference in the mean OKS for consultants compared to supervised senior trainees (Consultant: 37.70; SD 8.0 vs. Trainee: 36.30; SD 8.4; $p<0.001$) and supervised junior trainees (Consultant: 37.70; SD 8.0 vs. Trainee: 36.80; SD 7.9; $p=0.02$). However, the authors acknowledge that their study is limited by the lack of preoperative scores and question the significance of these findings with reference to a minimal clinically important difference (MCID) in OKS of 1 to 5 points (Storey et al., 2018, Clement et al., 2014).

The functional outcomes of trainee- compared to consultant-performed TKRs are summarised in a recent systematic review and meta-analysis of 92,309 knee replacements, of which 11,654 were performed by trainees (Madanipour et al., 2021). Madanipour et al. included seven studies in their meta-analysis of PROMs at 6-18 months follow-up. They found evidence of an association between consultant-performed surgery and superior patient-reported outcome scores compared to trainees (weighted mean difference: -1.26; 95% CI -1.44 to -1.07; $p<0.01$). However, multiple different

PROM instruments (OKS, KSS, AKSK, EQ-5D, WOMAC) were compared on the same scale and preoperative scores were not included. The clinical significance of this finding is unclear.

1.8.2.3 Complication rates

In their recent systematic review on this subject, Madanipour et al. found no evidence of an association between trainee-performed TKR and an increased risk of neurological deficit, VTE, or blood transfusion compared to consultant-performed TKR (Singh et al., 2019). However, trainee-performed surgery was associated with lower odds of infection compared to consultant-performed surgery (OR 0.75; 95% CI 0.58 to 0.97; p=0.03) (Madanipour et al., 2021).

1.8.2.4 Other outcomes

In an analysis of operative duration for 76,472 TKRs, Storey et al. found that consultants were significantly faster than trainees when performing TKR (82 mins vs. 105 mins) (Storey et al., 2018). In contrast, a recent meta-analysis, which did not include the aforementioned NZJR study, found no evidence of an association between surgeon grade and operative duration (Madanipour et al., 2021). Studies of radiologically evaluated implant alignment have suggested that trainees can achieve comparable TKR alignment to consultants using both traditional techniques and computer navigated technology (Mahaluxmivala et al., 2001, Theelen et al., 2018, Khakha et al., 2015).

1.8.3 Unicompartmental knee replacement

1.8.3.1 Overview

The published literature on the outcomes of trainee-performed UKR is limited. Bottomley et al. conducted a single-centre cohort study of 1,084 UKRs, of which 673 (62.1%) were performed by trainees. Trainees were supervised by a scrubbed consultant in 48% of trainee-performed cases. They reported no significant difference in implant survival between the groups, with 9-year cumulative survival estimates of 93.9% (95% CI 90.2 to 97.6) and 93.0% (95% CI 90.3 to 95.7) for consultants and trainees, respectively (Bottomley et al., 2016).

Storey et al. included 8,854 UKRs, of which 304 (3.4%) were performed by trainees. They found no significant difference between the revision rate per 100 component years for UKRs performed by consultants (1.19, 95% CI 1.10 to 1.28) compared to supervised senior trainees (1.20, 95% CI 0.74 to 1.86). Furthermore, supervised senior trainees achieved comparable functional outcomes (OKS) to consultants at 6 months. The authors state that they had insufficient data for meaningful interpretation of their results for junior and unsupervised trainees (Storey et al., 2018).

1.9 Aims and objectives

1.9.1 Aims

The aim of this thesis is to investigate the association between surgeon grade, the supervision of trainees and implant survival following primary THR, TKR, and UKR. An additional aim was to explore the association between surgeon grade and the number of THRs, TKRs, and UKRs revised for different indications. Surgeon grade was treated as a binary exposure variable, where the operating surgeon was either a trainee or a consultant.

1.9.2 Objectives

1) To conduct a preliminary study using methods of evidence synthesis to explore the existing literature on the association between surgeon grade and implant survival outcomes following primary hip and knee replacement.
2) To use NJR data for England and Wales to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following primary THR, TKR, and UKR.
3) To use local single-centre observational data to investigate the association between surgeon grade and a range of clinical, functional, and radiological outcome measures.

Chapter 2 Association between surgeon grade and implant survival following hip and knee replacement: a systematic review and meta-analysis

2.1 Overview

Building on the narrative review of exiting literature in Chapter 1, the aim of this chapter was to conduct evidence synthesis using openly available published data to explore the association between surgeon grade (trainee vs. consultant) and implant survival following hip and knee replacement. This chapter establishes the baseline of evidence on this subject prior to this thesis. This article has been published in BMJ Open (Fowler et al., 2021) and included in this thesis according to the University's guidance on the integration of publications as chapters within a dissertation.

2.1.1 Contributors

TF, AB, AS and MW conceived and designed the study; TF and AA independently screened the articles and performed the data extraction; TF and AS were responsible for data analysis; all authors were responsible for interpreting the data; TF drafted the manuscript; AB, AA, AS and MW revised the article critically for important intellectual content; all authors reviewed the final version of the manuscript and gave approval for submission for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

2.1.2 Declaration of conflicts of interest

TF has no conflicts of interest to declare. AB and MW report a contract held by the University of Bristol for membership of HQIP Lot 2 Statistical Analysis team for the National Joint Registry, and royalties from Taylor & Francis. AA reports a part-time employment contract for Virti Ltd, and royalties from BPP Holdings Ltd. MW also reports an independently conducted research grant held by the University of Bristol with CeramTec, teaching payments from Heraeus, and research committee membership on the British Orthopaedic Association and the British Hip Society.

2.2 Abstract

2.2.1 Background

The association between the surgeon grade and the survival of hip and knee replacements is poorly understood and there is limited evidence in the existing literature. Our aim was to investigate the association between surgeon grade (trainee vs. consultant) and implant survival following primary hip and knee replacement.

2.2.2 Methods

We searched MEDLINE and Embase from inception until 6 October 2021. The participants were adult patients undergoing either a primary hip or knee replacement, predominantly for osteoarthritis. The intervention was whether the surgeon recorded as performing the procedure was a trainee or not. The primary outcome was net implant survival reported as a Kaplan-Meier survival estimate, and the secondary outcome was crude revision rate. Both outcomes were reported according to surgeon grade. Two independent reviewers screened 1,812 citations and identified 57 articles for full text review. Data were independently extracted in duplicate and fixed effects models were used for meta-analysis.

2.2.3 Results

Nine cohort studies capturing 4,066 total hip replacements (THRs), 936 total knee replacements (TKRs), and 1,357 unicompartmental knee replacements (UKRs) were included (five THR studies, two TKR studies, and two UKR studies). The pooled net implant survival estimates for THRs at 5 years were 97.9% (95% CI 96.6 to 99.2) for trainees and 98.1% (95% CI 97.1 to 99.2) for consultants. The relative risk of revision of THRs at 5 and 10 years was 0.88 (95% CI 0.46 to 1.70) and 0.68 (95% CI 0.37 to 1.26), respectively. For TKRs, the net implant survival estimates at 10 years were 96.2% (95% CI 94.0 to 98.4) for trainees and 95.1% (95% CI 93.0 to 97.2) for consultants. We report a narrative summary of UKR outcomes.

2.2.4 Conclusions

There is no strong evidence in the existing literature that trainee surgeons have worse outcomes compared to consultants, in terms of the net survival or crude revision rate of hip and knee replacements at 5 to 10 years follow-up. These findings are limited by the quality of the existing published data and are applicable to countries with established orthopaedic training programmes.

2.3 Background

Hip and knee replacements are effective surgical interventions for the treatment of end-stage degenerative conditions of the hip and knee (Ferguson et al., 2018, Price et al., 2018). More than 200,000 are performed per year in the United Kingdom alone (The National Joint Registry, 2020). These procedures are performed by surgeons at various stages in their training, with varying levels of senior supervision. Contemporary training practices must ensure a balance between protecting development opportunities for the next generation of surgeons, while limiting the exposure of patients to unnecessary risk during the training process.

Implant survival, which is determined by the absence of revision surgery, is an important and commonly used measure of surgical performance (Evans et al., 2019a, Evans et al., 2019b). Net survival estimates are calculated using statistical methods of survival analysis (e.g. Kaplan-Meier analysis), which look at time to a defined failure 'event' (e.g. revision) and account for censored data that arise due to incomplete follow-up, or death (Sayers et al., 2018). Another commonly reported metric is crude revision rate, which is defined as the observed number of failure events in a specified period of time.

The survival of hip and knee replacements according to surgeon grade is poorly understood. Higher rates of complications and longer operative times have been identified in orthopaedic procedures performed by trainees (Hedlundh et al., 1996, Schoenfeld et al., 2013). Radiographic studies comparing trainee and consultant joint replacement have identified differences in acetabular anteversion (Moran et al., 2004), hip centre of rotation (Kim et al., 2017), and various measures of knee replacement component positioning (Kazarian et al., 2019). However, the relative impact of these findings on implant survival has not been established. It has been suggested that when trainees are appropriately supervised, they can obtain comparable functional outcomes and implant survivorship to their consultant colleagues when performing total hip replacement (THR) (Palan et al., 2009, Reidy et al., 2016, Beattie et al., 2018), total knee replacement (TKR) (Faulkner et al., 2018), and unicompartmental knee replacement (UKR) (Bottomley et al., 2016).

The aim of this study was to conduct a systematic review and meta-analysis using the existing literature on the association between surgeon grade (trainee vs. consultant) and implant survival outcomes in hip and knee replacement surgery. We aimed to answer the question: do trainees achieve equivalent implant survival outcomes to consultants when performing primary hip and knee replacement?

2.4 Methods

This review was conducted using methods described in the Cochrane Handbook for Systematic Reviews of Interventions, with reporting in accordance with the Meta-analyses Of Observational Studies in Epidemiology (MOOSE) checklist, which is included in Appendix 5 (Higgins et al., 2008, Stroup et al., 2000). The study was registered with the PROSPERO database at inception (CRD42019150494).

2.4.1 Data sources and search strategy

We searched for cohort studies reporting implant survival estimates and/or revision rates for hip and knee replacements, according to surgeon grade. Separate searches were performed for hips and knees. We conducted searches of MEDLINE and Embase from inception until 6 October 2021. Searches used keywords and Medical Subject Headings (MeSH) terms relating to hip and knee replacement, implant survival, revision surgery and surgeon grade (see Appendix 6 for full search strategy). There were no language restrictions. Titles and abstracts of potentially relevant non-English language citations were translated. We manually screened the bibliographies of full text articles and used Web of Science citation tracking to identify additional relevant studies.

2.4.2 Eligibility criteria

We included studies if they involved predominantly unselected adult patients (≥ 18 years old) undergoing primary hip or knee replacement (including THR, TKR, UKR and hip resurfacing), predominantly for the treatment of osteoarthritis. Included articles needed to report the primary and/or secondary outcome measure for two different groups of surgeons defined according to their grade

(e.g. trainee vs. consultant). We defined a minimum follow-up of 5 years and articles that did not clearly define the length of follow-up were excluded. For example, we excluded studies reporting the revision rate 'per 100 component years', as these did not explicitly define the length of follow-up. We excluded studies in which the index operation was performed prior to 1990; thereby, including studies that are representative of contemporary training practices, but also allowing for inclusion of studies reporting in excess of 30 years of follow-up (Appendix 7).

2.4.3 Primary exposure

The primary exposure was whether the surgeon recorded as performing the procedure was a trainee or not. Surgeon grade is a measure of the designated level of surgical experience and seniority, which we considered to be a binary variable: either 'trainee', or 'consultant'. Consultant surgeons have completed their formal training in orthopaedic surgery and have been appointed to a senior position in which they can practice independently and supervise trainee surgeons. The term 'consultant' is used synonymously with 'attending surgeon' in many healthcare settings including the USA. Additional terms used to describe this variable were deemed eligible during screening (e.g. Trainee: registrar; resident; junior/young surgeon; fellow. Consultant: attending; senior surgeon; trainer).

2.4.4 Outcome measures

The primary outcome was net implant survival, reported as a Kaplan-Meier survival estimate. The secondary outcome measure was crude revision rate, which was defined as the observed number of revision events in a specified period of time.

2.4.5 Screening and data extraction

Two authors (TF and AA) independently screened all titles and abstracts of journal articles using Rayyan (Rayyan QCRI, Doha) (Ouzzani et al., 2016). Studies were initially screened for relevance according to information contained within the title and abstract. Cases of disagreement were resolved through re-review and consensus. Full texts of potentially relevant studies were reviewed in detail and disagreements on final inclusion were resolved through discussion with a senior author (MW).

Specific indications for exclusion were documented following full text review (Figure 2 and Appendix 8).

Data were extracted in duplicate using a standardised proforma. We recorded data on the following: healthcare setting, study period, implant type, age, sex, indication, level of supervision, crude revision rate, and net implant survival estimates (including confidence intervals [CIs]). Life tables were reviewed, and estimates were extracted for all available 5-year intervals of follow-up. Discrepancies in data collection were resolved through re-review and consensus. Where survival estimates, CIs and revision rates were incompletely reported, we contacted corresponding authors to request missing data.

2.4.6 Risk of bias and quality of evidence assessment

The risk of bias was assessed using the Cochrane ROBINS-I tool for the risk of bias in non-randomised cohort studies (Sterne et al., 2016). We assessed the quality of evidence for each outcome using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, which considers the imprecision, inconsistency, indirectness, and risk of bias in a body of evidence (Atkins et al., 2004).

2.4.7 Statistical analysis

Statistical analysis was performed using Stata (Version SE 15.1; StataCorp LP, USA). For the primary outcome measure of net implant survival, we performed separate meta-analyses for each implant type, by surgeon grade and length of follow-up. We pooled survival estimates, assuming that survivorship approximated risk, with fixed effects meta-analysis weighting each study on the overall pooled estimate according to its standard error, which was calculated from published CIs; an established method for the meta-analysis of implant survival estimates described by Evans et al. (Evans et al., 2019a, Evans et al., 2019b). The effect size (survival) for trainees and consultants was compared using a Wald test. For the secondary outcome measure, we derived and meta-analysed the relative risk (RR) of revision for each implant type by surgeon grade and length of follow-up. We used a fixed effects model using the Mantel-Haenszel method (Mantel and Haenszel, 1959). Heterogeneity was

assessed with chi-squared tests, with I^2 used to quantify inconsistency (Higgins and Thompson, 2002). Publication bias was assessed by inspecting funnel plot symmetry (Egger et al., 1997).

2.4.8 Patient and public involvement

There was no direct patient or public involvement in the design or conduct of this review.

2.5 Results

Separate searches for hip and knee replacements identified 1,178 and 634 articles, respectively. After removal of duplicates and abstract screening, 30 hip papers and 27 knee papers remained. Through review of full text articles, we identified five hip and four knee studies eligible for inclusion. This process of review is summarised as a flow diagram in Figure 2 and the characteristics of included studies are summarised in Table 1. Six studies were conducted in the UK, with the remaining three studies originating from France, Switzerland, and Japan.

Figure 2 - Study flow diagram (Chapter 2).

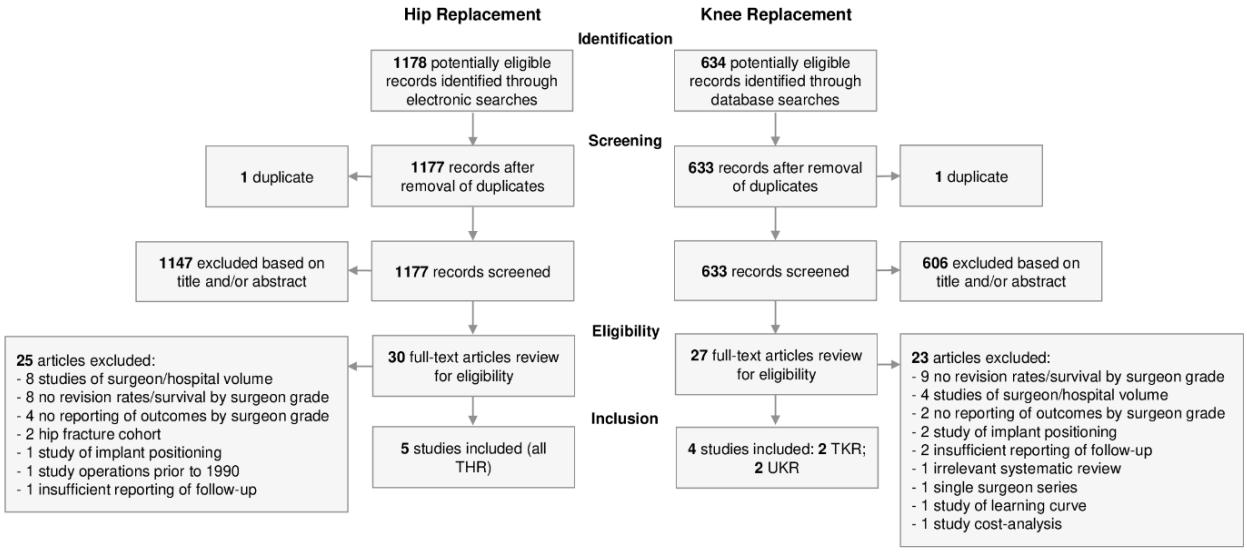


Table 1 - Characteristics of included studies.

Source, Year	Country	Study period	Study design	Implant	Surgeon grade terminology (exposure)	Follow-up (years)	Number of cases (trainee)	Implant brand (stem/cup if hip)	Sex (% female)	Mean age (SD or range)	Indication (% OA)	Supervision reported	Survival analysis	Revision rates reported	ROBINS-I overall risk of bias†
Hasegawa, 2015	Japan	2006-10	PC	THR	Trainee vs. instructor	5	483 (259)	Multiple	-	61.3 (SD 11.6)	-	No	Yes	No	Serious
Jain, 2018	UK	2005-12	RC	THR	Trainee vs. consultant	5, 10	1082 (348)	Corail/multiple	61.3	69.2 (21-94)	91.0	No	Yes (Add.)	Yes	Moderate
Muller, 2013	Switzerland	2005-06	RC	THR	Junior vs. senior	5	130 (43)	Quadra-H /Versafit-CC	52.0	64 (SD 12.36)	86.0	No	Yes	Yes	Serious
Palan, 2009	UK	1999-02	RC	THR	Trainee vs. consultant trainer	5	1501 (528)	Exeter/multiple	-	68.4 (21-94)	-	No	No	Yes	Moderate
Reidy, 2016	UK	2003-04	RC	THR	Trainee vs. consultant	10	870 (286)	Multiple	60.5	69.5 (37-94)	94.8	Yes	Yes (no CIs)	Yes	Moderate
Faulkner, 2018	UK	2003-04	RC	TKR	Trainee vs. consultant	5, 10	686 (236)	Multiple	-	69.9 (30-94)	93.1	No	Yes (Add.)	Yes	Moderate
Hernigou, 2009	France	1990-95	RC	TKR	Young (<30) vs. senior	10, 15	250 (150)	Ceraver Hermes	69.7	73 (46-88)	-	No	Yes	No	Serious
Bottomley, 2016	UK	1998-08	RC	UKR	Trainee vs. consultant	10	1084 (673)	Oxford	51.4	66.5 (SD 9.6)	100	Yes	Yes	Yes	Moderate
Alvand, 2021	UK	2009-15	RC	UKR	Trainee vs. consultant	5	273 (118)	Oxford	49.5	67.8 (SD 10.1)	98.2	Yes	No	Yes	Moderate

† See Appendix 9 for full risk of bias assessment. Add., additional data provided by author; OA, Osteoarthritis; PC, prospective cohort; RC, retrospective cohort; ROBINS-I, Risk of Bias in Non-randomised Studies – of Interventions; THR, total hip replacement; TKR, total knee replacement; UKR, unicompartmental knee replacement.

2.5.1 Risk of bias assessment

Appendix 9 provides a summary of the ROBINS-I assessment, which indicates a moderate to severe risk of bias in all studies. Funnel plot asymmetry and statistical tests for funnel plot asymmetry as a means of assessing publication bias were not applicable due to the small number of studies (Sterne et al., 2011).

2.5.2 Hip replacement

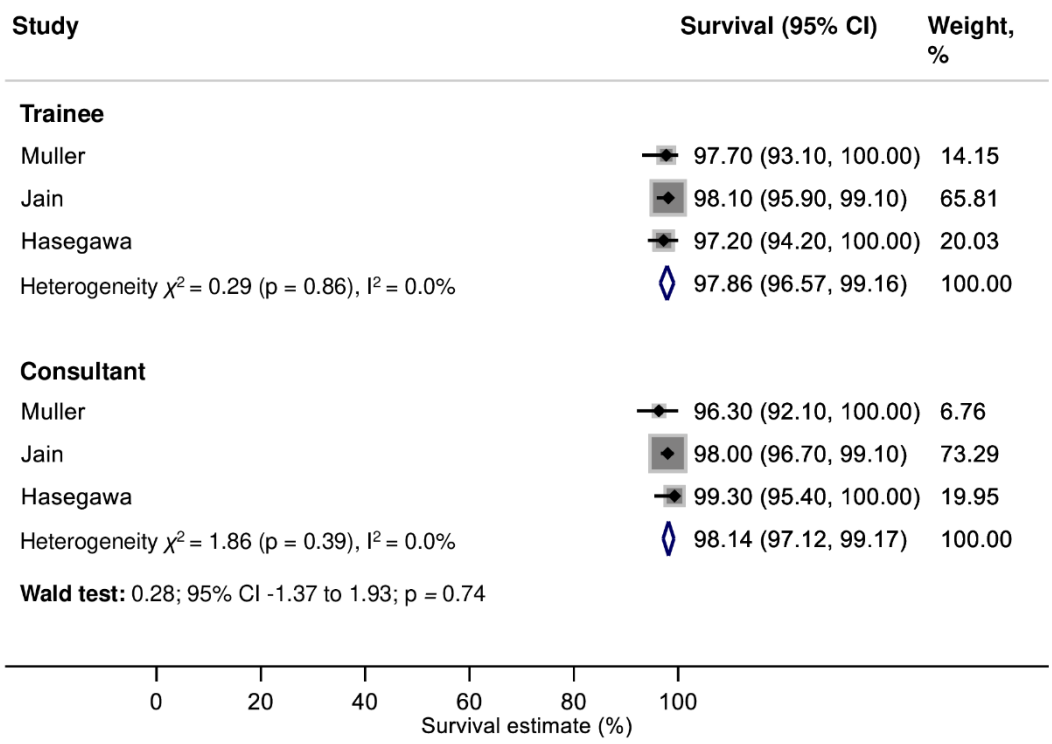
The five included hip studies represent 1,464 THRs performed by trainees and 2,602 THRs performed by consultants, with follow-up ranging from 5 to 10 years. Four studies were retrospective cohort studies (Reidy et al., 2016, Muller et al., 2014, Jain et al., 2018, Palan et al., 2009); one was a non-randomised prospective cohort study (Hasegawa and Amano, 2015). No articles on hip resurfacing met the inclusion criteria. One author provided additional unpublished data in the form of net survival estimates (Jain et al., 2018). Reidy et al. reported survival estimates, but no CIs (Reidy et al., 2016). Net survival estimates and corresponding CIs were thus extracted from three studies at 5 years and one study at 10 years. Crude revision rates were reported in three studies at 5 years and two studies at 10 years.

2.5.3 Primary outcome: Net implant survival (THR)

Meta-analysis showed net survivorship of 97.9% (95% CI 96.6 to 99.2) at 5 years for THRs performed by trainees, compared to 98.1% (95% CI 97.1 to 99.2) for THRs performed by consultants (Figure 3). There was no strong evidence of an association between surgeon grade and net implant survival at this interval of follow-up (Wald test: p=0.74).

Meta-analysis was not possible for the 10-year data given the availability of only one study for this timepoint. In a cohort of 1,082 reverse hybrid THRs, Jain et al. demonstrated overall 97.2% implant survival at 10 years. Additional data provided by the author indicate that they found no evidence of a difference in implant survival according to surgeon grade (Trainee: 98.1%, 95% CI 95.9 to 99.1; Consultant: 96.7%, 95% CI 94.7 to 97.9) (Jain et al., 2018).

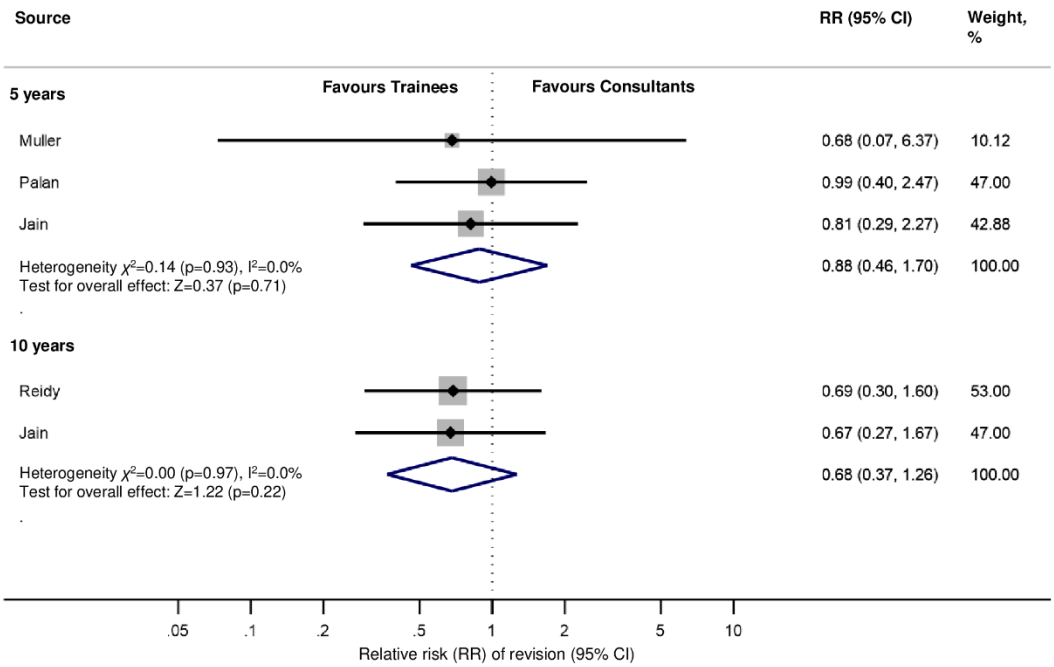
Figure 3 - Meta-analysis of net implant survival of THRs at 5 years according to surgeon grade.



2.5.4 Secondary outcome: Crude revision rate (THR)

Meta-analysis showed no strong evidence of an association between surgeon grade and the crude revision rate at 5, or 10 years. The RR of revision at 5 and 10 years was 0.88 (95% CI 0.46 to 1.70) and 0.68 (95% CI 0.37 to 1.26), respectively (Figure 4).

Figure 4 - Meta-analysis of the relative risk of revision of THRs at 5 and 10 years according to surgeon grade.



2.5.5 Knee replacement

The four knee studies represent 1,177 knee replacements (TKR n=386; UKR n=791) performed by trainees and 1,116 knee replacements (TKR n=550; UKR n=566) performed by consultants, with follow-up ranging from 5 to 15 years. All four were retrospective cohort studies (Bottomley et al., 2016, Faulkner et al., 2018, Hernigou et al., 2009, Alvand et al., 2021). Two studies reported on TKRs (Hernigou et al., 2009, Faulkner et al., 2018), and two studies reported on UKRs (Bottomley et al., 2016, Alvand et al., 2021).

With regards to the two TKR studies, Faulkner et al. provided additional unpublished survival data from which we calculated corresponding CIs for their published survival estimates (Faulkner et al.,

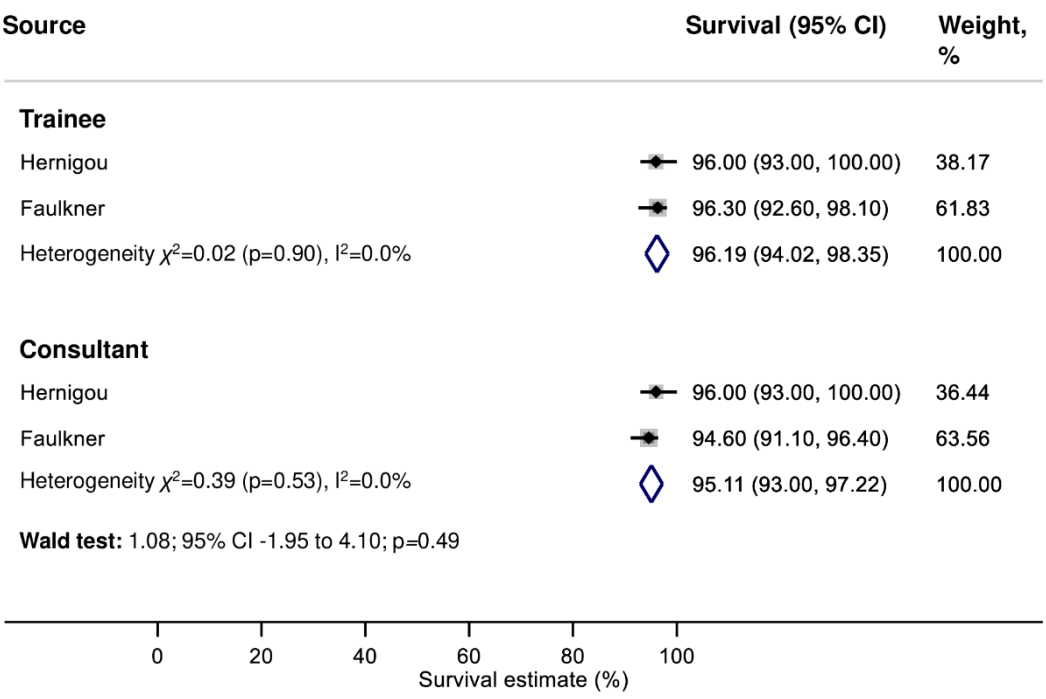
2018). Net survival estimates and CIs were thus extracted from both TKR studies at 10 years, which permitted meta-analysis of this primary outcome measure. Crude revision rates were only available from one TKR study at each 5-year interval of follow-up.

With regards to the two UKR papers, net survival estimates were only available from one study (Bottomley et al., 2016). Crude revision rates were available from one study at 5 years and one study at 10 years (Alvand et al., 2021, Bottomley et al., 2016). Meta-analysis was not feasible; thus we provide a narrative summary of UKR outcomes.

2.5.6 Primary outcome: Net implant survival (TKR)

Meta-analysis showed net survivorship of 96.2% (95% CI 94.0 to 98.4) at 10 years for TKRs performed by trainees, compared to 95.1% (95% CI 93.0 to 97.2) for TKRs performed by consultants (Figure 5). There was no strong evidence of an association between surgeon grade and net implant survival at this interval of follow-up (Wald test: p=0.49).

Figure 5 - Meta-analysis of net implant survival of TKRs at 10 years according to surgeon grade.



2.5.7 Secondary outcome: Crude revision rate (TKR)

Two studies reported crude revision rates according to surgeon grade; however, with data from only one study available at each interval of follow-up, meta-analysis was not feasible. Instead, we provide a narrative summary. Faulkner et al. provided additional unpublished data, which indicated crude revision rates at 5 years for trainees and consultants of 2.1% and 4.4%, respectively (Faulkner et al., 2018). This rises to 3.4% (trainees) and 5.8% (consultants) at 10 years. These data represent a RR of revision of 0.49 (95% CI 0.19 to 1.28) at 5 years and 0.60 (95% CI 0.28 to 1.31) at 10 years. Hernigou published crude revision rates at 15 years of 2.7% for junior surgeons and 4.0% for senior surgeons, which represents a RR of revision of 0.68 (95% CI 0.17 to 2.64) (Hernigou et al., 2009).

2.5.8 Unicompartmental knee replacement (UKR)

Both UKR studies were conducted in the same centre but capture separate cohorts of patients (Bottomley et al., 2016, Alvand et al., 2021). Bottomley et al. conducted a retrospective cohort study of 1,084 consecutive UKRs performed between 1998 and 2008. They demonstrated that consultants and trainees had cumulative 9-year survival estimates of 93.9% and 93.0%, respectively. They found no strong evidence of a difference in implant survival between the groups (log rank: p=0.30) (Bottomley et al., 2016). These data represent crude revision rates at 10 years of 4.6% and 3.6% for trainees and consultants, respectively (RR 1.26, 95% CI 0.69 to 2.31). Trainees were supervised by a scrubbed consultant in 48% of cases. Alvand et al. reported a series of 273 UKRs performed between 2009 and 2015. They did not report net survival estimates according to surgeon grade. However, they reported crude revision rates at 5 years of 0.8% and 2.6% for trainees and consultants, respectively. These data represent a RR of revision of 0.33 (95% CI 0.04 to 2.90). Trainees were supervised by a scrubbed consultant in 100% of cases.

2.5.9 Assessment of the quality of evidence

The GRADE assessment of the quality of evidence for each outcome indicates a low, or very low quality of evidence for all outcomes (Table 2).

Table 2 - GRADE Summary of Findings Table.

Outcomes	Follow-up (years)	Trainee revision/cases†, n	Consultant revisions/cases†, n	Net survival/relative risk (95% CI)	Participants (studies), n	Quality of Evidence	Comments
THR: net implant survival	5	650	1,045	NS: Trainee 97.9% (96.6 to 99.2) NS: Consultant 98.1% (97.1 to 99.2)	1,695 (3)	Very low	Serious ROB, indirectness, and imprecision
	10	348	734	NS: Trainee 98.1% (95.9 to 99.1) NS: Consultant 96.7% (94.7 to 97.9)	1,082 (1)	Low	Serious indirectness and imprecision
THR: crude revision rate	5	13/919	29/1794	RR: 0.88 (0.46 to 1.70)	2,713 (3)	Very low	Serious ROB, indirectness, and imprecision
	10	13/634	40/1318	RR: 0.68 (0.37 to 1.26)	1,952 (2)	Low	Serious indirectness and imprecision
TKR: net implant survival	5	236	450	NS: Trainee 97.9% (95.0 to 99.2) NS: Consultant 95.4% (93.0 to 97.0)	686 (1)	Low	Serious imprecision
	10	386	550	NS: Trainee 96.2% (94.0 to 98.4) NS: Consultant 95.1% (93.0 to 97.2)	936 (2)	Very low	Serious inconsistency and imprecision
	15	150	100	NS: Trainee 91.0% (85.0 to 97.0) NS: Consultant 92.0% (90.0 to 94.0)	250 (1)	Very low	Serious inconsistency and very serious imprecision
TKR: crude revision rate	5	5/236	20/450	RR: 0.47 (0.18 to 1.25)	686 (1)	Low	Serious imprecision
	10	8/236	26/450	RR: 0.58 (0.27 to 1.27)	686 (1)	Low	Serious imprecision
	15	4/150	4/100	RR: 0.67 (0.17 to 2.60)	250 (1)	Very low	Serious inconsistency and very serious imprecision
UKR: net implant survival	10	673	411	NS: Trainee 93.0% (90.3 to 95.7) NS: Consultant 93.9% (90.2 to 97.6)	1,084 (1)	Low	Serious imprecision
UKR: crude revision rate	5	1/118	4/155	RR: 0.33 (0.04 to 2.90)	273 (1)	Low	Serious imprecision
	10	31/673	15/411	RR: 1.26 (0.69 to 2.31)	1,084 (1)	Low	Serious imprecision

GRADE, Grading of Recommendations Assessment, Development and Evaluation; CI, confidence interval; NS, net survival; RR, relative risk; †, number of revisions not reported for net implant survival; ROB, risk of bias

2.6 Discussion

The results of this study suggest that, in the context of contemporary practice, trainees do not achieve worse hip and knee replacement survival outcomes compared to their consultant colleagues at 5 to 10 years follow-up. We found no strong evidence of an association between surgeon grade and the net survival of THRs at 5 years (trainees: 97.9% vs. consultants: 98.1%). There was no association between surgeon grade and the crude revision rate of THRs at either 5-, or 10-years follow-up. Furthermore, we found no strong evidence of an association between surgeon grade and the net survival of TKRs at 10 years (trainees: 96.2% vs. consultants: 95.1%). Our narrative summary of two studies highlights that there is no evidence in the existing literature of an association between trainee-performed UKR and an increased risk of revision.

2.6.1 Strengths and limitations

This review has a number of strengths. We conducted a comprehensive systematic review with an exhaustive search according to current best practice guidelines and published the protocol for the methodology at inception. However, the data captured by this review have several limitations, which we have attempted to address through quality of evidence assessment and risk of bias analysis. The GRADE assessment indicates a low to very low quality of evidence for each outcome. Furthermore, the ROBINS-I assessment indicates a moderate to severe risk of bias in the included studies. These findings are generally consistent with the predominantly retrospective design of the included studies. The conclusions of this review are therefore limited by the strength and quality of the existing published data, which originate from a relatively small number of observational studies.

Meta-analysis of outcome measures was only possible at 5 and 10 years for THRs and 10 years for TKRs, which limits the generalisability of our findings to these short and medium-term intervals of follow-up. Therefore, this review does not capture any differences in early failure rates that might exist between trainee and consultant cohorts before 5 years. The included studies originated from the UK, France, Switzerland, and Japan, which limits the generalisability of the findings to countries with established orthopaedic training programmes.

Formal orthopaedic training is a long process (lasting up to 10 years in some countries); therefore, individual trainees have varying levels of experience, which are not captured by the binary variables used in this study, or in the existing literature. The included studies did not provide sufficient data to perform meaningful adjustment or sensitivity analysis according to specific training grade, or the level of senior supervision. Furthermore, our study captures cases performed between 1990 and 2015 (Table 1) and we were unable to adjust for variations in training practice (such as the level of senior supervision) that may have occurred over this 25-year period.

Implant survival is a key determinant of good outcome in joint replacement surgery and is the sole variable considered in the current benchmarking strategies for the assessment of implant components. However, this review did not consider other factors that may be important when evaluating surgical outcomes, such as patient-reported outcome measures, or complications other than failure, which have previously been found to occur in higher rates when joint replacements are performed by less experienced surgeons (Hedlundh et al., 1996, Schoenfeld et al., 2013).

Published literature did not consistently report age, sex, comorbidities, implant design, or the level of senior supervision; making it very difficult to adjust for these variables. Methods of categorising the procedural complexity of a hip or knee replacement are not widely used in the orthopaedic literature and were not reported by any of the studies included in this review. Therefore, it was not possible to adjust for this factor. It is reasonable to suggest that the predominantly superior survival outcomes observed in the trainee cohorts are a product of patient selection and close senior supervision, with good trainers selecting appropriately complex cases for their trainees.

Our conclusion for UKRs is limited by the strength and quality of existing evidence, which is based on a relatively small number of cases. The two UKR studies included in this review represent separate cohorts originating from the same designer centre, which is a potential source of bias (Alvand et al., 2021, Bottomley et al., 2016). A total of 1,357 UKRs were included, of which 58% (n=791) were performed by trainees. This high proportion of trainee-performed UKRs is not representative of wider practice in England and Wales, where only 4% of UKRs are performed by trainees (see Chapter 5).

The outcomes of this single institution may not be generalisable to wider practice and lower volume centres, which may have less experience in training surgeons to perform UKR.

2.6.2 Comparison with other studies

A single study was excluded because the THRs under follow-up were performed prior to 1990 (Marston et al., 1996); thus not considered representative of contemporary training practices. The authors of this 10-year study of 413 THRs reported a significantly higher rate of revision for trainees, with 15 of 16 revised hips performed by trainees. Inclusion of this study in our meta-analysis of 10-year THR crude revision rates increases the RR of revision to 1.12 (95% CI 0.66 to 1.92), in favour of THRs performed by consultants. One explanation for this is that the model of training in the UK at the time differed, with trainees more often operating without appropriate senior supervision.

Our findings are consistent with those of the New Zealand Joint Registry (Storey et al., 2018, Inglis et al., 2013). In a cohort of 35,415 THRs, of which 4,049 were performed by trainees, the authors reported no significant difference in the revision rate between surgeon groups (Inglis et al., 2013). In a further cohort of 79,671 TKRs and 8,854 UKRs, of which approximately 10% were performed by trainees, they reported no significant difference in the revision rates of knee replacements performed by trainees and consultants (Storey et al., 2018). These studies were not included in this meta-analysis because the authors did not report net survival estimates and revision rates were reported ‘per 100 component years’, rather than for clearly defined periods of follow-up, which cannot be calculated from the data presented.

2.6.3 Implications

There is a delicate balance between ensuring optimal outcomes for patients and the necessity to train the next generation of surgeons. Reidy et al. suggest that the availability of surgeon level registry data as a means of benchmarking performance, may lead to a desire to avoid perceived poor performance and thus a reluctance among consultants to let trainees operate (Reidy et al., 2016, Faulkner et al., 2018). However, the findings of this review are encouraging and support the notion that in the context of contemporary practice, in countries with established and regulated orthopaedic training

programmes, trainees can achieve implant survival outcomes equivalent to their consultant colleagues. The senior supervision of trainees was inconsistently reported in the studies included in this review but is likely to play an important role in the successful outcome of trainee-performed hip and knee replacements.

An adequately powered non-inferiority RCT with 10 years follow-up assuming an acceptable revision rate of 5% and a 1% absolute non-inferiority delta (α=0.05; power=0.80; 1:1 allocation ratio), would require a sample size of 6,400 patients (Sayers et al., 2017). However, factors inherent to the training process, such as variation amongst trainees, the need for case selection according to complexity and varying levels of supervision based on a trainee's experience, may preclude an inclusive and therefore generalisable RCT. Further investigation should focus on the associations between surgeon grade, the senior supervision of trainees, and the risk of revision following trainee-performed hip and knee replacements. Future work should also investigate the risk of early revision and the specific indications for revision following trainee-performed procedures. The analysis of patient data recorded in a mandatory national joint replacement registry would be an appropriate means of further investigation.

2.6.4 Conclusion

In conclusion, there is no strong evidence in the existing literature that trainee surgeons have worse outcomes than their consultant surgeon colleagues, in terms of the net survival, or crude revision rate of hip and knee replacements at 5 to 10 years follow-up. This may mean that there is no difference, or that appropriate case mix selection and supervision of trainees is currently employed and is safe to continue. Our results are concordant with published registry data and represent the best available evidence but are limited by the quality of the existing published studies.

Chapter 3 The association between surgeon grade and risk of revision following total hip replacement: an analysis of National Joint Registry data

3.1 Overview

Building on the findings of Chapter 2, this chapter seeks to overcome some of the limitations of the existing literature. The aim of this chapter was to use NJR data for England and Wales to examine the association between surgeon grade, the supervision of trainees, and the risk of revision following THR. This study benefits from the use of generalisable NJR data and is significantly larger than any previous study on this subject. This study has been published in The Bone & Joint Journal (Fowler et al., 2021) and included in this thesis according to the University's guidance on the integration of publications as chapters within a dissertation.

3.1.1 Contributors

TF was responsible for study concept, design, literature review, data analysis, interpretation of the results, and writing the manuscript. AA contributed to the literature review and was responsible for interpretation of the results and review of the manuscript. AB, MR and MW were responsible for study concept, design, interpretation of the results, and review of the manuscript. AS was responsible for study concept, design, data analysis, interpretation of the results, and review of the manuscript.

3.1.2 Declaration of conflicts of interest

TF has no conflicts of interest to declare. MR reports an institutional grant to support the Bone and Joint Infection Registry from Stryker, Zimmer Biomet, Heraeus, Link, Depuy, Smith & Nephew, and Implantcast, a contract with Heraeus for management time on a RCT, and educational grant and speaker payments from Zimmer Biomet. AB and MW report a contract held by the University of Bristol for membership of HQIP Lot 2 Statistical Analysis team for the National Joint Registry, and royalties from Taylor & Francis. AA reports a part-time employment contract for Virti Ltd, and

royalties from BPP Holdings Ltd. MW also reports an independently conducted research grant held by the University of Bristol with CeramTec, teaching payments from Heraeus, and research committee membership on the British Orthopaedic Association and the British Hip Society.

3.2 Abstract

3.2.1 Background

Total hip replacements (THRs) are performed by surgeons at various stages in training with varying levels of supervision, but we do not know if this is safe practice with comparable outcomes to consultant-performed THR. Our aim was to examine the association between surgeon grade, the senior supervision of trainees, and the risk of revision following THR.

3.2.2 Methods

We performed an observational study using National Joint Registry (NJR) data. We included adult patients who underwent primary THR for osteoarthritis (OA), recorded in the NJR between 2003 and 2016. Exposures were operating surgeon grade (consultant or trainee) and whether or not trainees were directly supervised by a scrubbed consultant. Outcomes were all-cause revision and the indication for revision up to 10 years. We used methods of survival analysis, adjusted for patient, operation, and healthcare setting factors.

3.2.3 Results

We included 603,474 THRs, of which 58,137 (9.6%) procedures were performed by a trainee. There was no association between surgeon grade and all-cause revision up to 10 years (crude hazard ratio (HR) 1.00, 95% CI 0.94 to 1.07; p=0.966); a finding which persisted with adjusted analysis. Fully adjusted analysis demonstrated an association between trainees operating without scrubbed consultant supervision and an increased risk of all-cause revision (HR 1.10, 95% CI 1.00 to 1.21; p=0.045). There was an association between trainee-performed THR and revision for instability (HR 1.14, 95% CI 1.01 to 1.30; p=0.039). However, this was not observed in adjusted models, or when trainees were supervised by a scrubbed consultant.

3.2.4 Conclusion

Within the current training system in England and Wales, appropriately supervised trainees achieve comparable THR survival to consultants. Trainees who are supervised by a scrubbed consultant

achieve superior outcomes compared to trainees who are not supervised by a scrubbed consultant, particularly in terms of revision for instability.

3.3 Background

Total hip replacement (THR) is a common, effective intervention for the treatment of arthritic conditions of the hip (Ferguson et al., 2018). Surgeons undergo extensive training to acquire the skills to perform THR safely and independently. These procedures are performed by surgeons at various stages in their training, with varying levels of senior supervision. There is a balance between protecting training opportunities in order to develop a competent future generation of surgeons and limiting the exposure of patients to undue risk during the training process.

The survival of a joint replacement, defined by the absence of revision surgery, is the principal measure used for comparing the longevity of implant components and is a commonly used measure of surgical performance (Evans et al., 2019a). Revision is defined as any operation to add, remove, or modify one or more components of an implant construct (The National Joint Registry, 2020). Net survival estimates are calculated using statistical methods of survival analysis (e.g. Kaplan-Meier), which look at time to a defined failure event (e.g. revision) and account for censored data that arise due to administrative conveyance, incomplete follow-up, or death (Sayers et al., 2018).

A recent meta-analysis of observational studies found that there is no strong evidence in the existing literature that trainees achieve worse implant survival compared to consultants at 5 to 10 years follow-up (Fowler et al., 2021). This study is based on low quality evidence, a low number of patients, and gives limited insight into the importance of scrubbed consultant supervision. A New Zealand Joint Registry (NZJR) study of 35,415 primary THRs found no significant difference in the revision rates, or indication for revision of THRs performed by consultants, supervised trainees, and unsupervised trainees (Inglis et al., 2013). However, this study did not report survival estimates and the overall number of trainee cases in the cohort was relatively low (n=4,049). The survival of THRs according to surgeon grade and supervision remains poorly understood.

We aimed to use a national mandated prospective registry to investigate the association between surgeon grade, the senior supervision of trainees, and the risk of all-cause revision following THR. An

additional aim of this study was to investigate the association between surgeon grade and the indication for revision.

3.4 Methods

3.4.1 Patients and data sources

We performed an observational study using prospectively collected data from the National Joint Registry (NJR). The base dataset was 1,230,989 linked hip procedures performed in England and Wales between 1 April 2003, and 31 December 2016. We included primary THRs in adult patients (aged ≥18 years) undertaken to treat osteoarthritis (OA), performed by either a trainee or a consultant surgeon. The indication for surgery is an important source of case complexity. Primary THRs performed for indications other than OA (e.g. inflammatory arthropathy, dysplasia, malignancy, or multiple indications) are generally considered to be more technically challenging and may be more likely to be performed by a consultant than a trainee. In an attempt to standardise variations in case complexity relating to the indication for surgery, our analyses were restricted to procedures performed for OA only. This approach and its limitations are discussed in more detail in Section 7.3.2.4.

Procedures were included in the primary analyses if the operating surgeon grade was recorded as any of the following: Foundation Year 1 (F1) to Specialty Trainee Year 2 (ST2); ST3-ST8 (i.e. registrar); fellow; or consultant. Surgeons of all grades (F1-ST2, ST3-ST8, SAS, consultant, and 'other') were included for sensitivity analysis. The process of mapping grade classifications to account for different versions of the NJR Minimum Data Set (MDS) form is detailed in Appendix 10.

Resurfacing procedures were excluded. Procedures with a metal-on-metal bearing combination were also excluded, as such prosthesis constructs are known to have a high failure rate and their use no longer reflects contemporary practice (Smith et al., 2012b). Cases were not selected according to the funding source or type of hospital (e.g. NHS, treatment centre, or private). This approach means that the cohort is representative and generalisable to consultant practice in all healthcare settings in England and Wales. However, private sector cases may be considered a separate group of patients, with no exposure to trainee-performed surgery. An alternative approach would be to exclude

privately-funded cases and base the analsyses on NHS-funded cases performed in NHS hospitals, which is the primary setting for orthopaedic training in England and Wales. However, the results of this approach would not be generalisable to wider consultant practice.

Due to the use of existing permissions for data access and linkage, it was not feasible to link NJR data to Hospital Episode Statistics (HES) or the National Health Service (NHS) England Patient-Reported Outcome Measure (PROM) database. The HES database captures additional patient-level information, including comorbidity data. It also provides additional operation-level data that are not recorded in the NJR, including the length of admission, postoperative medical complications, blood transfusion, readmission, and reoperations other than revision. Linkage to HES would have provided additional variables for confounding adjustment, which may have reduced residual confounding relating to unobserved differences in patient characteristics between the consultant and trainee groups. It would also provide additional outcome data, that were not included in the current study. Since 2009, the NHS PROM database has collected preoperative and 6-month postoperative PROM data (OHS and EQ-5D) for all NHS-funded hip replacements performed in England. However, complete data are only available for a relatively small proportion of NJR records (Sayers et al., 2020b, Liddle et al., 2015, Baker et al., 2012). Linkage to the NHS PROM database would have provided additional outcome data for a proportion of the cases included in this study.

3.4.2 Data access and processing

Access to the data was facilitated under existing NJR permissions for the study of training and volume in the context of hip and knee replacement, which are held by a supervisor of this thesis and senior author of this study (AS) (The National Joint Registry, 2022a). The base dataset used in the current study is based on the same cut of NJR data used in a previous NJR study (Sayers et al., 2020a). The steps taken in data processing are summarised in the flow diagram in Figure 6 and illustrated in greater detail in Figure 7. Figure 7 gives details of sequential exclusions from the base dataset, some of which were performed by AS prior to the initiation of the current study. All exclusions are consistent with the exclusion criteria of this study and the stage at which these occurred is clearly

denoted. Complete-case analysis was used in all analyses and records with missing data in any confounding variable field used in subsequent statistical models were excluded from the relevant model. Details of missing data, the number of cases excluded, and the reasons for exclusion are summarised in Figure 6 and illustrated in greater detail in Figure 8. The demographic characteristics of included cases and cases excluded due to missing data are summarised in Appendix 11.

3.4.3 Ethics

This study was approved by the NJR Research Committee (The National Joint Registry, 2022a). The NJR supports public health surveillance and wider clinical decision making and holds data that are anonymous to the researchers who use it. NHS Health Research Authority guidance dictates that the secondary use of such data for research does not require approval by a research ethics committee (Health Research Authority, 2022). Therefore, separate research ethics committee approval was not required for the NJR studies included in this thesis. Patients are consented for inclusion in the NJR according to standard practice, with permission under the Health Service (Control of Patient Information) Regulations, otherwise referred to as Section 251 support (The National Joint Registry, 2022b).

Figure 6 - Study flow diagram (Chapter 3).

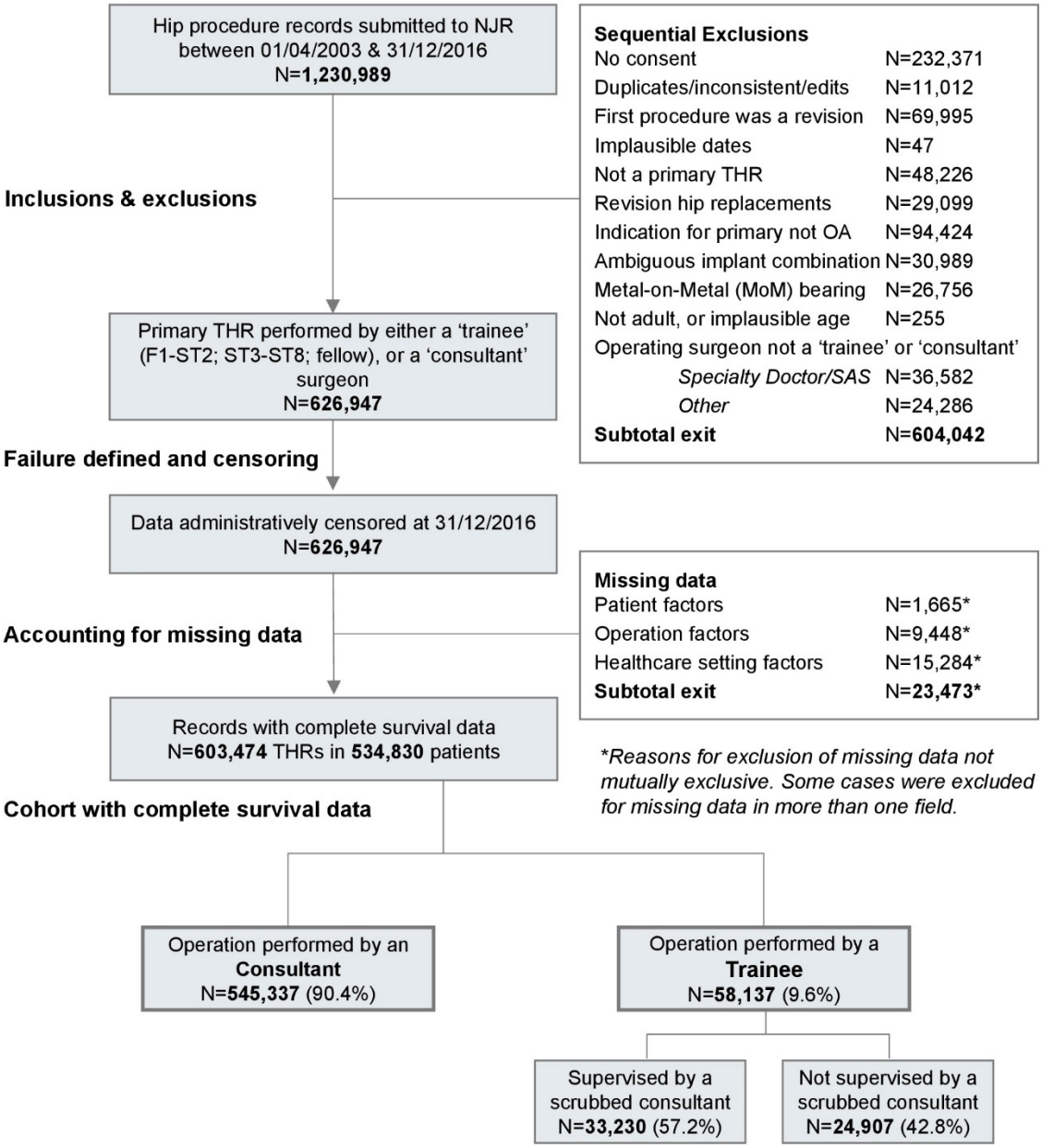
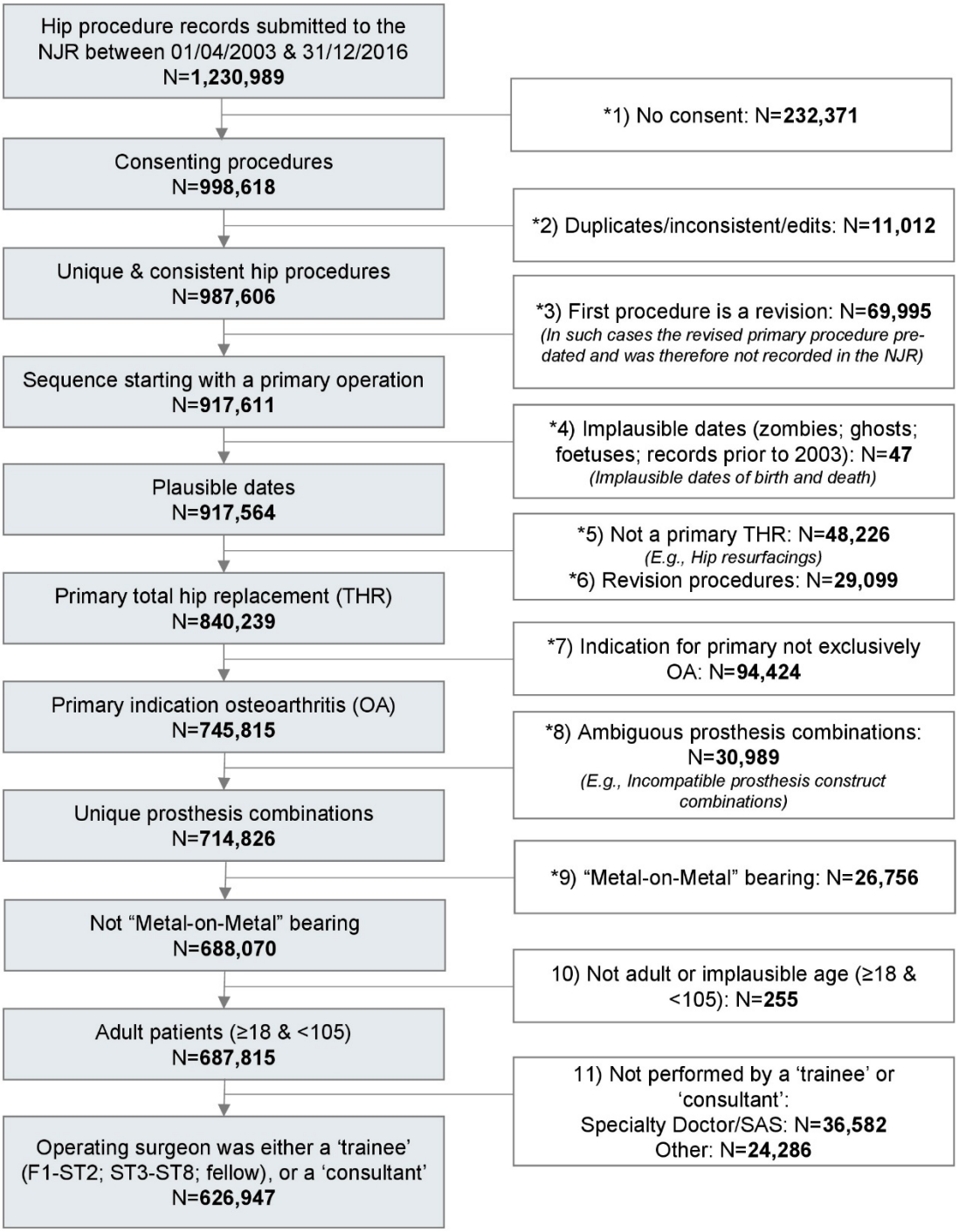
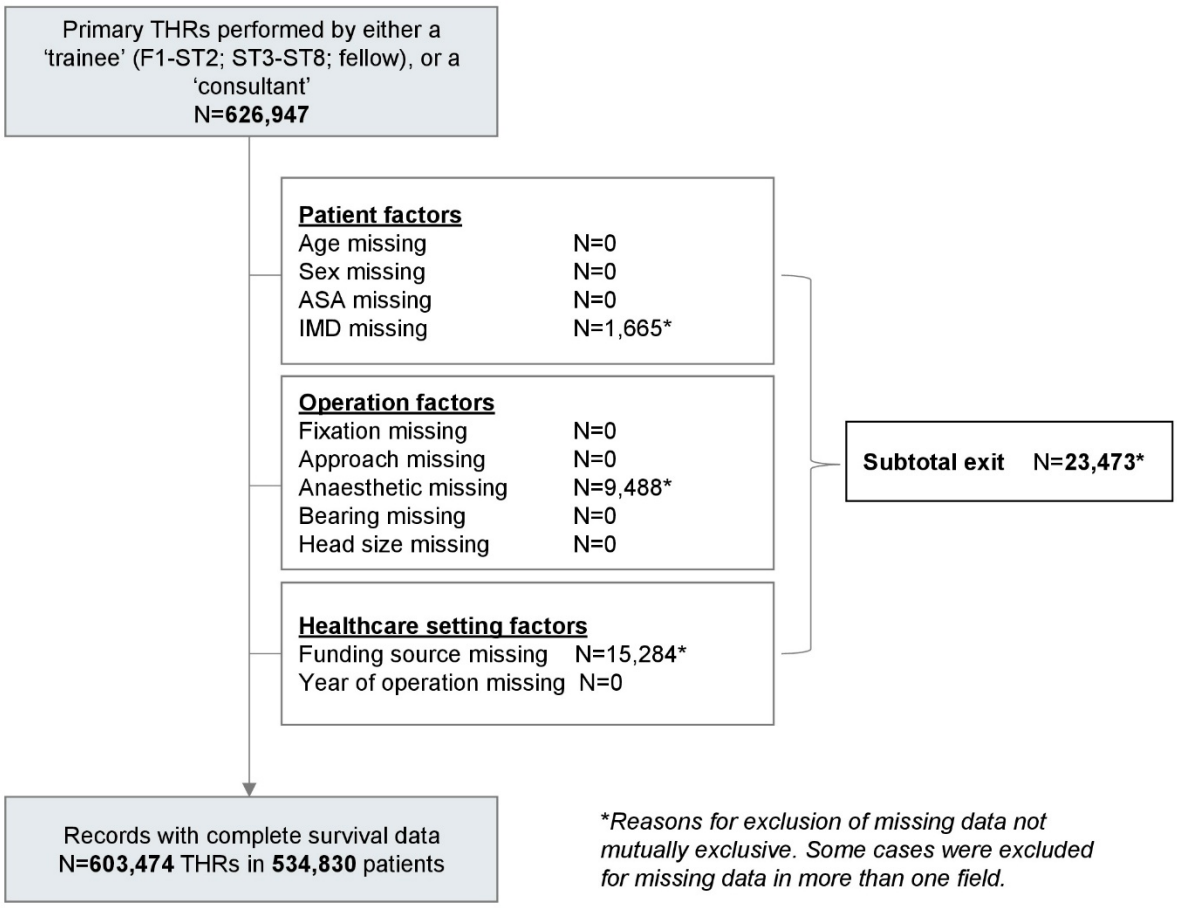


Figure 7 - Detailed study flow diagram showing sequential exclusions (Chapter 3).



***NB:** Sequential exclusions 1 to 9 had already been applied to the base dataset by AS, prior to the initiation of this study. Steps 10 and 11 were completed by TF.*

Figure 8 - Detailed study flow diagram showing exclusion of missing data (Chapter 3).



3.4.4 Exposures

The primary exposure (exposure A) was surgeon grade, which was categorised into two groups according to the grade of the operating surgeon: 1) consultant, or 2) trainee. Procedures performed by surgeons of the following grades were categorised under the variable 'trainee': F1-ST2; ST3-ST8; and fellow. Consultant surgeons have completed their formal training in orthopaedic surgery and have been appointed to a senior position in which they can practice independently and supervise trainees. The term 'consultant' is synonymous with 'attending', and the term 'registrar' synonymous with 'resident' in many healthcare settings including the USA.

We were interested in a secondary exposure (exposure B), which was whether or not the trainee was directly supervised by a scrubbed consultant during the procedure. Therefore, trainee cases were subcategorised as follows: 1) trainee supervised by a scrubbed consultant; 2) trainee not supervised by a scrubbed consultant.

Given the variability in the level of experience among surgical trainees, we performed a sensitivity analysis through recategorising procedures according to the specific training grade of the operating surgeon (exposure C: consultant; F1-ST2; ST3-ST8; SAS; other), with further subcategorisation according to level of scrubbed consultant supervision.

3.4.5 Outcomes of interest

The primary outcome was all-cause revision up to 10 years (i.e. any procedure to add, remove, or modify one or more components of an implant construct for any reason). We were also interested in the following indications for revision surgery (up to 10 years): infection; periprosthetic fracture; aseptic loosening; instability; and revision for 'other' causes. This follow-up period is consistent with the revision rate standard at 10 years set out by National Institute for Health and Care Excellence (NICE) guidance (Kandala et al., 2015, National Institute of Health and Care Excellence, 2021), as well as the analysis period used by the NJR (The National Joint Registry, 2021b). An additional outcome of interest was the temporal variation in the risk of revision (i.e. were THRs performed by trainees more likely to fail earlier than those performed by consultants?).

3.4.6 Statistical analysis

3.4.6.1 Overview

Frequencies and percentages were used to describe categorical variables. The mean, standard deviation (SD), and interquartile range (IQR) were used to describe continuous variables. We used methods of survival analysis to account for censoring due to incomplete follow-up, or death. Episodes were either administratively censored on 31 December 2016, or the date of death, whichever was earliest. Unadjusted net implant failure was estimated using the Kaplan-Meier (KM) method.

For the primary analyses, the associations between surgeon grade, supervision, and implant failure were modelled using Cox regression, with incremental adjustment for patient-, operation-, and healthcare setting-level confounding variables. A detailed description of the confounding variables included in the models and levels of adjustment is documented in Section 3.4.6.2. We performed separate analyses for each indication for revision, which were examined as separate survival endpoints. We tested and satisfied the proportional hazards (PH) assumption using likelihood ratio testing to compare proportional and non-proportional hazards models, which justified the use of Cox models for the primary and sensitivity analyses. The process of model selection, construction and justification is described in more detail in Section 3.4.6.3.

In a secondary analysis, we used flexible parametric survival modelling (FPM), described by Lambert and Royston and previously used in NJR analyses (Lambert and Royston, 2009b, Smith et al., 2012b, Hunt et al., 2018, Blom et al., 2020), to explore any temporal variation in the hazard ratio (HR) for revision. We used restricted cubic spline functions to model the baseline cumulative hazard of the primary effect and confirmed the superiority of our final model to preceding iterations through likelihood ratio testing. Our final FPM was fully adjusted for patient-, operation-, and healthcare setting-level confounding variables (Model 4: see Section 3.4.6.2). We repeated the analysis with the sequential exclusion of revisions for specific indications in order to examine the contribution of each mode of failure to the baseline hazard. Statistical analysis was performed using Stata (Version SE 15.1; StataCorp LP, USA).

3.4.6.2 Confounding variables

The analyses were incrementally adjusted for the following confounding variables. Model 1 was unadjusted. Model 2 was adjusted for patient-level factors (age, sex, American Society of Anaesthesiologists [ASA] grade, and index of multiple deprivation [IMD] decile). Model 3 was further adjusted for operation-level factors (anaesthetic, surgical approach, fixation, bearing surface, and head size). Model 4 was further adjusted for healthcare setting factors (funding source, and year of operation). In each case, the baseline category was the most frequently occurring (as detailed in Appendix 12).

Patient factors included age, sex, ASA grade, and IMD decile. All confounding variables were categorical, other than age which was continuous. ASA grade is a five-category system for the assessment of a patient's preoperative fitness according to the severity of medical comorbidities (Dripps et al., 1961). Patients are categorised according to the following definitions. ASA I – a normal healthy patient; ASA II – a patient with mild systemic disease; ASA III – a patient with severe systemic disease; ASA IV – a patient with severe systemic disease that is a constant threat to life; ASA V – a moribund patient who is not expected to survive without the operation. In this study, we used the following three categories: ASA I, ASA II and ASA ≥III. The index of multiple deprivation (IMD) is a geographical area-based measure of relative social deprivation. In England and Wales, small areas are ranked according to seven domains of deprivation (income, employment, education, health, crime, housing and services, and living environment). Our models were adjusted for IMD decile (1 being the most deprived, 10 being the least deprived) (Noble et al., 2019).

Operation-level factors included anaesthetic, surgical approach, fixation, bearing surface, and head size. Cases were categorised according to the primary method of anaesthesia used: general anaesthetic, spinal, epidural, nerve block, or 'other'. The surgical approach reflects the anatomical approach through which a joint is replaced. This was treated as a binary variable according to whether or not the procedure was performed via a posterior approach. Fixation describes the method by which the femoral and acetabular components are fixed within bone, and cases were categorised as follows:

cemented, uncemented, hybrid, or reverse hybrid. The bearing surface is defined by the materials used in the articulation between the femoral head and acetabular component. Cases were categorised as follows: metal-on-polyethylene, ceramic-on-polyethylene, ceramic-on-ceramic, or 'other'. The size of the femoral head used in each case was categorised as follows: <32mm, 32mm, or ≥36mm. These operative confounding factors are described in more detail in Section 1.2.3.

Healthcare setting factors included the funding source and year of operation. Cases were categorised according to whether the procedure was privately funded or funded by the NHS. Finally, cases were categorised according to the year of operation: either 2003-2012, or 2013-2016.

3.4.6.3 Model selection, construction and justification

The structured approach to model selection and construction used in this study was based around an in-depth assessment of the proportionality of hazard functions. Where data satisfied the PH assumption (i.e. the ratio of hazard functions was constant) or further analysis demonstrated that there was no significant advantage to be gained from using FPM analysis, the Cox model was preferred for primary analysis. This approach is summarised here:

Step 1: Schoenfeld residuals tests and were applied to incrementally adjusted Cox models for 'surgeon grade' to assess the PH assumption at each level of adjustment.

1) Cox PH model for surgeon grade (Model 1: unadjusted)
    a. Schoenfeld residual test: p=0.01
2) Cox PH model for surgeon grade (Model 2: adjusted for patient factors)
    a. Schoenfeld residual test: p<0.01
3) Cox PH model for surgeon grade (Model 3: adjusted for patient & operation factors)
    a. Schoenfeld residual test: p<0.01
4) Cox PH model for surgeon grade (Model 4: adjusted for patient, operation & healthcare setting factors)
    a. Schoenfeld residual test: p<0.01

The Schoenfeld residuals tests suggested that 'surgeon grade' may have a time-dependent effect and may not satisfy the PH assumption. FPM was used to investigate this further. FPM uses restricted cubic spline functions to model the baseline hazard function and account for time-dependent effects of specified variables (Royston and Lambert, 2011, Lambert and Royston, 2009b). This allows researchers to overcome limitations of the Cox model relating to its semiparametric nature (see Section 1.6.1).

Step 2: FPMs were constructed according to the methods described by Royston and Lambert (Royston and Lambert, 2011). We constructed the following two models:

1) A non-proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors)
    a. A non-PH model using the stpm2 command with 'surgeon grade' treated as having a time-dependent effect.
    b. Graphical assessment, Akaike information criteria (AIC), and Bayes information criteria (BIC) were used to optimise the fit and complexity of the model. These methods were used to determine the degrees of freedom with which to model hazard functions, as well as the optimal number and location of knots.
    c. We confirmed the superiority of our final model to preceding iterations using likelihood ratio testing.
    d. The final model was as follows: The baseline hazard was modelled with 7 degrees of freedom. Surgeon grade had a time-dependent effect and was modelled with 6 degrees of freedom. Confounding variables were modelled with fixed effects.
2) A proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors)
    a. A PH model using the stpm2 command, equivalent to the Cox model.
    b. This model is identical to the non-PH FPM model described above, but time-dependent effects were not included.

Step 3: Likelihood ratio tests were used to compare PH and non-PH models at each level of adjustment:

1) PH FPM compared to non-PH FPM (Model 1): p=0.10
2) PH FPM compared to non-PH FPM (Model 2): p=0.09
3) PH FPM compared to non-PH FPM (Model 3): p=0.10
4) PH FPM compared to non-PH FPM (Model 4): p=0.07

Summary: The PH (i.e. Cox) and non-PH models were parsimonious in this context. There was no significant advantage to using FPM with time-dependent effects. In view of these findings, the Cox model was used for the primary analysis of the association between surgeon grade and the risk of THR revision. The non-PH FPM model described in Step 2 was used in a secondary analysis to examine the temporal variation in the risk of revision.

3.5 Results

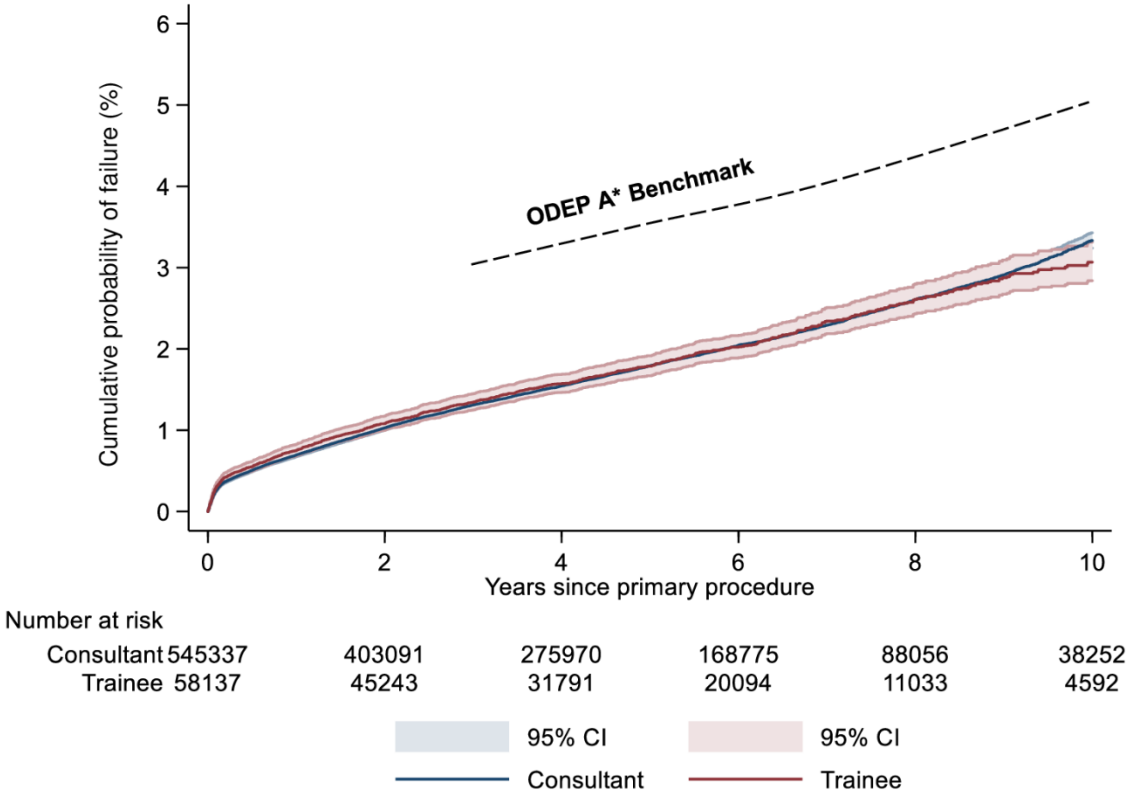
We included 603,474 THR procedures in 534,830 patients, of which 58,137 (9.6%) procedures were performed by a trainee. Trainees were supervised by a scrubbed consultant in 57.2% (n=33,230) of trainee-performed cases (Figure 6). Mean follow-up was 4.5 years (SD 3.0; IQR 1.9 to 6.8 years) and 10,194 hips were revised at a mean of 2.5 years (SD 2.6; IQR 0.3 to 3.9 years).

The mean age of patients operated on by trainees was 2.2 years older than patients operated on by consultants (71.7 vs. 69.5 years). Trainees operated on a lower proportion of ASA I patients (10.3% vs. 15.6%) and a higher proportion of ASA ≥III patients (19.6% vs. 14.3%). A higher proportion of trainee procedures used cemented fixation (46.8% vs. 35.2%), a metal-on-polyethylene bearing (74.3% vs. 63.3%), and a <32mm head (55.9% vs. 47.7%). Absolute numbers are reported in Table 3.

Table 3 - Descriptive statistics for patient, operation, and healthcare setting factors for included THRs.

	Surgeon grade and supervision			
Variable	**Consultant (n=545,337)**	**Trainee (overall) (n=58,137)**	**Trainee supervised by consultant (n=33,230)**	**Trainee not supervised by consultant (n=24,907)**
Mean (SD) age (years)	69.5 (10.3)	71.7 (9.5)	71.3 (9.7)	72.1 (9.2)
Female (%)	331,440 (60.8)	36,306 (62.5)	20,624 (62.1)	15,682 (63.0)
Side (%)				
Right	302,473 (55.5)	32,276 (55.5)	18,430 (55.5)	13,846 (55.6)
Body Mass Index (kg/m^2)				
Mean (SD)	28.7 (5.2)	28.8 (5.2)	28.8 (5.2)	28.7 (5.2)
ASA grade (%)				
ASA I	84,793 (15.6)	5,989 (10.3)	3,544 (10.7)	2,445 (9.8)
ASA II	382,525 (70)	40,779 (70)	23,012 (69.3)	17,767 (71.1)
ASA ≥III	78,019 (14.3)	11,369 (19.6)	6,674 (20.1)	4,695 (18.9)
Index of multiple deprivation decile (%)				
1	29,700 (5.5)	3,451 (5.9)	2,221 (6.7)	1,230 (4.9)
2	33,099 (6.1)	4,052 (7.0)	2,451 (7.4)	1,601 (6.4)
3	38,749 (7.1)	4,550 (7.8)	2,626 (7.9)	1,924 (7.7)
4	47,706 (8.8)	5,393 (9.3)	3,026 (9.1)	2,367 (9.5)
5	56,671 (10.4)	6,128 (10.5)	3,417 (10.3)	2,711 (10.9)
6	63,610 (11.7)	6,763 (11.6)	3,881 (11.7)	2,882 (11.6)
7	66,686 (12.2)	6,841 (11.8)	3,859 (11.6)	2,982 (12.0)
8	68,016 (12.5)	6,809 (11.7)	3,849 (11.6)	2,960 (11.9)
9	69,627 (12.8)	7,070 (12.2)	3,920 (11.8)	3,150 (12.7)
10	71,472 (13.1)	7,080 (12.2)	3,980 (12.0)	3,100 (12.5)
Anaesthetic (%)				
Spinal	278,270 (51)	28,224 (48.6)	17,652 (53.1)	10,572 (42.5)
General	230,153 (42.2)	24,842 (42.7)	12,860 (38.7)	11,982 (48.1)
Epidural	20,296 (3.7)	2,714 (4.7)	1,587 (4.8)	1,127 (4.5)
Nerve block	16,036 (2.9)	2,293 (3.9)	1,096 (3.3)	1,197 (4.8)
Other	582 (0.1)	64 (0.1)	35 (0.1)	29 (0.1)
Approach (%)				
Posterior	325,822 (59.8)	35,229 (60.6)	19,270 (58.0)	15,959 (64.1)
Other	219,515 (40.3)	22,908 (39.4)	13,960 (42.0)	8,948 (35.9)
Fixation (%)				
Cemented	191,736 (35.2)	27,182 (46.8)	15,302 (46.1)	11,880 (47.7)
Uncemented	229,040 (42)	15,349 (26.4)	9,345 (28.1)	6,004 (24.1)
Hybrid	108,501 (19.9)	14,108 (24.3)	7,561 (22.8)	6,547 (26.3)
Reverse hybrid	16,060 (2.9)	1,498 (2.6)	1,022 (3.1)	476 (1.9)
Bearing (%)				
Metal-on-polyethylene	345,157 (63.3)	43,218 (74.3)	23,939 (72.0)	19,279 (77.4)
Ceramic-on-polyethylene	101,748 (18.7)	7,889 (13.6)	4,889 (14.7)	3,000 (12.0)
Ceramic-on-ceramic	96,477 (17.7)	6,883 (11.8)	4,337 (13.1)	2,546 (10.2)
Other	1955 (0.3)	147 (0.3)	65 (0.2)	82 (0.4)
Head size (%)				
<32mm	259,879 (47.7)	32,511 (55.9)	18,638 (56.1)	13,873 (55.7)
32mm	164,354 (30.1)	16,273 (28.0)	9,095 (27.4)	7,178 (28.8)
≥36mm	121,104 (22.2)	9,353 (16.1)	5,497 (16.5)	3,856 (15.5)
Funding source (%)				
NHS	455,550 (83.5)	58,035 (99.8)	33,191 (99.9)	24,844 (99.8)
Private	89,787 (16.5)	102 (0.2)	39 (0.1)	63 (0.2)
Year of operation (%)				
2003-2012	300,459 (55.1)	35,233 (60.6)	18,446 (55.5)	16,787 (67.4)
2013-2016	244,878 (44.9)	22,904 (39.4)	14,784 (44.5)	8,120 (32.6)
ASA, American Society of Anaesthesiologists; NHS, National Health Service. Data are n (%) or mean (SD)				

Figure 9 - Kaplan-Meier plot (one minus survival) demonstrating the cumulative probability of THR failure (i.e. all-cause revision) according to surgeon grade (exposure A).



Reference line is the ODEP A benchmark at 3, 5, 7, and 10 years.*

Table 4 - The unadjusted cumulative probability of all-cause failure of THRs according to surgeon grade and supervision.

Follow-up (years)	ODEP A* (%)	Consultant			Trainee (overall)			Trainee supervised by a scrubbed consultant			Trainee not supervised by a scrubbed consultant		
		Number at risk**	Number of revisions	% Failure (95% CI)	Number at risk**	Number of revisions	% Failure (95% CI)	Number at risk**	Number of revisions	% Failure (95% CI)	Number at risk**	Number of revisions	% Failure (95% CI)
1	n/a	545,337	3,651	0.72 (0.69 to 0.74)	58,137	423	0.77 (0.70 to 0.85)	33,230	240	0.77 (0.68 to 0.88)	24,907	183	0.76 (0.66 to 0.88)
3	3	403,091	2,526	1.33 (1.30 to 1.36)	45,243	274	1.37 (1.27 to 1.47)	24,656	140	1.33 (1.20 to 1.47)	20,587	134	1.41 (1.26 to 1.57)
5	3.5	275,970	1,354	1.81 (1.77 to 1.86)	31,791	145	1.81 (1.69 to 1.94)	16,639	82	1.81 (1.65 to 1.98)	15,152	63	1.82 (1.64 to 2.01)
7	4.0	168,775	863	2.31 (2.26 to 2.36)	20,094	112	2.36 (2.21 to 2.53)	9,905	52	2.33 (2.12 to 2.56)	10,189	60	2.40 (2.18 to 2.65)
10	5.0	58,564	774	3.34 (3.25 to 3.43)	7,460	72	3.10 (2.85 to 3.32)	3,744	30	2.95 (2.65 to 3.29)	3,716	42	3.21 (2.88 to 3.57)

Data are the number at risk, the number of revision events, the unadjusted cumulative probability of failure and the 95% CI. The Orthopaedic Data Evaluation Panel (ODEP) A* benchmark is documented alongside for comparison. **Number at risk at the beginning of interval.

3.5.1 All-cause revision

The unadjusted cumulative probability of failure at 10 years was 3.34% (95% CI 3.25 to 3.43) for consultants, 3.10% (95% CI 2.85 to 3.32) for trainees overall, 2.95% (95% CI 2.65 to 3.29) for trainees supervised by a scrubbed consultant, and 3.21% (95% CI 2.88 to 3.57) for trainees who were not supervised by a scrubbed consultant (Table 4). The cumulative probability of failure according to surgeon grade is graphically displayed in Figure 9.

In the unadjusted Cox regression analysis comparing THRs performed by consultants and trainees, surgeon grade (exposure A) was not associated with all-cause revision (HR 1.00, 95% CI 0.94 to 1.07; p=0.966). This finding persisted despite incremental adjustment for patient, operation, and healthcare setting factors (Table 5).

Further analysis was performed according to the level of senior supervision (exposure B). Unadjusted analysis indicated no association between the supervision of trainees and all-cause revision. However, the fully adjusted analysis (Model 4) demonstrated an association between trainees operating without scrubbed consultant supervision and an increased risk of all-cause revision (HR 1.10, 95% CI 1.00 to 1.21; p=0.045), in contrast to trainees who were supervised by a scrubbed consultant (HR 1.03, 95% CI 0.95 to 1.13; p=0.451) (Table 5).

Sensitivity analysis was performed following further subcategorisation of procedures according to specific training grade (exposure C) and consultant supervision. We found no evidence of an association between any specific training grade and an increased risk of all-cause revision, regardless of the level of supervision (Table 6).

Table 5 - Results of Cox models according to surgeon grade (exposure A) and supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Model 1 (unadjusted) n=603,474			Model 2 (adjusted for †) n=603,474			Model 3 (adjusted for †, ‡) n=603,474			Model 4 (adjusted for †, ‡, §) n=603,474		
			HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
All cause revision (10,194 revisions)	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	1.00	0.94 to 1.07	0.966	1.02	0.96 to 1.09	0.544	1.05	0.99 to 1.13	0.108	1.06	1.00 to 1.14	0.065
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	0.98	0.90 to 1.07	0.616	0.99	0.91 to 1.08	0.838	1.02	0.94 to 1.12	0.588	1.03	0.95 to 1.13	0.451
		Trainee not supervised	1.02	0.93 to 1.12	0.629	1.06	0.96 to 1.16	0.249	1.09	1.00 to 1.20	0.062	1.10	1.00 to 1.21	0.045
Revision for infection (1,863 revisions)*	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	1.10	0.95 to 1.27	0.209	1.10	0.95 to 1.28	0.192	1.08	0.94 to 1.26	0.263	1.09	0.94 to 1.26	0.275
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	1.12	0.93 to 1.36	0.239	1.11	0.92 to 1.35	0.273	1.10	0.90 to 1.33	0.350	1.09	0.90 to 1.32	0.372
		Trainee not supervised	1.07	0.86 to 1.32	0.533	1.09	0.88 to 1.35	0.429	1.08	0.87 to 1.34	0.494	1.08	0.87 to 1.34	0.488
Revision for fracture (1,613 revisions)*	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	0.98	0.83 to 1.15	0.812	0.91	0.78 to 1.07	0.281	0.97	0.83 to 1.15	0.750	0.98	0.83 to 1.16	0.813
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	0.94	0.75 to 1.17	0.591	0.88	0.71 to 1.10	0.266	0.94	0.76 to 1.18	0.605	0.94	0.76 to 1.18	0.613
		Trainee not supervised	1.03	0.82 to 1.29	0.820	0.95	0.76 to 1.20	0.672	1.00	0.80 to 1.27	0.938	1.02	0.81 to 1.29	0.848
Revision for aseptic loosening (2,600 revisions)*	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	0.93	0.82 to 1.06	0.277	1.01	0.88 to 1.15	0.931	1.07	0.94 to 1.22	0.307	1.08	0.95 to 1.24	0.236
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	0.94	0.79 to 1.12	0.487	1.00	0.84 to 1.19	0.974	1.04	0.88 to 1.24	0.622	1.06	0.89 to 1.27	0.488
		Trainee not supervised	0.92	0.76 to 1.11	0.374	1.02	0.84 to 1.23	0.866	1.10	0.91 to 1.33	0.307	1.11	0.92 to 1.34	0.289
Revision for instability (2,378 revisions)*	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	1.14	1.01 to 1.30	0.039	1.11	0.98 to 1.26	0.099	1.09	0.96 to 1.24	0.189	1.10	0.97 to 1.25	0.151
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	1.07	0.90 to 1.27	0.446	1.04	0.88 to 1.24	0.625	1.05	0.88 to 1.24	0.611	1.06	0.89 to 1.26	0.547
		Trainee not supervised	1.23	1.03 to 1.47	0.021	1.20	1.00 to 1.43	0.047	1.14	0.95 to 1.36	0.149	1.15	0.96 to 1.38	0.123
Revision for other causes (2,199 revisions)*	A	Consultant	1.00			1.00			1.00			1.00		
		Trainee (overall)	0.84	0.72 to 0.98	0.022	0.92	0.79 to 1.07	0.261	1.00	0.86 to 1.16	0.993	1.01	0.87 to 1.18	0.867
	B	Consultant	1.00			1.00			1.00			1.00		
		Trainee supervised	0.79	0.64 to 0.97	0.022	0.85	0.69 to 1.04	0.108	0.91	0.74 to 1.12	0.374	0.92	0.75 to 1.14	0.473
		Trainee not supervised	0.90	0.73 to 1.11	0.335	1.01	0.82 to 1.24	0.948	1.11	0.90 to 1.37	0.310	1.12	0.91 to 1.38	0.287

Data are number of revisions for each indication, hazard ratio, 95% CI, or p-value. †**Patient factors:** age; sex; ASA grade; index of multiple deprivation decile. ‡**Operation factors:** bearing material; fixation; head size; approach; anaesthetic. §**Healthcare setting factors:** funding; year of operation. *Some cases revised for more than one indication.

Table 6 - Sensitivity analysis: Results of Cox models for all-cause revision according to specific training grade (exposure C) and supervision.

Exposure	Number of cases	Number of revisions	Complete cases (n=661,276) *		
			HR	95% CI	p-value
Model 1 (unadjusted)					
Consultant	545,337	9,168	1.00		
F1-ST2 supervised by scrubbed consultant	332	3	0.49	0.16 to 1.51	0.212
F1-ST2 not supervised by scrubbed consultant	312	6	0.95	0.43 to 2.11	0.891
ST3-ST8 supervised by scrubbed consultant	32,258	526	0.99	0.91 to 1.08	0.817
ST3-ST8 not supervised by scrubbed consultant	22,118	403	1.02	0.92 to 1.13	0.689
SAS supervised by scrubbed consultant	7,128	108	0.95	0.78 to 1.14	0.562
SAS not supervised by scrubbed consultant	31,512	545	0.93	0.85 to 1.01	0.091
Other supervised by scrubbed consultant	4,827	69	0.76	0.60 to 0.97	0.025
Other not supervised by scrubbed consultant	17,452	356	1.01	0.91 to 1.12	0.847
Model 2 (adjusted for †)					
Consultant	545,337	9,168	1.00		
F1-ST2 supervised by scrubbed consultant	332	3	0.51	0.16 to 1.57	0.237
F1-ST2 not supervised by scrubbed consultant	312	6	0.97	0.44 to 2.17	0.944
ST3-ST8 supervised by scrubbed consultant	32,258	526	1.01	0.92 to 1.10	0.906
ST3-ST8 not supervised by scrubbed consultant	22,118	403	1.05	0.95 to 1.16	0.314
SAS supervised by scrubbed consultant	7,128	108	0.97	0.80 to 1.17	0.747
SAS not supervised by scrubbed consultant	31,512	545	0.96	0.88 to 1.05	0.365
Other supervised by scrubbed consultant	4,827	69	0.77	0.61 to 0.98	0.031
Other not supervised by scrubbed consultant	17,452	356	1.05	0.95 to 1.17	0.369
Model 3 (adjusted for †, ‡)					
Consultant	545,337	9,168	1.00		
F1-ST2 supervised by scrubbed consultant	332	3	0.50	0.16 to 1.55	0.229
F1-ST2 not supervised by scrubbed consultant	312	6	0.96	0.43 to 2.14	0.927
ST3-ST8 supervised by scrubbed consultant	32,258	526	1.04	0.95 to 1.13	0.422
ST3-ST8 not supervised by scrubbed consultant	22,118	403	1.08	0.98 to 1.19	0.136
SAS supervised by scrubbed consultant	7,128	108	0.96	0.79 to 1.16	0.669
SAS not supervised by scrubbed consultant	31,512	545	0.94	0.87 to 1.03	0.202
Other supervised by scrubbed consultant	4,827	69	0.79	0.62 to 1.00	0.052
Other not supervised by scrubbed consultant	17,452	356	1.07	0.96 to 1.19	0.227
Model 4 (adjusted for †, ‡, §)					
Consultant	545,337	9,168	1.00		
F1-ST2 supervised by scrubbed consultant	332	3	0.51	0.16 to 1.57	0.236
F1-ST2 not supervised by scrubbed consultant	312	6	0.97	0.44 to 2.16	0.941
ST3-ST8 supervised by scrubbed consultant	32,258	526	1.05	0.96 to 1.14	0.322
ST3-ST8 not supervised by scrubbed consultant	22,118	403	1.09	0.98 to 1.20	0.103
SAS supervised by scrubbed consultant	7,128	108	0.97	0.80 to 1.17	0.741
SAS not supervised by scrubbed consultant	31,512	545	0.95	0.87 to 1.04	0.274
Other supervised by scrubbed consultant	4,827	69	0.80	0.63 to 1.01	0.060
Other not supervised by scrubbed consultant	17,452	356	1.07	0.97 to 1.20	0.188

*†**Patient factors:** age; sex; ASA grade; index of multiple deprivation decile. ‡**Operation factors:** bearing material; mode of fixation; head size; approach; anaesthetic. §**Healthcare setting factors:** funding; year of operation. *Greater number of complete cases due to the re-inclusion of cases performed by surgeons of the following grades: SAS, and 'other'.*

3.5.2 Indication for revision

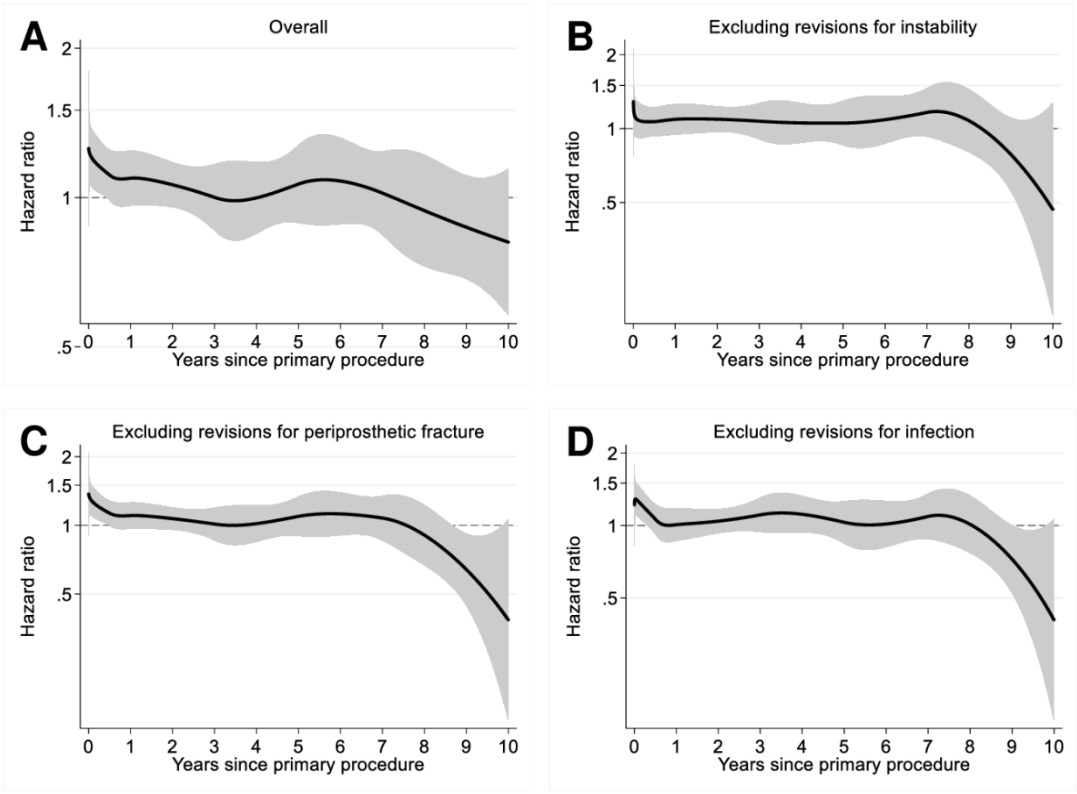
Unadjusted analyses demonstrated associations between trainee-performed procedures (exposure A) and an increased risk of revision for instability (HR 1.14, 95% CI 1.01 to 1.30; p=0.039), as well as a reduced risk of revision for 'other' causes (HR 0.84, 95% CI 0.72 to 0.98; p=0.022). However, adjusted analyses demonstrated no evidence of an association between surgeon grade (exposure A) and any specific indication for revision (including infection, periprosthetic fracture, aseptic loosening, instability, and other causes) (Table 5).

Further analysis was performed according to the level of supervision (exposure B). For procedures performed by trainees supervised by a scrubbed consultant, we found no evidence of an increased risk of revision for any indication. However, crude analysis demonstrated that procedures performed by trainees without scrubbed consultant supervision were associated with an increased risk of revision for instability (HR 1.23, 95% CI 1.03 to 1.47; p=0.021). This association was not observed in the fully adjusted model (Table 5).

3.5.3 Temporal variation in risk of revision

Our FPM analysis suggests that there may be a higher risk of failure within the first 6 months following THR performed by a trainee. There was no evidence of a higher risk of failure beyond 6 months. This is displayed in the form of a HR plot (Figure 10A), in which the lower 95% CI is above the baseline of 1.00 (for consultant-performed procedures) up to, but not exceeding, 6 months following the primary procedure. We see this effect disappear when cases that failed due to instability are excluded from the analysis (Figure 10B); a finding that was not clearly replicated following the exclusion of procedures that failed due to other early indications for revision, such as infection and periprosthetic fracture (Figure 10C-D).

Figure 10 - Results of flexible parametric survival models (FPMs) displayed as hazard ratio plots.



A) Overall model illustrating the hazard ratio for the all-cause revision of THRs performed by trainees compared to a baseline of consultant-performed THR. B) Excluding cases (n=2,378) revised for instability. C) Excluding cases (n=1,613) revised for periprosthetic fracture. D) Excluding cases (n=1,863) revised for infection. Plots B-D represent the sequential rather than cumulative exclusion of cases for each indication.

3.6 Discussion

We have demonstrated that when comparing THRs performed by consultants and trainees, there was no evidence of an association between the surgeon grade (exposure A) and all-cause revision; this was the case in both crude and adjusted models. However, further analysis indicated that THRs performed by trainees who were not supervised by a scrubbed consultant (exposure B), were associated with an increased risk of all-cause revision in the fully adjusted model only (adjusted for age, sex, ASA grade, index of multiple deprivation, anaesthetic, surgical approach, fixation, bearing surface, head size, funding source, and year of operation).

There was no evidence that trainees supervised by a scrubbed consultant conferred an increased risk of revision for any specific indication (including infection, periprosthetic fracture, aseptic loosening, instability, and other causes). However, procedures performed by trainees without scrubbed consultant supervision were associated with an increased risk of revision for instability in crude analysis; an association which was not observed in the fully adjusted model.

We have also demonstrated that there may be an increased risk of early revision up to, but not exceeding, 6 months following THR performed by trainees. Further analysis demonstrated that this effect may be attributable to early failures due to instability (Figure 10B).

The Orthopaedic Data Evaluation Panel (ODEP) is responsible for benchmarking implant components according to their revision rates at intervals of follow-up ranging from 3 to 15 years (Kandala et al., 2015). The A* benchmark is the highest at each interval and represents an internationally recognised acceptable limit. THR components with an A* rating at 10 years (revision rate of 5% or less) are recommended for use in end-stage arthritis by NICE (National Institute of Health and Care Excellence, 2021). The unadjusted KM analysis demonstrates that the cumulative probability of failure of THRs performed by both supervised and unsupervised trainees is below the ODEP A* threshold at all intervals of follow-up (Table 4), which indicates safe and acceptable implant survival for THRs performed by trainees in England and Wales. This is consistent with the results of the unadjusted Cox regression analysis, which shows no difference in the risk of all-cause revision for

THRs performed by both supervised and unsupervised trainees compared to consultants (Table 5). However, the results of our adjusted analyses suggest that improvements can be made through higher levels of supervision. The comparison of implant survival according to surgeon grade to the ODEP A* threshold represents a novel application of this benchmark.

3.6.1 Strengths and limitations

Our study has notable strengths. We analysed data for more than half a million patients recorded in the world's largest arthroplasty registry which, to our knowledge, makes this the largest study of the association between surgical training and hip replacement outcomes. We analysed patient data recorded in a mandatory, nationwide prospective database. This reduces sampling bias and improves the external validity and generalisability of our findings. We used a range of comprehensively adjusted statistical models, which have identified previously unknown associations.

Our study has limitations. Registry data are observational, and patients are not randomly allocated to intervention groups, thus causation cannot be attributed. Implant survival is a key determinant of successful operative outcome and is the sole outcome considered in the current benchmarking strategies for the assessment of implant components (Orthopaedic Data Evaluation Panel, 2021a). However, as they are not currently reported by the NJR and linkage to HES or the NHS England PROMs database was not feasible, we did not consider other factors that are important when evaluating surgical outcomes, such as PROMs, or postoperative complications other than failure.

The binary term 'trainee' does not account for the variability in the level of experience of individual trainees. We have attempted to account for this through sensitivity analysis by subcategorising procedures according to the specific training grade of the surgeon (exposure C); however, further work is required to stratify procedures according to individual surgeon experience. Similarly, supervision is recorded as a binary variable, which does not account for the spectrum of supervision that is inherent to the training process. For example, a consultant may vary the level of supervision based on their judgement of an individual trainee's level of experience and competence. Furthermore, a trainee may be indirectly supervised (e.g. by an unscrubbed consultant in the operating theatre). NJR

data entry are regularly audited to ensure the quality of data capture (The National Joint Registry, 2020). However, there are limitations to the accuracy and reliability with which surgeon grade and supervision are recorded. For example, operating surgeon grade is a categorical variable and does not account for procedures that may have been part-performed by a trainee.

3.6.2 Comparison with other studies

Our literature review identified no randomised controlled trials (RCTs) on this topic (Fowler et al., 2021). A study of 35,415 THRs using data from the NZJR reported no significant difference in the revision rate of THRs performed by consultants compared to those performed by supervised and unsupervised trainees (Inglis et al., 2013). They also found no significant difference in the indication for revision between the groups (including infection, fracture, aseptic loosening, instability and pain). This study reported crude revision rate (per 100 component years) rather than a formal survival analysis and their model was only adjusted for age and ASA grade. While they found no evidence of an association between unsupervised trainees and an increased risk of all-cause revision or revision for instability, this may reflect the weakness of the true effects, which were only identified in our study due to the high number of cases and comprehensive adjustment for confounding variables. It is encouraging to note the overall similarity of their findings to our own, which supports the generalisability of our findings to other countries with established orthopaedic training programmes. In particular, they found no evidence of an association between surgeon grade and revision for infection, facture, or aseptic loosening, which is consistent with our own findings. The remaining published literature on this topic is limited to a small number of retrospective studies, none of which have reported a negative association between THRs performed by trainees and implant failure (Reidy et al., 2016, Jain et al., 2018, Hasegawa and Amano, 2015, Palan et al., 2009).

The relationship between surgical experience and the dislocation of THRs has been the focus of previous investigation. Individual cohort studies (Hedlundh et al., 1996, Moran et al., 2004, Kim et al., 2017), and meta-analyses have drawn contradictory conclusions (Kunutsor et al., 2019, Singh et al., 2019). A recent meta-analysis found that procedures performed by surgeons with a ‘higher level

of experience' are associated with a lower risk of dislocation (Kunutsor et al., 2019). However, another recent meta-analysis that directly compared trainee and consultant procedures found no overall difference in the rates of dislocation, or deep infection according to surgeon grade (Singh et al., 2019). These results should be interpreted with caution as the terminology used to describe the exposures 'surgical experience' and 'surgeon grade' is inconsistent and often used interchangeably.

In our own study, we observed an association between THRs performed by trainees overall and revision for instability (in an unadjusted analysis only). With subcategorisation of trainee cases according to supervision, this was observed in the unsupervised trainee group, but not in the supervised trainee group (Table 5). Furthermore, we found that trainee-performed THRs may have an increased risk of revision for instability in the first 6 months (Figure 10B). Possible explanations for this observation are: 1) trainees are more likely to use a smaller head size (Table 3); and 2) trainees lack consistency in acetabular component positioning (Kim et al., 2017, Moran et al., 2004). Both are established risk factors for dislocation (Kunutsor et al., 2019, Lewinnek et al., 1978, Lopez-Lopez et al., 2017). Moran et al. found evidence of inconsistent cup positioning and reduced acetabular anteversion in THRs performed by trainees (Moran et al., 2004). Registry data are observational, and causation cannot be attributed. However, given the importance of these parameters to achieving optimal functional acetabular component alignment (Halawi and Haddad, 2018), component malposition may explain the failures due to instability that were observed when trainees were not supervised by a scrubbed consultant. We propose that scrubbed consultant supervision may reduce the risk of instability by helping trainees to choose appropriate implants and position them correctly.

3.6.3 Implications

Our findings have implications for clinical practice. In general, trainers appear to select appropriately complex cases for their trainees. While there was no overall difference in the survival of THRs performed by trainees and consultants, implant survival outcomes were superior when trainees were supervised by a scrubbed consultant. Nearly 43% of trainee-performed THRs recorded in the NJR were performed by trainees who were not directly supervised by a scrubbed consultant during the

procedure. We propose that trainees should be supervised by a scrubbed consultant when performing primary THR, particularly during the early years of training and during critical stages of the procedure, such as implant selection and when positioning implant components. The primary aim of training is to develop surgeons who are capable of safe independent practice. Therefore, it would seem prudent that when senior trainees operate unsupervised that careful case selection is undertaken, and that scrubbed consultant supervision is readily available if required.

3.6.4 Conclusion

This nationwide study of THRs with up to 10 years follow-up found that appropriately supervised trainees achieve comparable implant survival to consultants. However, there is some evidence that trainees who are not supervised by a scrubbed consultant achieve inferior results to consultants, particularly in terms of revision for instability. Our findings support the safety of the current methods by which surgeons are trained to perform THRs in England and Wales and may be generalisable to other countries with comparable orthopaedic training programmes. However, trainees achieve their best outcomes when supervised by a scrubbed consultant.

Chapter 4 The association between surgeon grade and risk of revision following total knee replacement: an analysis of National Joint Registry data

4.1 Overview

Building on the findings of the preceding chapters, this chapter seeks to extend this work to TKRs. The aim of this chapter was to use NJR data for England and Wales to examine the association between surgeon grade, the supervision of trainees, and the risk of revision following TKR. A more recent and thus larger NJR dataset was used compared to Chapter 3. While similar modelling techniques have been employed, inherent differences in the data dictate the need for an alternative approach to presenting the results.

4.1.1 Contributors

TF was responsible for study concept, design, literature review, data analysis, interpretation of the results, and writing the manuscript. AB, and MW were responsible for study concept, design, interpretation of the results, and review of the manuscript. AS was responsible for study concept, design, data analysis, interpretation of the results, and review of the manuscript.

4.1.2 Declaration of conflicts of interest

TF and AS have no conflicts of interest to declare. AB and MW report a contract held by the University of Bristol for membership of HQIP Lot 2 Statistical Analysis team for the National Joint Registry, and royalties from Taylor & Francis. MW also reports an independently conducted research grant held by the University of Bristol with CeramTec, teaching payments from Heraeus, and research committee membership on the British Orthopaedic Association and the British Hip Society.

4.2 Abstract

4.2.1 Background

Total knee replacements (TKRs) are performed by surgeons at different stages in training with varying levels of supervision, but we do not know if this is a safe practice or whether trainees achieve equivalent outcomes to consultant-performed TKR. This study aimed to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following TKR.

4.2.2 Methods

We included 953,081 cases in 788,288 adult patients who underwent primary TKR for osteoarthritis (OA), recorded in the National Joint Registry (NJR) between 2003 and 2019. Exposures were surgeon grade (consultant or trainee) and the level of scrubbed consultant supervision of trainees. The primary outcome was all-cause revision, and the secondary outcome was the number of procedures revised for the following indications: aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability. Flexible parametric survival models (FPM) were incrementally adjusted for patient, operation, and healthcare setting factors.

4.2.3 Results

Trainees performed 96,544 (10.1%) TKRs and were directly supervised by a scrubbed consultant in 63.2% of trainee-performed cases. Trainees achieved comparable outcomes to consultants in terms of the unadjusted cumulative probability of all-cause failure. Adjusted FPM analysis indicated evidence of an association between trainee-performed TKR and a small increased risk of early all-cause revision up to, but not exceeding, 4 years follow-up. This effect was not explained by the level of supervision. Further analysis suggested that this effect may be attributable to revisions for aseptic loosening/lysis, infection, and progression of OA (i.e. subsequent patellar resurfacing).

4.2.4 Conclusion

Trainees in England and Wales achieve safe and acceptable all-cause TKR implant survival, with comparable outcomes to consultants. However, adjusted analyses suggest an association between

trainee-performed TKR and a small transient increase in the risk of all-cause revision. This effect may be attributable to factors including aseptic loosening, infection, and progression of OA. Current training practices for TKR in England and Wales are safe in terms of equivalence of all-cause implant survival to consultant-performed TKR, but we have identified areas for potential improvement in trainee outcomes.

4.3 Background

Total knee replacement (TKR) is a clinically and cost-effective treatment for end-stage osteoarthritis (OA). It is one of the most common elective surgical procedures worldwide with over 100,000 performed annually in the UK, and over 700,000 performed annually in the USA (Price et al., 2018). TKRs are performed by surgeons at different stages in training, with varying levels of supervision. There is a balance between ensuring that trainees get adequate operative experience to develop the skills required for independent consultant practice, while ensuring safe and acceptable outcomes for patients. This is particularly relevant in the recovery of elective orthopaedic services in the wake of the COVID-19 pandemic (The Royal College of Surgeons of England, 2022).

There are challenges when quantifying the safety and efficacy of surgical interventions, but objective clinical outcomes can be useful metrics of success. The revision-free survival of a joint replacement is a commonly used objective clinical outcome measure. Our current understanding of the survival of TKRs in the context of surgical training is limited.

Studies of TKR outcomes have shown that trainees can achieve comparable results to consultants in terms of implant alignment (Mahaluxmivala et al., 2001, Kazarian et al., 2019, Khakha et al., 2015), patient-reported outcome measures (PROMs) (Beattie et al., 2018, Khakha et al., 2015, Faulkner et al., 2018, Storey et al., 2018), and complication rates (Theelen et al., 2018, Haughom et al., 2014). A recent meta-analysis, which included 936 TKRs, found no strong evidence of an association between surgeon grade and the net survival of TKRs at 10 years (Fowler et al., 2021). Furthermore, New Zealand Joint Registry (NZJR) evidence has suggested that implant survival is not compromised in trainee-performed TKR (Storey et al., 2018). However, this study did not investigate the indication for revision, and limitations in the statistical modelling employed leave scope for further investigation.

The aim of this study was to use National Joint Registry (NJR) data, to investigate the association between surgeon grade (consultant or trainee), the supervision of trainees, and the risk of revision following TKR. An additional aim of this study was to investigate the association between surgeon grade and the risk of revision for specific indications.

4.4 Methods

4.4.1 Patients and data sources

The base dataset was 1,502,564 linked knee procedures recorded in the NJR between 1 April 2003 and 31 December 2019. The study end date was limited to predate the anomalous period of elective orthopaedic practice during the COVID-19 pandemic (The National Joint Registry, 2021a). Patients were consented for inclusion in the NJR according to standard practice (The National Joint Registry, 2021c). We included primary TKRs in adult patients (≥18 years) performed in England and Wales.

The dataset used in this study was based on the analysis cohort used for the for the NJR 17th Annual Report. While the NJR records data for procedures performed in Northern Ireland, the States of Guernsey and the Isle of Man, these records are not currently included in the analysis cohort for the Annual Report. Cases from Northern Ireland and the States of Guernsey are excluded due to issues relating to tracing mortality, and cases from the Isle of Man are excluded as it is not possible to audit them against local hospital data (The National Joint Registry, 2020).

In the interest of relative standardisation of procedural complexity, we included cases with an indication of OA only. Cases were included if the operating surgeon grade was recorded as any of the following: Foundation Year 1 (F1) to Specialty Trainee Year 2 (ST2); ST3-ST8; fellow; or consultant. F1-ST2 is the most junior training category recorded in the NJR, followed by ST3-ST8. The process of mapping grade classifications to account for variations in terminology used in different versions of the Minimum Data Set (MDS) form is detailed in Appendix 10.

4.4.2 Data access and processing

Access to the data was facilitated under existing NJR permissions for the study of training and volume in the context of hip and knee replacement, which are held by a supervisor of this thesis and senior author of this study (AS) (The National Joint Registry, 2022a). The base dataset used in the current study is based on the same cut of NJR data that is used in the 17th Annual Report (The National Joint Registry, 2020). The steps taken in data processing and are summarised in the study flow diagram in Figure 11 and illustrated in greater detail in Figure 12. Figure 12 gives details of sequential exclusions

from the base dataset, some of which were performed by the authors of the 17th Annual Report prior to the initiation of the current study. All exclusions are consistent with the exclusion criteria of this study and the stage at which these occurred is clearly documented. Complete-case analysis was used in all analyses and records with missing data in any confounding variable field used in subsequent statistical models were excluded from the relevant model. Details of missing data, the number of cases excluded, and the reasons for exclusion are documented in Figure 13. The demographic characteristics of included cases and cases excluded due to missing data are summarised in Appendix 13.

Figure 11 - Study flow diagram (Chapter 4).

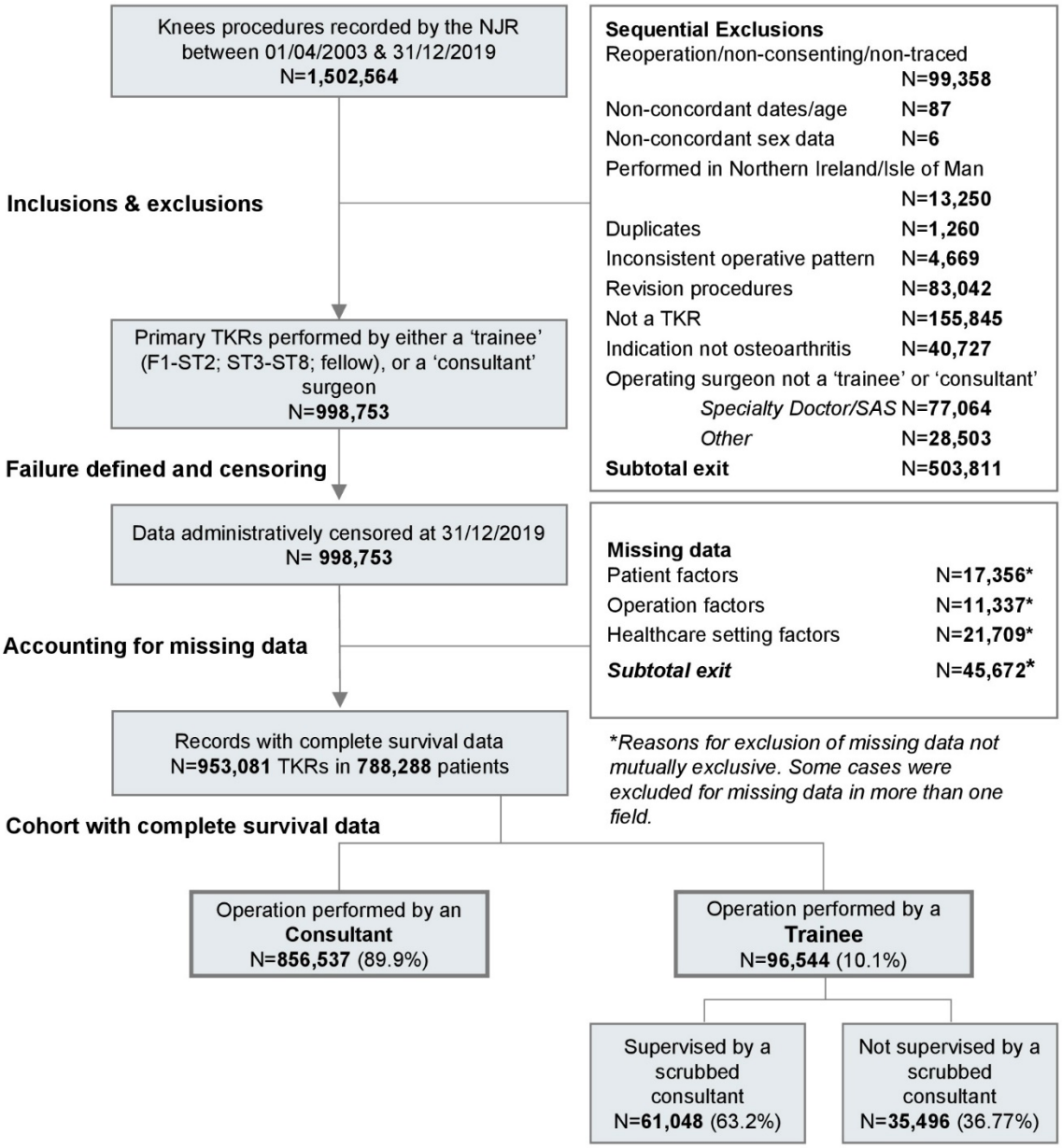
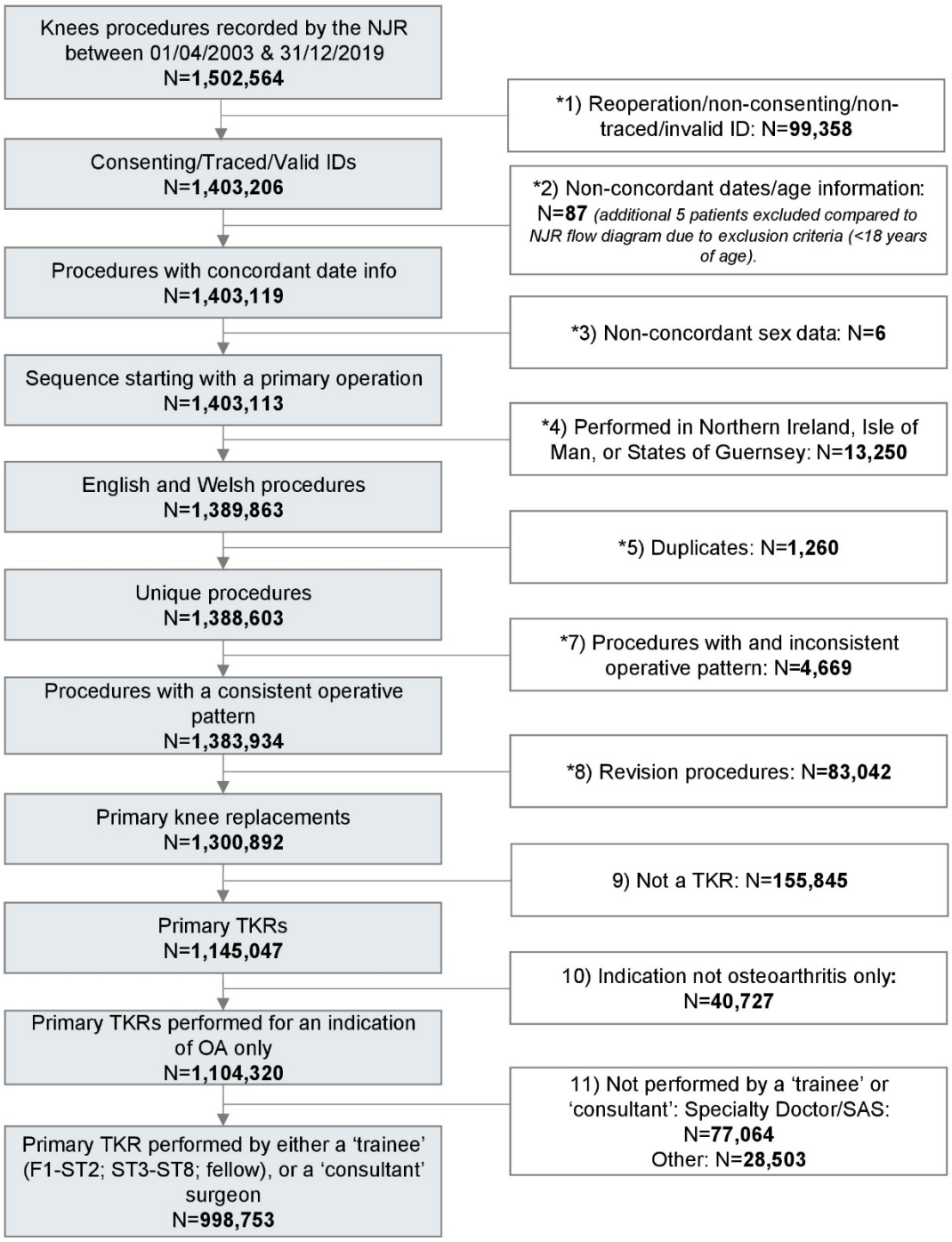
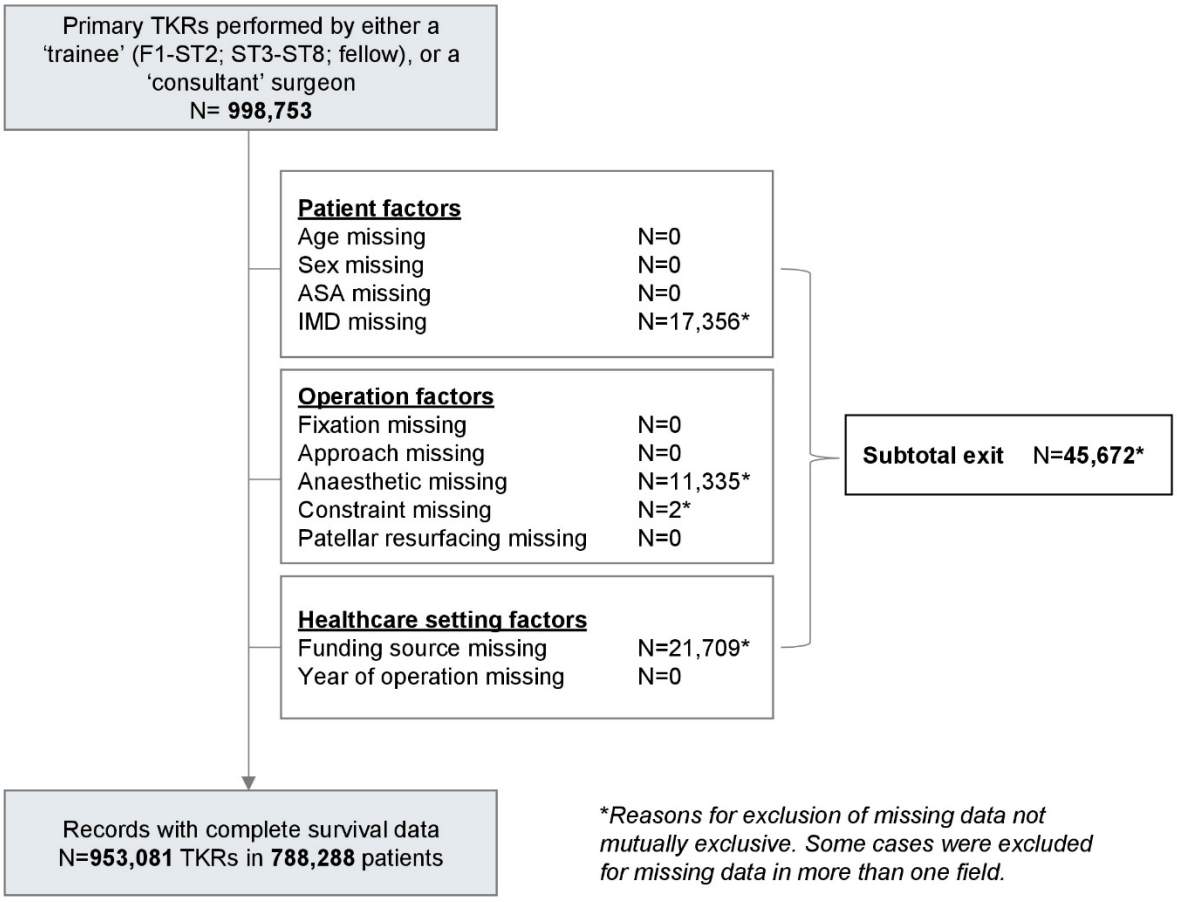


Figure 12 - Detailed study flow diagram showing sequential exclusions (Chapter 4).



***NB:** *Sequential exclusions 1 to 8 had already been applied to the base dataset by the authors of the NJR annual report, prior to the initiation of this study. Steps 9 to 11 were completed by TF.*

Figure 13 - Detailed study flow diagram showing exclusion of missing data (Chapter 4).



4.4.3 Exposures

The primary exposure (exposure A) was surgeon grade, which was categorised into two groups according to the grade of the operating surgeon: 1) consultant, or 2) trainee. Procedures performed by surgeons of the following grades were categorised under the variable 'trainee': F1-ST2; ST3-ST8; and fellow. Consultant surgeons have been awarded a Certificate of Completion of Training (CCT) in orthopaedic surgery and have been appointed to a senior position in which they can supervise trainees. The term 'consultant' is synonymous with 'attending' in several healthcare settings including the USA.

The secondary exposure was whether or not trainees were directly supervised by a scrubbed consultant during the procedure (exposure B). Trainee cases were subcategorised accordingly: 1) trainee supervised by a scrubbed consultant, or 2) trainee not supervised by a scrubbed consultant. A trainee was deemed to have been 'supervised by a scrubbed consultant' if a consultant was recorded as the first assistant. To account for variations in the level of experience between surgical trainees, we performed an additional analysis in which procedures were categorised according to the specific training grade of the operating surgeon (exposure C: consultant, F1-ST2, ST3-ST8, or fellow).

4.4.4 Outcomes of interest

The primary outcome measure was all-cause revision, which is defined by the NJR as any procedure to add, remove, or modify one or more components of an implant construct for any reason (The National Joint Registry, 2021a). The secondary outcome measure was the number of primary procedures revised for specific indications. We considered the following indications, which are listed as the five most common indications for knee replacement revision by the NJR: aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability (The National Joint Registry, 2020). Each indication for revision was examined as a separate survival endpoint and other events were administratively censored.

4.4.5 Statistical analysis

4.4.5.1 Overview

Categorical variables were described with frequencies and percentages. Continuous variables were described with the mean, standard deviation (SD), and interquartile range (IQR). Records were either administratively censored on 31 December 2019, or the date of death, whichever was earliest. Estimates of unadjusted net implant failure were calculated using Kaplan-Meier (KM) analysis.

We performed an initial analysis using Cox regression. A combination of Schoenfeld residual plots/tests and likelihood ratio testing (comparing proportional and non-proportional hazards models) was used to assess the proportional hazards (PH) assumption at each level of adjustment and assess the time-dependent effects of variables (Royston and Lambert, 2011). Neither crude nor adjusted analyses satisfied the PH assumption, which was explained by time-dependent effect of surgeon grade (exposure A).

To account for this we used flexible parametric survival modelling (FPM), which uses restricted cubic spline functions to model the baseline hazard and account for the time-dependent effects of specified variables (Lambert and Royston, 2009a, Royston and Lambert, 2011). We used graphical assessment, AIC and BIC, and likelihood ratio testing to select the most parsimonious model (Royston and Lambert, 2011). The process of model selection, construction and justification is explained in greater depth in Section 4.4.5.3. All analyses were performed using Stata (Version SE 15.1; StataCorp LP, USA).

4.4.5.2 Confounding variables

The analyses were incrementally adjusted for the following categorical confounding variables. Model 1 was unadjusted. Model 2 was adjusted for patient-level factors (age, sex, American Society of Anaesthesiologists [ASA] grade, and index of multiple deprivation [IMD] decile). Model 3 was further adjusted for operation-level factors (anaesthetic, approach, fixation, constraint and whether or not the patella was resurfaced). Model 4 was further adjusted for healthcare setting factors (funding

source, and year of operation). In each case, the baseline category was the most frequently occurring (as detailed in Appendix 14).

Patient factors included age, sex, ASA grade, and IMD decile. Age was categorised into five categories (<55; 55-64; 65-74; 75-84; and ≥85 years). ASA grade is a five-category system for the assessment of a patient's preoperative fitness according to the severity of medical comorbidities (Dripps et al., 1961). Patients are categorised according to the following definitions. ASA I – a normal healthy patient; ASA II – a patient with mild systemic disease; ASA III – a patient with severe systemic disease; ASA IV – a patient with severe systemic disease that is a constant threat to life; ASA V – a moribund patient who is not expected to survive without the operation. In this study, the following three categories were used: ASA I, ASA II, and ASA ≥III. The index of multiple deprivation (IMD) is a geographical area-based measure of relative social deprivation. In England and Wales, small areas are ranked according to seven domains of deprivation (income, employment, education, health, crime, housing and services, and living environment). Models were adjusted for IMD decile (1 being the most deprived, 10 being the least deprived) (Noble et al., 2019).

Operation-level factors included anaesthetic, approach, fixation, constraint and whether or not the patella was resurfaced. Cases were categorised according to whether or not the patient had any of the following forms of anaesthesia: general anaesthetic, nerve block, epidural, or spinal. The surgical approach reflects the anatomical approach through which a joint is replaced. The following categories were used: lateral parapatellar, medial parapatellar, mid-vastus, sub-vastus, and 'other'. Fixation describes the method by which implant components are fixed to bone, and cases were categorised as follows: cemented, uncemented, or hybrid. Constraint is a spectrum, which describes the level of stability and the freedom of movement conferred by the articulation between the tibial and femoral components of a TKR. For example, the level of constraint varies depending on whether the posterior cruciate ligament is preserved or sacrificed during the operation. The following categories were used: constrained condylar, monobloc polyethylene tibia, posterior-stabilised (fixed bearing), posterior-stabilised (mobile bearing), preassembled/hinged/linked, unconstrained (fixed bearing), and unconstrained (mobile bearing). The patella may be selectively resurfaced during TKR. Therefore,

cases were categorised according to whether or not the patella was resurfaced. These operative confounding are described in more detail in Section 1.3.3.1.

Healthcare setting factors included the funding source and year of operation. Cases were categorised according to whether the procedure was privately funded or funded by the National Health Service (NHS). Cases were categorised according to the year of operation: either 2003-2011, or 2012-2019.

4.4.5.3 Model selection, construction and justification

The structured approach to model selection and construction used in this study was based around an in-depth assessment of the proportionality of hazard functions. Where data did not satisfy the PH assumption (i.e. the ratio of hazard functions was not constant) and further analysis demonstrated that FPM was superior to the Cox model, FPM was preferred for primary analysis. This approach is summarised here:

Step 1: We performed an initial analysis using Cox regression. Schoenfeld residuals plots/tests were applied to incrementally adjusted Cox models for 'surgeon grade' to assess the PH assumption.

1) Cox PH model for surgeon grade (Model 1: unadjusted)
    a. Schoenfeld residual test: p=0.005
2) Cox PH model for surgeon grade (Model 2: adjusted for patient factors)
    a. Schoenfeld residual test: p<0.001
3) Cox PH model for surgeon grade (Model 3: adjusted for patient & operation factors)
    a. Schoenfeld residual test: p<0.001
4) Cox PH model for surgeon grade (Model 4: adjusted for patient, operation & healthcare factors)
    a. Schoenfeld residual test: p<0.001

The Schoenfeld residuals tests suggested that 'surgeon grade' may have a time-dependent effect. This was supported by the results of likelihood ratio testing, in which PH and non-PH models for 'surgeon grade' were compared at each level of adjustment. FPM was used to investigate this further.

Step 2: FPMs were constructed according to the methods described by Royston and Lambert (Royston and Lambert, 2011). We constructed the following two models:

1) A non-proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors)
    a. A non-PH model using the stpm2 command in Stata, with 'surgeon grade' specified as having a time-dependent effect.
    b. Graphical assessment, Akaike information criteria (AIC), and Bayes information criteria (BIC) were used to optimise the fit and complexity of the model. These methods were used to determine the degrees of freedom with which to model hazard functions, as well as the optimal number and location of knots.
    c. We confirmed the superiority of our final model to preceding iterations using likelihood ratio testing.
    d. The final model was as follows: The baseline hazard was modelled with 8 degrees of freedom. Surgeon grade had a time-dependent effect and was modelled with 2 degrees of freedom. Confounding variables were modelled with fixed effects.
2) A proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors)
    a. A PH model using the stpm2 command, equivalent to the Cox model.
    b. This model is identical to the non-PH FPM model described above, but surgeon grade is not specified as having a time-dependent effect.

Step 3: Likelihood ratio tests were used to compare PH and non-PH models at each level of adjustment:

1) PH FPM compared to non-PH FPM (Model 1): p=0.003
2) PH FPM compared to non-PH FPM (Model 2): p=0.003
3) PH FPM compared to non-PH FPM (Model 3): p=0.006
4) PH FPM compared to non-PH FPM (Model 4): p=0.006

Summary: Non-PH FPMs were superior to PH (i.e. Cox) models in this context, as PH models did not adequately account for time-dependent effects. Therefore, FPM was used as the primary method of adjusted analysis.

4.4.5.4 Sensitivity analysis

An unadjusted sensitivity analysis was performed using KM to investigate the influence of surgeon grade on the cumulative probability of failure, compared to other operative factors, including fixation, patellar resurfacing, and constraint.

4.5 Results

4.5.1 Descriptive analysis

We included 953,081 TKRs in 788,288 patients, with a maximum duration of follow-up of 16.8 years. Trainees performed 96,544 TKRs (10.1%) and were supervised by a scrubbed consultant in 61,048 (63.2%) of these cases (Figure 11). Mean follow-up was 6.3 years (SD 3.9; IQR 3.0 to 9.2 years) for trainee-performed TKRs and 5.8 years (SD 3.9; IQR 2.5 to 8.5 years) for consultant-performed TKRs. A total of 21,572 knees were revised at a mean of 3.4 years (SD 3.0; IQR 1.3 to 4.7 years).

Demographic data and summary statistics for confounding variables are listed in Table 7. The mean age of patients operated on by trainees was 1.2 years older than patients operated on by consultants (70.6 vs. 69.4 years). Trainees operated on a lower proportion of ASA I patients (7.9% vs. 10.6%) and a higher proportion of ASA ≥III patients (21.7% vs. 16.3%) compared to consultants. A higher proportion of trainee procedures utilised cemented implants (96.2% vs. 95.1%) and a lower proportion of trainee-performed cases included patellar resurfacing compared to consultant cases (34.5% vs. 38.9%). A higher proportion of trainee procedures have been supervised by a scrubbed consultant since 2012 (64.0% since 2012 vs. 36.0% pre-2012).

Table 7 - Descriptive statistics for patient, operation, and healthcare setting factors for included TKRs.

	Surgeon grade and supervision (n=953,081)			
Variable	**Consultant (n=856,537)**	**Trainee (overall) (n=96,544)**	**Trainee supervised by scrubbed consultant (n=61,048)**	**Trainee not supervised by scrubbed consultant (n=35,496)**
Mean age (SD) [years]	69.4 (9.2)	70.6 (8.9)	70.5 (8.9)	70.8 (8.7)
Age groups (%)				
<55	51,856 (6.1)	4,001 (4.1)	2,654 (4.4)	1,347 (3.8)
55-64	198,108 (23.1)	19,667 (20.4)	12,617 (20.7)	7,050 (19.9)
65-74	339,617 (39.7)	38,671 (40.1)	24,349 (39.9)	14,322 (40.4)
75-84	232,999 (27.2)	29,637 (30.7)	18,581 (30.4)	11,056 (31.2)
>85	33,957 (4.0)	4,568 (4.7)	2,847 (4.7)	1,721 (4.9)
Female (%)	488,057 (57.0)	55,688 (57.7)	35,082 (57.5)	20,606 (58.1)
Side (% right)	451,738 (52.7)	50,582 (52.4)	31,961 (52.4)	18,621 (52.5)
IMD decile (%)				
1-2 (most deprived)	121,773 (14.2)	15,709 (16.3)	10,497 (17.2)	5,212 (14.7)
3-4	150,768 (17.6)	18,198 (18.9)	11,397 (18.7)	6,801 (19.2)
5-6	186,143 (21.7)	21,167 (21.9)	13,226 (21.7)	7,941 (22.4)
7-8	199,725 (23.3)	21,309 (22.1)	13,220 (21.7)	8,089 (22.8)
9-10 (least deprived)	198,128 (23.1)	20,161 (20.9)	12,708 (20.8)	7,453 (21.0)
BMI (kg/m^2)				
<19 (underweight)	1,332 (0.2)	126 (0.1)	81 (0.1)	45 (0.1)
19-24.9 (normal)	57,186 (6.7)	5,657 (5.9)	3,620 (5.9)	2,037 (5.7)
25-29.9 (overweight)	197,325 (23.0)	19,836 (20.6)	13,172 (21.6)	6,664 (18.8)
>30 (obese)	332,930 (38.9)	36,095 (37.4)	24,077 (39.4)	12,018 (33.9)
Missing	267,764 (31.3)	34,830 (36.1)	20,098 (32.9)	14,732 (41.5)
ASA grade (%)				
ASA I	91,060 (10.6)	7,644 (7.9)	4,831 (7.9)	2,813 (7.9)
ASA II	625,871 (73.1)	67,980 (70.4)	42,506 (69.6)	25,474 (71.8)
ASA ≥III	139,606 (16.3)	20,920 (21.7)	13,711 (22.5)	7,209 (20.3)
Anaesthetic (%)				
Spinal	580,362 (67.8)	62,254 (64.5)	41,480 (68.0)	20,774 (58.5)
General	312,453 (36.5)	35,447 (36.7)	20,198 (33.1)	15,249 (43.0)
Epidural	46,068 (5.4)	7,242 (7.5)	4,143 (6.8)	3,099 (8.7)
Nerve block	122,671 (14.3)	17,145 (17.8)	9,606 (15.7)	7,539 (21.2)
Approach (%)				
Lateral parapatellar	7,301 (0.9)	689 (0.7)	343 (0.6)	346 (1.0)
Medial parapatellar	801,543 (93.6)	90,145 (93.4)	57,009 (93.4)	33,136 (93.4)
Mid-vastus	21,446 (2.5)	2,482 (2.6)	1,705 (2.8)	777 (2.2)
Sub-vastus	9,671 (1.1)	1,016 (1.1)	546 (0.9)	470 (1.3)
Other	16,576 (1.9)	2,212 (2.3)	1,445 (2.4)	767 (2.2)
Fixation (%)				
Cemented	814,854 (95.1)	92,896 (96.2)	59,009 (96.7)	33,887 (95.5)
Uncemented	34,999 (4.1)	2,820 (2.9)	1,681 (2.8)	1,139 (3.2)
Hybrid	6,684 (0.8)	828 (0.9)	358 (0.6)	470 (1.3)
Constraint (%)				
Constrained condylar	6,906 (0.8)	455 (0.5)	280 (0.5)	175 (0.5)
Monobloc poly tibia	12,902 (1.5)	2,664 (2.8)	2,151 (3.5)	513 (1.5)
Posterior stabilised, fixed	196,523 (22.9)	20,635 (21.4)	13,197 (21.6)	7,438 (21.0)
Posterior stabilised, mobile	10,007 (1.2)	943 (1.0)	670 (1.1)	273 (0.8)
Preassembled/hinged/linked	994 (0.1)	80 (0.1)	61 (0.1)	19 (0.1)
Unconstrained, fixed	579,257 (67.6)	68,013 (70.5)	41,960 (68.7)	26,053 (73.4)
Unconstrained, mobile	49,948 (5.8)	3,754 (3.9)	2,729 (4.5)	1,025 (2.9)
Patellar resurfacing (%)				
Patella resurfaced	333,035 (38.9)	33,309 (34.5)	21,037 (34.5)	12,272 (34.6)
Funding source (%)				
NHS	749,953 (87.6)	96,432 (99.9)	60,979 (99.9)	35,453 (99.9)
Private	106,584 (12.4)	112 (0.1)	69 (0.1)	43 (0.1)
Year of operation (%)				
2003-2011	301,105 (35.2)	41,112 (42.6)	21,949 (36.0)	19,163 (54.0)
2012-2019	555,432 (64.8)	55,432 (57.4)	39,099 (64.0)	16,333 (46.0)
ASA, American Society of Anaesthesiologists; NHS, National Health Service; IMD, Index of Multiple Deprivation; BMI, Body Mass Index. Data are n (%) or mean (SD); denoted where applicable.				

Figure 14 - Kaplan-Meier plots (one minus survival) demonstrating the cumulative probability of TKR failure (i.e. all-cause revision) according to surgeon grade (A) and supervision (B).

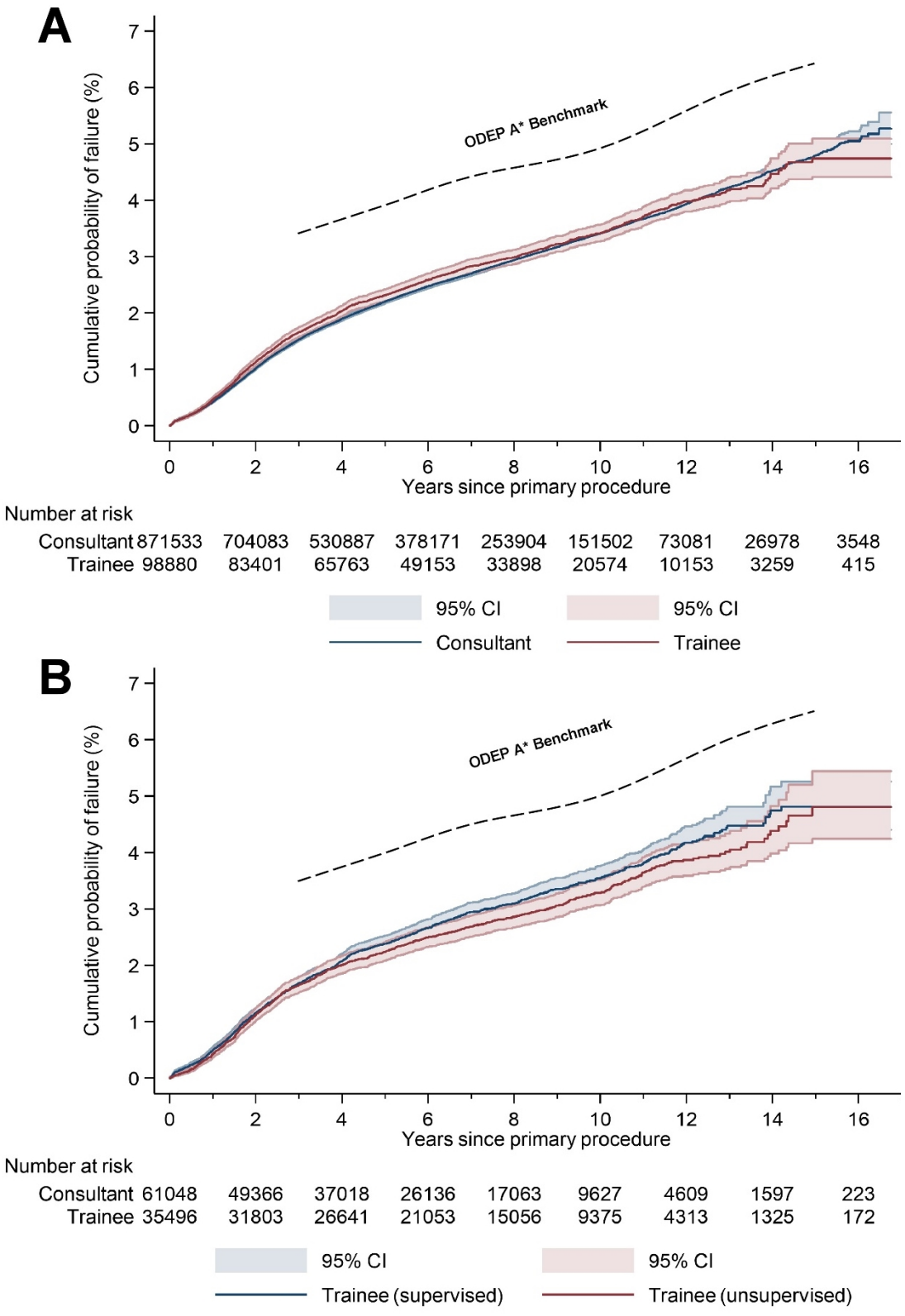


Table 8 - The unadjusted cumulative probability of all-cause failure of TKRs according to surgeon grade (exposure A) and supervision (exposure B).

Follow-up (years)	ODEP A* (%)	Consultant			Trainee (overall)			Trainee supervised by a scrubbed consultant			Trainee not supervised by a scrubbed consultant		
		Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)
1	N/A	856,537	3,460	0.42 (0.41 to 0.44)	96,544	434	0.47 (0.42 to 0.51)	61,048	287	0.49 (0.44 to 0.55)	35,496	147	0.42 (0.36 to 0.50)
3	3.5	689,676	7,756	1.53 (1.51 to 1.56)	81,169	993	1.67 (1.59 to 1.76)	49,366	598	1.68 (1.58 to 1.80)	31,803	395	1.65 (1.52 to 1.79)
5	4.0	517,132	3,580	2.21 (2.17 to 2.24)	63,659	430	2.33 (2.22 to 2.44)	37,018	268	2.39 (2.25 to 2.53)	26,641	162	2.24 (2.08 to 2.41)
7	4.5	365,234	1,892	2.70 (2.66 to 2.75)	47,189	249	2.84 (2.72 to 2.96)	26,136	151	2.95 (2.79 to 3.11)	21,053	98	2.69 (2.51 to 2.88)
10	5.0	188,465	1,602	3.42 (3.36 to 3.47)	25,143	174	3.34 (3.29 to 3.59)	13,038	92	3.54 (3.34 to 3.75)	12,105	82	3.29 (3.07 to 3.52)
13	6.0	64,200	701	4.23 (4.14 to 4.31)	8,922	103	4.25 (4.03 to 4.48)	4,609	55	4.44 (4.13 to 4.78)	4,313	48	4.02 (3.71 to 4.34)
15	6.5	24,663	153	4.79 (4.67 to 4.92)	2,922	17	4.75 (4.43 to 5.10)	1,597	6	4.74 (4.35 to 5.16)	1,325	11	4.76 (4.23 to 5.35)

Data are the number at risk, the number of revision events, the unadjusted cumulative probability of failure and the 95% CI. ODEP A* benchmark for comparison. *Number at risk at the beginning of interval.

4.5.2 All-cause revision: unadjusted Kaplan-Meier (KM) analysis

The unadjusted cumulative probability of all-cause failure according to surgeon grade (exposure A) and supervision (exposure B) is documented in Table 8 and displayed as a KM plot (one minus survival) in Figure 14. The Orthopaedic Data Evaluation Panel (ODEP) A* thresholds are presented alongside for comparison to an internationally recognised benchmark for implant component longevity.

The KM plot in Figure 14A shows that the cumulative probability of failure of trainee-performed TKRs follows a very similar trend to that of consultant-performed TKRs. There is a subtle divergence in the probability of failure between 1 and 4 years, with separation between the confidence intervals noted at the 3-year interval. However, the confidence intervals for consultant- and trainee-performed TKRs overlap at all other intervals of follow-up (Table 8). The upper confidence interval of the cumulative probability of failure of trainee-performed TKRs is below the ODEP A* threshold at all intervals of follow-up.

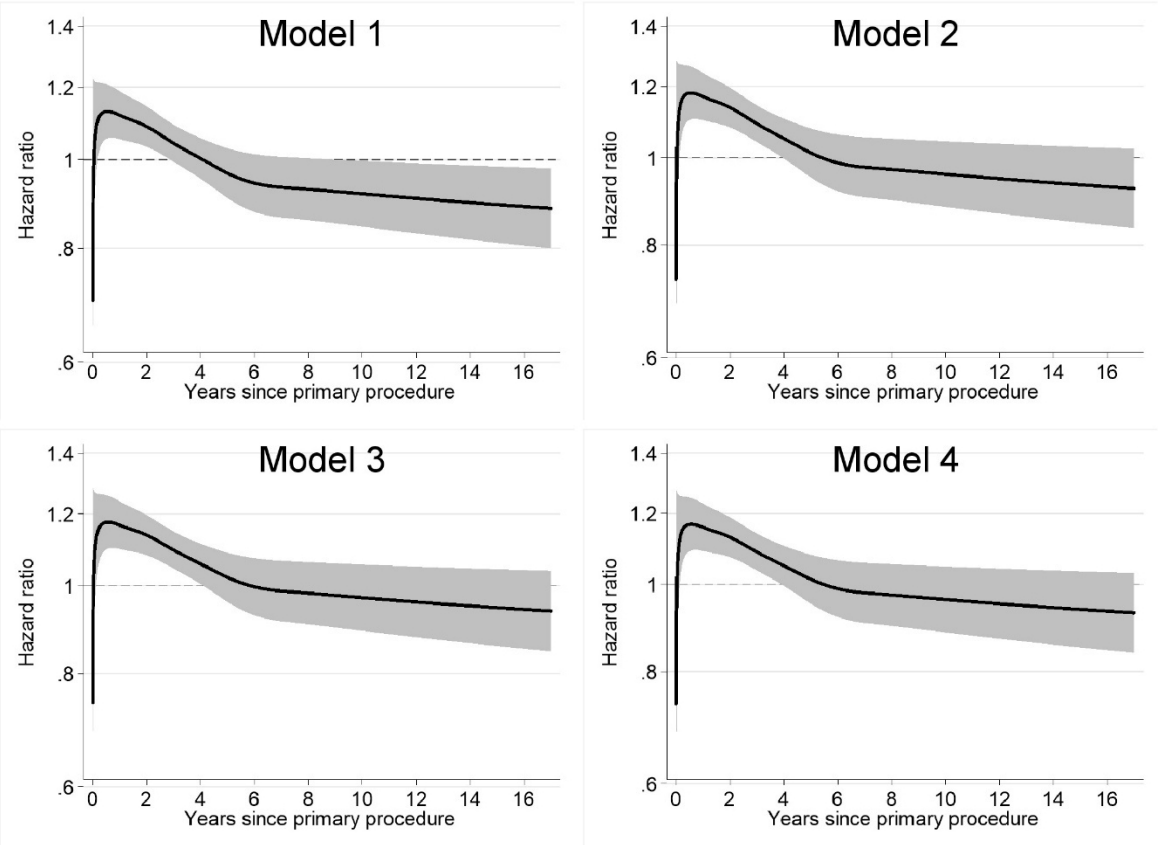
The KM plot in Figure 14B shows that the cumulative probability of failure of trainee-performed TKRs follows a similar trend, regardless of the level of scrubbed consultant supervision (exposure B). The upper confidence interval of both plots is below the ODEP A* threshold at all intervals of follow-up. The confidence intervals overlap at all time points (Figure 14B and Table 8), which suggests that there is no difference in the cumulative probability of failure between the two trainee groups.

An unadjusted sensitivity analysis demonstrates that surgeon grade has a relatively minor influence on implant survival compared to other operative factors, such as fixation, patellar resurfacing, and constraint (Appendix 15).

4.5.3 All-cause revision: flexible parametric survival modelling (FPM)

Surgeon grade had a significant time-dependent effect, and it is an oversimplification to report single numeric HR. The results of this analysis, which compare trainee-performed TKRs to a baseline of consultant-performed TKRs, are therefore displayed as HR plots (Figure 15).

Figure 15 - Risk of all-cause revision of TKRs according to surgeon grade (exposure A).



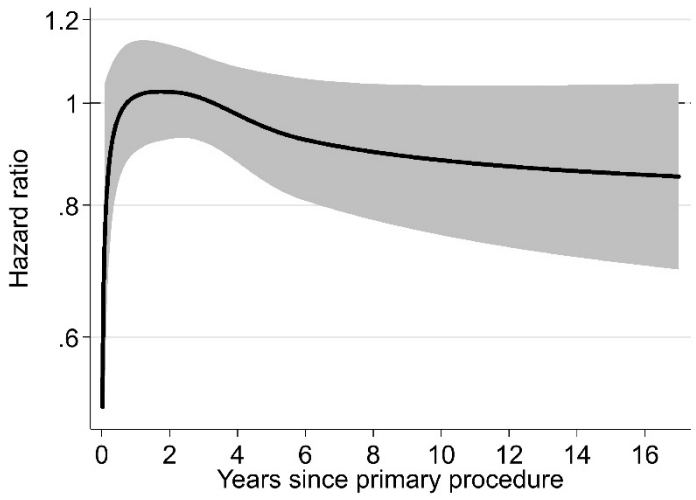
HR plots for incrementally adjusted flexible parametric survival models (FPMs), which show the risk of revision for trainee-performed TKR compared to baseline of consultant-performed TKR. (Model 1) Unadjusted. (Model 2) Adjusted for patient factors. (Model 3) Adjusted for patient and operation factors. (Model 4) Adjusted for patient, operation and healthcare-setting factors. The solid black line represents the HR, and the shaded grey area represents the 95% CI. Separation between the upper or lower 95% CI and the baseline of 1.00 (representing consultant-performed procedures) is suggestive of an association.

A near-linear reduction in HR was observed in the unadjusted model (Model 1) between 1- and 4-years follow-up, from 1.12 (95% CI 1.05 to 1.19) at 1 year to 1.00 (95% CI 0.95 to 1.06) at 4 years, with a flattening of risk beyond this point to 0.89 (95% CI 0.81 to 0.98) at 16 years. Despite extensive

adjustment for confounding factors, the marginal association between trainee-performed TKR and risk of revision in the first 4 years remains consistent. In the fully adjusted Model 4, the HR declines from 1.16 (95% CI 1.09 to 1.23) at 1 year to 1.05 (95% CI 0.99 to 1.10) at 4 years.

Further analysis was performed to compare the risk of revision of TKRs performed by trainees who were supervised by a scrubbed consultant to TKRs performed by trainees who were not supervised by a scrubbed consultant (exposure B). There was no evidence of an association between the level of scrubbed consultant supervision of trainee-performed TKRs and the risk of all-cause revision (Figure 16).

Figure 16 - Risk of all-cause revision of TKRs according to the level of supervision of trainees (exposure B).



HR plot for a fully adjusted flexible parametric survival model (Model 4). Represents TKRs performed by trainees supervised by a scrubbed consultant (baseline) compared to TKRs performed by trainees not supervised by a scrubbed consultant (exposure B).

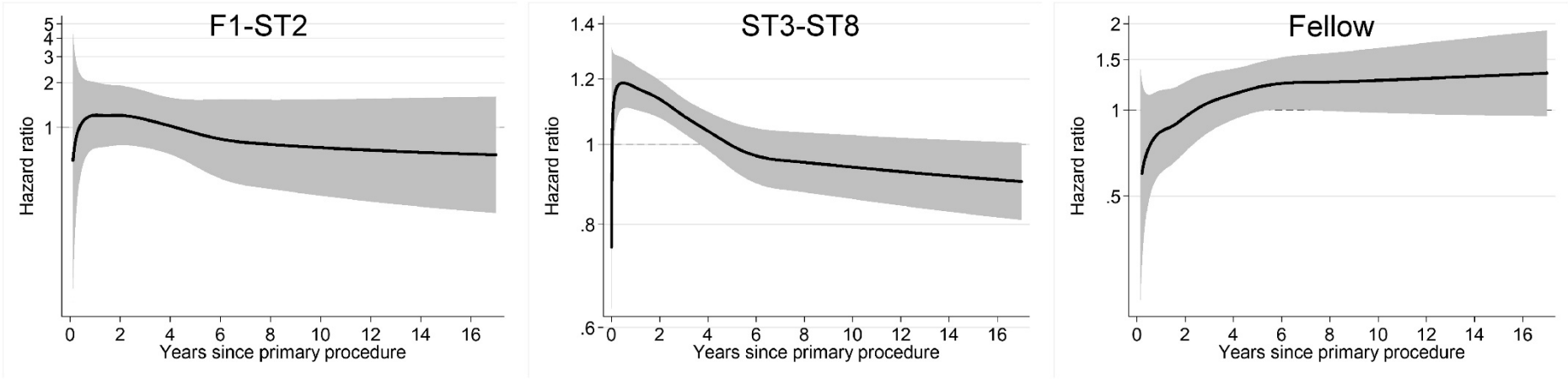
An additional analysis was performed following further subcategorisation of cases according to the specific training grade of the operating surgeon (exposure C). There was no evidence of an association between F1-ST2- or fellow-performed TKR and the risk of all-cause revision, noting that only a small proportion of trainee procedures were performed by surgeons of these grades (F1-ST2: 1,104/96,544 [1.1%]; Fellow: 3,301/96,544 [3.4%]) (Figure 17). The majority of trainee cases were

performed by ST3-ST8 surgeons (92,139/96,544 [95.4%]). As such, the HR plot for this group resembles our analysis for trainees overall (Figure 15).

4.5.4 Indication for revision: flexible parametric survival modelling (FPM)

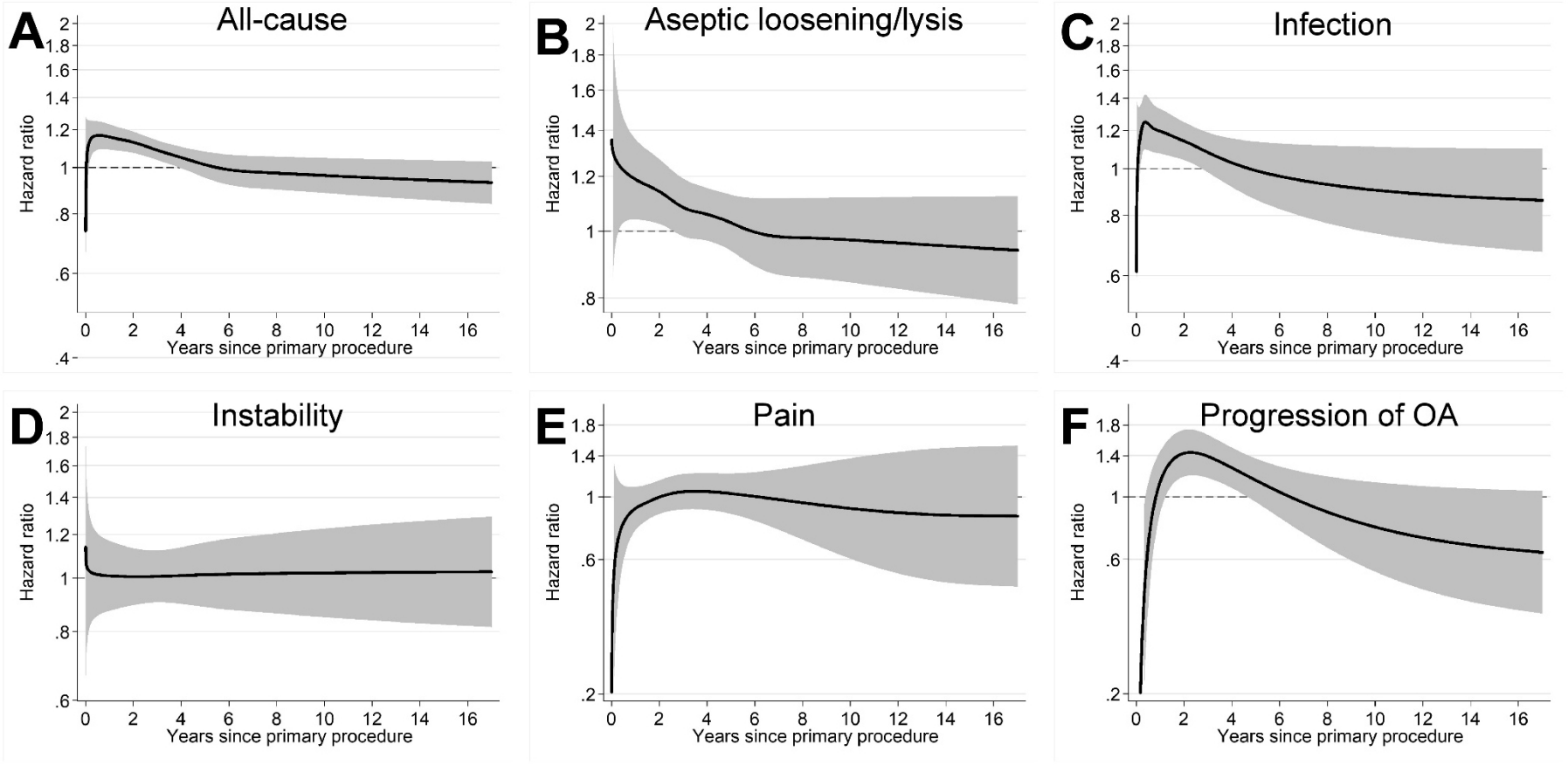
The three most common indications for revision in this cohort were aseptic loosening/lysis (n=6,244), infection (n=5,683) and instability (n=4,361). Fully adjusted models are displayed in Figure 18 and indicate marginal associations between trainee-performed TKR and early revision for aseptic loosening/lysis (up to 3 years), infection (up to 3 years), and progression of OA (up to 5 years). There was no evidence of an association between surgeon grade and revision for instability, or unexplained pain.

Figure 17 - Risk of all-cause revision of TKRs according to specific training grade (exposure C).



HR plot for fully adjusted flexible parametric survival models (Model 4). Risk of all-cause revision according to specific training grade (F1-ST2, ST3-ST8, Fellow) compared to baseline of consultant-performed TKR. F1-ST2: cases = 1,104; revisions = 29. ST3-ST8: cases = 92,139; revisions = 2,248. Fellow: cases = 3,301; revisions = 123.

Figure 18 - The indication for TKR revision according to surgeon grade (exposure A).



HR plots for fully adjusted flexible parametric survival models (Model 4). Each plot (A-F) represents the risk of revision for trainee-performed TKR compared to a baseline of consultant-performed TKR. (A) All-cause revision [consultant revisions = 19,172; trainee revisions = 2,400]. (B) Aseptic loosening/lysis [consultant revisions = 5,563; trainee revisions = 681]. (C) Infection [consultant revisions = 5,025; trainee revisions = 658]. (D) Instability [consultant revisions = 3,917; trainee revisions = 444]. (E) Unexplained pain [consultant revisions = 2,853; trainee revisions = 327]. (F) Progression of OA [consultant revisions = 1,597; trainee revisions = 219]. N.B. Some cases were revised for more than one indication.

4.6 Discussion

This study of nearly one million primary TKRs with over 16 years of follow-up provides novel insight into the association between surgeon grade, the supervision of trainees, and TKR survival. Our unadjusted KM analysis indicates that trainees achieve comparable outcomes to consultants, in terms of all-cause revision of TKRs. Trainees achieve implant survivorship within the ODEP A* threshold, which is an internationally recognised benchmark for the best performing implant components (Orthopaedic Data Evaluation Panel, 2021b). This supports the interpretation that trainees in England and Wales achieve safe and acceptable TKR implant survival. However, our adjusted FPM analyses identify areas for potential improvement in the outcomes of trainee-performed TKR.

The use of FPM gives insight into the temporal variation in the risk of revision of TKRs. Our adjusted FPM analyses demonstrate that trainee-performed TKRs may be associated with a small increased risk of all-cause revision up to, but not exceeding, 4 years after the index procedure. Further analysis indicates that this effect is not explained by the level of scrubbed consultant supervision and may be attributable to early revisions for aseptic loosening/lysis, infection, and progression of OA.

The adjusted analysis shown in Figure 16 shows that we found no evidence of an association between the level of supervision of trainees and the risk of all-cause TKR revision. This might be because, in general, trainees who perform TKR without scrubbed consultant supervision are more experienced than trainees who are directly supervised by a scrubbed consultant. However, we were unable to quantify this in the current study. In practice, any observed difference in the risk of revision between consultant- and trainee-performed TKRs is very small and surgeon grade has a relatively minor influence on implant survival compared to other operative factors, such as fixation, patellar resurfacing, and constraint (Appendix 15).

Performing TKR without patellar resurfacing is associated with an increased risk of revision, compared to TKR with patellar resurfacing (Hunt et al., 2021b). As such, patellar resurfacing is a cost-effective intervention, and National Institute for Health and Care Excellence (NICE) guidance suggests that it should be offered to patients undergoing an elective TKR (National Institue of Health

and Care Excellence, 2021). Our results show that a lower proportion of trainee-performed TKRs included patellar resurfacing compared to consultant-performed TKRs (34.5% vs. 38.9%), which may explain the early risk of revision for progression of OA that was observed in the trainee cohort. We propose that trainers should consider patellar resurfacing, particularly when the trainee performs the TKR.

NZJR data suggest that mean operative duration is longer for trainee-performed TKR compared to consultant-performed TKR (Storey et al., 2018). Large observational studies have identified an association between prolonged operative duration and prosthetic joint infection (Anis et al., 2019, Namba et al., 2013). While causation cannot be attributed, the prolonged operative times previously observed in trainee-performed TKR, might explain the early risk of revision for infection observed in Figure 18C. Operative duration is not recorded in the NJR, which precluded analysis of this variable in the current study.

4.6.1 Strengths and limitations

We analysed data for nearly one million knees recorded in the world's largest joint registry, which makes this the largest study of the association between surgeon grade and knee replacement outcomes (Fowler et al., 2021, Madanipour et al., 2021). We analysed data recorded in a mandatory national registry, which reduces sampling bias and improves the external validity and generalisability of our findings. Despite limiting our study period to predate the anomalous period of elective orthopaedic practice during the COVID-19 pandemic, our findings are current and represent TKRs with up to 16.8 years of follow-up. OA was the only indication, which in addition to comprehensive confounding adjustment, reasonably accounts for measurable variations in case complexity and case-mix selection between the groups in the context of this observational study. We used FPM to account for non-proportionality by modelling the time-dependent effect of surgeon grade on the hazard function. This methodology gives unique insight into the temporal variation in the risk of revision according to surgeon grade.

Despite the strengths of this study, it has limitations. The data are observational, and patients were not randomly allocated to the intervention groups; therefore, causation cannot be attributed. Furthermore, we used implant survival as a standalone outcome measure and did not consider other metrics that may be important when evaluating the successful outcome of a TKR, such as PROMs (Wylde and Blom, 2011).

The binary term 'surgeon grade' does not capture variations in the level of experience between trainees. We have attempted to account for this by recategorising cases according to the specific training grade of the surgeon (exposure C: F1-ST2; ST3-ST8; fellow); however, this categorical variable has similar limitations. Similarly, supervision is recorded by the NJR as a binary variable according to the grade of the first assistant, which does not capture the wider spectrum of supervision, e.g. the supervision of trainees by unscrubbed consultants (ISCP, 2021).

NJR data entry is audited on a rolling monthly basis to ensure data quality. Recent estimates suggest that over 96% of primary knee replacements and 91% of revision knee replacements are captured by the NJR (The National Joint Registry, 2020). However, there are likely to be limitations to the accuracy with which surgeon grade is recorded. For example, it is not currently possible to record more than one operating surgeon per case in the NJR, which does not account for procedures that have been part-performed by a trainee.

4.6.2 Comparison with other studies

The results of our unadjusted KM analysis are consistent with the findings of previous observational studies, which are summarised in a recent systematic review and meta-analysis on this subject. Evidence synthesis was performed on 936 TKRs and suggested net implant survival estimates at 10 years of 96.2% (95% CI 94.0 to 98.4) for trainee-performed TKRs and 95.1% (95% CI 93.0 to 97.2) for consultant-performed TKRs (Fowler et al., 2021); comparable to the ten-year net failure estimates reported here (Table 8).

A NZJR study of 79,671 TKRs, which reported revision rates per 100 component years rather than net survival, found no significant difference in the revision rate of TKRs performed by consultants

compared to TKRs performed by supervised and unsupervised trainees (Storey et al., 2018). While these findings are generally consistent with the results of our unadjusted KM analysis, the high case numbers and statistical methods used in our study give additional insight into previously unknown associations between surgeon grade and the risk of revision following TKR. The same proportion of TKRs were performed by trainees in the NJR (10%) and the NZJR (10%).

4.6.3 Implications

The findings of our unadjusted KM analyses suggest that trainees achieve comparable all-cause implant survival to consultants. Trainees achieve implant survivorship within the ODEP A* threshold, which indicates safe and acceptable implant survival for TKRs performed by trainees in England and Wales. In general, trainers select cases of appropriate complexity for their trainees and permit trainees to operate without scrubbed supervision when they have reached a threshold of expertise that was not quantifiable in the current study. However, our adjusted analyses suggest that trainee-performed TKRs may be susceptible to an increased risk of early revision for aseptic loosening, infection, and progression of OA. We recommend that consultants and trainees should recognise and take appropriate measures to reduce the risk of failure from these indications. Trainees should ideally be supervised by a scrubbed consultant when performing TKR, particularly when junior and at critical stages of the procedure such as implant selection, balancing, fixation, and deciding whether or not to resurface the patella.

4.6.4 Conclusion

This nationwide study of nearly one million TKRs demonstrates that trainees in England and Wales achieve safe and acceptable all-cause TKR implant survival, with comparable outcomes to consultants. However, we have identified areas for potential further improvement in trainee outcomes. Trainee-performed TKRs may be susceptible to early revision for some specific causes. Careful patient selection, measures to prevent infection, and surgical decisions such as routine patella resurfacing may mitigate against the small transient increase in risk of revision for infection, aseptic loosening, and progression of OA in trainee-performed TKRs.

Chapter 5 The association between surgeon grade and risk of revision following unicompartmental knee replacement: an analysis of National Joint Registry data

5.1 Overview

The findings of the systematic review and meta-analysis in Chapter 2 highlight the paucity of evidence on the survival of UKRs in the context of surgical training. This chapter uses NJR data to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following UKR. Similar to preceding chapters, we also explore the risk of revision for different indications following trainee-performed surgery. While similar modelling techniques have been used, inherent differences in the data dictate the need for an alternative approach to statistical analysis and presentation of the results.

5.1.1 Contributors

TF was responsible for study concept, design, literature review, data analysis, interpretation of the results, and writing the manuscript. AB, and MW were responsible for study concept, design, interpretation of the results, and review of the manuscript. AS was responsible for study concept, design, data analysis, interpretation of the results, and review of the manuscript.

5.1.2 Declaration of conflicts of interest

TF and AS have no conflicts of interest to declare. AB and MW report a contract held by the University of Bristol for membership of HQIP Lot 2 Statistical Analysis team for the National Joint Registry, and royalties from Taylor & Francis. MW also reports an independently conducted research grant held by the University of Bristol with CeramTec, teaching payments from Heraeus, and research committee membership on the British Orthopaedic Association and the British Hip Society.

5.2 Abstract

5.2.1 Background

Unicompartmental knee replacements (UKRs) are performed by surgeons at various stages in training with varying levels of supervision, but we do not know if this is a safe practice with comparable outcomes to consultant-performed UKR. We used NJR data to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following UKR.

5.2.2 Methods

We included 106,206 cases in 91,626 patients, who underwent primary UKR for osteoarthritis (OA) between 2003 and 2019. Exposures were surgeon grade (consultant or trainee), and the level of scrubbed consultant supervision of trainees. The primary outcome was all-cause revision, and the secondary outcome was the number of procedures revised for the following indications: aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability. Flexible parametric survival models were adjusted for patient, operation, and healthcare setting factors.

5.2.3 Results

Trainees performed 4,382 (4.1%) UKR procedures and were supervised by a scrubbed consultant in 66.1% of cases. There was no association between surgeon grade and all-cause revision in either crude or adjusted models (adjusted HR 1.01, 95% CI 0.90 to 1.13; p=0.88). Trainees achieved comparable all-cause survival to consultants, regardless of the level of scrubbed consultant supervision (supervised: adjusted HR 0.99, 95% CI 0.87 to 1.14; p=0.94; unsupervised: adjusted HR 1.03, 95% CI 0.87 to 1.22; p=0.74). Unsupervised trainee cases were associated with an increased risk of revision for unexplained pain compared to consultant-performed UKRs, in all but the fully adjusted model (adjusted HR 1.34, 95% CI 0.95 to 1.89; p=0.09).

5.2.4 Conclusion

In the context of current orthopaedic training practices in England and Wales, trainees achieve comparable all-cause UKR survival to consultants.

5.3 Background

Unicompartmental knee replacement (UKR) is an alternative to total knee replacement (TKR) in patients with symptomatic osteoarthritis (OA) isolated to a single compartment (Price et al., 2018). Proposed advantages of UKR over TKR include superior functional outcomes, reduced length of stay, fewer medical complications, greater cost-effectiveness, and lower mortality (Wilson et al., 2019, Beard et al., 2019). However, UKR revision rates are considerably higher than primary TKR revision rates (Evans et al., 2019b, Hunt et al., 2021a). Previous studies have suggested that UKRs performed by low-volume surgeons are associated with an increased risk of revision compared to UKRs performed by experienced higher volume surgeons (Baker et al., 2013, Liddle et al., 2016). This raises the question of whether or not it is safe for these procedures to be performed by trainees.

The survival of a joint replacement, defined as the absence of revision surgery over time, is the principal metric used for comparing the longevity of implant components and is a commonly used measure of surgical performance. Our current understanding of the survival of UKRs in the context of surgical training is based on a small number of observational studies (Fowler et al., 2021).

Bottomley et al. conducted an observational study of 1,084 UKRs, of which 673 (62.1%) were performed by trainees. They reported no significant difference in implant survival between the groups, with 9-year cumulative survival estimates of 93.9% and 93.0% for consultants and trainees, respectively (Bottomley et al., 2016). A New Zealand Joint Registry (NZJR) study of 8,854 UKRs, of which 304 (3.4%) were performed by trainees, reported no difference in the revision rates of supervised senior trainees compared to attending surgeons (Storey et al., 2018). This study did not report survival estimates and the overall number of trainee cases in the cohort was insufficient to facilitate meaningful comparison between the supervised and unsupervised trainee groups. The survival of UKRs according to surgeon grade and supervision remains poorly understood. It is not clear if current training practices are safe, or whether trainees achieve comparable outcomes to consultants.

The aim of this research was to use National Joint Registry (NJR) data from England and Wales to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following UKR.

5.4 Methods

5.4.1 Patients and data sources

We performed an observational study using prospectively collected data recorded in the NJR. The initial NJR dataset was 1,502,564 linked knee procedures performed between 1 April 2003 and 31 December 2019. We included primary UKRs in adult patients (aged ≥18 years) performed for an indication of OA only. Patellofemoral joint replacements were excluded. Cases were included if the operating surgeon grade was recorded as any of the following: Foundation Year 1 (F1) to Specialty Trainee Year 2 (ST2); ST3-ST8; fellow; or consultant. The process of mapping grade classifications to account for variations in terminology used in different versions of the NJR Minimum Data Set (MDS) form is outlined in Appendix 10.

5.4.2 Data access and processing

Access to the data was facilitated under existing NJR permissions for the study of training and volume in the context of hip and knee replacement, which are held by a supervisor of this thesis and senior author of this study (AS) (The National Joint Registry, 2022a). The base dataset used in the current study is based on the same cut of NJR data that is used in the 17th Annual Report (The National Joint Registry, 2020). The steps taken in data processing and are summarised in the study flow diagram in Figure 19 and illustrated in greater detail in Figure 20. Figure 20 gives details of sequential exclusions from the base dataset, some of which were performed by the authors of the 17th Annual Report prior to the initiation of the current study. All exclusions are consistent with the exclusion criteria of this study and the stage at which these occurred is clearly documented. Complete-case analysis was used in all analyses and records with missing data in any confounding variable field used in subsequent statistical models were excluded from the relevant model. Details of missing data, the number of cases

excluded, and the reasons for exclusion are documented in Figure 21. The demographic characteristics of included cases and cases excluded due to missing data are summarised in Appendix 16.

Figure 19 - Study flow diagram (Chapter 5).

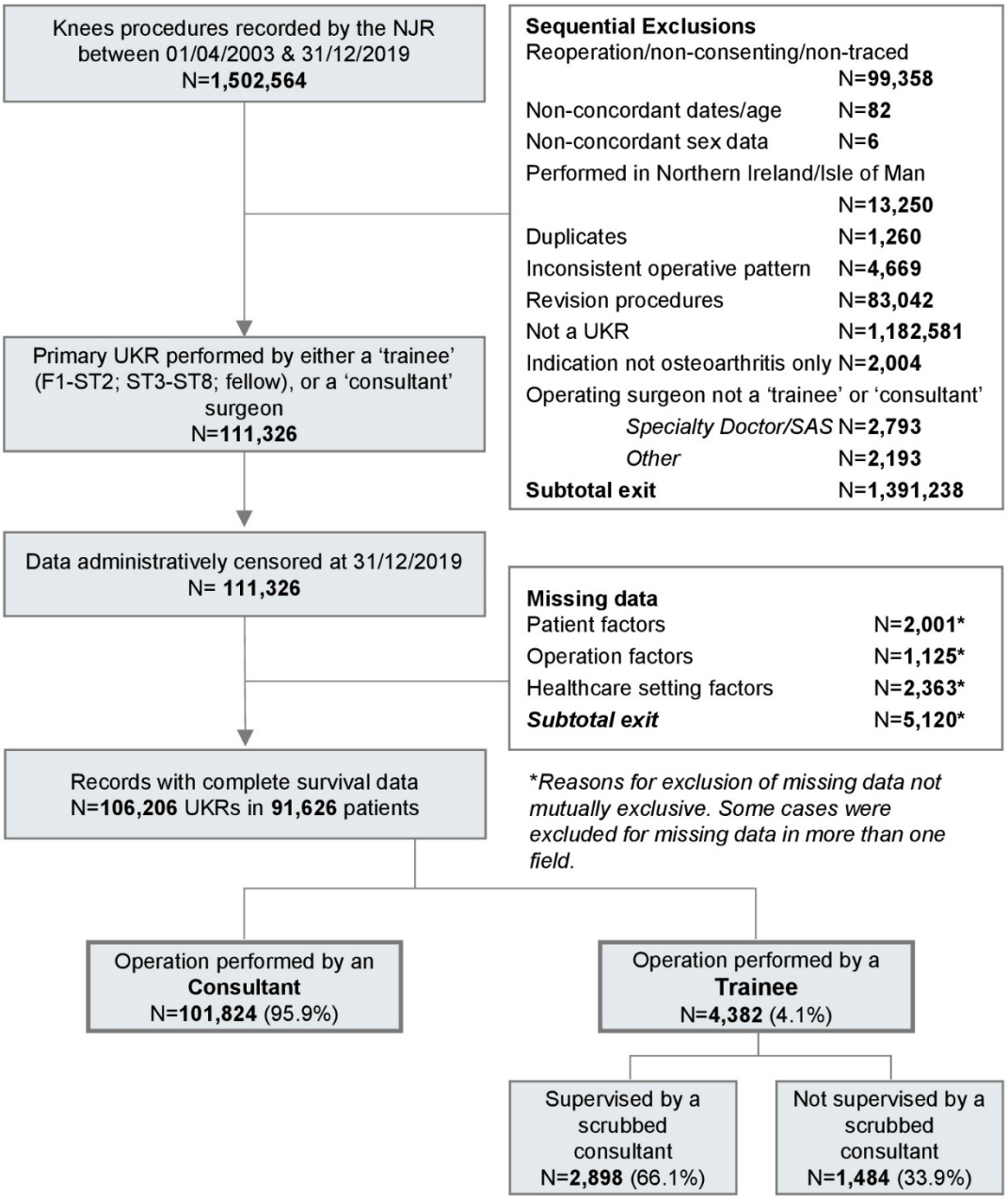
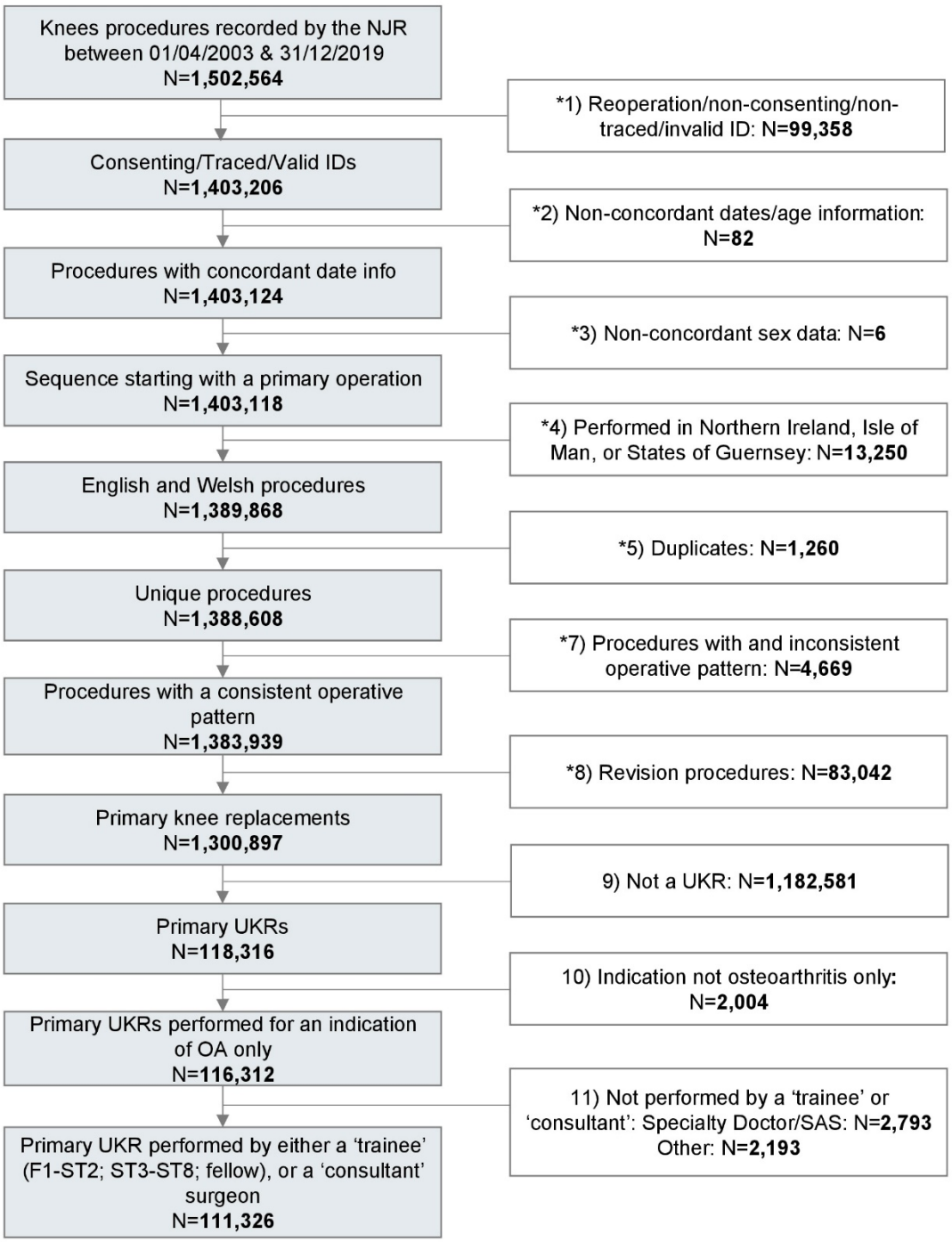
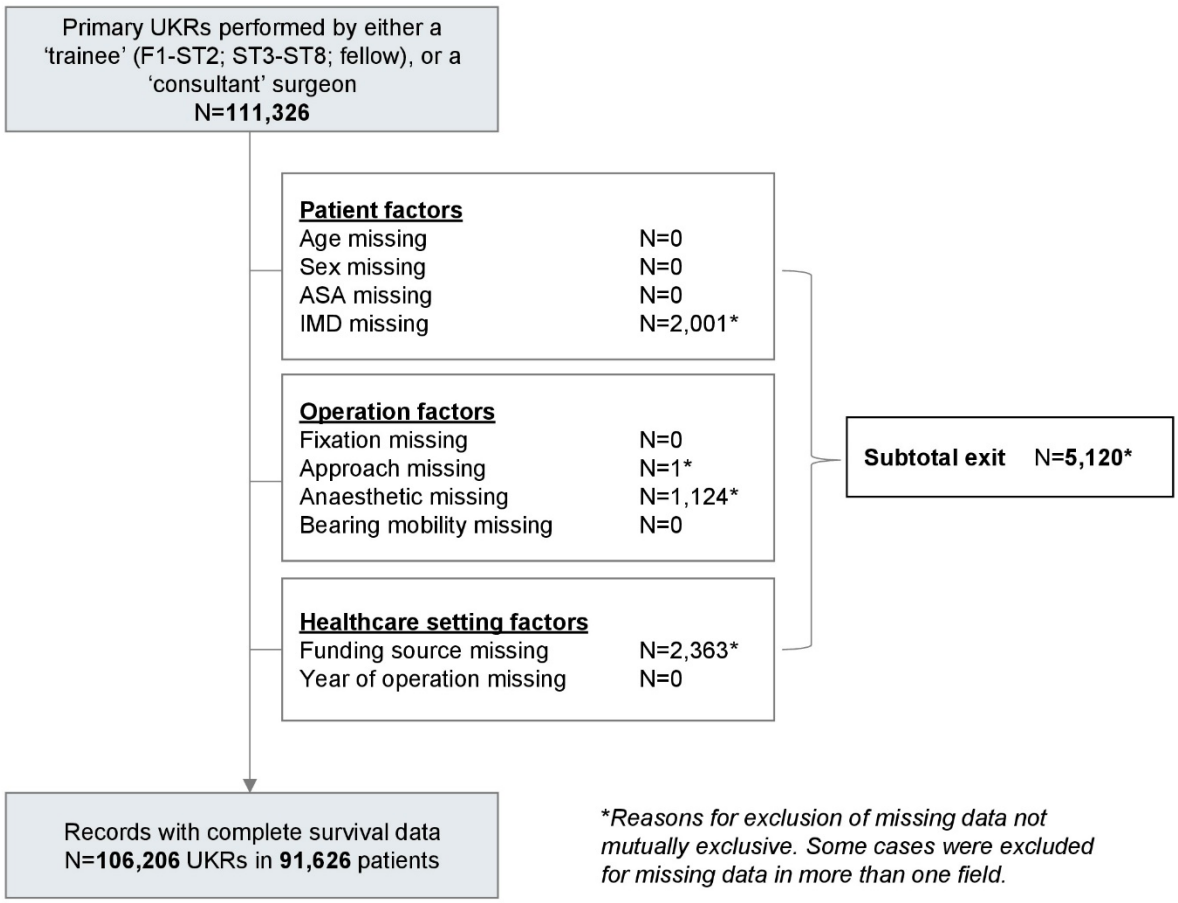


Figure 20 - Detailed study flow diagram showing sequential exclusions (Chapter 5).



***NB:** *Sequential exclusions 1 to 8 had already been applied to the base dataset by the authors of the NJR annual report, prior to the initiation of this study. Steps 9 to 11 were completed by TF.*

Figure 21 - Detailed study flow diagram showing exclusion of missing data (Chapter 5).



5.4.3 Exposures

The primary exposure (exposure A) was surgeon grade. This is a binary variable, which was categorised according to the grade of the operating surgeon: 1) consultant, or 2) trainee. Procedures performed by surgeons of the following grades were categorised under the variable 'trainee': F1-ST2; ST3-ST8; and fellow. Consultants have completed their formal training in orthopaedic surgery and been appointed to a senior position in which they can practice independently and supervise trainees. The term 'consultant' is synonymous with 'attending', and the term 'registrar' is synonymous with 'resident' in many healthcare settings including the USA.

The secondary exposure was whether or not trainees were supervised by a scrubbed consultant during the procedure (exposure B). Therefore, trainee cases were subcategorised as follows: 1) trainee supervised by a scrubbed consultant, or 2) trainee not supervised by a scrubbed consultant. Cases were categorised as 'supervised by a scrubbed consultant' if the first assistant was recorded as a consultant.

Given the variability in the level of experience between individual trainees, we performed a sensitivity analysis by recategorising cases according to the specific training grade of the operating surgeon (exposure C: consultant; F1-ST2; ST3-ST8; or fellow). Cases were further subcategorised according to the level of scrubbed consultant supervision.

5.4.4 Outcomes of interest

The primary outcome was all-cause revision, which was defined as any procedure to add, remove, or modify one or more components of an implant construct for any reason (The National Joint Registry, 2020). The secondary outcome measure was the number of procedures revised for the following specific indications, which are listed as the five most common indications for knee replacement revision by the NJR: aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability (The National Joint Registry, 2020).

5.4.5 Statistical analysis

5.4.5.1 Overview

Frequencies and percentages were used to describe categorical variables. The mean, standard deviation (SD), and interquartile range (IQR) were used to describe continuous variables. Unrevised cases were either administratively censored on 31 December 2019, or the date of death, depending on which was earliest. Unadjusted estimates of net implant failure were calculated using the Kaplan-Meier (KM) method.

We performed a comprehensive exploratory analysis using Cox regression. A combination of graphical plots, Schoenfeld residuals, and likelihood ratio testing (comparing proportional and non-proportional hazards models) were used to assess the proportional hazards (PH) assumption at each level of adjustment and to assess the time-dependent effects of each confounding variable (Royston and Lambert, 2011). Adjusted analyses did not satisfy the PH assumption, which was due to the time-dependent effects of multiple confounding variables included in the models (age, sex, IMD, approach, fixation, bearing mobility, year of operation, and funding source). Surgeon grade (exposures A) did not demonstrate a time-dependent effects.

To account for non-proportionality, we used flexible parametric survival modelling (FPM) (Lambert and Royston, 2009a, Royston and Lambert, 2011), which has been used in previous NJR analyses (Smith et al., 2012b, Blom et al., 2021, Blom et al., 2020, Hunt et al., 2018). This method uses restricted cubic spline functions to model the baseline hazard and account for the time-dependent effects of specified variables. Graphical assessment, AIC, BIC, and likelihood ratio testing were used to optimise the fit and complexity of the final model (Royston and Lambert, 2011). This process of model selection, construction and justification is described in greater detail in Section 5.4.5.3.

We performed separate analyses for all-cause revision and the five specific indications for revision (aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability), which were examined as separate survival endpoints. Separate FPM analyses were performed for each exposure

and analyses were incrementally adjusted for confounding variables. All analyses were performed using Stata (Version SE 15.1; StataCorp LP, USA).

5.4.5.2 Confounding variables

Analyses were adjusted for categorical confounding variables in the following manner. Model 1 was unadjusted. Model 2 was adjusted for patient-level factors (age, sex, American Society of Anaesthesiologists [ASA] grade, and index of multiple deprivation [IMD] decile). Model 3 was further adjusted for operation-level factors (anaesthetic, approach, fixation, and bearing mobility). Model 4 was further adjusted for healthcare setting factors (funding source, and year of operation). In each case, the baseline category was the most frequently occurring (as detailed in Appendix 17).

Patient factors included age, sex, ASA grade, and IMD decile. Age was categorised into five categories (<55; 55-64; 65-74; 75-84; and ≥85 years). ASA grade is a five-category system for the assessment of a patient's preoperative fitness according to the severity of medical comorbidities (Dripps et al., 1961). Patients are categorised according to the following definitions. ASA I – a normal healthy patient; ASA II – a patient with mild systemic disease; ASA III – a patient with severe systemic disease; ASA IV – a patient with severe systemic disease that is a constant threat to life; ASA V – a moribund patient who is not expected to survive without the operation. In this study, the following three categories were used: ASA I, ASA II and ASA ≥III. The index of multiple deprivation (IMD) is a geographical area-based measure of relative social deprivation. In England and Wales, small areas are ranked according to seven domains of deprivation (income, employment, education, health, crime, housing and services, and living environment). Models were adjusted for IMD decile (1 being the most deprived, 10 being the least deprived) (Noble et al., 2019).

Operation-level factors included anaesthetic, approach, fixation, and bearing mobility. Cases were categorised according to whether or not the patient had any of the following forms of anaesthesia: general anaesthetic, nerve block, epidural, or spinal. The surgical approach reflects the anatomical approach through which a joint is replaced. The following categories were used: lateral parapatellar, medial parapatellar, mid-vastus, sub-vastus, and 'other'. Fixation describes the method by which

implant components are fixed to the bone, and cases were categorised as follows: cemented, uncemented, or hybrid. An additional operative confounding factor was bearing mobility, which describes whether or not the polyethylene insert is fixed to the tibial base plate. These operative confounding factors are described in more detail in Section 1.3.3.2.

Healthcare setting factors included the funding source and year of operation. Cases were categorised according to whether the procedure was privately funded or funded by the National Health Service (NHS). Finally, cases were categorised according to the year of operation: either 2003-2011, or 2012-2019.

5.4.5.3 Model selection, construction and justification

The structured approach to model selection and construction used in this study was based around an in-depth assessment of the proportionality of hazard functions. Where data did not satisfy the PH assumption (i.e. the ratio of hazard functions was not constant) and further analysis demonstrated that FPM was superior to the Cox model, FPM was preferred for primary analysis. This approach is summarised here:

Step 1: Schoenfeld residuals tests were applied to incrementally adjusted Cox models for 'surgeon grade' to assess the PH assumption.

1) Cox PH model for surgeon grade (Model 1: unadjusted)
    a. Schoenfeld residual test: p=0.143
2) Cox PH model for surgeon grade (Model 2: adjusted for patient factors)
    a. Schoenfeld residual test: p<0.001
3) Cox PH model for surgeon grade (Model 3: adjusted for patient & operation factors)
    a. Schoenfeld residual test: p<0.001
4) Cox PH model for surgeon grade (Model 4: adjusted for patient, operation & healthcare factors)
    a. Schoenfeld residual test: p<0.001

Schoenfeld residuals tests suggested that 'surgeon grade' did not have a significant time-dependent effect in this context. The lack of proportionality in adjusted Cox models was explained by the time-dependent effects of confounding variables. A range of methods were employed to investigate this further, including Schoenfeld residuals test/plots, and likelihood ratio testing comparing PH and non-PH models for each variable. We found evidence that the following variables had time-dependent effects: age; sex; IMD decile; approach; fixation; bearing mobility; funder and year of operation.

Step 2: The following FPMs were constructed to model the hazard function of 'surgeon grade':

1) A non-proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors)
    a. A non-PH model using the stpm2 command in Stata, with the following variables specified as having a time-dependent effect: age, sex, IMD decile, approach, fixation, bearing mobility, funder, and year of operation.
    b. Graphical assessment, Akaike information criteria (AIC), and Bayes information criteria (BIC) were used to optimise the fit and complexity of the model. These methods were used to determine the degrees of freedom with which to model hazard functions, as well as the optimal number and location of knots.
    c. We confirmed the superiority of our final model to preceding iterations using likelihood ratio testing.
    d. The final model was as follows: The baseline hazard was modelled with 6 degrees of freedom (df). 'Surgeon grade' was modelled with fixed effects. The following variables were found to have time-dependent effects and the degrees of freedom used for each is denoted: age (1 df); sex (1 df); IMD decile (1 df); approach (1 df); fixation (3 df); bearing mobility (2 df); funder (3 df); year of operation (2 df). Remaining confounding variables were modelled with fixed effects.
2) A proportional hazards FPM (Model 4: adjusted for patient, operation & healthcare factors):
    a. A PH model using the stpm2 command in Stata, equivalent to the Cox model.
    b. The same as the non-PH FPM above, but with no time-dependent effects.

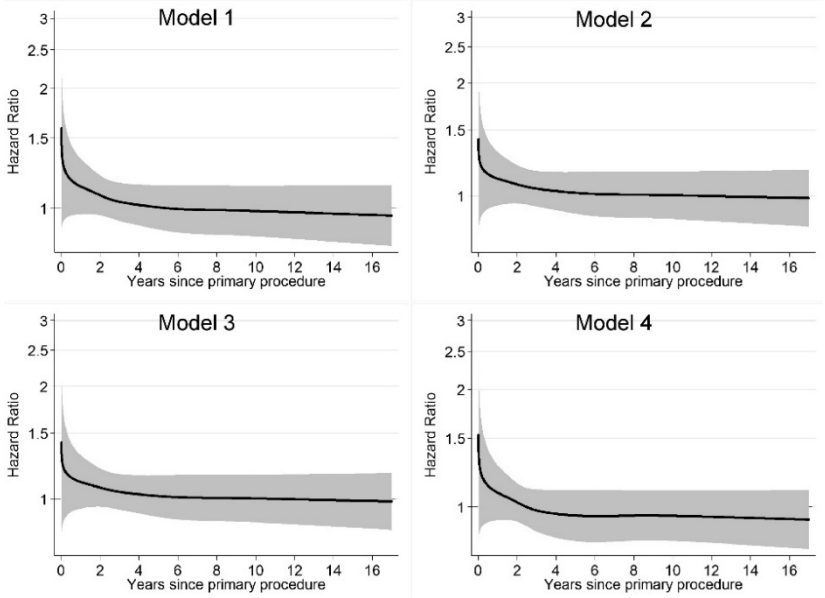
Step 3: Likelihood ratio tests were used to compare PH and non-PH models at each level of adjustment:

1) PH FPM compared to non-PH FPM (Model 1): N/A
2) PH FPM compared to non-PH FPM (Model 2): $p<0.001$
3) PH FPM compared to non-PH FPM (Model 3): $p<0.001$
4) PH FPM compared to non-PH FPM (Model 4): $p<0.001$

There was strong evidence to support the superiority of non-PH FPMs over PH models (i.e. Cox).

Step 4: Hazard ratio (HR) plots were produced using non-PH FPMs at each level of adjustment (Models 1-4) to graphically represent the hazard function for 'surgeon grade' (Figure 22). Figure 22 shows that there is no significant deflection of the HR above or below the baseline of one with any level of adjustment. Therefore, it was appropriate to quote numeric HRs rather than display HR plots, as in Chapter 4.

Figure 22 – HR plots for non-PH FPMs at each level of adjustment. The risk of all-cause revision of UKRs according to surgeon grade (exposure A).



Summary: We found strong evidence to support the superiority of non-PH FPMs over PH models (i.e. Cox) in this context. It was appropriate to present the results as numeric HRs.

5.5 Results

5.5.1 Descriptive analysis

We included 106,206 UKR procedures in 91,626 patients, of which 4,382 (4.1%) were performed by trainees. Trainees performed 5.3% of NHS-funded cases and were supervised by a scrubbed consultant in 66.1% (n=2,898) of trainee-performed cases (Table 9 and Figure 19).

The mean age of patients operated on by trainees was 1.7 years older than patients operated on by consultants (65.5 vs. 63.8 years). Trainees operated on a lower proportion of ASA I patients (15.7% vs. 21.3%) and a higher proportion of ASA ≥III patients (13.4% vs. 8.4%). A higher proportion of trainee procedures utilised uncemented implants (23.6% vs. 19.9%) and a mobile bearing (72.0% vs. 60.9%) (Table 9).

The maximum duration of follow-up was 16.8 years. Mean follow-up was 6.5 years (SD 4.3; IQR 2.6 to 10.1 years) for trainee UKRs and 5.6 years (SD 4.00; IQR 2.2 to 8.6 years) for consultant UKRs. A total of 6,920 UKRs were revised at a mean of 4.3 years (SD 3.5; IQR 1.5 to 6.5 years).

Table 9 - Descriptive statistics for patient, operation, and healthcare setting factors for included UKRs.

Variable	Surgeon grade and supervision (n=106,206)			
	Consultant (n=101,824)	**Trainee (overall) (n=4,382)**	**Trainee supervised by a scrubbed consultant (n=2,898)**	**Trainee not supervised by a scrubbed consultant (n=1,484)**
Mean age (SD)	63.8 (9.7)	65.5 (9.6)	65.4 (9.7)	65.7 (9.5)
Age groups (%)				
<55	18,562 (18.2)	594 (13.6)	408 (14.1)	186 (12.5)
55-64	35,656 (35.0)	1,416 (32.3)	938 (32.4)	478 (32.2)
65-74	33,057 (32.5)	1,541 (35.2)	1,014 (35.0)	527 (35.5)
75-84	13,132 (12.9)	745 (17.0)	474 (16.4)	271 (18.3)
>85	1,417 (1.4)	86 (2.0)	64 (2.2)	22 (1.5)
Female (%)	46,972 (46.1)	2,105 (48.0)	1,410 (48.7)	695 (46.8)
Side (%)				
Right	50,989 (50.1)	2,138 (48.8)	1,396 (48.2)	742 (50.0)
IMD decile (%)				
1-2 (most deprived)	9,778 (9.6)	496 (11.3)	356 (12.3)	140 (9.4)
3-4	14,969 (14.7)	712 (16.3)	501 (17.3)	211 (14.2)
5-6	22,054 (21.7)	886 (20.2)	588 (20.3)	298 (20.1)
7-8	25,655 (25.2)	1,013 (23.1)	660 (22.8)	353 (23.8)
9-10 (least deprived)	29,368 (28.8)	1,275 (29.1)	793 (27.4)	482 (32.5)
BMI (kg/m^2)				
<19 (underweight)	140 (0.1)	5 (0.1)	3 (0.1)	2 (0.1)
19-24.9 (normal)	8,049 (7.9)	264 (6.0)	190 (6.6)	74 (5.0)
25-29.9 (overweight)	27,948 (27.5)	1,032 (23.6)	699 (24.1)	333 (22.4)
>30 (obese)	37,431 (36.8)	1,571 (35.9)	1,101 (38.0)	470 (31.7)
Missing	28,256 (27.8)	1,510 (34.5)	905 (31.2)	605 (40.8)
ASA grade (%)				
ASA I	21,663 (21.3)	686 (15.7)	466 (16.1)	220 (14.8)
ASA II	71,562 (70.3)	3,107 (70.9)	2,021 (69.7)	1,086 (73.2)
ASA ≥III	8,599 (8.4)	589 (13.4)	411 (14.2)	178 (12.0)
Anaesthetic (%)				
Spinal	57,928 (56.9)	2,193 (50.1)	1,544 (53.3)	649 (43.7)
General	47,812 (47.0)	2,164 (49.4)	1,380 (47.6)	784 (52.8)
Epidural	4,290 (4.2)	331 (7.6)	174 (6.0)	157 (10.6)
Nerve block	16,847 (16.6)	948 (21.6)	607 (21.0)	341 (23.0)
Approach (%)				
Lateral parapatellar	3,310 (3.3)	111 (2.5)	89 (3.1)	22 (1.5)
Medial parapatellar	90,149 (88.5)	3,982 (90.9)	2,593 (89.5)	1,389 (93.6)
Mid-vastus	3,968 (3.9)	131 (3.0)	109 (3.8)	22 (1.5)
Sub-vastus	1,595 (1.6)	44 (1.0)	29 (1.0)	15 (1.0)
Other	2,802 (2.8)	114 (2.6)	78 (2.7)	36 (2.4)
Fixation (%)				
Cemented	79,206 (77.8)	3,208 (73.2)	2,123 (72.2)	1,085 (73.1)
Uncemented	20,209 (19.9)	1,036 (23.6)	666 (23.0)	370 (24.9)
Hybrid	2,409 (2.4)	138 (3.2)	109 (3.8)	29 (2.0)
Bearing mobility (%)				
Fixed	34,268 (33.7)	912 (20.8)	695 (24.0)	217 (14.6)
Mobile	62,011 (60.9)	3,153 (72.0)	1,978 (68.3)	1,175 (79.2)
Monobloc poly tibia	5,545 (5.5)	317 (7.2)	225 (7.8)	92 (6.2)
Funding source (%)				
NHS	77,595 (76.2)	4,370 (99.7)	2,893 (99.8)	1,477 (99.5)
Private	24,229 (23.8)	12 (0.3)	5 (0.2)	7 (0.5)
Year of operation (%)				
2003-2011	35,054 (34.4)	2,080 (47.5)	1,241 (42.8)	839 (56.5)
2012-2019	66,770 (65.6)	2,302 (52.5)	1,657 (57.2)	645 (43.5)

ASA, American Society of Anaesthesiologists; NHS, National Health Service; IMD, Index of Multiple Deprivation; BMI, Body Mass Index. Data are n (%) or mean (SD); denoted where applicable.

Figure 23 - Kaplan-Meier plot (one minus survival) demonstrating the cumulative probability of UKR failure (i.e. all-cause revision) according to surgeon grade (exposure A).

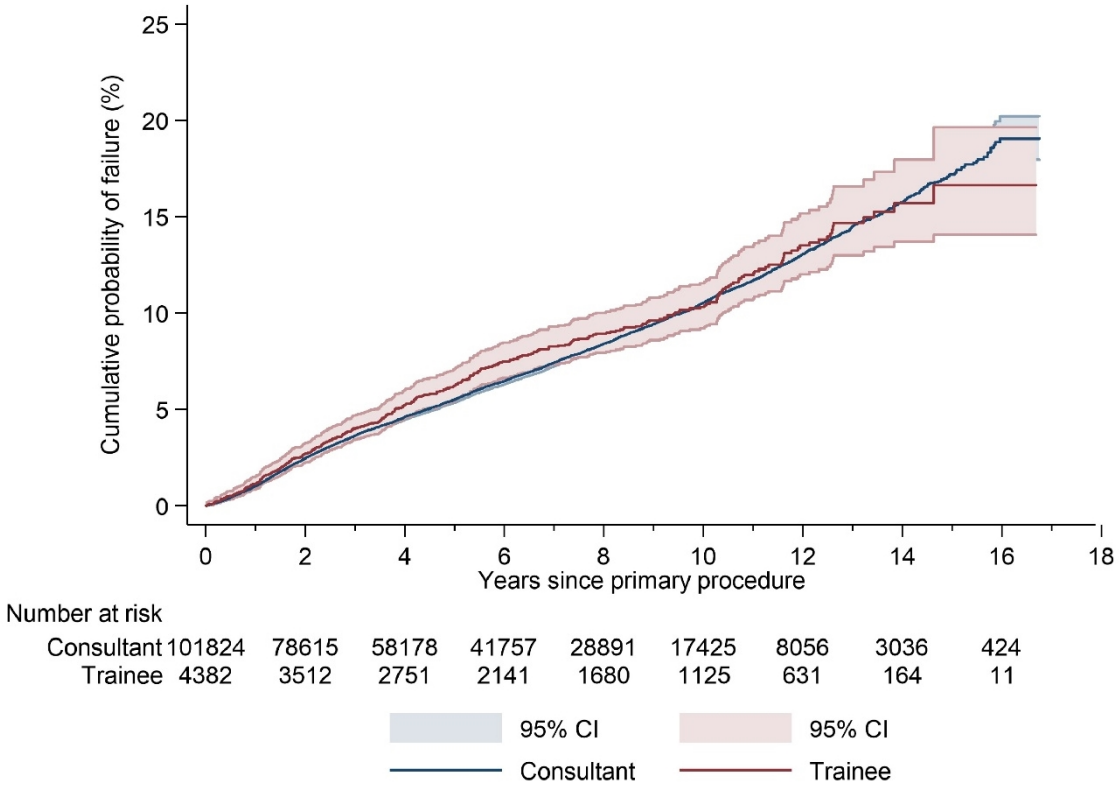


Table 10 - The unadjusted cumulative probability of all-cause failure of UKRs according to surgeon grade (exposure A) and supervision (exposure B).

Follow-up (years)	Consultant			Trainee (overall)			Trainee supervised by a scrubbed consultant			Trainee not supervised by a scrubbed consultant		
	Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)	Number at risk*	Number of revisions	% Failure (95% CI)
1	101,824	986	1.02 (0.96 to 1.10)	4,382	47	1.12 (0.84 to 1.49)	2,898	29	1.05 (0.73 to 0.15)	1,484	18	1.26 (0.80 to 2.00)
3	90,264	2,121	3.63 (3.51 to 3.76)	3,954	105	4.01 (3.43 to 4.68)	2,605	64	3.74 (3.06 to 4.57)	1,349	41	4.52 (3.52 to 5.80)
5	67,809	1,150	5.50 (5.34 to 5.67)	3,119	65	6.22 (5.46 to 7.08)	2,009	47	6.29 (5.34 to 7.40)	1,110	18	6.16 (4.95 to 7.65)
7	49,530	867	7.43 (7.23 to 7.64)	2,475	49	8.28 (7.36 to 9.32)	1,532	24	7.92 (6.80 to 9.21)	943	25	8.93 (7.34 to 10.77)
10	35,199	874	10.52 (10.24 to 10.81)	1,916	35	10.35 (9.23 to 11.59)	1,180	21	10.10 (8.70 to 11.71)	736	14	10.86 (9.08 to 12.96)
13	17,425	461	14.44 (13.99 to 14.91)	1,125	39	14.74 (13.05 to 16.63)	629	23	14.87 (12.62 to 17.49)	496	16	14.76 (12.30 to 17.67)
15	5,117	98	17.13 (16.44 to 17.85)	369	4	16.42 (14.09 to 19.08)	208	2	15.98 (13.36 to 19.07)	161	2	17.32 (13.24 to 22.50)

Data are the number at risk, the number of revision events, the unadjusted cumulative probability of failure and the 95% CI. ODEP A* benchmark for comparison. *Number at risk at the beginning of interval.

Table 11 - Results of flexible parametric models (FPMs) according to surgeon grade (exposure A) and supervision (exposure B).

Indication for revision	Exposure subgroup	Exposure	Revisions (n)*	Model 1 (unadjusted)			Model 2 (adjusted for †)			Model 3 (adjusted for †, ‡)			Model 4 (adjusted for †, ‡, §)		
				n=106,206			n=106,206			n=106,206			n=106,206		
				HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value	HR	95% CI	p-value
All-cause revision	A	Consultant	6,576	1.00			1.00			1.00			1.00		
		Trainee (overall)	344	1.05	0.94 to 1.17	0.40	1.09	0.98 to 1.21	0.13	1.05	0.94 to 1.17	0.40	1.01	0.90 to 1.13	0.88
	B	Consultant	6,576	1.00			1.00			1.00			1.00		
		Trainee supervised	210	1.02	0.89 to 1.17	0.75	1.05	0.92 to 1.21	0.46	1.03	0.90 to 1.18	0.70	0.99	0.87 to 1.14	0.94
		Trainee unsupervised	134	1.09	0.92 to 1.29	0.32	1.15	0.97 to 1.36	0.12	1.08	0.91 to 1.29	0.36	1.03	0.87 to 1.22	0.74
Progression of OA	A	Consultant	2,161	1.00			1.00			1.00			1.00		
		Trainee (overall)	110	0.96	0.79 to 1.16	0.64	0.98	0.81 to 1.19	0.84	0.97	0.80 to 1.17	0.77	0.99	0.82 to 1.21	0.95
	B	Consultant	2,161	1.00			1.00			1.00			1.00		
		Trainee supervised	77	1.10	0.87 to 1.37	0.43	1.11	0.88 to 1.40	0.36	1.11	0.88 to 1.39	0.38	1.13	0.90 to 1.42	0.31
		Trainee unsupervised	33	0.74	0.52 to 1.04	0.08	0.77	0.54 to 1.08	0.13	0.76	0.54 to 1.07	0.11	0.78	0.55 to 1.10	0.16
Aseptic loosening/lysis	A	Consultant	1,877	1.00			1.00			1.00			1.00		
		Trainee (overall)	95	1.02	0.83 to 1.25	0.86	1.07	0.87 to 1.32	0.52	1.03	0.84 to 1.27	0.78	0.96	0.78 to 1.19	0.72
	B	Consultant	1,877	1.00			1.00			1.00			1.00		
		Trainee supervised	60	1.03	0.79 to 1.33	0.84	1.08	0.83 to 1.39	0.57	1.05	0.81 to 1.36	0.70	0.99	0.77 to 1.29	0.96
		Trainee unsupervised	35	1.00	0.72 to 1.41	0.97	1.06	0.76 to 1.48	0.74	0.99	0.71 to 1.39	0.97	0.92	0.65 to 1.28	0.61
Unexplained pain	A	Consultant	1,236	1.00			1.00			1.00			1.00		
		Trainee (overall)	72	1.24	0.98 to 1.57	0.08	1.25	0.98 to 1.58	0.07	1.20	0.95 to 1.53	0.13	1.08	0.85 to 1.37	0.54
	B	Consultant	1,236	1.00			1.00			1.00			1.00		
		Trainee supervised	38	1.02	0.74 to 1.41	0.90	1.03	0.75 to 1.43	0.85	1.01	0.73 to 1.39	0.98	0.92	0.66 to 1.27	0.60
		Trainee unsupervised	34	1.62	1.15 to 2.28	0.01	1.63	1.16 to 2.29	0.01	1.54	1.10 to 2.17	0.01	1.34	0.95 to 1.89	0.09
Instability	A	Consultant	1,052	1.00			1.00			1.00			1.00		
		Trainee (overall)	44	0.86	0.64 to 1.16	0.33	0.92	0.68 to 1.25	0.59	0.86	0.63 to 1.16	0.31	0.80	0.59 to 1.09	0.16
	B	Consultant	1,052	1.00			1.00			1.00			1.00		
		Trainee supervised	27	0.83	0.57 to 1.22	0.35	0.89	0.61 to 1.31	0.56	0.85	0.58 to 1.24	0.40	0.80	0.54 to 1.17	0.25
		Trainee unsupervised	17	0.90	0.56 to 1.46	0.67	0.97	0.60 to 1.57	0.90	0.87	0.54 to 1.41	0.57	0.81	0.50 to 1.32	0.41
Infection	A	Consultant	359	1.00			1.00			1.00			1.00		
		Trainee (overall)	22	1.31	0.85 to 2.02	0.22	1.32	0.86 to 2.04	0.20	1.30	0.84 to 2.00	0.24	1.30	0.84 to 2.01	0.25
	B	Consultant	359	1.00			1.00			1.00			1.00		
		Trainee supervised	13	1.21	0.69 to 2.09	0.51	1.22	0.70 to 2.13	0.48	1.21	0.70 to 2.11	0.50	1.22	0.70 to 2.13	0.49
		Trainee unsupervised	9	1.50	0.77 to 2.90	0.23	1.51	0.78 to 2.93	0.22	1.44	0.74 to 2.79	0.29	1.43	0.73 to 2.79	0.30

Data are the number of revisions for each indication, hazard ratio, 95% CI, or p-value. †**Patient factors:** age; sex; ASA; IMD decile. ‡**Operation factors:** anaesthetic; approach; fixation; bearing mobility. §**Healthcare setting factors:** funding; year of operation. *Some cases were revised for more than one indication.

5.5.2 All-cause revision

The unadjusted cumulative probability of failure at 15 years was 17.13% (95% CI 16.44 to 17.55) for consultants, 16.42% (95% CI 14.09 to 19.08) for trainees overall, 15.98% (95% CI 13.36 to 19.07) for trainees supervised by a scrubbed consultant, and 17.32% (95% CI 13.24 to 22.50) for trainees not supervised by a scrubbed consultant. Failure estimates (one minus survival) for all intervals of follow-up are presented in Table 10, and graphically displayed as a one minus KM plot in Figure 23.

Unadjusted FPM analysis comparing UKRs performed by consultants and trainees (exposure A), indicated that surgeon grade was not associated with the risk of all-cause revision (Model 1: HR 1.05, 95% CI 0.94 to 1.17; p=0.40). This finding, which is documented in Table 11, persisted despite incremental adjustment for patient, operation, and healthcare setting factors (Model 4: HR 1.01, 95% CI 0.90 to 1.13; p=0.88). Further analysis was performed according to the level of senior supervision (exposure B). Neither crude nor adjusted models demonstrated an association between the level of supervision of trainees and the risk of all-cause revision (Table 11).

Sensitivity analysis was performed following further subcategorisation of cases according to specific training grade (exposure C) and supervision. There was no evidence of an association between any specific training grade (F1-ST2, ST3-ST8, or fellow) and an increased risk of all-cause revision, regardless of the level of supervision (Table 12). It should be noted that very few UKRs were performed by surgeons in the most junior category (F1-ST2).

5.5.3 Indication for revision

The three most common indications for revision in this cohort were progression of OA (n=2,271), aseptic loosening/lysis (n=1,972) and unexplained pain (n=1,308). Crude and adjusted analyses demonstrated no evidence of an association between surgeon grade (exposure A) and an increased risk of revision for any indication, including aseptic loosening/lysis, infection, progression of OA, unexplained pain, or instability (Table 11).

Further analysis was performed according to the level of trainee supervision (exposure B). We found no evidence of an increased risk of revision for any indication when trainees were supervised by a scrubbed consultant. However, both crude and adjusted analyses (Models 1-3) demonstrated that procedures performed by trainees without scrubbed consultant supervision were associated with an increased risk of revision for unexplained pain, compared to procedures performed by consultants (Model 1: HR 1.62, 95% CI 1.15 to 2.28; p=0.01). This was not observed in the fully adjusted model (Model 4: HR 1.34, 95% CI 0.95 to 1.89; p=0.09) (Table 11).

Table 12 - Sensitivity analysis: Results of flexible parametric models (FPMs) for all-cause revision according to the specific training grade (exposure C) and supervision.

Exposure	Number of cases	Number of revisions	Complete cases (n=106,206)		
			HR	95% CI	p-value
Model 1 (unadjusted)					
Consultant	101,824	6,576	1.00		
F1-ST2 supervised by scrubbed consultant	25	2	2.16	0.70 to 6.71	0.18
F1-ST2 not supervised by scrubbed consultant	19	2	1.17	0.29 to 4.69	0.82
ST3-ST8 supervised by scrubbed consultant	2,746	193	1.04	0.90 to 1.20	0.59
ST3-ST8 not supervised by scrubbed consultant	1,244	99	1.07	0.87 to 1.30	0.52
Fellow supervised by scrubbed consultant	127	14	0.76	0.45 to 1.28	0.31
Fellow not supervised by scrubbed consultant	221	33	1.16	0.83 to 1.64	0.39
Model 2 (adjusted for †)					
Consultant	101,824	6,576	1.00		
F1-ST2 supervised by scrubbed consultant	25	2	2.08	0.67 to 6.45	0.21
F1-ST2 not supervised by scrubbed consultant	19	2	1.15	0.29 to 4.64	0.84
ST3-ST8 supervised by scrubbed consultant	2,746	193	1.08	0.93 to 1.24	0.32
ST3-ST8 not supervised by scrubbed consultant	1,244	99	1.12	0.92 to 1.37	0.25
Fellow supervised by scrubbed consultant	127	14	0.75	0.45 to 1.27	0.29
Fellow not supervised by scrubbed consultant	221	33	1.23	0.87 to 1.73	0.23
Model 3 (adjusted for †, ‡)					
Consultant	101,824	6,576	1.00		
F1-ST2 supervised by scrubbed consultant	25	2	1.99	0.64 to 6.17	0.23
F1-ST2 not supervised by scrubbed consultant	19	2	1.08	0.27 to 4.34	0.91
ST3-ST8 supervised by scrubbed consultant	2,746	193	1.06	0.91 to 1.21	0.46
ST3-ST8 not supervised by scrubbed consultant	1,244	99	1.08	0.88 to 1.32	0.45
Fellow supervised by scrubbed consultant	127	14	0.71	0.42 to 1.19	0.19
Fellow not supervised by scrubbed consultant	221	33	1.11	0.79 to 1.56	0.56
Model 4 (adjusted for †, ‡, §)					
Consultant	101,824	6,576	1.00		
F1-ST2 supervised by scrubbed consultant	25	2	1.93	0.62 to 6.01	0.25
F1-ST2 not supervised by scrubbed consultant	19	2	1.03	0.26 to 4.12	0.97
ST3-ST8 supervised by scrubbed consultant	2,746	193	1.02	0.89 to 1.18	0.74
ST3-ST8 not supervised by scrubbed consultant	1,244	99	1.03	0.84 to 1.26	0.77
Fellow supervised by scrubbed consultant	127	14	0.66	0.39 to 1.12	0.12
Fellow not supervised by scrubbed consultant	221	33	1.04	0.74 to 1.46	0.84

†Patient factors: age; sex; ASA; IMD decile.
‡Operation factors: anaesthetic; approach; fixation; bearing mobility.
§Healthcare setting factors: funding; year of operation.

F1=Foundation Year 1; ST=Specialty Trainee (number denotes year of training). F1-ST2 is the most junior category, followed by ST3-ST8.

5.6 Discussion

This analysis of 106,206 knees with over 16 years follow-up represents the largest study to date of UKR outcomes in the context of surgical training. We have demonstrated that when comparing UKRs performed by consultants and trainees, there was no evidence of an association between surgeon grade and the risk of all-cause revision. Trainees achieved comparable outcomes to consultants regardless of the level of scrubbed supervision. There was no evidence that UKRs performed by trainees who were supervised by a scrubbed consultant were associated with an increased risk of revision for any specific indication (including aseptic loosening/lysis, infection, progression of OA, unexplained pain, and instability) compared to consultant-performed UKRs. We found evidence that UKRs performed by trainees who were not supervised by a scrubbed consultant, were more likely to be revised for unexplained pain compared to consultant-performed UKRs. However, this was not observed in the fully adjusted model. Revision for unexplained pain following UKR has previously been attributed to low-volume surgeons, but not unsupervised trainees (Baker et al., 2013).

5.6.1 Strengths and limitations

We included over 100,000 knees, which makes this significantly larger than any previous study of the association between surgeon grade and UKR outcomes (Fowler et al., 2021, Madanipour et al., 2021). Despite limiting our study period to predate the anomalous period of elective orthopaedic practice during the COVID-19 pandemic, our findings are current and represent UKRs with over 16 years of follow-up. The data were recorded in a mandatory, nationwide prospective register, which improves the external validity and generalisability of our findings by reducing sampling bias. We employed FPM to model the time-dependent effects of confounding variables and account for non-proportionality. Furthermore, our incremental approach to confounding adjustment increases transparency by demonstrating the relative contribution of patient, operation, and healthcare setting factors to the adjusted results.

Despite these strengths, our study has limitations. Implant survival is an important objective metric of success. However, as they are not currently reported by the NJR, we did not consider other measures

that may be relevant when evaluating the success of a joint replacement, such as patient-reported outcome measures, or postoperative complications other than failure. OA was the only indication, which along with adjustment for confounding variables, accounts for measurable variations in case complexity and case-mix selection between the groups. However, our findings remain susceptible to residual confounding. For example, we did not adjust for BMI which, consistent with other NJR studies, was missing in a high proportion of records (Sayers et al., 2020b). We performed multiple testing for various reasons for revision which may account for the association between unsupervised trainees and revision for unexplained pain that attenuated with adjustment.

The binary variable 'surgeon grade' does not capture variations in the level of experience between individual trainees. We have attempted to address this through sensitivity analysis, by categorising cases according to the specific training grade of the surgeon; however, this categorical variable has similar limitations. Furthermore, supervision is recorded by the NJR as a binary variable according to the grade of the first assistant, which does not capture the spectrum of supervision that is necessary in the training process (ISCP, 2021). Thus, these categorical variables do not account for procedures that may have been part-performed by a trainee, or in which a trainee was supervised by an unscrubbed consultant.

5.6.2 Comparison with other studies

Our literature review identified a small number of observational studies relating to this subject (Fowler et al., 2021). In their NZJR study, Storey et al. found no significant difference in the revision rate of UKRs performed by supervised senior trainees (n=276) compared to attending surgeons (n=8,550). They also reported that supervised senior trainees achieved comparable functional outcomes (Oxford Knee Score) to attending surgeons at 6 months. With only 14 cases in each group, the authors acknowledge that they had insufficient data for any meaningful analysis of the outcomes of UKRs performed by supervised junior trainees and unsupervised senior trainees. Furthermore, the indication for revision was not reported, and the description of the statistical methodology employed

is limited (Storey et al., 2018). Of note, a similarly low proportion of UKRs are recorded as performed by trainees in the NZJR (3.3%) and NJR (4.1%).

Bottomley et al. conducted a single-centre retrospective study of 1,084 Oxford medial UKRs (Zimmer Biomet, Swindon, UK). Trainees performed 673 UKRs (62.1%) and were supervised by a scrubbed consultant in 48% of cases. They reported no difference in implant survival between the groups, with 9-year cumulative survival estimates of 93.9% (95% CI 90.2 to 97.6) and 93.0% (95% CI 90.3 to 95.7) for consultants and trainees, respectively. In a subgroup analysis they showed that trainees who had performed fewer than ten UKRs had a failure rate of 5.1% compared to a failure rate of 4.7% in those who had undertaken more than ten UKRs; a difference that was not statistically significant (Bottomley et al., 2016).

In comparison to the existing literature, the current study is significantly larger, has methodological advantages, longer follow-up, and provides novel insight into the importance of scrubbed consultant supervision. Our findings are generally concordant with published data from another national joint registry (Storey et al., 2018), which suggests that our findings might be generalisable to other countries.

5.6.3 Implications

Our findings suggest that current training practices for UKR in England and Wales are safe, when defined by equivalence of survival outcomes. However, only a small proportion of UKRs in these countries are performed by trainees and it should be noted that very few UKRs were performed by surgeons of the most junior specific training grade (F1-ST2). It is likely that UKRs are typically performed by more experienced, senior trainees. However, we were unable to quantify this in the current study, due to the broad categories used by the NJR to record the grade of the operating surgeon.

It is presumed that trainers select appropriate cases for their trainees and permit trainees to operate without scrubbed supervision only when they have reached a subjective threshold of expertise commensurate with safe independent surgical practice. Our study suggests that in this context,

trainees achieve comparable all-cause UKR survival to consultant surgeons. In terms of revision for unexplained pain, trainees might achieve their best outcomes when supervised by a scrubbed consultant. However, this association was not observed in the fully adjusted analysis. We propose that trainees should ideally be supervised by a scrubbed consultant when performing UKR, particularly during the early stages of training. When experienced senior trainees operate without scrubbed supervision, careful case selection is required, and scrubbed consultant supervision should be readily available.

The findings of this study are reassuring and support the current methods by which surgeons are trained to perform UKR in England and Wales. This is of particular importance in the wake of the COVID-19 pandemic, which has been detrimental to trainee case numbers (Clements et al., 2021, Munro et al., 2021).

5.6.4 Conclusion

This nationwide study of UKRs with over 16 years follow-up demonstrates that trainees in England and Wales achieve comparable all-cause implant survival to consultants. Our findings support the current methods by which surgeons in England and Wales are trained to perform UKR.

Chapter 6 A comparison of clinical and radiological outcomes between trainee- and consultant-performed total hip replacement: a retrospective cohort study of 530 hips with up to 15 years follow-up

6.1 Overview

The preceding chapters have focused on implant survival according to surgeon grade and supervision, which is consistent with the primary aim of this thesis. As discussed in Section 1.5.1, there are limitations to using implant survival as a standalone outcome measure. Therefore, this chapter aims to investigate the association between surgeon grade and a range of clinical, radiological, and functional outcomes measures following THR. This chapter is based on single-centre observational data.

6.1.1 Contributors

TF was responsible for study concept, design, literature review, data collection, data analysis, interpretation of the results, and writing the manuscript. JB contributed to data collection. AB, and MW were responsible for study concept, design, interpretation of the results, and review of the manuscript. AS was responsible for study concept, design, data analysis, interpretation of the results, and review of the manuscript.

6.1.2 Declaration of conflicts of interest

TF, AS and JB have no conflicts of interest to declare. AB and MW report a contract held by the University of Bristol for membership of HQIP Lot 2 Statistical Analysis team for the National Joint Registry, and royalties from Taylor & Francis. MW also reports an independently conducted research grant held by the University of Bristol with CeramTec, teaching payments from Heraeus, and research committee membership on the British Orthopaedic Association and the British Hip Society.

6.2 Abstract

6.2.1 Background

The long-term clinical, radiological, and functional outcomes of trainee- as compared to consultant-performed total hip replacement (THR) are poorly understood. We aimed to use observational data to examine these outcomes.

6.2.2 Methods

This retrospective cohort study is based on a consecutive series of 530 primary THRs performed at a single institution between 2005 and 2009. Exposures were surgeon grade (consultant or trainee) and whether or not trainees were supervised by a scrubbed consultant. Outcomes of interest were PROMs (OHS and SF-12) at minimum 10 years follow-up, the rate of surgical complications, radiographic outcomes including Barrack grade and periprosthetic radiolucency, and all-cause revision.

6.2.3 Results

Eleven patients were lost to follow-up leaving 519 THRs in 480 patients for further analysis. Trainees performed 400 (77%) THRs and were supervised by a scrubbed consultant in 51% of these cases. Trainees achieved comparable outcomes to consultants in terms of the rate of complications, the quality of femoral cementation, radiological periprosthetic lucency, and the OHS at minimum 10 years of follow-up. There were differences in the SF-12 scores (MCS and PCS) between the groups. However, trainees who were supervised by a scrubbed consultant achieved equivalent PROMs to consultants (OHS and SF-12). There was no association between surgeon grade, or the supervision of trainees and the risk of all-cause revision in either crude or adjusted models (trainees overall: crude HR 0.54, 95% CI 0.18 to 1.62; p=0.28; fully adjusted HR 0.51, 95% CI 0.15 to 1.71; p=0.28).

6.2.4 Conclusion

These findings give reassuring insight into the outcomes of trainee-performed THR in the UK. Trainees who were supervised by a scrubbed consultant achieved comparable PROMs, complication rates, radiological outcomes, and all-cause implant survival to consultants.

6.3 Background

Approximately 10% of primary total hip replacements (THRs) in England and Wales are performed by trainees (Fowler et al., 2022). Surgical training programmes must balance the necessity to train enough surgeons to meet evolving workforce requirements while ensuring safe and acceptable outcomes for patients (Centre for Workforce Intelligence, 2014, The British Orthopaedic Association, 2021). The Royal College of Surgeons of England has recommended that every planned NHS procedure should include a surgical trainee to help compensate for lost training opportunities during the COVID-19 pandemic (The Royal College of Surgeons of England, 2021).

The National Joint Registry (NJR) study presented in Chapter 3 demonstrates that supervised trainees achieve comparable THR implant survival to consultants, but THRs performed by trainees who are not supervised by a scrubbed consultant may have an increased risk of revision for instability (Fowler et al., 2022). However, studies based on NJR data are limited as they do not report patient-reported outcome measures (PROMs), radiographic analysis, or postoperative complications other than revision surgery.

There is contradictory evidence on the association between surgeon grade and the risk of postoperative complications. It has been suggested that surgeon inexperience and trainee-performed THR are risk factors for dislocation and prosthetic joint infection (Kunutsor et al., 2019, Smith et al., 2018, Hedlundh et al., 1996). However, a recent systematic review and meta-analysis that directly compared trainee- and consultant-performed THRs, found no difference in the rates of dislocation or infection between the groups (Singh et al., 2019).

Observational studies have suggested that trainees achieve comparable patient-reported outcomes to consultants at various intervals of follow-up ranging from 6 months to 10 years (Palan et al., 2009, Reidy et al., 2016, Weber et al., 2017, Moran et al., 2004). However, these findings are contradicted by 6-month registry data (Inglis et al., 2013), and the functional outcomes of trainee-performed THR beyond 10 years are unknown (Singh et al., 2019).

Moran et al. reported inconsistent cup positioning and significantly reduced acetabular anteversion in THRs performed by supervised trainees (Moran et al., 2004). A more recent study of THRs performed for hip fracture reported that supervised trainees achieved equivalent radiological outcomes to consultants in terms of acetabular inclination, leg length, and Barrack grade (MacDonald et al., 2020). The radiological outcomes of trainee-performed THR remain poorly understood. The clinical and radiological outcomes of trainee-performed THR warrant further investigation.

The aim of this study was to investigate the association between surgeon grade (consultant or trainee), the supervision of trainees, and the following clinical and radiological outcomes of THR: 1) PROMs at minimum 10 years; 2) the rate of surgical complications; 3) radiographic outcomes relating to the quality of femoral cementation and periprosthetic radiolucency, and 4) all-cause revision.

6.4 Methods

6.4.1 Patients and data sources

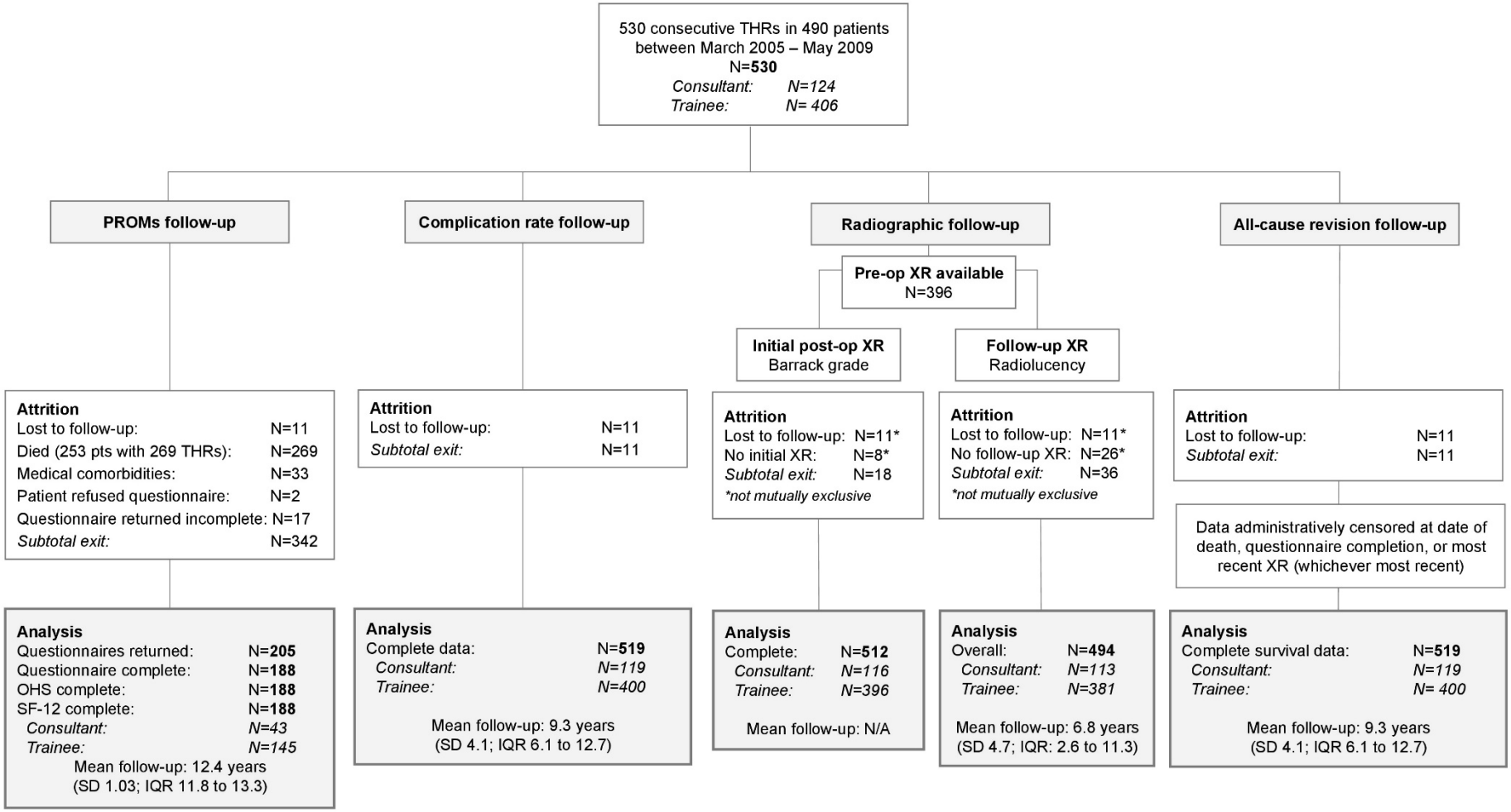
This retrospective cohort study is based on a consecutive series of 530 primary THRs in 490 patients performed at North Bristol NHS Trust between March 2005 and May 2009. Institutional research committee approval was granted prior to the initiation of this study (CE10851). Reporting follows the recommendations outlined in the STROBE checklist (Appendix 18).

All cases used a cemented C-stem AMT femoral component (Depuy Synthes, Leeds, UK) and were performed under the care of three consultant surgeons. Electronic patient records were reviewed for all patients to identify information about exposures, demographic data, episodes of revision, and surgical complications. Surviving patients completed postal questionnaires, including the Oxford Hip Score (OHS) and Short Form-12 Health Survey (SF-12), at a minimum of 10 years follow-up. Patients were asked directly about subsequent hip surgery to identify cases of revision surgery performed in other units. Preoperative OHS and SF-12 data were unavailable as these were not routinely recorded at preoperative assessment. Routine postoperative radiographs were reviewed to assess: 1) The quality of the cement mantle on the immediate postoperative radiograph, which was assessed using the method described by Barrack et al. (Barrack et al., 1992); 2) Evidence of radiolucency in the seven

zones described by Gruen et al. on the most recent anteroposterior (AP) pelvic radiograph (Gruen et al., 1979). These outcome measures are described in more detail in Section 6.4.3.3. Patients with missing data for a given outcome were excluded from the relevant analysis. The process of exclusions and attrition is summarised in Figure 24.

The dataset used in the current study is based on a smaller consecutive cohort of 415 primary THRs performed at the same institution between March 2005 and May 2008, from which a shorter-term follow-up study has previously been published (Berstock et al., 2014). The cohort presented here has been expanded to 530 THRs and up-to-date follow-up data have been collected for all patients. The original dataset was established and collected by JB. Expansion of the dataset and the most recent round of data collection were conducted by TF. TF reviewed electronic patient records and operation notes for all patients in order to confirm the grade of the operating surgeon and the level of supervision. TF collected up-to-date follow-up data for all patients and the following outcome measures: PROMs, complication rates, periprosthetic radiolucency, and all-cause revision. Regarding the quality of femoral cementation on the initial postoperative radiograph, data collection for the first 415 cases in the cohort was performed by JB. Data collection for the most recent 115 cases was performed by TF.

Figure 24 - Study flow diagram (Chapter 6).



6.4.2 Exposures

The primary exposure (exposure A) was surgeon grade, which was a binary variable dependent on the grade of the operating surgeon: 1) consultant, or 2) trainee. THRs performed by surgeons of the following grades were categorised under the variable 'trainee': F1-ST2; ST3-ST8; and fellow. Consultant surgeons have completed their formal training in orthopaedic surgery, been awarded a Certificate of Completion of Training, and have been appointed to a senior position in which they can supervise trainees. The term 'consultant' is equivalent to the term 'attending'. The grade category ST3-ST8 is traditionally referred to as 'registrar' and is equivalent to 'resident', which is used in other healthcare settings including North America. UK training terminology has evolved over the past 20 years, thus the process of accounting for variations in training terminology over the course of this study is documented in Appendix 19. Trainee cases were subcategorised according to the level of consultant supervision, which was recorded on the operation note (exposure B): 1) trainee supervised by a scrubbed consultant; or 2) trainee not supervised by a scrubbed consultant. Given the variability in experience between surgical trainees, we performed a sensitivity analysis for all-cause revision by recategorising procedures according to the specific training grade of the operating surgeon (exposure C): 1) consultant; 2) F1-ST2; 3) ST3-ST8; 4) fellow.

6.4.3 Outcomes of interest

6.4.3.1 PROMs

The OHS and SF-12 were assessed at a minimum of 10 years follow-up. The combined use of these two instruments is established in the context of THR (Ostendorf et al., 2004, Murray et al., 2007). The OHS is a 12-item joint-specific questionnaire developed to assess the perception of pain and function in patients undergoing THR (Dawson et al., 1996a). It was designed to be completed by patients in order to reduce potential bias introduced by surgeon-based outcome tools (Murray et al., 2007).

The OHS is internally consistent, valid, reproducible, and sensitive to change (Dawson et al., 1996a, Dawson et al., 1996b). It provides data that correlate strongly with patient satisfaction (Fitzpatrick et al., 2000). Its responsiveness, or sensitivity to clinically important change following THR, is superior

to a range of disease-specific and generic measures of health status, including the WOMAC (Ostendorf et al., 2004), SF-36 (Dawson et al., 1996b), and EQ-5D (Dawson et al., 2001). The OHS was originally developed for use in randomised controlled trials but has become extensively used in day-to-day clinical practice, observational studies, and national joint registries (Bohm et al., 2021, Murray et al., 2007). Since 2009, the OHS has been routinely collected for all NHS-funded THRs performed in England (Devlin et al., 2010).

The questionnaire consists of 12 questions regarding the severity of pain and disability experienced by the respondent over the past four weeks. Each question has five response levels, which are scored from 0 to 4. The sum of the 12 individual scores produces an overall score ranging from 0 to 48, with 48 representing the best possible outcome. The original scoring system produced a score between 12 and 60, with 60 representing the worst possible outcome. Consistent with the recommendation of Murray et al., the current study uses the 0 to 48 scoring system (Murray et al., 2007). The results of the OHS should be interpreted in the context of an estimated minimal clinically important difference (MCID) of 3 to 5 points (Murray et al., 2007). The preoperative OHS is an important determinant of outcome after THR (Hajat et al., 2002). Therefore, where possible, preoperative and postoperative scores should be reported.

The SF-12 is a generic measure of health status, designed to give a global assessment of a patient's health-related quality of life (HRQoL). It is a shorter version of the SF-36, which can be printed on a single side and self-administered in approximately two minutes. This reduces respondent burden compared to the SF-36 and makes it suitable for larger studies, particularly when participants are required to complete multiple questionnaires (Ware et al., 1996).

The developer used psychometric methods to select a 12-item subset from the original 36-item survey (Ware et al., 1996). This process of item selection has been validated in multiple health settings (Gandek et al., 1998). The resulting SF-12 survey assesses the same eight HRQoL dimensions as the SF-36 (physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional and mental health). However, as four of the eight dimensions are assessed by a single

item, only Physical Component Summary (PCS) and Mental Component Summary (MCS) scores are calculated.

The SF-12 is valid, reliable, and demonstrates strong correlations with SF-36 summary scores (Ware et al., 1998). SF-12 PCS and MCS scores account for more than 90% of the variance in the corresponding SF-36 summary scores (Ware et al., 1996). A strong correlation between SF-12 and SF-36 PCS and MCS scores has been observed in both general health studies and THR cohorts (Ware et al., 1996, Jenkinson and Layte, 1997, Ostendorf et al., 2004). In the context of joint replacement, Ostendorf et al. found that the SF-12 PCS demonstrates high levels of responsiveness in patients undergoing THR (Ostendorf et al., 2004).

The SF-12 PCS and MCS have a minimum score of 0 and a maximum score of 100. These scores are transformed to have a mean of 50 and a standard deviation of 10 in the general population (Gandhi et al., 2001). Due to the strong correlation between SF-12 and SF-36 scores, the developers recommended that SF-36 interpretation guidelines may be useful when interpreting the SF-12 (Ware et al., 1995, Ware et al., 1994).

6.4.3.2 Complications

We were interested in the overall complication rate and the rate of the following specific surgical complications: dislocation; intraoperative periprosthetic fracture; postoperative periprosthetic fracture; nerve injury; symptomatic heterotrophic ossification (HO); superficial surgical site infection (SSI) and deep infection (both defined according to CDC criteria) (Centers for Disease Control and Prevention, 2021). The rate of complications was assessed through a comprehensive review of electronic patient records. Furthermore, postal questionnaires included a free-text question that directly asked patients about subsequent hip surgery.

6.4.3.3 Radiographic outcomes

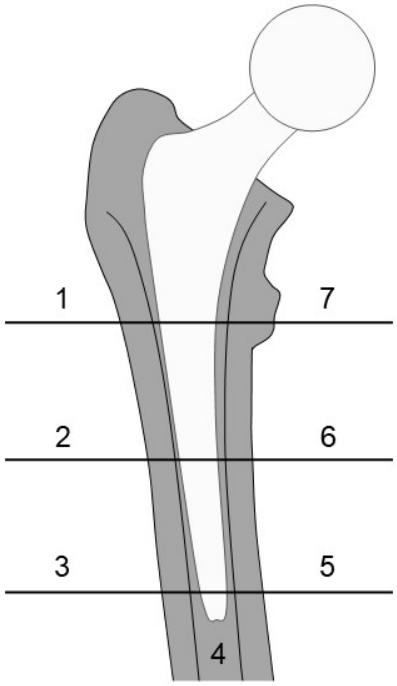
Radiographic outcome measures included the Barrack grade for the quality of cementation on the initial postoperative radiograph, and the presence of femoral periprosthetic lucency on the most recent

follow-up radiograph. The Barrack grading system for the radiographic assessment of initial cement mantle quality is summarised in Table 13 (Barrack et al., 1992). An association between Barrack grade C and D cementation, loosening and failure has been described (Chambers et al., 2001). The presence of femoral periprosthetic radiolucency was assessed using the method described by Gruen et al., which subdivides the proximal femur into seven zones. The distribution of these zones on an AP radiograph is illustrated in Figure 25.

Table 13 – The Barrack grading system of initial cement mantle quality (Barrack et al., 1992).

Grade	Description of radiographic cement mantle quality
A	Complete filling of the medullary cavity by cement, with 'white-out' of the cement-bone interface
B	Radiolucency involving <50% of the cement-bone interface
C	Radiolucency involving 50-99% of the cement-bone interface
D	Radiolucency at the cement-bone interface of 100% in any projection, or a failure to fill the canal such that the tip of the stem is not covered with cement.

Figure 25 - Diagram illustrating the distribution of Gruen zones on an AP radiograph (Gruen et al., 1979).



6.4.3.4 All-cause revision

All-cause revision was defined as any procedure to add, remove, or modify one or more components of an implant construct for any reason (The National Joint Registry, 2020). Implant survival data were collected by reviewing electronic patient records, and information regarding the timing and reason for revision was recorded. Postal questionnaires included a free-text question asking about subsequent hip surgery to identify episodes of revision performed in other units. This outcome measure is discussed in more detail in Section 1.5.1.

6.4.4 Statistical analysis

Frequencies and percentages were used to describe categorical variables. The mean, standard deviation (SD) and interquartile range (IQR) were used to describe continuous variables. The Shapiro-Wilk test was used to assess the normality of continuous data. Wilcoxon rank sum and chi-squared tests were used, where appropriate, to assess equivalence between the consultant and trainee groups.

PROMs and the assessment of radiolucency on a linear scale (number of affected zones from 0-7) were analysed using linear regression models. Complication rates, Barrack grade and the presence of radiolucency (≥1 zone) were analysed using logistic regression models.

Estimates of unadjusted net implant failure were calculated using the Kaplan-Meier (KM) method. Patients with surviving implants were censored at either the time of the most recent questionnaire, radiograph, or death. Adjusted survival analyses used Cox regression and Schoenfeld residual testing was used to assess the proportional hazards assumption at each level of adjustment. Separate survival analyses were performed for each exposure (A-C) as specified in Appendix 20. The proportionality of THR survival according to surgeon grade has previously been demonstrated using NJR data (Chapter 3) (Fowler et al., 2022).

The level of incremental confounding adjustment is specified for each analysis. Model 1 was unadjusted. Model 2 was adjusted for patient-level factors (age, sex, American Society of Anaesthesiologists [ASA] grade, and indication). Model 3 was further adjusted for operation-level

factors (cup fixation, cup brand, head size, and bearing surface). Statistical analysis was performed using Stata (Version SE 15.1; StataCorp LP, USA).

6.5 Results

6.5.1 Descriptive analysis

From a consecutive series of 530 THRs in 490 patients, 11 cases (2.1%) were lost to follow-up. We included 519 THRs, of which 400 (77.1%) were performed by trainees. Trainees were supervised by a scrubbed consultant in 51.0% of trainee cases. A higher proportion of trainee-performed procedures utilised a fully cemented construct compared to consultant-performed procedures (54.5% vs. 38.6%; chi-squared: $p<0.01$), but demographic and operation-level factors were otherwise equivalent between the consultant and trainee groups (Table 14). Summary data for each outcome measure are presented in Table 15.

Table 14 - Demographic and operation-level data for included THRs.

	Surgeon grade and supervision (n=519)			
Variable	**Consultant (n=119)**	**Trainee (overall) (n=400)**	**Trainee supervised by a scrubbed consultant (n=204)**	**Trainee not supervised by a scrubbed consultant (n=196)**
Mean age (SD) years	73.6 (8.4)	74.6 (8.0)	75.0 (7.9)	74.1 (8.1)
Sex (%female)	78 (65.6)	269 (67.3)	143 (70.1)	126 (64.3)
Side (%right)	65 (54.6)	221 (55.3)	111 (54.4)	110 (56.1)
ASA (%)				
1	8 (6.7)	14 (3.5)	4 (2.0)	10 (5.1)
2	81 (68.1)	289 (72.3)	152 (74.5)	137 (69.9)
≥3	30 (25.2)	97 (24.3)	48 (23.5)	49 (25.0)
Indication (%)				
OA	113 (95.0)	384 (96.0)	196 (96.1)	188 (95.9)
RA	2 (1.7)	7 (1.8)	4 (2.0)	3 (1.5)
AVN	1 (0.8)	6 (1.5)	2 (1.0)	4 (2.0)
NOF	1 (0.8)	2 (0.5)	1 (0.5)	1 (0.5)
Other	2 (1.7)	1 (0.3)	1 (0.5)	0 (0.0)
Bearing surface (%)				
MoP	99 (83.2)	327 (81.8)	169 (82.8)	158 (80.6)
MoM	14 (11.8)	32 (8.0)	9 (4.4)	23 (11.7)
CoP	1 (0.8)	4 (1.0)	3 (1.5)	1 (0.5)
CoC	3 (2.5)	32 (8.0)	20 (9.8)	12 (6.1)
CoM	2 (1.7)	5 (1.2)	3 (1.5)	2 (1.0)
Cup brand (%)				
Pinnacle	65 (54.6)	176 (44.0)	93 (45.6)	83 (42.4)
Ogee	46 (38.7)	218 (54.5)	108 (52.9)	110 (56.1)
Duraloc	5 (4.2)	2 (0.5)	0 (0.0)	2 (1.0)
Omnifit	0 (0.0)	2 (0.5)	1 (0.5)	1 (0.5)
Trident	2 (1.7)	1 (0.3)	1 (0.5)	0 (0.0)
ASR	1 (0.8)	1 (0.3)	1 (0.5)	0 (0.0)
Cup fixation (%)				
Hybrid	73 (61.3)	182 (45.5)	96 (47.1)	86 (43.9)
Cemented	46 (38.7)	218 (54.5)	108 (52.9)	110 (56.1)
Head size (%)				
28mm	93 (78.2)	322 (80.5)	163 (79.9)	159 (81.1)
36mm	21 (17.7)	71 (17.8)	36 (17.7)	35 (17.9)
Other	5 (4.2)	7 (1.8)	5 (2.5)	2 (1.0)

Data are the frequency (n), proportion (%), mean, or standard deviation (SD); denoted where applicable. OA, osteoarthritis; RA, rheumatoid arthritis; AVN, avascular necrosis; NOF, neck of femur fracture; MoP, metal-on-polyethylene; MoM, metal-on-metal; CoP, ceramic-on-polyethylene; CoC, ceramic-on-ceramic; CoM, ceramic-on-metal.

Table 15 - Summary of outcome data according to surgeon grade and supervision

Outcome	Surgeon grade and supervision			
	Consultant	**Trainee (overall)**	**Trainee supervised by a scrubbed consultant**	**Trainee not supervised by a scrubbed consultant**
PROMs				
OHS, mean (SD)	36.3 (11.1)	37.3 (10.8)	38.5 (10.1)	36.0 (11.4)
SF12-PCS, mean (SD	39.3 (11.2)	35.3 (12.1)	36.1 (12.4)	34.5 (11.7)
SF12-MCS, mean (SD)	46.3 (12.1)	49.3 (11.4)	49.9 (11.4)	48.6 (11.4)
Complications, n				
Nerve injury	0	1	0	1
Superficial infection (SSI)	2	0	0	0
Symptomatic HO	2	1	1	0
Periprosthetic fracture (intra-op)	1	3	1	2
Deep infection	2	3	0	3
Periprosthetic fracture (post-op)	1	7	5	2
Dislocation	2	8	4	4
Overall complications	10	23	11	12
Barrack grade, n (%)				
A	78 (65.6)	251 (62.8)	132 (64.7)	119 (60.7)
B	36 (30.3)	140 (35.0)	66 (32.4)	74 (37.8)
C	2 (1.7)	5 (1.3)	4 (2.0)	1 (0.5)
Lucency (number of Gruen zones affected), n (%)				
0	36 (31.9)	128 (33.6)	70 (36.5)	58 (30.7)
1	22 (19.5)	58 (15.2)	32 (16.7)	26 (13.8)
2	17 (15.0)	55 (14.4)	25 (13.0)	30 (15.9)
3	8 (7.1)	39 (10.2)	18 (9.4)	21 (11.1)
4	7 (6.2)	34 (8.9)	17 (8.9)	17 (9.0)
5	5 (4.4)	26 (6.8)	11 (5.7)	15 (7.9)
6	12 (10.6)	34 (8.9)	16 (8.3)	18 (9.5)
7	6 (5.3)	7 (1.8)	3 (1.6)	4 (2.1)
Revisions (n)	5	9	2	7

6.5.2 PROMs

Complete ≥10-year PROM data were available for 72% (n=188) of surviving patients (Figure 24). Mean PROM follow-up was 12.4 years (SD 1.03; IQR 11.8 to 13.3). The linear regression analyses of PROMs were either unadjusted or adjusted for patient-level confounding factors. Additional adjustment for operative factors was not conducted due to the small number of patients. The characteristics of responders and non-responders are summarised in Appendix 22 and the results are presented in Table 16.

We observed no evidence of an association between either surgeon grade, or the level of scrubbed supervision and the OHS. There was evidence of an association between trainee-performed THR, and a lower SF-12 PCS score compared to consultant-performed surgery, which is suggestive of lower physical function (unadjusted coefficient -4.09, 95% CI -8.18 to -0.16; p=0.05). However, this was not observed in the adjusted model. There was evidence in both crude and adjusted models that cases performed by trainees who were not supervised by a scrubbed consultant were associated with a lower SF-12 PCS, compared to cases performed by consultants (adjusted coefficient -5.13, 95% CI -9.72 to -0.55; p=0.03). There was evidence of an association between THRs performed by trainees who were supervised by a scrubbed consultant and a higher SF-12 MCS score, compared to cases performed by consultants (adjusted coefficient 5.07, 95% CI 0.62 to 9.53; p=0.03).

Table 16 - Linear regression models for PROMs data (OHS, SF-12 MCS, SF-12 PCS) according to surgeon grade (exposure A) and supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Responses (n)	Model 1 (unadjusted)			Model 2 (adjusted for †)		
				n=188			n=188		
				Coef.	**95% CI**	**p-value**	**Coef.**	**95% CI**	**p-value**
OHS	A	Consultant*	43	36.26**	32.99 to 39.52	-	34.93	31.26 to 38.60	-
		Trainee (overall)	145	1.07	-2.65 to 4.78	0.57	1.44	-2.34 to 5.21	0.45
	B	Consultant*	43	36.26**	32.99 to 39.52	-	34.64	30.99 to 38.30	-
		Trainee supervised by a scrubbed consultant	75	2.28	-1.81 to 6.36	0.27	3.21	-0.96 to 7.39	0.13
		Trainee not supervised by a scrubbed consultant	70	-0.23	-4.36 to 3.91	0.91	-0.20	-4.31 to 3.91	0.93
SF-12-PCS	A	Consultant*	43	39.36**	35.79 to 42.93	-	38.14	34.08 to 42.20	-
		Trainee (overall)	145	-4.09	-8.18 to -0.16	0.05	-3.95	-8.14 to 0.23	0.06
	B	Consultant*	43	39.36**	35.79 to 42.93	-	37.95	33.88 to 42.01	-
		Trainee supervised by a scrubbed consultant	75	-3.24	-7.74 to 1.26	0.16	-2.68	-7.33 to 1.96	0.26
		Trainee not supervised by a scrubbed consultant	70	-5.01	-9.57 to -0.45	0.03	-5.13	-9.72 to -0.55	0.03
SF-12-MCS	A	Consultant*	43	46.65**	43.18 to 50.12	-	44.19	40.30 to 48.07	-
		Trainee (overall)	145	2.62	-1.35 to 6.59	0.19	3.87	-0.14 to 7.88	0.06
	B	Consultant*	43	46.65**	43.18 to 50.12	-	44.00	40.11 to 47.90	-
		Trainee supervised by a scrubbed consultant	75	3.25	-1.13 to 7.63	0.15	5.07	0.62 to 9.53	0.03
		Trainee not supervised by a scrubbed consultant	70	1.94	-2.50 to 6.39	0.39	2.75	-1.64 to 7.14	0.22

Data are the coefficient (Coef.), 95% CI, or p-value. OHS, Oxford Hip Score; SF-12, Short Form-12; MCS, mental component summary; PCS, physical component summary. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). *Reference group. **Represents the mean score. §Represents the baseline category used for each confounding variable, which was the most frequently occurring.

6.5.3 Complication rates

A total of 33 surgical complications were recorded. The overall complication rate was 8.4% in the consultant cohort (10/119) and 5.8% in the trainee cohort (23/400). The logistic regression analysis of complication rates was unadjusted due to the low number of events in each group (Table 17). We found no evidence of a significant association between surgeon grade and any specific surgical complication (including SSI, deep infection, dislocation, symptomatic HO, nerve injury, or periprosthetic fracture), or the overall complication rate (OR 0.66, 95% CI 0.31 to 1.44; p=0.30).

Table 17 - Summary of surgical complications with logistic regression analysis according to surgeon grade (exposure A).

Complication	**Overall incidence (n)**	**Consultant incidence (n)**	**Trainee incidence (n)**	**Model 1 (unadjusted) (n=519)**	
				OR (95%CI)	**p-value**
Nerve injury	1	0	1	1	-
Superficial infection (SSI)	2	2	0	1	-
Symptomatic HO	3	2	1	0.15 (0.01 to 1.63)	0.12
Periprosthetic fracture (intra-op)	4	1	3	0.89 (0.09 to 8.65)	0.92
Deep infection	5	2	3	0.44 (0.07 to 2.68)	0.37
Periprosthetic fracture (post-op)	8	1	7	2.10 (0.26 to 17.26)	0.49
Dislocation	10	2	8	1.19 (0.25 to 5.70)	0.82
Total	33	10	23	0.66 (0.31 to 1.44)	0.30

6.5.4 Radiographic outcomes

6.5.4.1 Femoral cementation quality (Barrack grade)

The results for this outcome measure are documented in Table 18. Immediate postoperative radiographs were available for 512 hips. Barrack grade A was observed in 329 (64.3%) cases, grade B was observed in 176 (34.4%) cases, and grade C was observed in only 7 (1.4%) cases. The breakdown according to surgeon grade is documented in Table 15. Barrack grade was analysed as a binary variable: whether or not grade A was achieved. Logistic regression models were adjusted for patient-level factors only, as the available operation factors are unrelated to the initial quality of femoral stem cementation. We observed no evidence of an association between surgeon grade, or the scrubbed supervision of trainees and Barrack grade A cementation (Table 18).

6.5.4.2 Radiolucency

The mean duration of radiographic follow-up was 6.8 years (SD 4.7; IQR 2.6 to 11.3 years). Radiographic lucency was analysed using both logistic and linear regression models, which were incrementally adjusted for patient and operation factors. We found no evidence of an association between surgeon grade, or supervision and the presence of ≥1 zone of radiolucency (Table 19). Furthermore, there was no evidence of an association between surgeon grade, or supervision and the number of affected Gruen zones (Table 20).

Table 18 - Logistic regression models for Barrack cementation grade according to surgeon grade (exposure A) and supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Cases (n)	Barrack A (n)	Model 1 (unadjusted)			Model 2 (adjusted for †)		
					n=512			n=512		
					OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Barrack grade A cementation achieved (Y/N)	A	Consultant*	116	78	1.00	-	-	1.00	-	-
		Trainee (overall)	396	251	0.84	0.54 to 1.31	0.45	0.82	0.52 to 1.28	0.39
	B	Consultant*	116	78	1.00	-	-	1.00	-	-
		Trainee supervised by a scrubbed consultant	202	132	0.92	0.57 to 1.49	0.73	0.91	0.55 to 1.50	0.70
		Trainee not supervised by a scrubbed consultant	194	119	0.77	0.48 to 1.25	0.30	0.74	0.45 to 1.22	0.23

Data are odds ratio (OR), 95% CI, or p-value. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). *Reference group. §Represents the baseline category for each confounding variable.

Table 19 - Logistic regression models for femoral stem radiolucency according to surgeon grade (exposure A) and the level of senior supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Cases (n)	Lucency in ≥1 zone (n)	Model 1 (unadjusted)			Model 2 (adjusted for †)			Model 3 (adjusted for †, ‡)		
					n=494			n=494			n=494		
					OR	95% CI	p-value	OR	95% CI	p-value	OR	95% CI	p-value
Lucency in ≥1 Gruen zone (Y/N)	A	Consultant*	113	77	1.00	-	-	1.00	-	-	1.00	-	-
		Trainee (overall)	381	253	0.92	0.59 to 1.45	0.73	0.86	0.54 to 1.37	0.53	0.81	0.50 to 1.31	0.39
	B	Consultant*	113	77	1.00	-	-	1.00	-	-	1.00	-	-
		Trainee supervised by a scrubbed consultant	192	122	0.81	0.50 to 1.33	0.42	0.71	0.43 to 1.20	0.20	0.68	0.40 to 1.15	0.15
		Trainee not supervised by a scrubbed consultant	189	131	1.06	0.64 to 1.74	0.83	1.04	0.62 to 1.73	0.89	0.97	0.57 to 1.66	0.91

Data are odds ratio (OR), 95% CI, or p-value. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). ‡**Operation factors:** cup fixation (cemented§); head size (28mm§), cup brand (Ogee§), bearing (MoP§). *Reference group. §Represents the baseline category for each confounding variable.

Table 20 - Linear regression models for femoral stem radiolucency according to surgeon grade (exposure A) and the level of senior supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Cases (n)	Model 1 (unadjusted)			Model 2 (adjusted for †)			Model 3 (adjusted for †, ‡)		
				n=494			n=494			n=494		
				Coef.	95% CI	p-value	Coef.	95% CI	p-value	Coef.	95% CI	p-value
Lucency (number of Gruen zones affected)	A	Consultant*	113	2.19**	1.79 to 2.58	0.00	2.52	2.08 to 2.96	0.00	2.55	2.03 to 3.07	0.00
		Trainee (overall)	381	-0.08	-0.53 to 0.38	0.74	-0.14	-0.59 to 0.31	0.54	-0.10	-0.57 to 0.37	0.67
	B	Consultant*	113	2.19**	1.79 to 2.58	0.00	2.53	2.09 to 2.97	0.00	2.55	2.03 to 3.06	0.00
		Trainee supervised by a scrubbed consultant	192	-0.23	-0.73 to 0.27	0.37	-0.34	-0.84 to 0.16	0.18	-0.28	-0.79 to 0.24	0.29
		Trainee not supervised by a scrubbed consultant	189	0.08	-0.42 to 0.58	0.76	0.06	-0.44 to 0.56	0.81	0.08	-0.44 to 0.59	0.76

Data are the coefficient (Coef.), 95% CI, or p-value. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). ‡**Operation factors:** cup fixation (cemented§); head size (28mm§), cup brand (Ogee§), bearing (MoP§). *Reference group. **Represents the mean number of affected Gruen zones. §Represents the baseline category for each confounding variable.

6.5.5 All-cause revision

The mean duration of follow-up for the survival analysis was 9.3 years (SD 4.1; IQR 6.1 to 12.7 years; maximum 15.4 years). A total of 14 hips, the details of which are summarised in Table 21, were revised at a mean of 5.5 years (SD 4.1; IQR 1.4 to 8.3 years). The unadjusted cumulative probability of failure according to surgeon grade is displayed as a one minus KM plot in Figure 26 and as a corresponding life table in Appendix 21. The cumulative probability of failure at 10 years was 3.4% (95% CI 1.1 to 10.1) for consultants and 2.7% (95% CI 1.4 to 5.3) for trainees. The four THRs that failed at <2 years were all performed by trainees (two supervised, two unsupervised) and all revised for instability (Table 21).

The association between surgeon grade and implant survival was assessed using Cox regression. Schoenfeld residual testing confirmed that the proportional hazards assumption was met in adjusted models for both surgeon grade (exposure A) and supervision (exposure B). We observed no evidence of an association between surgeon grade, or the supervision of trainees and the risk of all-cause revision (Model 3: trainees overall: HR 0.51, 95% CI 0.15 to 1.71; p=0.28; supervised trainees: HR 0.19, 95% CI 0.03 to 1.11; p=0.07; unsupervised trainees: HR 0.85, 95% CI 0.25 to 2.96; p=0.80). This was the case in both crude and adjusted models (Table 22). However, the crude models should be interpreted with caution due to a lack of proportionality, which relates to a higher early failure rate in the trainee cohort compared to the consultant cohort.

A sensitivity analysis was performed with subcategorisation of trainee cases according to specific training grade (exposure C). There was no evidence of an association between THRs performed by surgeons of any specific training grade (F1-ST2, ST3-ST8, or fellow) and an increased risk of revision (Table 23). Only a small proportion of THRs was performed by the most junior trainees (F1-ST2: n=35; 6.7%) and no episodes of revision were recorded in this group.

Table 21 - Summary of revised THRs.

Age at primary	Sex	Bearing	Fixation	Head size	Surgeon grade (specific)	Scrubbed supervision	Indication for revision	Follow-up time (years)
78	F	MoP	Cemented	28	Trainee (fellow)	Yes	Instability	0.1
58	F	CoC	Hybrid	36	Trainee (fellow)	No	Instability	0.1
60	F	CoM	Hybrid	36	Trainee (fellow)	Yes	Instability	0.2
80	F	MoP	Cemented	28	Trainee (ST3-8)	No	Instability	1.4
59	F	CoC	Hybrid	36	Trainee (ST3-8)	No	Infection	2.0
62	F	MoP	Hybrid	36	Trainee (fellow)	No	Infection	4.8
73	M	MoP	Hybrid	28	Consultant	N/A	Fracture	6.1
70	F	MoP	Cemented	28	Consultant	N/A	Aseptic	6.2
57	F	MoM	Hybrid	36	Consultant	N/A	Infection	7.4
75	M	MoP	Hybrid	28	Trainee (fellow)	No	Infection	7.5
66	M	MoM	Hybrid	36	Trainee (fellow)	No	Instability	8.3
71	F	MoP	Hybrid	28	Trainee (fellow)	No	Aseptic	9.5
71	M	MoP	Hybrid	28	Consultant	N/A	Infection	10.2
78	F	MoP	Cemented	28	Consultant	N/A	Fracture	12.7

MoP, metal-on-polyethylene; MoM, metal-on-metal; CoC, ceramic-on-ceramic; CoM, ceramic-on-metal; N/A, not applicable

Figure 26 - Kaplan-Meier plot (one minus survival) demonstrating the cumulative probability of THR failure (i.e. all-cause revision) according to surgeon grade (exposure A).

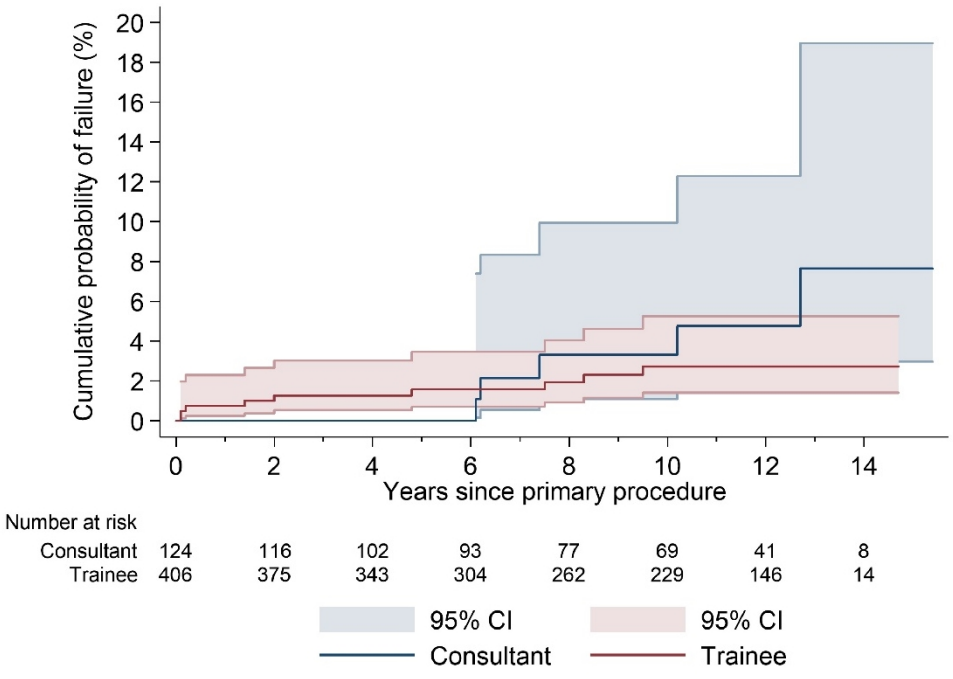


Table 22 - Results of Cox models for all-cause revision according to surgeon grade (exposure A) and supervision (exposure B).

Outcome	Exposure subgroup	Exposure	Model 1 (unadjusted)			Model 2 (adjusted for †)			Model 3 (adjusted for †, ‡)		
			n=519			n=519			n=519		
			HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
All-cause revision	A	Consultant*	1.00	-	-	1.00	-	-	1.00	-	-
		Trainee (overall)	0.54	0.18 to 1.62	0.28	0.57	0.19 to 1.72	0.32	0.51	0.15 to 1.71	0.28
	B	Consultant*	1.00	-	-	1.00	-	-	1.00	-	-
		Trainee supervised by a scrubbed consultant	0.24	0.05 to 1.25	0.09	0.25	0.05 to 1.33	0.11	0.19	0.03 to 1.11	0.07
		Trainee not supervised by a scrubbed consultant	0.87	0.28 to 2.75	0.82	0.94	0.29 to 3.02	0.92	0.85	0.25 to 2.96	0.80

Data are hazard ratio (HR), 95% CI, or p-value. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). ‡**Operation factors:** cup fixation (cemented§); head size (28mm§), cup brand (Ogee§), bearing (MoP§). *Reference group. §Represents the baseline category for each confounding variable.

Table 23 - Sensitivity analysis: Results of Cox models for all-cause revision according to specific training grade (exposure C).

Outcome	Exposure subgroup	Exposure	Number of cases (n)	Number of revisions (n)	Model 1 (unadjusted)			Model 2 (adjusted for †)			Model 3 (adjusted for †, ‡)		
					n=519			n=519			n=519		
					HR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
All-cause revision	C	Consultant*	119	5	1.00	-	-	1.00	-	-	1.00	-	-
		F1-ST2**	35	0	-	-	-	-	-	-	-	-	-
		ST3-ST8	176	2	0.26	0.05 to 1.36	0.11	0.27	0.05 to 1.44	0.13	0.27	0.05 to 1.52	0.14
		Fellow	189	7	0.93	0.30 to 2.94	0.91	0.90	0.29 to 2.86	0.86	0.84	0.24 to 3.00	0.79

Data are hazard ratio (HR), 95% CI, or p-value. †**Patient factors:** age (centred§); sex (female§); ASA (2§); indication (OA§). ‡**Operation factors:** cup fixation (cemented§); head size (28mm§), cup brand (Ogee§), bearing (MoP§). *Reference group. §Represents the baseline category for each confounding variable.

6.6 Discussion

This study, which is based on a consecutive series of 530 THRs with up to 15 years follow-up, offers new insight into the clinical, radiological, and functional outcomes of trainee-performed THR. Trainees in this cohort achieved comparable outcomes to consultants in terms of the OHS at minimum 10 years of follow-up, the rate of surgical complications, the quality of femoral cementation, radiological periprosthetic lucency, and all-cause revision. We observed differences in the SF-12 scores (MCS and PCS) between the groups. However, trainees who were supervised by a scrubbed consultant achieved equivalent PROMs to consultants.

6.6.1 Strengths and limitations

Strengths of this study include the use of a range of clinical, functional, and radiological outcome measures that are not currently reported by the NJR. We report PROMs and implant survival data for trainee-performed procedures with more than 10 years follow-up, which is longer than previously reported (Fowler et al., 2021, Singh et al., 2019). Furthermore, the study is based on a consecutive series of primary THRs for unselected indications, which increases the generalisability of the findings.

This study has methodological limitations, many of which relate to its observational design and the study population, which is small compared to existing registry studies on this subject (Inglis et al., 2013, Fowler et al., 2022). It is based on a selected cohort of THRs that used a cemented C-stem AMT femoral component. Therefore, the findings may not be generalisable to constructs that utilise alternative stem designs, or uncemented femoral fixation. The high proportion of trainee-performed THRs (77%) in this single-centre study is not representative of wider practice in England and Wales, where only 10% of THRs are performed by trainees (Fowler et al., 2022). The outcomes of this institution may not be generalisable to centres where a lower proportion of procedures are performed by trainees. The binary term ‘trainee’ does not account for the variation in the level of experience between individual trainees, or for procedures that may have been part-performed by a trainee. Similarly, the binary variable used to describe the level of supervision does not account for the

spectrum of supervision that is inherent to the training process, for example, procedures in which trainees are supervised by an unscrubbed consultant.

The PROMs analysis is limited by the lack of preoperative data, as neither the OHS nor SF-12 were routinely recorded at preoperative assessment. Differences in SF-12-MCS and PCS scores were observed between the groups, but the clinical significance of this finding is not clear without comparison to a preoperative baseline. While up-to-date electronic patient records were reviewed for all patients, we are unable to account for complications that may have been treated in other units. We attempted to account for this with the postal questionnaire, which included a free-text area for patients to document further surgery to the replaced hip. However, this does not account for complications that may have occurred in non-responders or complications that were managed in other units but did not require surgical intervention. Radiographic outcome data were collected by two authors (as documented in Section 6.4.1). There is existing evidence to suggest that the radiographic measures used in this study demonstrate limited interobserver and intraobserver reliability (McCaskie et al., 1996, Smith et al., 2011). Our radiographic analysis did not include measures of interobserver or intraobserver reliability, which is a potential source of bias.

Although our models were adjusted for a range of patient and operation factors, we are unable to fully account for variations in case complexity, which leaves the potential for residual confounding. Consultants may have carried out selection so that they operated on more complex cases, which was not accounted for by the measured confounding variables used in our analyses. Preoperative radiographs were available for 74.7% (n=396) of cases (Figure 24), but these were not included in our radiographic analysis. Methods of classifying procedural complexity based on the assessment of preoperative radiographs have been described (D'Antonio et al., 1989, D'Antonio et al., 1993). The evaluation of preoperative radiographs could have provided additional information regarding differences in case complexity between the consultant and trainee groups.

6.6.2 Comparison with other studies

There was no evidence of an association between surgeon grade (exposure A) or the level of supervision (exposure B) and the risk of all-cause revision of THRs with up to 15 years follow-up (Table 22). This finding is concordant with the shorter-term results of national registry studies (Fowler et al., 2022, Inglis et al., 2013). However, descriptive data presented in Table 21 highlight that all five THRs that were revised for instability were performed by trainees; three of whom were not supervised by a scrubbed consultant. The NJR study presented in Chapter 3 found evidence of an association between THRs performed by trainees who were not supervised by a scrubbed consultant and an increased risk of early revision for instability, which is consistent with the results of the current study. Importantly, NJR data suggest that THRs performed by trainees who are supervised by a scrubbed consultant are not associated with an increased risk of revision for instability (Fowler et al., 2022).

There was no significant association between surgeon grade and the overall complication rate, or the likelihood of specific surgical complications including dislocation and deep infection (Table 17). This is consistent with the findings of a recent systematic review (Singh et al., 2019).

Trainees achieved comparable outcomes to consultants in terms of the Barrack grade on the initial postoperative radiograph, and the extent of femoral periprosthetic radiolucency on the most recent follow-up radiograph (Tables 18-20). This is comparable to the findings of Moran et al. who found no difference in the Barrack grade for cemented THRs performed by trainees and consultants (Moran et al., 2004). They performed an extensive radiological analysis, which included measurements of cup and stem alignment in addition to the quality of cementation. They found that trainees inserted the acetabular component in less anteversion compared to consultants, but that this was not associated with a higher rate of dislocation. The relationships between surgeon grade, implant alignment and instability may warrant further investigation. However, the low overall number of dislocations and seven different stem-cup combinations would preclude meaningful analysis of this relationship in the current study.

6.6.3 Implications & Conclusion

Trainees in this cohort achieved comparable outcomes to consultants in terms of the OHS at minimum 10 years of follow-up, the rate of surgical complications, the quality of femoral cementation, radiological periprosthetic lucency, and all-cause revision. Trainees who were supervised by a scrubbed consultant achieved comparable PROMs (OHS and SF-12) to consultants. The results of this study support the current methods by which surgeons are trained to perform THR in the UK. These findings are concordant with registry evidence and give reassuring insight into the outcomes of THR performed by appropriately supervised trainees in the UK setting.

Chapter 7 Concluding discussion

7.1 Overview

NJR data show that approximately 10% of THRs, 10% of TKRs, and 4% of UKRs in England and Wales are performed by trainees. The data presented in this thesis suggest that over 159,000 primary hip and knee replacements have been performed by trainees in England and Wales since 2003, with varying levels of consultant supervision. However, until now, our understanding of implant survival in the context of surgical training has been poorly understood. We have not known whether trainees in England and Wales achieve comparable outcomes to fully-trained consultant surgeons, and the importance of supervision in this context has not been quantified. Our understanding of this subject has been based on a small number of low-quality studies, and previous registry studies on this subject have methodological limitations. The work presented here represents a substantial novel contribution to the subject.

A range of research methods have been employed to meet the objectives of this thesis. Evidence synthesis has been used to establish estimates of implant survival for consultant- and trainee-performed procedures. Comprehensive analyses of NJR data have been conducted to investigate the association between surgeon grade, the supervision of trainees, and the risk of revision following THR, TKR, and UKR. The research question in these NJR studies is consistent, but different methodological approaches were required to account for inherent differences in the data. The primary outcome throughout this thesis is all-cause revision. However, the indication for revision has also been extensively investigated in order to explore whether there are variations in outcome between the groups, which may be demonstrated by different rates of revision for specific indications. Single-centre data have been used to gain additional understanding of the functional and radiological outcomes of trainee-performed THR. A single question unifies the five studies included in this thesis – do primary hip and knee replacements last as long when a trainee performs the operation compared to a consultant? The answer to this question helps us define if trainees performing joint replacements

in the current context of training and supervision is safe, defined by equivalence of revision outcomes. This concluding chapter aims to summarise the principal findings of this thesis, discuss the limitations of the included studies, draw comparisons to the existing literature, and outline potential areas for future research.

7.2 Summary of findings

This thesis consists of five main research chapters, which employ a range of methodologies to explore the associations between surgeon grade, the supervision of trainees, and the risk of revision following THR, TKR and UKR. A summary of each chapter is discussed here.

7.2.1 Association between surgeon grade and implant survival following hip and knee replacement: a systematic review and meta-analysis

The aim of this study was to conduct evidence synthesis using existing published data on the association between surgeon grade and implant survival outcomes following hip and knee replacement. Nine observational studies capturing 4,066 THRs, 936 TKRs, and 1,357 UKRs were included. There was no strong evidence in the existing literature that trainees achieve worse outcomes compared to consultants, in terms of the net survival or crude revision rate of hip and knee replacements at 5 to 10 years follow-up. These findings are limited by the quality of the included studies, small sample sizes, and are only applicable to countries with established orthopaedic training programmes from which the studies were drawn.

7.2.2 The association between surgeon grade and risk of revision following total hip replacement: an analysis of National Joint Registry data

This study included 603,474 primary THRs performed in England and Wales between 2003 and 2016 for an indication of OA only. Trainees performed 58,137 (9.6%) THRs and were supervised by a scrubbed consultant in 57.2% of trainee-performed cases. Outcomes of interest were all-cause revision and the indication for revision up to 10 years. The unadjusted cumulative probability of failure at 10 years was 3.34% (95% CI 3.25 to 3.43) for consultants and 3.10% (95% CI 2.85 to 3.32)

for trainees. The upper confidence interval for trainee-performed THRs was below the ODEP A* threshold at all intervals of follow-up, regardless of the level of scrubbed supervision.

Adjusted analysis was primarily by means of Cox regression models, which were incrementally adjusted for patient, operation, and healthcare setting factors. There was no evidence of an association between surgeon grade and the risk of all-cause revision. Furthermore, we found no evidence of an association between THRs performed by trainees supervised by a scrubbed consultant and an increased risk of revision for any indication, including all-cause revision, infection, periprosthetic fracture, aseptic loosening, instability, and other causes. THRs performed by trainees who were not supervised by a scrubbed consultant were associated with an increased risk of all-cause revision (in the fully adjusted model only) and revision for instability (in the unadjusted model and with adjustment for patient factors, but not in the fully adjusted model).

Further analysis using FPM demonstrated that there might be an increased risk of early revision up to, but not exceeding, 6 months after trainee-performed THR; an effect which is predominantly attributable to cases of early revision for instability. This study of THRs with up to 10 years of follow-up showed that appropriately supervised trainees achieve comparable implant survival to consultants. However, there is some evidence that trainees who are not supervised by a scrubbed consultant achieve inferior results to consultants, particularly in terms of revision for instability.

7.2.3 The association between surgeon grade and risk of revision following total knee replacement: an analysis of National Joint Registry data

This study included nearly one million primary TKRs with up to 16.8 years of follow-up. Trainees performed 96,544 (10.1%) procedures and were supervised by a scrubbed consultant in 63.2% of trainee cases. Unadjusted KM analysis indicated that trainees achieve comparable all-cause implant survival to consultants. The upper confidence interval for trainee-performed TKRs was below the ODEP A* threshold at all intervals of follow-up, regardless of the level of scrubbed consultant supervision.

Adjusted FPM analysis gave additional insight into the temporal variation in the risk of revision. There was evidence of an association between trainee-performed TKR and an increased risk of all-cause revision within the first 4 years of follow-up, which was not explained by the level of supervision. Further FPM analysis identified marginal associations between trainee-performed TKR and early revision for aseptic loosening (up to 3 years), infection (up to 3 years), and progression of OA (up to 5 years).

In practice, any absolute difference in the risk of revision between consultant- and trainee-performed TKRs is very small. Overall, trainees achieve comparable implant survival estimates to consultants, which are within an internationally recognised acceptable limit. However, trainers should take appropriate measures to mitigate the risk of early revision for indications, such as infection, aseptic loosening and progression of OA.

7.2.4 The association between surgeon grade and risk of revision following unicompartmental knee replacement: an analysis of National Joint Registry data

This NJR study included 106,206 primary UKRs performed between 2003 and 2019 for an indication of OA only. Trainees performed 4,382 (4.1%) UKR procedures and were supervised by a scrubbed consultant in 66.1% of cases. Adjusted analyses used FPM, with confounding adjustment for patient, operation, and healthcare setting factors. There was no association between surgeon grade and all-cause revision in crude or adjusted models. Trainees achieved comparable all-cause survival to consultants, regardless of the level of scrubbed consultant supervision. We found evidence that UKRs performed by unsupervised trainees were associated with an increased risk of revision for unexplained pain, but this was not observed in the fully adjusted model.

7.2.5 A comparison of clinical and radiological outcomes between trainee- and consultant-performed total hip replacement: a retrospective cohort study of 530 hips with up to 15 years follow-up

The preceding chapters further our understanding of the outcomes of trainee-performed hip and knee replacements. However, the conclusions of these studies are limited as they are based on survival data alone. An additional study was conducted using a single-centre consecutive series of 530 primary

THRs with up to 15 years of follow-up, which investigated a range of clinical and radiological outcome measures.

A large proportion of cases in this series were performed by trainees (77.1%), and a scrubbed consultant directly supervised 51% of trainee-performed cases. Trainees in this cohort achieved comparable outcomes to consultants in terms of all-cause revision, the rate of surgical complications, the quality of femoral cementation, periprosthetic radiolucency, and the >10-year OHS. There were differences in the SF-12 scores (MCS and PCS) between the groups. However, trainees who were supervised by a scrubbed consultant achieved equivalent PROMs to consultants (OHS and SF-12).

7.2.6 Summary

The collective findings of this thesis should be reassuring for patients and the key stakeholders involved in orthopaedic training in England and Wales. In general, the findings support the current methods by which surgeons in England and Wales are trained to perform primary THR, TKR, and UKR. Our results provide novel insight that may be used to improve the outcomes of trainee-performed surgery. Of note, we have identified several situations in which procedures performed by trainees who are directly supervised by a scrubbed consultant are associated with superior implant survival outcomes, compared to procedures performed by trainees without scrubbed consultant supervision.

7.3 Strengths and limitations

7.3.1 Strengths

The collective body of work included in this thesis has several strengths, notably the large number of patients, the use of generalisable registry data, the length of follow-up, and robust statistical methodology. Furthermore, the presented work is consistent with current research priorities in this field, including "the preoperative, intraoperative, and postoperative factors that can be modified to influence outcome following hip and knee replacement (James Lind Alliance, 2021)."

A range of methods have been used, including evidence synthesis, registry-based survival analysis, and a local single-centre observational study. On a personal level, this approach has allowed me to establish a deep understanding of the subject while gaining invaluable experience in a range of research methodologies. Over 1.6 million hip and knee replacements were included, mainly in the form of NJR data, which gives statistical power, limits selection bias, and increases the generalisability of results. The statistical modelling techniques used, and the completeness of the data have facilitated comprehensive adjustment for confounding variables. Where NJR data lacked proportionality, FPM allowed us to overcome the limitations of the Cox model assumptions and explore the time-varying effects of the variables included in the models. This thesis significantly furthers our understanding of this subject.

7.3.2 Limitations

Each of the studies included in this thesis has limitations that must be considered when interpreting the findings, drawing conclusions, and making recommendations based on this work. Several limitations are consistent throughout this thesis and warrant further discussion.

7.3.2.1 Implant survival as an outcome

A limitation of this thesis is the use of implant survival as an outcome in the absence of additional outcome data, such as PROMs. Time-to-event data, in which all-cause revision was the primary endpoint, were used to calculate implant survival (or failure) estimates and assess the outcomes of hip and knee replacements performed by trainees compared to consultants. Specific indications for revision were examined using secondary survival endpoints. However, this objective outcome measure does not account for patients with unfavourable outcomes who have not undergone revision surgery. A patient may perceive their joint replacement to have failed but been deemed ineligible or chosen not to undergo revision surgery. Thus estimates of implant survival can overestimate the success of a joint replacement, particularly when reported in the absence of alternative measures, such as PROMs or complication rates.

The limitations of implant survival as an outcome measure are discussed in an editorial on this subject. Wylde et al. draw attention to the proportion of patients who experience unfavourable long-term pain outcomes after THR and TKR (Wylde and Blom, 2011). They cite several studies in which composite measures of success (combining objective and subjective measures) illustrate the shortcomings of implant survival when used as a standalone outcome measure (Murray and Frost, 1998, Bullens et al., 2001). For example, in a relatively small cohort of TKRs, Bullens et al. showed that success defined by implant survival alone was 96.7% at 5 years. However, when success was defined using a composite endpoint of revision, satisfaction and pain, the success rate fell to 68.8% (Bullens et al., 2001).

It is possible to link NHS England PROM data to the NJR, but complete preoperative and postoperative PROMs are only available for a relatively small proportion of NJR records; in the region of 10-25% (Sayers et al., 2020b, Liddle et al., 2015, Baker et al., 2012). While it was not feasible to include NHS England PROMs data in this thesis, we have used a local single-centre study to investigate a broader range of outcomes following trainee-performed THR. Outcome measures included all-cause revision, complication rates, PROMs (SF-12 and OHS), and radiological outcomes, including Barrack grade and periprosthetic radiolucency. This study was relatively small and preoperative PROMs were not available. Nonetheless, it gives additional insight into the outcomes of trainee-performed THR beyond implant survival.

7.3.2.2 Variations in NJR operating surgeon grade categories

The current version of the NJR data entry form (MDS version 7.0) allows the grade of the operating surgeon to be recorded using the following categories: F1-ST2; ST3-ST8; SAS; consultant; or ‘other’. However, these categories have changed over time to reflect changes in training terminology. For example, in previous versions of the form (until 2010), it was possible to record the operating surgeon as a ‘fellow’. Using MDS version 7.0, a procedure performed by a ‘fellow’ may now be recorded as performed by ‘other’. Cases in which the surgeon was recorded as ‘other’ were excluded from our primary NJR analyses, as it was not possible to determine whether or not a trainee performed the

procedure. A small proportion of trainee-performed procedures are likely to have been excluded in this process. The process used to account for variations in NJR grade classification is detailed in Appendix 10.

7.3.2.3 Binary exposure variables

The objective of this thesis was to compare the survival of hip and knee replacements performed by trainees compared to consultants. The use of the binary variable 'surgeon grade', which has previously been used in the orthopaedic literature (Bottomley et al., 2016, Palan et al., 2009, Jain et al., 2018, Faulkner et al., 2018, Reidy et al., 2016), is appropriate and informative in this context but does not reflect the wide variation in the level of experience between individual trainees. We have attempted to account for this through sensitivity analyses, in which cases were categorised according to the specific training grade of the operating surgeon (e.g. F1-ST2, ST3-ST8, fellow, or consultant). However, this approach also has limitations, as these broad categories do not necessarily correlate with experience in a specific procedure. For example, there are several possible scenarios in which an ST3 trainee may have performed more THRs, TKRs, or UKRs than an ST8 trainee.

The use of the binary variable 'surgeon grade' does not account for procedures that have been part-performed by a trainee. It is common for trainees to perform part of an operation, e.g. when learning a new procedure or when case complexity dictates the need for senior input. This is recognised in JCST guidelines for recording surgical training experience in the UK (Appendix 4). However, it is not currently possible to record more than one operating surgeon per case in the NJR.

The supervision of trainees during surgical procedures is a spectrum ranging from scrubbed consultant supervision (i.e. the consultant is scrubbed alongside the trainee providing guidance throughout) to a trainee operating independently without a consultant in the operating theatre. The UK T&O curriculum states there should be a gradual reduction in supervision and an increase in case complexity until the level of competence for independent practice is acquired (ISCP, 2021). This process requires the supervising consultant to make an informed judgement of a trainee's competence. However, when categorising the level of supervision using NJR data, it is only possible to categorise

supervision as a binary variable: 1) supervised by a scrubbed consultant; or 2) not supervised by a scrubbed consultant. This categorisation is based on whether the first assistant was recorded as 'consultant' or 'other'. This binary variable gives us valuable information about whether or not a trainee was directly supervised by a scrubbed consultant, but it does not capture information about the wider spectrum of supervision that is inherent to the training process.

7.3.2.4 Residual confounding

It is likely that in the context of training in England and Wales that complex cases are preferentially performed by consultants. We have attempted to account for variations in case-mix selection and case complexity between the groups, but the results remain susceptible to residual confounding. Complexity is a continuous spectrum, and differences between groups cannot be fully accounted for by adjusting for categorical confounding variables.

The indication for surgery is a source of case complexity. Primary THRs performed for inflammatory arthropathy, dysplasia, malignancy, or multiple indications are generally considered more technically challenging and may be more likely to be performed by a consultant than a trainee. Furthermore, when a trainee performs more complex cases, the trainee is generally more likely to be supervised by a scrubbed consultant, or only perform part of the procedure. In an attempt to standardise variations in case complexity relating to the indication for surgery, our NJR analyses were restricted to procedures performed for OA only. However, OA is a spectrum of disease that poses various technical challenges and is itself a source of residual confounding.

In the three NJR studies (Chapters 3-5), our statistical models were incrementally adjusted for patient, operation, and healthcare setting factors. Patient factors included age, sex, ASA grade, and IMD decile (a measure of social deprivation). The operation factors varied according to the implant. THRs were adjusted for anaesthetic, approach, fixation, head size, and bearing material. TKRs were adjusted for anaesthetic, approach, fixation, constraint, and patellar resurfacing. UKRs were adjusted for anaesthetic, approach, fixation, and bearing mobility. Finally, to account for variations in practice

over time and between public and private healthcare settings, we adjusted for the year of operation and funding source.

This incremental approach to adjusting for confounders increases transparency by demonstrating the relative contribution of patient, operation, and healthcare setting factors to the adjusted result. Adjustment for confounding variables accounts for some of the variations in case-mix selection between the groups, but there is likely to be residual confounding, which is a potential source of bias. For example, we did not adjust for Body Mass Index (BMI) as BMI is missing in a high proportion of NJR records (Sayers et al., 2020b). Furthermore, while the NJR regularly audits the process of capturing and linking primary and revision procedures, there is no robust system in place to audit the quality of confounding data (The National Joint Registry, 2021a, Konan and Haddad, 2013).

7.3.2.5 Lack of patient and public involvement

This thesis is limited by the lack of direct patient and public involvement in its design and implementation. The current "Top 10" research priorities in this field are published by the James Lind Alliance and include, "What (health service) preoperative, intraoperative, and postoperative factors can be modified to influence outcomes following hip and knee replacement?" The studies included in this thesis are consistent with this national research priority, which was defined by a priority setting partnership involving patient, carer and clinician groups (James Lind Alliance, 2021).

With regards to the NJR studies included in this thesis, patient representatives sit on the NJR committee and the research priorities of the NJR are defined and approved by this committee structure. However, patients were not directly involved in setting the research questions, defining the outcome measures, or designing and implementing the work included in this thesis. Collaboration with a thesis-specific patient and public involvement group would have been a valuable exercise that could have informed the objectives of this thesis, influenced study design and implementation, and been a source of advice regarding the communication of findings to patients and the public.

7.4 Comparison to the existing literature

7.4.1 Total hip replacement

The systematic review in Chapter 2 provides a summary of the existing literature on implant survival outcomes following trainee-performed THR. We conclude that there is no strong evidence that trainees achieve worse outcomes compared to consultants in terms of the net survival or crude revision rate of THRs with 5 to 10 years follow-up. However, this conclusion is based on low-quality evidence originating from a small number of observational studies. While another group has recently published a similar review on this subject (Singh et al., 2019), our study is the first to conduct a meta-analysis of THR net survival estimates according to surgeon grade (Fowler et al., 2021).

Chapter 3 uses NJR data to explore the association between surgeon grade and implant survival following THR. The main comparator to this work is the NZJR study of Inglis et al., which found no significant difference in the revision rate (per 100 component years) or indication for revision of THRs performed by consultants compared to supervised and unsupervised trainees (Inglis et al., 2013). We used a different statistical approach (KM, Cox regression, and FPM) in a comprehensive analysis of data for over half a million patients. Our study, which is significantly larger than previous work on this subject, suggests that appropriately supervised trainees achieve comparable implant survival to consultants. This is generally consistent with the findings of Inglis. However, we identified marginal associations between THRs performed by trainees without scrubbed consultant supervision and increased risks of all-cause revision and revision for instability. This may reflect the weakness of the true effects, which were only identified in our study due to the high number of cases and comprehensive confounding adjustment.

7.4.2 Total knee replacement

In Chapter 2, we performed a systematic review and meta-analysis of TKR implant survival outcomes according to surgeon grade. We found no strong evidence that trainees achieve inferior outcomes compared to consultants in terms of the net survival or crude revision rate of TKRs with 5 to 10 years follow-up. As with our findings for THR, this conclusion is based on low-quality evidence from a

small number of studies. Madanipour et al. have recently published a similar systematic review on this subject (Madanipour et al., 2021). Consistent with our findings, they found no significant difference in the revision rates of trainee- and consultant-performed TKRs. However, their study did not include net survival estimates (Fowler et al., 2021).

To our knowledge, there is only one other registry study on this subject. Storey et al. found no significant difference between the revision rate (per 100 component years) for TKRs performed by consultants compared to supervised senior trainees, unsupervised senior trainees, and supervised junior trainees (Storey et al., 2018). Our study in Chapter 4, which includes data for nearly one million primary TKRs recorded in the NJR, employs a statistical approach that is unique in the context of the relevant literature. FPM, with results reported as HR plots, accounts for the non-proportionality of the data and gives novel insight into the temporal variation in the risk of revision following trainee-performed TKR.

7.4.3 Unicompartmental knee replacement

Before this thesis, our understanding of the association between surgeon grade and the risk of revision following UKR had been principally based on two studies (Bottomley et al., 2016, Storey et al., 2018). In their NZJR study, Storey et al. found no significant difference in the revision rate (per 100 component years) for UKRs performed by supervised senior trainees compared to attending surgeons (Storey et al., 2018). However, insufficient data precluded any meaningful analysis of outcomes according to the level of supervision, and it is not clear what form of regression modelling was used. In a single-centre study that included 673 trainee UKRs, Bottomley et al. demonstrated the equality of the 9-year cumulative survival estimates for UKRs performed by trainees compared to consultants, but adjusted survival analysis was not conducted, and the study was restricted to a single implant design (Bottomley et al., 2016).

With over 100,000 UKRs included, our registry study (Chapter 5) is significantly larger than any previous study on this subject, has a longer follow-up, and is based on UKRs-in-general rather than a single brand. Our findings are concordant with the conclusion of Storey et al., that implant survival is

not compromised in trainee-performed UKR. Our statistical approach provides new insight on the indication for revision following trainee-performed UKR, and into the importance of scrubbed consultant supervision.

7.5 Implications and recommendations

The findings of this thesis further our understanding of the factors that contribute to the successful outcome of THRs, TKRs, and UKRs in the context of surgical training. Our findings have several implications for clinical practice and provide insight into factors that can be modified to improve outcomes following trainee-performed hip and knee replacements. The implications of our findings for THRs, TKRs, and UKRs are discussed in turn. In general, our recommendations involve increasing the proportion of trainee-performed procedures supervised by a scrubbed consultant.

When considering the implications of this thesis, it is essential to remember the training and healthcare settings in which the studies have been conducted. The findings of our systematic review and meta-analysis (Chapter 2) are based on nine studies (six from the UK), all conducted in countries with established orthopaedic training programmes. The remaining studies (Chapters 3-6) are based on over 1.6 million hip and knee replacements performed in England and Wales, which have one of the most stringent orthopaedic training programmes in the world (Tahir et al., 2021). While some of our findings are concordant with NZJR data and may apply to other countries with established orthopaedic training programmes, they are unlikely to be globally applicable. Based on the findings of this thesis, we are unable to make recommendations regarding the outcomes of trainee-performed hip and knee replacements in countries other than England and Wales.

It is also important to consider that trainees who perform THRs, TKRs, and UKRs in England and Wales are unlikely to have no prior operative experience. In the case of hip arthroplasty, junior trainees (e.g. F1-ST2) will typically gain experience in less complex orthopaedic procedures, e.g. hemiarthroplasty for fractured neck of femur, before proceeding to learn THR under supervision. Performing hemiarthroplasty in a closely supervised setting allows trainees to gain familiarity with surgical approaches to the hip joint, preparation of the femoral canal, and cementation of the femoral

component. It is generally a less complex procedure as it does not involve replacement of the acetabulum, and it is typically easier to restore stable alignment due to the larger head size. Thus, it should not be assumed that postgraduate surgical trainees with no prior orthopaedic operative experience can achieve comparable outcomes to consultants.

7.5.1 Total hip replacement

The findings of Chapter 3 suggest that appropriately supervised trainees achieve comparable implant survival to consultants. We demonstrated that the unadjusted cumulative probability of failure of THRs performed by both supervised and unsupervised trainees is below the ODEP A* threshold at intervals of follow-up ranging from 3 to 10 years. This demonstrates acceptable implant survival for THRs performed by trainees in England and Wales. However, our adjusted analyses suggest that trainees who are not supervised by a scrubbed consultant achieve inferior results compared to consultants, particularly in terms of revision for instability. Based on previous evidence (Moran et al., 2004), this observation may relate to suboptimal or inconsistent implant positioning when trainees are not supervised by a scrubbed consultant.

Chapter 6 is a smaller study based in a single institution, but it gives additional insight into outcome measures other than implant survival. Trainees in this cohort achieved comparable outcomes to consultants in a range of clinical and radiological outcomes measures. Trainees who were supervised by a scrubbed consultant achieved comparable PROMs to consultants (OHS and SF-12).

Approximately 43% of trainee-performed THRs recorded in the NJR since 2003 were performed by trainees who were not directly supervised by a scrubbed consultant. Our findings support our recommendation that trainees should be supervised by a scrubbed consultant when performing THR (Fowler et al., 2022). We suggest that this is particularly important during the early years of training and may reduce the risk of revision for instability by helping to ensure that trainees select appropriate implants and position them correctly. We acknowledge that the primary aim of surgical training in the UK is to develop surgeons capable of safe, independent practice at the level that is expected of a "day one consultant" (ISCP, 2021). Therefore, we recommend that if competent trainees are allowed to

operate without scrubbed supervision, careful case selection is undertaken and that a consultant is readily available to provide scrubbed supervision if required.

7.5.2 Total knee replacement

Chapter 4 compared the cumulative probability of failure of consultant- and trainee-performed TKRs to the ODEP A* benchmark. The upper confidence interval for trainee-performed TKRs was below the ODEP A* threshold at all intervals of follow-up ranging from 3 to 15 years, regardless of the level of scrubbed consultant supervision. This supports the interpretation that trainees in England and Wales achieve safe and acceptable TKR implant survival.

The findings of our adjusted FPM analyses suggest that trainee-performed TKRs may be susceptible to a small increased risk of early revision for aseptic loosening, infection, and progression of OA. Given these findings, we recommend that careful case selection should be undertaken, and consultants and trainees should take appropriate measures to mitigate against the small transient increase in risk of revision from these indications. Trainees should be supervised by a scrubbed consultant when performing TKR, particularly when junior and at critical stages of the procedure such as implant selection, balancing, fixation, and deciding whether or not to resurface the patella.

7.5.3 Unicompartmental knee replacement

The findings of our NJR study in Chapter 5 support the current training practices for UKR in England and Wales and suggest that trainees achieve comparable UKR survival to consultants. Our interpretation of these findings is that, in general, trainers select appropriate cases for their trainees and permit trainees to operate without scrubbed supervision only when they have reached a subjective threshold of expertise commensurate with safe independent surgical practice. We found evidence that UKRs performed by trainees without scrubbed consultant supervision may be associated with an increased risk of revision for unexplained pain compared to consultant-performed UKR. This was the case in the crude analysis and with adjustment for patient and operation-level factors but was not observed in the fully adjusted model. We recommend that trainees should ideally be supervised by a scrubbed consultant when performing UKR, particularly during the early phases of training. When

competent trainees are allowed to operate without scrubbed consultant supervision, cases should be carefully selected, and scrubbed supervision should be readily available.

7.5.4 Recommendations in the context of the COVID-19 pandemic

The studies included in this thesis do not include any operations performed after 31 December 2019, i.e. before the first UK death from COVID-19 was confirmed in March 2020 (Mahase, 2020). Cases performed after 31 December 2019 were deliberately excluded from our NJR studies on TKRs and UKRs to avoid the anomalous period of elective orthopaedic practice during the COVID-19 pandemic. The pandemic has significantly reduced the number of elective joint replacements performed in England and Wales and has left surgical training in crisis with a detrimental reduction in trainee case numbers (The National Joint Registry, 2021a, Munro et al., 2021, Clements et al., 2021).

To ensure the competency of the future orthopaedic workforce, it is more important than ever that trainees have the opportunity to perform operations. Our findings should be reassuring to patients and key stakeholders involved in orthopaedic training in England and Wales. Appropriately supervised trainees achieve comparable implant survival to consultants, but the recommendations discussed here have the potential to further improve the outcomes of trainee-performed hip and knee replacements during the recovery of elective orthopaedic services.

7.6 Future research

The presented work meets the aims and objectives of this thesis, but there is clear scope for further investigation. Three additional themes of research on this subject are outlined here.

7.6.1 Collaboration with other registries

Excluding Chapter 2, our findings are solely based on data from England and Wales. While some of our findings are concordant with NZJR data, the existing registry studies on this subject used markedly different statistical methodology (Storey et al., 2018, Inglis et al., 2013). A collaborative approach between international registries using consistent modelling techniques would increase the

generalisability of results to other healthcare settings and may help to identify training practices that lead to favourable outcomes for patients.

7.6.2 Patient-reported outcome measures

Previous attempts to use registry data to investigate the association between surgeon grade (or equivalent measure) and PROMs following primary hip and knee replacement are summarised in Section 1.8 (Storey et al., 2018, Inglis et al., 2013, Jolbäck et al., 2018). These existing studies are limited by several factors. In an analysis of SHAR data, Jolbäck et al. used adjusted regression models to analyse both preoperative and postoperative data (EQ-5D, EQ-5D VAS, Pain VAS, and postoperative satisfaction). Their study only included 6,713 THRs, trainee cases were not subcategorised according to the level of supervision, and the maximum duration of follow-up was 1-year (Jolbäck et al., 2018). Storey and Inglis included 88,525 knee replacements (79,671 TKRs and 8,854 UKRs) and 35,415 THRs, respectively. However, their findings are limited by the lack of preoperative data and the short duration of follow-up (6 months). Future attempts to use registry data to investigate the association between surgeon grade and PROMs following hip and knee replacement should include complete preoperative and postoperative data, subcategorise trainee cases according to the level of supervision, and use a combination of generic health and joint-specific instruments. While it was not within the scope of this thesis, it would be possible to perform such a study by linking NHS England PROMs data to the NJR. The duration of follow-up would be limited to 6 months, but there is evidence and precedent to support this (Liddle et al., 2015).

7.6.3 Alternative measures of trainee experience

As previously discussed, the binary variable ‘surgeon grade’ and the grade categories on the NJR MDS form do not reflect the wide variation in the level of experience between individual trainees. In Chapter 1, a range of measures that have previously been used to define surgical experience are described. It would be interesting to explore alternative measures of experience in the context of trainee-performed hip and knee replacement. For example, registry data could be used to investigate whether trainees exhibit the same volume effects as consultants, assuming that all of their procedures

are captured by single registries, and to explore whether supervising consultant volume is associated with the risk of revision following trainee-performed surgery.

7.7 Concluding remarks

The primary aim of this thesis was to investigate the association between surgeon grade and implant survival following hip and knee replacement. We have achieved this using a range of methodologies, and the findings further our understanding of the factors that contribute to successful outcomes following THR, TKR, and UKR in the context of surgical training. Our findings support the conclusion that appropriately supervised trainees in England and Wales achieve comparable implant survival to consultants. However, future studies are required to further investigate the association between surgeon grade, the supervision of trainees, and additional outcome measures, including PROMs. Further investigation is also needed to establish the extent to which our findings are generalisable to other countries and healthcare settings.

The current findings have implications for clinical practice. We have identified several areas of practice that could be modified to improve implant survival following trainee-performed hip and knee replacements. Most importantly, the balance between training surgeons and ensuring the best outcomes for patients could be improved by increasing the proportion of trainee-performed procedures that are supervised by a scrubbed consultant.

References

Abu Al-Rub, Z., Lamb, J. N., West, R. M., Yang, X., Hu, Y. & Pandit, H. G. 2020. Survivorship of fixed vs mobile bearing unicompartmental knee replacement: A systematic review and meta-analysis of sixty-four studies and National Joint Registries. *Knee,* 27**,** 1635-1644.

Aggarwal, R. & Darzi, A. 2006. Technical-skills training in the 21st century. *N Engl J Med,* 355**,** 2695-6.

Agency for Healthcare Research and Quality. 2018. *AHRQ: Healthcare Cost and Utilisation Project - Hospital Inpatient National Statistics* [Online]. Available: https://hcupnet.ahrq.gov/#setup [Accessed 26/01/2022 2022].

Alvand, A., Wilson, H. A., Sabah, S. A., Middleton, R., Bottomley, N., Jackson, W. F. M. & Price, A. J. 2021. New instrumentation system for cementless mobile-bearing unicompartmental knee arthroplasty improves surgical performance particularly for trainees. *Knee,* 31**,** 46-53.

Anis, H. K., Sodhi, N., Klika, A. K., Mont, M. A., Barsoum, W. K., Higuera, C. A. & Molloy, R. M. 2019. Is Operative Time a Predictor for Post-Operative Infection in Primary Total Knee Arthroplasty? *J Arthroplasty,* 34**,** S331-S336.

Atkins, D., Best, D., Briss, P. A., Eccles, M., Falck-Ytter, Y., Flottorp, S., Guyatt, G. H., Harbour, R. T., Haugh, M. C., Henry, D., Hill, S., Jaeschke, R., Leng, G., Liberati, A., Magrini, N., Mason, J., Middleton, P., Mrukowicz, J., O'connell, D., Oxman, A. D., Phillips, B., Schunemann, H. J., Edejer, T., Varonen, H., Vist, G. E., Williams, J. W., Jr., Zaza, S. & Group, G. W. 2004. Grading quality of evidence and strength of recommendations. *BMJ,* 328**,** 1490.

AOANJRR. 2020. *Australian Orthopaedic Association National Joint Replacement Registry: Annual Report 2020* [Online]. Available: https://aoanjrr.sahmri.com/documents/10180/689619/Hip%2C+Knee+%26+Shoulder+Arthroplasty+New/6a07a3b8-8767-06cf-9069-d165dc9baca7 [Accessed 29/11/2021 2021].

Bachmeier, C. J., March, L. M., Cross, M. J., Lapsley, H. M., Tribe, K. L., Courtenay, B. G., Brooks, P. M., Arthritis, C. & Outcome Project, G. 2001. A comparison of outcomes in osteoarthritis patients undergoing total hip and knee replacement surgery. *Osteoarthritis Cartilage,* 9**,** 137-46.

Baker, P., Jameson, S., Critchley, R., Reed, M., Gregg, P. & Deehan, D. 2013. Center and surgeon volume influence the revision rate following unicondylar knee replacement: an analysis of 23,400 medial cemented unicondylar knee replacements. *J Bone Joint Surg Am,* 95**,** 702-9.

Baker, P. N., Petheram, T., Jameson, S. S., Avery, P. J., Reed, M. R., Gregg, P. J. & Deehan, D. J. 2012. Comparison of patient-reported outcome measures following total and unicondylar knee replacement. *J Bone Joint Surg Br,* 94**,** 919-27.

Barrack, R. L., Mulroy, R. D., Jr. & Harris, W. H. 1992. Improved cementing techniques and femoral component loosening in young patients with hip arthroplasty. A 12-year radiographic review. *J Bone Joint Surg Br,* 74**,** 385-9.

Bayliss, L. E., Culliford, D., Monk, A. P., Glyn-Jones, S., Prieto-Alhambra, D., Judge, A., Cooper, C., Carr, A. J., Arden, N. K., Beard, D. J. & Price, A. J. 2017. The effect of patient age at intervention on risk of implant revision after total replacement of the hip or knee: a population-based cohort study. *Lancet,* 389**,** 1424-1430.

Beard, D. J., Davies, L. J., Cook, J. A., Maclennan, G., Price, A., Kent, S., Hudson, J., Carr, A., Leal, J., Campbell, H., Fitzpatrick, R., Arden, N., Murray, D., Campbell, M. K. & Group, T. S. 2019. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. *Lancet,* 394**,** 746-756.

Beattie, N., Maempel, J. F., Roberts, S., Waterson, H. B., Brown, G., Brenkel, I. J. & Walmsley, P. J. 2018. Surgery performed by supervised registrars does not adversely affect medium-term functional outcomes after total knee replacement. *Ann R Coll Surg Engl,* 100**,** 57-62.

Berstock, J. R., Whitehouse, M. R., Piper, D. C., Eastaugh-Waring, S. J. & Blom, A. W. 2014. A 5-8 year retrospective follow-up of the C-stem AMT femoral component: patient reported outcomes and survivorship analysis. *J Arthroplasty,* 29**,** 1753-7.

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

Bhandari, M., Smith, J., Miller, L. E. & Block, J. E. 2012. Clinical and economic burden of revision knee arthroplasty. *Clin Med Insights Arthritis Musculoskelet Disord,* 5**,** 89-94.

Blom, A. W., Hunt, L. P., Matharu, G. S., Reed, M. & Whitehouse, M. R. 2021. The effect of surgical approach in total knee replacement on outcomes. An analysis of 875,166 elective operations from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *Knee,* 31**,** 144-157.

Blom, A. W., Hunt, L. P., Matharu, G. S., Reed, M. R. & Whitehouse, M. R. 2020. The effect of surgical approach in total hip replacement on outcomes: an analysis of 723,904 elective operations from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *BMC Med,* 18**,** 242.

The British Orthopaedic Association. 2021. *BOA: Written evidence submitted by The British Orthopaedic Association (CBP0008)* [Online]. Available: https://committees.parliament.uk/writtenevidence/38332/pdf/ [Accessed 08/03/2022 2022].

Bohm, E. R., Kirby, S., Trepman, E., Hallstrom, B. R., Rolfson, O., Wilkinson, J. M., Sayers, A., Overgaard, S., Lyman, S., Franklin, P. D., Dunn, J., Denissen, G., A, W. D., Ingelsrud, L. H. & Navarro, R. A. 2021. Collection and Reporting of Patient-reported Outcome Measures in Arthroplasty Registries: Multinational Survey and Recommendations. *Clin Orthop Relat Res,* 479**,** 2151-2166.

Bottomley, N., Jones, L. D., Rout, R., Alvand, A., Rombach, I., Evans, T., Jackson, W. F., Beard, D. J. & Price, A. J. 2016. A survival analysis of 1084 knees of the Oxford unicompartmental knee arthroplasty: a comparison between consultant and trainee surgeons. *Bone Joint J,* 98-B**,** 22-27.

Bullens, P. H., Van Loon, C. J., De Waal Malefijt, M. C., Laan, R. F. & Veth, R. P. 2001. Patient satisfaction after total knee arthroplasty: a comparison between subjective and objective outcome assessments. *J Arthroplasty,* 16**,** 740-7.

Calman, K. C., Temple, J. G., Naysmith, R., Cairncross, R. G. & Bennett, S. J. 1999. Reforming higher specialist training in the United Kingdom--a step along the continuum of medical education. *Med Educ,* 33**,** 28-33.

Carr, N. J., Burke, M. M., Corbishley, C. M., Suarez, V. & Mccarthy, K. P. 2002. The autopsy: lessons from the National Confidential Enquiry into Perioperative Deaths. *J R Soc Med,* 95**,** 328-30.

Centers for Disease Control and Prevention. 2021. *CDC: NHSN Surgical Site Infection (SSI) Checklist* [Online]. Available: https://www.cdc.gov/nhsn/pdfs/checklists/ssi-checklist-508.pdf [Accessed 10/11/2021 2021].

Centre for Workforce Intelligence. 2014. *CfWI: Securing the future workforce supply: Trauma and orthopaedic surgery stocktake* [Online]. Available: https://recipeforworkforceplanning.hee.nhs.uk/Portals/0/HEWM_LinksAndResources/Trauma%20and%20Orthopaedic%20Surgery%20stocktake%20-%20Sept%202014.pdf?ver=sQluh6ReNd5f2LCxPlyLYw%3d%3d [Accessed 08/03/2022 2022].

Chambers, I. R., Fender, D., Mccaskie, A. W., Reeves, B. C. & Gregg, P. J. 2001. Radiological features predictive of aseptic loosening in cemented Charnley femoral stems. *J Bone Joint Surg Br,* 83**,** 838-42.

Charnley, J. 1961. Arthroplasty of the hip. A new operation. *Lancet,* 1**,** 1129-32.

Clement, N. D., Macdonald, D. & Simpson, A. H. 2014. The minimal clinically important difference in the Oxford knee score and Short Form 12 score after total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc,* 22**,** 1933-9.

Clements, J. M., Burke, J. R., Hope, C., Nally, D. M., Doleman, B., Giwa, L., Griffiths, G. & Lund, J. N. 2021. The quantitative impact of COVID-19 on surgical training in the United Kingdom. *BJS Open,* 5.

Collins, C. 1999. Surgical training, supervision, and service. *BMJ,* 318**,** 682-3.

Cox, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological),* 34**,** 187-202.

Culliford, D., Maskell, J., Judge, A., Cooper, C., Prieto-Alhambra, D., Arden, N. K. & Group, C. O. S. 2015. Future projections of total hip and knee arthroplasty in the UK: results from the UK Clinical Practice Research Datalink. *Osteoarthritis Cartilage,* 23**,** 594-600.

Culliford, D. J., Maskell, J., Kiran, A., Judge, A., Javaid, M. K., Cooper, C. & Arden, N. K. 2012. The lifetime risk of total hip and knee arthroplasty: results from the UK general practice research database. *Osteoarthritis Cartilage,* 20**,** 519-24.

D'antonio, J., Mccarthy, J. C., Bargar, W. L., Borden, L. S., Cappelo, W. N., Collis, D. K., Steinberg, M. E. & Wedge, J. H. 1993. Classification of femoral abnormalities in total hip arthroplasty. *Clin Orthop Relat Res***,** 133-9.

D'antonio, J. A., Capello, W. N., Borden, L. S., Bargar, W. L., Bierbaum, B. F., Boettcher, W. G., Steinberg, M. E., Stulberg, S. D. & Wedge, J. H. 1989. Classification and management of acetabular abnormalities in total hip arthroplasty. *Clin Orthop Relat Res***,** 126-37.

Daigle, M. E., Weinstein, A. M., Katz, J. N. & Losina, E. 2012. The cost-effectiveness of total joint arthroplasty: a systematic review of published literature. *Best Pract Res Clin Rheumatol,* 26**,** 649-58.

Dawson, J., Fitzpatrick, R., Carr, A. & Murray, D. 1996a. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br,* 78**,** 185-90.

Dawson, J., Fitzpatrick, R., Frost, S., Gundle, R., Mclardy-Smith, P. & Murray, D. 2001. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *J Bone Joint Surg Br,* 83**,** 1125-9.

Dawson, J., Fitzpatrick, R., Murray, D. & Carr, A. 1996b. Comparison of measures to assess outcomes in total hip replacement surgery. *Qual Health Care,* 5**,** 81-8.

De Steiger, R. N., Lorimer, M. & Solomon, M. 2015. What is the learning curve for the anterior approach for total hip arthroplasty? *Clin Orthop Relat Res,* 473**,** 3860-6.

Deere, K. C., Whitehouse, M. R., Porter, M., Blom, A. W. & Sayers, A. 2019. Assessing the non-inferiority of prosthesis constructs used in hip replacement using data from the National Joint Registry of England, Wales, Northern Ireland and the Isle of Man: a benchmarking study. *BMJ Open,* 9**,** e026685.

Devlin, N. J., Parkin, D. & Browne, J. 2010. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. *Health Econ,* 19**,** 886-905.

Dripps, R. D., Lamont, A. & Eckenhoff, J. E. 1961. The role of anesthesia in surgical mortality. *Jama,* 178**,** 261-266.

Egger, M., Davey Smith, G., Schneider, M. & Minder, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ,* 315**,** 629-34.

Engelbrecht, E., Siegel, A., Rottger, J. & Buchholz, H. W. 1976. Statistics of total knee replacement: partial and total knee replacement, design St. Georg: a review of a 4-year observation. *Clin Orthop Relat Res***,** 54-64.

Evans, J. T., Blom, A. W., Timperley, A. J., Dieppe, P., Wilson, M. J., Sayers, A. & Whitehouse, M. R. 2020. Factors associated with implant survival following total hip replacement surgery: A

registry study of data from the National Joint Registry of England, Wales, Northern Ireland and the Isle of Man. *PLoS Med,* 17**,** e1003291.

Evans, J. T., Evans, J. P., Walker, R. W., Blom, A. W., Whitehouse, M. R. & Sayers, A. 2019a. How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. *Lancet,* 393**,** 647-654.

Evans, J. T., Walker, R. W., Evans, J. P., Blom, A. W., Sayers, A. & Whitehouse, M. R. 2019b. How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. *Lancet,* 393**,** 655-663.

Faulkner, A., Reidy, M. J., Davies, P. S. E., Wang, W., Buddhadev, N., Johnston, L. & Clift, B. 2018. Balancing training and outcomes in total knee replacement: A ten-year review. *Surgeon,* 16**,** 227-231.

Fawsitt, C. G., Thom, H. H. Z., Hunt, L. P., Nemes, S., Blom, A. W., Welton, N. J., Hollingworth, W., Lopez-Lopez, J. A., Beswick, A. D., Burston, A., Rolfson, O., Garellick, G. & Marques, E. M. R. 2019. Choice of Prosthetic Implant Combinations in Total Hip Replacement: Cost-Effectiveness Analysis Using UK and Swedish Hip Joint Registries Data. *Value Health,* 22**,** 303-312.

Fennema, P. & Lubsen, J. 2010. Survival analysis in total joint replacement: an alternative method of accounting for the presence of competing risk. *J Bone Joint Surg Br,* 92**,** 701-6.

Ferguson, R. J., Palmer, A. J., Taylor, A., Porter, M. L., Malchau, H. & Glyn-Jones, S. 2018. Hip replacement. *Lancet,* 392**,** 1662-1671.

Fitzgerald, J. E., Giddings, C. E., Khera, G. & Marron, C. D. 2012. Improving the future of surgical training and education: consensus recommendations from the Association of Surgeons in Training. *Int J Surg,* 10**,** 389-92.

Fitzpatrick, R., Morris, R., Hajat, S., Reeves, B., Murray, D. W., Hannen, D., Rigge, M., Williams, O. & Gregg, P. 2000. The value of short and simple measures to assess outcomes for patients of total hip replacement surgery. *Qual Health Care,* 9**,** 146-50.

Fowler, J. L., Gie, G. A., Lee, A. J. & Ling, R. S. 1988. Experience with the Exeter total hip replacement since 1970. *Orthop Clin North Am,* 19**,** 477-89.

Fowler, T. J., Aquilina, A. L., Blom, A. W., Sayers, A. & Whitehouse, M. R. 2021. Association between surgeon grade and implant survival following hip and knee replacement: a systematic review and meta-analysis. *BMJ Open,* 11**,** e047882.

Fowler, T. J., Aquilina, A. L., Reed, M. R., Blom, A. W., Sayers, A. & Whitehouse, M. R. 2022. The association between surgeon grade and risk of revision following total hip arthroplasty : an analysis of National Joint Registry data. *Bone Joint J,* 104-B**,** 341-351.

Gandek, B., Ware, J. E., Aaronson, N. K., Apolone, G., Bjorner, J. B., Brazier, J. E., Bullinger, M., Kaasa, S., Leplege, A., Prieto, L. & Sullivan, M. 1998. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol,* 51**,** 1171-8.

Gandhi, S. K., Salmon, J. W., Zhao, S. Z., Lambert, B. L., Gore, P. R. & Conrad, K. 2001. Psychometric evaluation of the 12-item short-form health survey (SF-12) in osteoarthritis and rheumatoid arthritis clinical trials. *Clin Ther,* 23**,** 1080-98.

General Medical Council. 2021a. *GMC: The Specialist Register* [Online]. Available: https://www.gmc-uk.org/registration-and-licensing/the-medical-register/a-guide-to-the-medical-register/specialist-registration [Accessed 06/12/2021 2021].

General Medical Council. 2021b. *GMC: The Trauma and Orthopaedic Surgery Curriculum* [Online]. Available: https://www.gmc-uk.org/-/media/documents/trauma---orthopaedic-surgery-curriculum-aug-2021-approved-oct-20_pdf-84479267.pdf [Accessed 17/04/2022 2022].

Goodfellow, J. & O'connor, J. 1978. The mechanics of the knee and prosthesis design. *J Bone Joint Surg Br,* 60-B**,** 358-69.

Greenaway, D. 2013. *Shape of Training* [Online]. Available: https://www.gmc-uk.org/-/media/documents/shape-of-training-final-report_pdf-53977887.pdf [Accessed 26/01/2022 2022].

Greenwood, M. 1926. The" errors of sampling" of the survivorship tables. *Reports on public health and medical subjects*.

Gruen, T. A., Mcneice, G. M. & Amstutz, H. C. 1979. "Modes of failure" of cemented stem-type femoral components: a radiographic analysis of loosening. *Clin Orthop Relat Res***,** 17-27.

Gunston, F. H. 1971. Polycentric knee arthroplasty. Prosthetic simulation of normal knee movement. *J Bone Joint Surg Br,* 53**,** 272-7.

Hajat, S., Fitzpatrick, R., Morris, R., Reeves, B., Rigge, M., Williams, O., Murray, D. & Gregg, P. 2002. Does waiting for total hip replacement matter? Prospective cohort study. *J Health Serv Res Policy,* 7**,** 19-25.

Halawi, M. J. & Haddad, F. S. 2018. Functional acetabular component alignment: the final frontier. *Bone Joint J,* 100-B**,** 1262-1263.

Harris, K., Lim, C. R., Dawson, J., Fitzpatrick, R., Beard, D. J. & Price, A. J. 2017. The Oxford knee score and its subscales do not exhibit a ceiling or a floor effect in knee arthroplasty patients: an analysis of the National Health Service PROMs data set. *Knee Surg Sports Traumatol Arthrosc,* 25**,** 2736-2742.

Hasegawa, Y. & Amano, T. 2015. Surgical skills training for primary total hip arthroplasty. *Nagoya J Med Sci,* 77**,** 51-7.

Haughom, B. D., Schairer, W. W., Hellman, M. D., Yi, P. H. & Levine, B. R. 2014. Does resident involvement impact post-operative complications following primary total knee arthroplasty? An analysis of 24,529 cases. *J Arthroplasty,* 29**,** 1468-1472 e2.

Health, G. B. D. O., Northern Ireland. Department of Health, S. S., Safety, P., Department, S. S. E. H. & Assembly, W. N. 2003. Modernising medical careers The response of the four UK Health

Ministers to the consultation on Unfinished business: proposals for reform of the senior house officer grade.

Hedlundh, U., Ahnfelt, L., Hybbinette, C. H., Weckstrom, J. & Fredin, H. 1996. Surgical experience related to dislocations after total hip arthroplasty. *J Bone Joint Surg Br,* 78**,** 206-9.

Health Education England. 2021. *HEE: Core Surgical Training Person Specification* [Online]. Available: https://specialtytraining.hee.nhs.uk/portals/1/Content/Person%20Specifications/Core%20Surgical%20Training/CORE%20SURGICAL%20TRAINING%20-%20CT1%202022.pdf [Accessed 04/01/2022 2022].

Hernigou, P., Manicom, O., Flouzat-Lachaniete, C. H., Roussignol, X., Filippini, P. & Poignard, A. 2009. Fifteen year outcome of the ceraver hermes posterior-stabilized total knee arthroplasty: safety of the procedure with experienced and inexperienced surgeons. *Open Orthop J,* 3**,** 36-9.

Higgins, J. P. & Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Stat Med,* 21**,** 1539-58.

Higgins, J. P. T., Green, S. & Cochrane Collaboration. 2008. *Cochrane handbook for systematic reviews of interventions,* Chichester, England ; Hoboken, NJ, Wiley-Blackwell.

Howells, N. R., Gill, H. S., Carr, A. J., Price, A. J. & Rees, J. L. 2008. Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *J Bone Joint Surg Br,* 90**,** 494-9.

Health Research Authority. 2022. *HRA: Guidance for Using Patient Data* [Online]. Available: https://www.hra.nhs.uk/covid-19-research/guidance-using-patient-data/#research [Accessed 11/09/2022 2022].

Hunt, L. P., Blom, A. W., Matharu, G. S., Kunutsor, S. K., Beswick, A. D., Wilkinson, J. M. & Whitehouse, M. R. 2021a. Patients Receiving a Primary Unicompartmental Knee Replacement Have a Higher Risk of Revision but a Lower Risk of Mortality Than Predicted Had They Received a Total Knee Replacement: Data From the National Joint Registry for England, Wales, Northern Ireland, and the Isle of Man. *J Arthroplasty,* 36**,** 471-477 e6.

Hunt, L. P., Matharu, G. S., Blom, A. W., Howard, P. W., Wilkinson, J. M. & Whitehouse, M. R. 2021b. Patellar resurfacing during primary total knee replacement is associated with a lower risk of revision surgery. *Bone Joint J,* 103-B**,** 864-871.

Hunt, L. P., Whitehouse, M. R., Beswick, A., Porter, M. L., Howard, P. & Blom, A. W. 2018. Implications of Introducing New Technology: Comparative Survivorship Modeling of Metal-on-Metal Hip Replacements and Contemporary Alternatives in the National Joint Registry. *J Bone Joint Surg Am,* 100**,** 189-196.

Hunter, D. J. & Bierma-Zeinstra, S. 2019. Osteoarthritis. *Lancet,* 393**,** 1745-1759.

Hunter, S. & Mclaren, P. 1993. Specialist medical training and the Calman report. *BMJ,* 306**,** 1281-2.

Hurreiz, H. 2019. The evolution of surgical training in the UK. *Adv Med Educ Pract,* 10**,** 163-168.

Inglis, T., Dalzell, K., Hooper, G., Rothwell, A. & Frampton, C. 2013. Does orthopedic training compromise the outcome in total hip arthroplasty? *J Surg Educ,* 70**,** 76-80.

ISCP. 2021. *Intercollegiate Surgical Curriculum Programme: Trauma and Orthopaedic Surgery Curriculum* [Online]. Available: https://www.iscp.ac.uk/iscp/curriculum/trauma-orthopaedic-surgery-curriculum/1-introduction/ [Accessed 06/12/2021 2021].

Jain, S., Magra, M., Dube, B., Veysi, V. T., Whitwell, G. S., Aderinto, J. B., Emerton, M. E., Stone, M. H. & Pandit, H. G. 2018. Reverse hybrid total hip arthroplasty. *Bone Joint J,* 100-B**,** 1010-1017.

James Lind Alliance. 2021. *James Lind Alliance Priority Setting Partenership: Hip and Knee Replacement for Osteoarthritis* [Online]. NIHR. Available: https://www.jla.nihr.ac.uk/priority-setting-partnerships/hip-and-knee-replacement-for-osteoarthritis/top-10-priorities/ [Accessed 29/11/2021 2021].

JCST. 2020. *JCST: Terms of Reference* [Online]. Available: https://www.jcst.org/key-documents/ [Accessed 06/12/2021 2021].

Jenkinson, C. & Layte, R. 1997. Development and testing of the UK SF-12 (short form health survey). *J Health Serv Res Policy,* 2**,** 14-8.

Jolbäck, P., Rolfson, O., Cnudde, P., Odin, D., Malchau, H., Lindahl, H. & Mohaddes, M. 2019. High annual surgeon volume reduces the risk of adverse events following primary total hip arthroplasty: a registry-based study of 12,100 cases in Western Sweden. *Acta Orthop,* 90**,** 153-158.

Jolbäck, P., Rolfson, O., Mohaddes, M., Nemes, S., Karrholm, J., Garellick, G. & Lindahl, H. 2018. Does surgeon experience affect patient-reported outcomes 1 year after primary total hip arthroplasty? *Acta Orthop,* 89**,** 265-271.

Kandala, N. B., Connock, M., Pulikottil-Jacob, R., Sutcliffe, P., Crowther, M. J., Grove, A., Mistry, H. & Clarke, A. 2015. Setting benchmark revision rates for total hip replacement: analysis of registry evidence. *BMJ,* 350**,** h756.

Kaplan, E. L. & Meier, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association,* 53**,** 457-481.

Kazarian, G. S., Lawrie, C. M., Barrack, T. N., Donaldson, M. J., Miller, G. M., Haddad, F. S. & Barrack, R. L. 2019. The Impact of Surgeon Volume and Training Status on Implant Alignment in Total Knee Arthroplasty. *J Bone Joint Surg Am,* 101**,** 1713-1723.

Kelly, A. & Canter, R. 2007. A new curriculum for surgical training within the United Kingdom: context and model. *J Surg Educ,* 64**,** 10-9.

Khakha, R. S., Chowdhry, M., Sivaprakasam, M., Kheiran, A. & Chauhan, S. K. 2015. Radiological and Functional Outcomes in Computer Assisted Total Knee Arthroplasty Between Consultants and Trainees - A Prospective Randomized Controlled Trial. *J Arthroplasty,* 30**,** 1344-7.

Kim, S. C., Lim, Y. W., Kwon, S. Y., Jo, W. L., Ju, S. H., Park, C. J., Lee, C. W. & Kim, Y. S. 2017. Level of surgical experience is associated with change in hip center of rotation following cementless total hip arthroplasty: A radiographic assessment. *PLoS One,* 12**,** e0178300.

Konan, S. & Haddad, F. S. 2013. Joint Registries: a Ptolemaic model of data interpretation? *Bone Joint J,* 95-B**,** 1585-6.

Kunutsor, S. K., Barrett, M. C., Beswick, A. D., Judge, A., Blom, A. W., Wylde, V. & Whitehouse, M. R. 2019. Risk factors for dislocation after primary total hip replacement: a systematic review and meta-analysis of 125 studies involving approximately five million hip replacements. *The Lancet Rheumatology,* 1**,** e111-e121.

Kurtz, S., Ong, K., Lau, E., Mowat, F. & Halpern, M. 2007. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. *J Bone Joint Surg Am,* 89**,** 780-5.

Kurtz, S. M., Ong, K. L., Lau, E. & Bozic, K. J. 2014. Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021. *J Bone Joint Surg Am,* 96**,** 624-30.

Lacny, S., Wilson, T., Clement, F., Roberts, D. J., Faris, P. D., Ghali, W. A. & Marshall, D. A. 2015. Kaplan-Meier Survival Analysis Overestimates the Risk of Revision Arthroplasty: A Meta-analysis. *Clin Orthop Relat Res,* 473**,** 3431-42.

Lambert, P. C., Dickman, P. W., Nelson, C. P. & Royston, P. 2010. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med,* 29**,** 885-95.

Lambert, P. C. & Royston, P. 2009a. Further development of flexible parametric models for survival analysis. *Stata Journal,* 9**,** 265-290.

Lambert, P. C. & Royston, P. 2009b. Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal,* 9**,** 265-290.

Lampropoulou-Adamidou, K., Karachalios, T. S. & Hartofilakidis, G. 2018. Overestimation of the risk of revision with Kaplan-Meier presenting the long-term outcome of total hip replacement in older patients. *Hip Int,* 28**,** 246-253.

Learmonth, I. D., Young, C. & Rorabeck, C. 2007. The operation of the century: total hip replacement. *Lancet,* 370**,** 1508-19.

Lewinnek, G. E., Lewis, J. L., Tarr, R., Compere, C. L. & Zimmerman, J. R. 1978. Dislocations after total hip-replacement arthroplasties. *J Bone Joint Surg Am,* 60**,** 217-20.

Liddle, A. D., Judge, A., Pandit, H. & Murray, D. W. 2014. Adverse outcomes after total and unicompartmental knee replacement in 101,330 matched patients: a study of data from the National Joint Registry for England and Wales. *Lancet,* 384**,** 1437-45.

Liddle, A. D., Pandit, H., Judge, A. & Murray, D. W. 2015. Patient-reported outcomes after total and unicompartmental knee arthroplasty: a study of 14,076 matched patients from the National Joint Registry for England and Wales. *Bone Joint J,* 97-B**,** 793-801.

Liddle, A. D., Pandit, H., Judge, A. & Murray, D. W. 2016. Effect of Surgical Caseload on Revision Rate Following Total and Unicompartmental Knee Replacement. *J Bone Joint Surg Am,* 98**,** 1-8.

Lim, C. R., Harris, K., Dawson, J., Beard, D. J., Fitzpatrick, R. & Price, A. J. 2015. Floor and ceiling effects in the OHS: an analysis of the NHS PROMs data set. *BMJ Open,* 5**,** e007765.

Little, R. J. A. & Rubin, D. B. 1987. *Statistical analysis with missing data*, Wiley.

Logishetty, K., Gofton, W. T., Rudran, B., Beaule, P. E. & Cobb, J. P. 2020. Fully Immersive Virtual Reality for Total Hip Arthroplasty: Objective Measurement of Skills and Transfer of Visuospatial Performance After a Competency-Based Simulation Curriculum. *J Bone Joint Surg Am,* 102**,** e27.

Logishetty, K., Rudran, B. & Cobb, J. P. 2019. Virtual reality training improves trainee performance in total hip arthroplasty: a randomized controlled trial. *Bone Joint J,* 101-B**,** 1585-1592.

Lopez-Lopez, J. A., Humphriss, R. L., Beswick, A. D., Thom, H. H. Z., Hunt, L. P., Burston, A., Fawsitt, C. G., Hollingworth, W., Higgins, J. P. T., Welton, N. J., Blom, A. W. & Marques, E. M. R. 2017. Choice of implant combinations in total hip replacement: systematic review and network meta-analysis. *BMJ,* 359**,** j4651.

Lourie, J. 1998. Delegation of orthopaedic workload. *Ann R Coll Surg Engl,* 80**,** 260-4.

Lourie, J. 1999. At least 35% more orthopaedic surgeons are needed to ensure consultant based service. *BMJ,* 319**,** 256.

Macdonald, D. R. W., Dougall, T. W., Mitchell, M. & Farrow, L. 2020. Can Total Hip Arthroplasty for Hip Fracture Be Safely Performed by Trainees? A Retrospective Cohort Study. *J Arthroplasty,* 35**,** 1303-1306.

Macintosh, D. L. & Hunter, G. A. 1972. The use of the hemiarthroplasty prosthesis for advanced osteoarthritis and rheumatoid arthritis of the knee. *J Bone Joint Surg Br,* 54**,** 244-55.

Madanipour, S., Singh, P., Karia, M., Bhamra, J. S. & Abdul-Jabar, H. B. 2021. Trainee performed total knee arthroplasty is safe and effective: A systematic review and meta-analysis comparing outcomes between trainees and consultants. *Knee,* 30**,** 291-304.

Mahaluxmivala, J., Bankes, M. J., Nicolai, P., Aldam, C. H. & Allen, P. W. 2001. The effect of surgeon experience on component positioning in 673 Press Fit Condylar posterior cruciate-sacrificing total knee arthroplasties. *J Arthroplasty,* 16**,** 635-40.

Mahase, E. 2020. Covid-19: UK records first death, as world's cases exceed 100 000. *BMJ,* 368**,** m943.

Mantel, N. & Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst,* 22**,** 719-48.

Marmor, L. 1979. Marmor modular knee in unicompartmental disease. Minimum four-year follow-up. *J Bone Joint Surg Am,* 61**,** 347-53.

Marston, R. A., Cobb, A. G. & Bentley, G. 1996. Stanmore compared with Charnley total hip replacement. A prospective study of 413 arthroplasties. *J Bone Joint Surg Br,* 78**,** 178-84.

Matharu, G., Culliford, D., Blom, A. & Judge, A. 2021. Projections for primary hip and knee replacement surgery up to the year 2060: an analysis based on data from The National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *The Annals of The Royal College of Surgeons of England,* 0**,** bull.

Mccaskie, A. W., Brown, A. R., Thompson, J. R. & Gregg, P. J. 1996. Radiological evaluation of the interfaces after cemented total hip replacement. Interobserver and intraobserver agreement. *J Bone Joint Surg Br,* 78**,** 191-4.

Mcculloch, R. A., Howgate, D., Gibbs, V. N., Palmer, A., Taylor, A. H. & Kendrick, B. 2021. Assessing the performance and learning curve of orthopaedic surgical trainees in primary total hip arthroplasty. *Ann R Coll Surg Engl,* 103**,** 514-519.

Mckeever, D. C. 2005. The classic: Tibial plateau prosthesis.1960. *Clin Orthop Relat Res,* 440**,** 4-8; discussion 3.

Mohammad, H. R., Bullock, G. S., Kennedy, J. A., Mellon, S. J., Murray, D. & Judge, A. 2021. Cementless unicompartmental knee replacement achieves better ten-year clinical outcomes than cemented: a systematic review. *Knee Surg Sports Traumatol Arthrosc,* 29**,** 3229-3245.

Mohammad, H. R., Matharu, G. S., Judge, A. & Murray, D. W. 2020. The Effect of Surgeon Caseload on the Relative Revision Rate of Cemented and Cementless Unicompartmental Knee Replacements: An Analysis from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. *J Bone Joint Surg Am,* 102**,** 644-653.

Moran, M., Yap, S. L., Walmsley, P. & Brenkel, I. J. 2004. Clinical and radiologic outcome of total hip arthroplasty performed by trainee compared with consultant orthopedic surgeons. *J Arthroplasty,* 19**,** 853-7.

Muirhead-Allwood, S. K. 1998. Lessons of a hip failure. *BMJ,* 316**,** 644.

Muller, D. A., Zingg, P. O. & Dora, C. 2014. Anterior minimally invasive approach for total hip replacement: five-year survivorship and learning curve. *Hip Int,* 24**,** 277-83.

Munro, C., Burke, J., Allum, W. & Mortensen, N. 2021. Covid-19 leaves surgical training in crisis. *BMJ,* 372**,** n659.

Murray, D. W., Carr, A. J. & Bulstrode, C. 1993. Survival analysis of joint replacements. *J Bone Joint Surg Br,* 75**,** 697-704.

Murray, D. W., Fitzpatrick, R., Rogers, K., Pandit, H., Beard, D. J., Carr, A. J. & Dawson, J. 2007. The use of the Oxford hip and knee scores. *J Bone Joint Surg Br,* 89**,** 1010-4.

Murray, D. W. & Frost, S. J. 1998. Pain in the assessment of total knee replacement. *J Bone Joint Surg Br,* 80**,** 426-31.

Namba, R. S., Inacio, M. C. & Paxton, E. W. 2013. Risk factors associated with deep surgical site infections after primary total knee arthroplasty: an analysis of 56,216 knees. *J Bone Joint Surg Am,* 95**,** 775-82.

National Institue of Health and Care Excellence. 2021. *NICE: Evidence review for patella resurfacing. [NG157]* [Online]. Available: https://www.nice.org.uk/guidance/ng157/evidence/l-patella-resurfacing-pdf-315756469335 [Accessed 25/11/2021 2021].

National Institute of Health and Care Excellence. 2021. *NICE: Total hip replacement and resurfacing arthroplasty for end-stage arthritis of the hip . Technology appraisal guidance [TA304]* [Online]. National Institute of Health and Care Excellence. Available: https://www.nice.org.uk/guidance/ta304 [Accessed 25/11/2021 2021].

Nilsdotter, A. K., Lohmander, L. S., Klassbo, M. & Roos, E. M. 2003. Hip disability and osteoarthritis outcome score (HOOS)--validity and responsiveness in total hip replacement. *BMC Musculoskelet Disord,* 4**,** 10.

The National Joint Registry. 2020. *NJR: 17th Annual Report* [Online]. Available: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2017th%20Annual%20Report%202020.pdf [Accessed 26/01/2022 2022].

The National Joint Registry. 2021a. *NJR: 18th Annual Report* [Online]. Available: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2018th%20Annual%20Report%202021.pdf [Accessed 26/01/2022 2022].

The National Joint Registry. 2022a. *NJR: 19th Annual Report* [Online]. Available: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2019th%20Annual%20Report%202022.pdf [Accessed 25/09/2022 2022].

The National Joint Registry. 2022b. *NJR: Collecting Patient Consent* [Online]. Available: https://www.njrcentre.org.uk/healthcare-providers/collecting-patient-consent/ [Accessed 06/10/2022 2022].

The National Joint Registry. 2021b. *NJR: Executive Summary* [Online]. Available: https://reports.njrcentre.org.uk/headline-summaries [Accessed 25/11/2021 2021].

The National Joint Registry. 2021c. *NJR: Patient Consent* [Online]. Available: https://www.njrcentre.org.uk/njrcentre/Healthcare-providers/Collecting-data/Patient-consent [Accessed 25/10/2021 2021].

The National Joint Registry. 2018. *NJR: Patient Reported Outcome Measures* [Online]. Available: https://www.njrcentre.org.uk/njrcentre/Research/NJR-PROMs [Accessed 03/01/2022 2022].

Noble, S., Noble, M. & Mclennan, D. 2019. *English Indices of Deprivation 2019 : Research report*.

Nzeako, O. & Back, D. 2016. Learning Curves in Arthroplasty in Orthopedic Trainees. *J Surg Educ,* 73**,** 689-93.

The New Zealand Joint Registry. 2020. *NZJR: Twenty-one Year Report* [Online]. Available: https://www.nzoa.org.nz/sites/default/files/DH8426_NZJR_2020_Report_v5_30Sep.pdf [Accessed 29/11/2021 2021].

Orthopaedic Data Evaluation Panel. 2021a. *ODEP: Criteria for Hips* [Online]. Available: https://www.odep.org.uk/Methodology/ForHips.aspx [Accessed 25/11/2021 2021].

Orthopaedic Data Evaluation Panel. 2021b. *ODEP: Criteria for Total Knee Replacements (TKRs)* [Online]. Available: https://www.odep.org.uk/Methodology/ForKnees/TotalKneeReplacements.aspx [Accessed 26/01/2022 2022].

OECD. 2021. *OECD: Health at a Glance 2021* [Online]. Available: https://www.oecd-ilibrary.org/content/publication/ae3016b9-en [Accessed 26/01/2022 2022].

Ostendorf, M., Van Stel, H. F., Buskens, E., Schrijvers, A. J., Marting, L. N., Verbout, A. J. & Dhert, W. J. 2004. Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. *J Bone Joint Surg Br,* 86**,** 801-8.

Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. 2016. Rayyan-a web and mobile app for systematic reviews. *Syst Rev,* 5**,** 210.

Pabinger, C., Lothaller, H. & Geissler, A. 2015. Utilization rates of knee-arthroplasty in OECD countries. *Osteoarthritis Cartilage,* 23**,** 1664-73.

Palan, J., Gulati, A., Andrew, J. G., Murray, D. W., Beard, D. J. & Group, E. S. 2009. The trainer, the trainee and the surgeons' assistant: clinical outcomes following total hip replacement. *J Bone Joint Surg Br,* 91**,** 928-34.

Patel, A., Pavlou, G., Mujica-Mota, R. E. & Toms, A. D. 2015. The epidemiology of revision total knee and hip arthroplasty in England and Wales: a comparative analysis with projections for the United States. A study using the National Joint Registry dataset. *Bone Joint J,* 97-B**,** 1076-81.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., Mcpherson, K., Peto, J. & Smith, P. G. 1977. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer,* 35**,** 1-39.

Polce, E. M., Kunze, K. N., Williams, B. T., Krivicich, L. M., Maheshwer, B., Beletsky, A., Cohn, M. R., Kogan, M. & Chahla, J. 2020. Efficacy and Validity of Orthopaedic Simulators in Surgical Training: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *J Am Acad Orthop Surg,* 28**,** 1027-1040.

Price, A. J., Alvand, A., Troelsen, A., Katz, J. N., Hooper, G., Gray, A., Carr, A. & Beard, D. 2018. Knee replacement. *Lancet,* 392**,** 1672-1682.

Purcell Jackson, G. & Tarpley, J. L. 2009. How long does it take to train a surgeon? *BMJ,* 339**,** b4260.

Ranstam, J. & Robertsson, O. 2017. The Cox model is better than the Fine and Gray model when estimating relative revision risks from arthroplasty register data. *Acta Orthop,* 88**,** 578-580.

Ravi, B., Jenkinson, R., Austin, P. C., Croxford, R., Wasserstein, D., Escott, B., Paterson, J. M., Kreder, H. & Hawker, G. A. 2014. Relation between surgeon volume and risk of complications after total hip arthroplasty: propensity score matched cohort study. *BMJ,* 348**,** g3284.

The Royal College of Surgeons of England. 2001. *RCS: An Investigation of the Performance of the 3M™ Capital™ Hip System* [Online]. Available: file:///C:/Users/xi18408/Downloads/An%20Investigation%20of%20the%20Performance%20of%20the%203M%20Capital%20Hip%20System%20Report%202001.pdf [Accessed 26/01/2022 2022].

The Royal College of Surgeons of England. 2021. *RCS: Every operation should include a surgical trainee* [Online]. Available: https://www.rcseng.ac.uk/news-and-events/media-centre/press-releases/every-operation-should-include-a-surgical-trainee/ [Accessed 26/01/2022 2022].

The Royal College of Surgeons of England. 2015. *RCS: Improving Surgical Training Report* [Online]. Available: https://www.rcseng.ac.uk/-/media/files/rcs/careers-in-surgery/improving_surgical_training_text.pdf [Accessed 03/12/2021 2021].

The Royal College of Surgeons of England. 2022. *RCS: Recovery of surgical services during and after COVID-19* [Online]. Available: https://www.rcseng.ac.uk/coronavirus/recovery-of-surgical-services/#s7 [Accessed 14/01/2022 2022].

Reidy, M. J., Faulkner, A., Shitole, B. & Clift, B. 2016. Do trainee surgeons have an adverse effect on the outcome after total hip arthroplasty?: a ten-year review. *Bone Joint J,* 98-B**,** 301-6.

Riordan, P., Bickler, G. & Lyons, C. 1998. Lessons of a hip failure. Registers of joint replacement operations should be set up. *BMJ,* 316**,** 1985.

Roos, E. M. & Toksvig-Larsen, S. 2003. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. *Health Qual Life Outcomes,* 1**,** 17.

Royston, P. & Lambert, P. C. 2011. *Flexible parametric survival analysis using stata : beyond the Cox model,* College Station, Tex., Stata.

Sayers, A., Crowther, M. J., Judge, A., Whitehouse, M. R. & Blom, A. W. 2017. Determining the sample size required to establish whether a medical device is non-inferior to an external benchmark. *BMJ Open,* 7**,** e015397.

Sayers, A., Evans, J. T., Whitehouse, M. R. & Blom, A. W. 2018. Are competing risks models appropriate to describe implant failure? *Acta Orthop,* 89**,** 256-258.

Sayers, A., Steele, F., Whitehouse, M. R., Price, A., Ben-Shlomo, Y. & Blom, A. W. 2020a. Association between surgical volume and failure of primary total hip replacement in England and Wales: findings from a prospective national joint replacement register. *BMJ Open,* 10**,** e033045.

Sayers, A., Whitehouse, M. R., Judge, A., Macgregor, A. J., Blom, A. W. & Ben-Shlomo, Y. 2020b. Analysis of change in patient-reported outcome measures with floor and ceiling effects using the multilevel Tobit model: a simulation study and an example from a National Joint Register using body mass index and the Oxford Hip Score. *BMJ Open,* 10**,** e033646.

Schoenfeld, A. J., Serrano, J. A., Waterman, B. R., Bader, J. O. & Belmont, P. J., Jr. 2013. The impact of resident involvement on post-operative morbidity and mortality following orthopaedic procedures: a study of 43,343 cases. *Arch Orthop Trauma Surg,* 133**,** 1483-91.

Scott, R. D. 2005. Lateral unicompartmental replacement: a road less traveled. *Orthopedics,* 28**,** 983-4.

Sedgwick, P. & Joekes, K. 2015. Interpreting hazard ratios. *BMJ,* 351**,** h4631.

Sher, J. L., Reed, M. R., Calvert, P., Wallace, W. A. & Lamb, A. 2005. Influencing the national training agenda. The UK & Ireland orthopaedic eLogbook. *J Bone Joint Surg Br,* 87**,** 1182-6.

Singh, P., Madanipour, S., Fontalis, A., Bhamra, J. S. & Abdul-Jabar, H. B. 2019. A systematic review and meta-analysis of trainee-versus consultant surgeon-performed elective total hip arthroplasty. *Efort Open Reviews,* 4**,** 44-55.

Smith-Petersen, M. N. 2006. Evolution of mould arthroplasty of the hip joint. 1948. *Clin Orthop Relat Res,* 453**,** 17-21.

Smith, A. J., Dieppe, P., Porter, M., Blom, A. W., National Joint Registry Of, E. & Wales 2012a. Risk of cancer in first seven years after metal-on-metal hip replacement compared with other bearings and general population: linkage study between the National Joint Registry of England and Wales and hospital episode statistics. *BMJ,* 344**,** e2383.

Smith, A. J., Dieppe, P., Vernon, K., Porter, M., Blom, A. W., National Joint Registry Of, E. & Wales 2012b. Failure rates of stemmed metal-on-metal hip replacements: analysis of data from the National Joint Registry of England and Wales. *Lancet,* 379**,** 1199-204.

Smith, J. O., Frampton, C. M. A., Hooper, G. J. & Young, S. W. 2018. The Impact of Patient and Surgical Factors on the Rate of Postoperative Infection After Total Hip Arthroplasty-A New Zealand Joint Registry Study. *J Arthroplasty,* 33**,** 1884-1890.

Smith, T. O., Williams, T. H., Samuel, A., Ogonda, L. & Wimhurst, J. A. 2011. Reliability of the radiological assessments of radiolucency and loosening in total hip arthroplasty using PACS. *Hip Int,* 21**,** 577-82.

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hrobjartsson, A., Kirkham, J., Juni, P., Loke, Y. K., Pigott, T. D., Ramsay, C. R., Regidor, D., Rothstein, H. R., Sandhu, L., Santaguida, P. L., Schunemann, H. J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J. C., Waddington, H., Waters, E., Wells, G. A., Whiting, P. F. & Higgins, J. P. 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ,* 355**,** i4919.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D. & Higgins, J. P. 2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ,* 343**,** d4002.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. & Carpenter, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ,* 338**,** b2393.

Storey, R., Frampton, C., Kieser, D., Ailabouni, R. & Hooper, G. 2018. Does Orthopaedic Training Compromise the Outcome in Knee Joint Arthroplasty? *J Surg Educ,* 75**,** 1292-1298.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A. & Thacker, S. B. 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA,* 283**,** 2008-12.

Tahir, M., Rahman, U. & Gulati, A. 2021. An international comparison of competency-based orthopaedic curricula and minimum operative experience - Review article. *Int J Surg,* 94**,** 106125.

Theelen, L., Bischoff, C., Grimm, B. & Heyligers, I. C. 2018. Current practice of orthopaedic surgical skills training raises performance of supervised residents in total knee arthroplasty to levels equal to those of orthopaedic surgeons. *Perspect Med Educ,* 7**,** 126-132.

Thygesen, L. C. & Ersboll, A. K. 2014. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur J Epidemiol,* 29**,** 551-8.

Ware, J., Jr., Kosinski, M. & Keller, S. D. 1996. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med Care,* 34**,** 220-33.

Ware, J. E., Kosinski, M. & Keller, S. D. 1994. *SF-36 physical and mental health summary scales : a user's manual,* Boston, Mass., Health Institute, New England Medical Center.

Ware, J. E., Kosinski, M. & Keller, S. D. 1995. *SF-12 : how to score the SF-12 physical and mental health summary scales,* Boston, Mass., Health Institute, New England Medical Center.

Ware, J. E., Kosinski, M. & Keller, S. D. 1998. *SF-12 : how to score the SF-12 physical and mental health summary scales,* Lincoln, R.I., QualityMetric Inc.

Weber, M., Benditz, A., Woerner, M., Weber, D., Grifka, J. & Renkawitz, T. 2017. Trainee Surgeons Affect Operative Time but not Outcome in Minimally Invasive Total Hip Arthroplasty. *Sci Rep,* 7**,** 6152.

White, I. R. & Carlin, J. B. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med,* 29**,** 2920-31.

Wiles, P. 1958. The surgery of the osteoarthritic hip. *Br J Surg,* 45**,** 488-97.

Wilson, H. A., Middleton, R., Abram, S. G. F., Smith, S., Alvand, A., Jackson, W. F., Bottomley, N., Hopewell, S. & Price, A. J. 2019. Patient relevant outcomes of unicompartmental versus total knee replacement: systematic review and meta-analysis. *BMJ,* 364**,** l352.

Wilson, J. A. 1997. Unsupervised surgical training: questionnaire study. *BMJ,* 314**,** 1803-4.

Wylde, V. & Blom, A. W. 2009. Assessment of outcomes after hip arthroplasty. *Hip Int,* 19**,** 1-7.

Wylde, V. & Blom, A. W. 2011. The failure of survivorship. *J Bone Joint Surg Br,* 93**,** 569-70.

Wylde, V., Hewlett, S., Learmonth, I. D. & Dieppe, P. 2011. Persistent pain after joint replacement: prevalence, sensory qualities, and postoperative determinants. *Pain,* 152**,** 566-572.

List of abbreviations

AIC	Akaike Information Criteria
ARCP	Annual Review of Competence Progression
ASA	American Society of Anaesthesiologists
AVN	Avascular necrosis
BIC	Bayes Information Criteria
BMI	Body Mass Index
CBD	Case-Based Discussion
CCT	Certificate of Completion of Training
CDC	Centers for Disease Control and Prevention
CEX	Clinical Evaluation Exercise
CI	Confidence Interval
CiP	Capabilities in Practice
CR	Competing Risk
CRPD	Clinical Research Practice Datalink
CT1-CT2	Core Surgical Training Year 1 - Core Surgical Training Year 2
EQ-5D	EuroQol-Five-Dimension
EWTD	European Working Time Directive

FPM	Flexible Parametric Survival Modelling
FRCS	Fellow of the Royal College of Surgeons
F1-F2	Foundation Year 1 – Foundation Year 2
GDP	Gross Domestic Product
GPC	Generic Professional Capabilities
GMC	General Medical Council
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HHS	Harris Hip Score
HOOS	Hip Disability and Osteoarthritis Outcome Score
HR	Hazard Ratio
HRQOL	Health-Related Quality of Life
IMD	Index of Multiple Deprivation
ISAR	International Society of Arthroplasty Registries
ISCP	Intercollegiate Surgical Curriculum Programme
IQR	Interquartile Range
JCST	Joint Committee on Surgical Training
KM	Kaplan-Meier
KOOS	Knee Injury and Osteoarthritis Outcome Score
KSS	Knee Society Score

MCID	Minimal Clinically Important Difference
MDS	Minimum Data Set Form (NJR)
MeSH	Medical Subject Headings
MMC	Modernising Medical Careers
MOOSE	Meta-analyses Of Observational Studies in Epidemiology
MRCS	Member of the Royal College of Surgeons
NHS	National Health Service
NICE	National Institute of Health and Care Excellence
NJR	National Joint Registry
NTN	National Training Number
NZJR	New Zealand Joint Registry
OA	Osteoarthritis
ODEP	Orthopaedic Data Evaluation Panel
OECD	Organisation for Economic Co-operation and Development
OHS	Oxford Hip Score
OKS	Oxford Knee Score
OR	Odds Ratio
PBA	Procedure-Based Assessment
PH	Proportional Hazards

PMMA	Polymethylmethacrylate
PROMs	Patient-Reported Outcome Measures
RCT	Randomised Controlled Trial
ROBINS-I	Risk of Bias in Non-Randomised Studies of Interventions
RR	Relative Risk
SAC	Specialty Advisory Committee
SAS	Specialty Doctor/Associate Specialist
SD	Standard Deviation
SF-12	Short Form-12 Health Survey
SHAR	Swedish Hip Arthroplasty Register
ST3-ST8	Specialty Training Year 3 - Specialty Training Year 8
TD	Time Dependent Effect
THR	Total Hip Replacement
TKR	Total Knee Replacement
T&O	Trauma and Orthopaedics
UK	United Kingdom
UKR	Unicompartmental Knee Replacement
USA	United States of America
VTE	Venous Thromboembolism

WBA	Workplace-Based Assessment
WOMAC	Western Ontario and McMaster University Osteoarthritis Index

Contributors

AA	Alexander Aquilina[1]
AB	Ashley Blom[1,2]
AS	Adrian Sayers[1]
JB	James Berstock[3]
MR	Mike Reed[4]
MW	Michael Whitehouse[1,2]
TF	Timothy Fowler[1]

Affiliations:

1) Musculoskeletal Research Unit, Translational Health Sciences, Bristol Medical School, Southmead Hospital, Bristol, UK
2) National Institute for Health Research Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust, University of Bristol, Bristol, UK
3) Royal United Hospital, Combe Park, Bath, UK
4) Northumbria Healthcare NHS Foundation Trust, Department of Trauma and Orthopaedics, Wansbeck General Hospital, Northumbria, UK

Appendices

Appendix 1 - GPCs and CiP defined by the UK T&O Curriculum.

The UK T&O curriculum states the following: "Simply put, the Capabilities in Practice (CiPs) and Generic Professional Capabilities (GPCs) are the constituent parts of the role of a consultant T&O surgeon. Each part is as important as the next and doctors are required to be capable in all parts of the role in order to be able to practice independently. In order to complete training and to be recommended to the GMC for certification and entry on to the specialist register, the doctor must demonstrate that they are capable of unsupervised practice in all the CiPs and GPCs (ISCP, 2021)."

Assessment domain	**Description**
Generic Professional Capabilities (GPCs)	Professional values and behaviours
	Professional skills (practical/communication/dealing with complexity/clinical)
	Professional knowledge
	Capabilities in health promotion and illness prevention
	Capabilities in leadership and team working
	Capabilities in patient safety and quality improvement
	Capabilities in safeguarding vulnerable groups
	Capabilities in education and training
	Capabilities in research and scholarship
Capabilities in Practice	Manages an out-patient clinic
	Manages the unselected emergency take
	Manages ward rounds and the on-going care of in-patients
	Manages an operating list
	Manages multi-disciplinary working
Adapted from: trauma-orthopaedic-surgery-curriculum-aug-2021-approved-oct-20.pdf (iscp.ac.uk)	

Appendix 2 - Example of a Procedure Based Assessment (PBA) for primary THR.

Assessment domain	Global and task-specific items
Preoperative planning	Demonstrates recognition of anatomical and pathological abnormalities (and relevant co-morbidities) and selects appropriate operative strategies/techniques to deal with these e.g. nutritional status
	Demonstrates ability to make reasoned choice of appropriate equipment, materials or devices (if any) taking into account appropriate investigations e.g. x-rays or devices
	Checks materials, equipment and device requirements with operating room staff
	Ensures the operation site is marked where applicable
	Checks patient records, personally reviews investigations
Preoperative preparation	Checks in theatre that consent has been obtained
	Gives effective briefing to theatre team
	Ensures proper and safe positioning of the patient on the operating table
	Demonstrates careful skin preparation
	Demonstrates careful draping of the patient's operative field
	Ensures general equipment and materials are deployed safely (e.g. catheter, diathermy)
	Ensures appropriate drugs administered
	Arranges for and deploys supporting specialist equipment (e.g. image intensifiers) effectively
Exposure and Closure	Demonstrates knowledge of optimum skin incision / portal / access
	Achieves an adequate exposure through purposeful dissection in correct tissue planes and identifies all structures correctly
	Completes a sound wound repair where appropriate
	Protects the wound with dressings, splints and drains where appropriate
Intra-operative technique	Follows an agreed, logical sequence or protocol for the procedure
	Consistently handles tissue well with minimal damage
	Controls bleeding promptly by an appropriate method
	Demonstrates a sound technique of knots and sutures/staples
	Uses instruments appropriately and safely
	Proceeds at appropriate pace with economy of movement
	Anticipates and responds appropriately to variation e.g. anatomy
	Deals calmly and effectively with untoward events/complications
	Uses assistant(s) to the best advantage at all times
	Communicates clearly and consistently with the scrub team
	Dislocates hip safely
	Cuts femoral neck appropriately to match design of implant
	Demonstrates familiarity and understanding of acetabular preparation including osteophyte trimming medially and at rim
	Broaches the femur properly and prepares the bony surface
	Uses trials and checks component orientation properly
	Fix acetabular component appropriately
	Implants femoral component appropriately
	Performs final reduction and checks for stability
Postoperative management	Ensures the patient is transferred safely from the operating table to bed
	Constructs a clear operation note
	Records clear and appropriate post operative instructions
	Deals with specimens. Labels and orientates specimens appropriately
Adapted from: https://www.iscp.ac.uk/forms/pba.aspx	

PBA rating:

Task-specific rating: N/A; development required; satisfactory

Overall rating:

Level 1a: able to assist with guidance

Level 1b: able to assist without guidance

Level 2a: guidance required for most/all of procedure

Level 2b: guidance required for key steps only

Level 3a: procedure performed with minimal guidance

Level 3b: procedure performed competently without guidance but lacked fluidity

Level 4a: procedure performed fluently without guidance

Level 4b: as 4a, but able to anticipate/avoid/deal with complications.

Appendix 3 - Index procedure indicative numbers on the UK T&O curriculum.

Procedure		**Indicative number**
Elective	Major joint replacement (hip, knee, shoulder, ankle)	80
	Osteotomy	20
	Nerve decompression	20
	Arthroscopy	50
Trauma	Compression hip screw for neck of femur fracture	40
	Hemiarthroplasty for neck of femur fracture	40
	Application of limb external fixator	5
	Tendon repair for trauma	10
	Intramedullary nailing for fracture fixation	30
	Plate fixation of fracture or arthrodesis	40
	Tension band wire fixation for fracture or arthrodesis	5
	K-wire fixation for fracture or arthrodesis	20
	Fixation of paediatric supracondylar fracture	5
Adapted from: trauma-orthopaedic-surgery-curriculum-aug-2021-approved-oct-20.pdf (iscp.ac.uk)		

Supervision Code Help Guide

The following supervision definitions have been approved by JCST for usage in recording surgical training experience in the UK & Ireland. The respective contribution of the trainee and trainer should be considered at the end of the procedure and the recorded supervision code should reflect this discussion. Logbook record should be validated by the trainer.

Assisting (A):

- The trainer completes the procedure from start to finish.
- The trainee performs the approach and closure of the wound.
- The trainer performs the key components of the procedure.

Supervised - trainer scrubbed (S-TS):

- The trainee performs key components of the procedure (as defined in the relevant PBA) with the trainer scrubbed.
- S-TS is equivalent to the trainee performing the operation while the trainer is scrubbed.

To further clarify common issues which arise:

a) The trainer will be able to offer advice and carry out limited parts of the procedure (e.g. demonstrate a dissection or suturing technique) before allowing the trainee to continue.
b) A junior trainee may carry out some parts of the operation which the senior trainee is clearly competent to perform (e.g. opening or closing the incision).
c) Some operations are commonly divided into different components (e.g. Whipple's, femoro-distal bypass) and are carried out by two trained surgeons. If the trainee takes the place of one of those surgeons and performs their component(s) as described above, then it is appropriate to record the case as supervised - trainer scrubbed even if all components were not carried out by the trainee.

Supervised - trainer unscrubbed (S-TU):

- The trainee completes the procedure from start to finish.
- The trainer is unscrubbed and is:
    - in the operating theatre throughout.
    - in the operating theatre suite and regularly enters the operating theatre during the procedure (70% of the duration of the procedure).

Performed (P):

- The trainee completes the procedure from start to finish.
- The trainer is present for <70% of the duration of the procedure.
- The trainer is not in the operating theatre and is:
    - scrubbed in the adjacent operating theatre.
    - not in the operating suite but is in the hospital.

Training more junior trainee (T):

- A non-consultant grade surgeon training a junior trainee.

Observed (O):

- Procedure observed by an unscrubbed trainee.

Reproduced from: https://client.elogbook.org/docs/eLogbook_Supervision_codes.pdf

Appendix 5 - Meta-analyses of Observational Studies in Epidemiology (MOOSE) Checklist

Item No	Recommendation	Reported on Page No
Reporting of background should include		
1	Problem definition	35
2	Hypothesis statement	36
3	Description of study outcome(s)	35-37
4	Type of exposure or intervention used	37
5	Type of study designs used	36
6	Study population	36
Reporting of search strategy should include		
7	Qualifications of searchers (e.g., librarians and investigators)	36
8	Search strategy, including time period included in the synthesis and key words	36, Appendix 6
9	Effort to include all available studies, including contact with authors	38
10	Databases and registries searched	36
11	Search software used, name and version, including special features used (e.g., explosion)	36-37
12	Use of hand searching (e.g., reference lists of obtained articles)	36
13	List of citations located and those excluded, including justification	Appendix 8
14	Method of addressing articles published in languages other than English	36
15	Method of handling abstracts and unpublished studies	36-37
16	Description of any contact with authors	38, 41, 43
Reporting of methods should include		
17	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	36-37
18	Rationale for the selection and coding of data (e.g., sound clinical principles or convenience)	36-37
19	Documentation of how data were classified and coded (e.g., multiple raters, blinding and interrater reliability)	37-38
20	Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate)	40
21	Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results	38
22	Assessment of heterogeneity	38-39
23	Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated	38-39
24	Provision of appropriate tables and graphics	39-46
Reporting of results should include		
25	Graphics summarizing individual study estimates and overall estimates	42-44
26	Table giving descriptive information for each study included	40
27	Results of sensitivity testing (e.g., subgroup analysis)	48 (limitations)
28	Indication of statistical uncertainty of findings	46-47

Reference: Stroup DF, Berlin JA, Morton SC, et al, for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of Observational Studies in Epidemiology. A Proposal for Reporting. *JAMA*. 2000;283(15):2008-2012. doi: 10.1001/jama.283.15.2008.

Appendix 6 - Search strategy.

Hip Search:

Hip replacement

Hip Prosthesis/ OR Arthroplasty, Replacement, Hip/ OR

(hip adj2 arthroplast$.mp) OR (hip adj2 replacement?.mp) OR (hip adj2 prosthes$.mp) OR

THA.mp OR THR.mp OR (TJR$.mp AND hip$.mp)

AND

Training

exp Education, Medical/ OR exp Inservice Training/ OR Clinical Competence/ OR

training.mp OR trainee.mp OR

experience.mp OR

junior.mp OR

senior$.mp OR

(surgeon adj2 grade).mp OR

consultant.mp OR attending?.mp OR registrar.mp OR SpR.mp OR StR.mp OR ST?.mp OR residen$.mp OR fellow$.mp OR intern.mp OR

(house adj2 officer).mp OR (foundation adj2 doctor).mp

AND

Survival

exp Prosthesis Failure/ OR exp Survival Analysis/ OR Reoperation/ OR

cox.mp OR proportional?hazard?.mp OR proportional hazard?.mp OR

cumulative?incidence?function.mp OR cumulative incidence function.mp OR CIF.mp OR

failure.mp OR

survival.mp OR survivor?ship.mp OR

revision?.mp OR

re?operation.mp OR re operation.mp OR

Kaplan?meier.mp OR Kaplan meier.mp OR KM.mp OR

product?limit?method.mp OR product limit method.mp

AND

Case-series

exp Cohort Studies/ OR Controlled Clinical Trials

follow?up.mp OR follow up.mp OR series.mp OR cohort.mp OR observational.mp OR longitudinal.mp OR prospective.mp OR retrospective.mp OR registry.mp OR registries.mp

Knee search:

Knee replacement

Knee Prosthesis/ OR Arthroplasty, Replacement, Knee/ OR

(knee adj2 arthroplast$.mp) OR (knee adj2 replacement?.mp) OR (knee adj2 prosthes$.mp) OR

TKA.mp OR TKR.mp OR (TJR$.mp AND knee$.mp) OR

UKA.mp OR UKR.mp

AND

Training

exp Education, Medical/ OR exp Inservice Training/ OR Clinical Competence/ OR

training.mp OR trainee.mp OR

experience.mp OR

junior.mp OR

senior$.mp OR

(surgeon adj2 grade).mp OR

consultant.mp OR attending?.mp OR registrar.mp OR SpR.mp OR StR.mp OR ST?.mp OR residen$.mp OR fellow$.mp OR intern.mp OR

(house adj2 officer).mp OR (foundation adj2 doctor).mp

AND

Survival

exp Prosthesis Failure/ OR exp Survival Analysis/ OR Reoperation/ OR

cox.mp OR proportional?hazard?.mp OR proportional hazard?.mp OR

cumulative?incidence?function.mp OR cumulative incidence function.mp OR CIF.mp OR

failure.mp OR

survival.mp OR survivor?ship.mp OR

revision?.mp OR

re?operation.mp OR re operation.mp OR

Kaplan?meier.mp OR Kaplan meier.mp OR KM.mp OR

product?limit?method.mp OR product limit method.mp

AND

Case-series

exp Cohort Studies/ OR Controlled Clinical Trials

follow?up.mp OR follow up.mp OR series.mp OR cohort.mp OR observational.mp OR longitudinal.mp OR prospective.mp OR retrospective.mp OR registry.mp OR registries.mp

Appendix 7 - Clarification of eligibility criteria.

Inclusion criteria:

- Study of predominantly adult patients (≥ 18 years old) undergoing primary hip or knee replacement (including THR, TKR, UKR and hip resurfacing), predominantly for the treatment of osteoarthritis.
- Included articles needed to report the primary and/or secondary outcome measure for two different groups of surgeons defined according to their grade (e.g. trainee vs. consultant). Additional terms used to describe this variable were deemed eligible during screening:
    - **Trainee:** registrar; resident; junior/young surgeon; fellow.
    - **Consultant:** attending; senior surgeon; trainer.
- Minimum follow-up of 5 years with clearly defined length of follow-up.

Exclusion criteria:

- Index operation performed prior to 1990.
- Follow-up not clearly defined.
- Irrelevant study design, or outcomes (therefore not meeting inclusion criteria above).

Specific examples for exclusion (documented in Appendix 8):

- Principally a study of surgeon/hospital volume
- Principally a study of implant positioning
- No revision rates/survival analysis reported according to surgeon grade
- No reporting of outcomes according to surgeon grade
- Insufficient reporting of follow-up
- Study of operations performed prior to 1990
- Hip fracture cohort
- Single surgeon series
- Irrelevant systematic review
- Study of cost-analysis

Appendix 8 - Reasons for exclusion.

Reasons for Exclusion – Hip Papers	
First author/Year of study	**Reason for Exclusion**
De Vries, 2011	Principally a study of surgeon/hospital volume
Fender, 2003	Principally a study of surgeon/hospital volume
Hooper, 2009	Principally a study of surgeon/hospital volume
Johnsson, 1994	Principally a study of surgeon/hospital volume
Namba, 2012	Principally a study of surgeon/hospital volume
Ravi, 2014	Principally a study of surgeon/hospital volume
Canadian Arthroplasty Soc., 2013	Principally a study of surgeon/hospital volume
MacBride, 2010	Principally a study of surgeon/hospital volume
Enocson, 2009	No revision rates/survival analysis reported according to surgeon grade
Field, 2006	No revision rates/survival analysis reported according to surgeon grade
Leguerrand, 2018	No revision rates/survival analysis reported according to surgeon grade
Moran, 2004	No revision rates/survival analysis reported according to surgeon grade
Smith, 2018	No revision rates/survival analysis reported according to surgeon grade
Wilson, 2016	No revision rates/survival analysis reported according to surgeon grade
Wroblewski, 1998	No revision rates/survival analysis reported according to surgeon grade
Schoenfeld, 2013	No revision rates/survival analysis reported according to surgeon grade
Inglis, 2013	Insufficient reporting of follow-up
Marston, 1996	Study of operations performed prior to 1990
Khatod, 2014	No reporting of outcomes according to surgeon grade
Whitehouse, 2014	No reporting of outcomes according to surgeon grade
Williams, 2002	No reporting of outcomes according to surgeon grade
Zwartele, 2005	No reporting of outcomes according to surgeon grade
Kim, 2017	Principally a study of implant positioning
MacDonald, 2020	Hip fracture cohort
DeAngelis, 2020	Hip fracture cohort
N.B. Multiple reasons for some papers	

Reasons for Exclusion – Knee Papers	
First author/Year of study	**Reason for Exclusion**
Bini, 2013	Principally a study of surgeon/hospital volume
Namba, 2012	Principally a study of surgeon/hospital volume
Zambianchi, 2014	Principally a study of surgeon/hospital volume
Rissolio, 2021	Principally a study of surgeon/hospital volume
Liddle, 2014	No revision rates/survival analysis reported according to surgeon grade
Beattie, 2016	No revision rates/survival analysis reported according to surgeon grade
Haughom, 2014	No revision rates/survival analysis reported according to surgeon grade
Khakha, 2015	No revision rates/survival analysis reported according to surgeon grade
Schoenfeld, 2013	No revision rates/survival analysis reported according to surgeon grade
Windisch, 2017	No revision rates/survival analysis reported according to surgeon grade
Wilson, 2016	No revision rates/survival analysis reported according to surgeon grade
Woolson, 2007	No revision rates/survival analysis reported according to surgeon grade
Atrey, 2014	No reporting of outcomes according to surgeon grade
Back, 2000	No reporting of outcomes according to surgeon grade
Singh, 2021	No reporting of outcomes according to surgeon grade
Gaillard, 2016	Principally a study of implant positioning
Mahaluxmivala, 2001	Principally a study of implant positioning
Storey, 2018	Insufficient reporting of follow-up
Theelen, 2018	Insufficient reporting of follow-up
Jasper, 2016	Irrelevant systematic review
Lacko, 2018	Single surgeon series
Matas-Diez, 2018	Principally a study of learning curve
Lavernia, 2000	Study of cost-analysis
N.B. Multiple reasons for some papers	

Appendix 9 - Risk of Bias (ROBINS-I) assessment.

ROBINS-I	**Alvand, 2021**	**Bottomley, 2016**	**Faulkner, 2017**	**Hernigou, 2009**	**Hasegawa, 2015**	**Jain, 2018**	**Muller, 2013**	**Palan, 2009**	**Reidy, 2016**
Bias due to confounding	⊕⊕	⊕⊕	⊕	⊕⊕	⊕⊕	⊕	⊕⊕	⊕	⊕
Bias in selection of patients	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Bias in classification of interventions	⊖	⊖	⊕	⊕	⊕⊕	⊕	⊕⊕	⊕	⊕
Bias due to deviations from interventions	⊕	⊕	⊕	⊕⊕	⊕⊕	⊕	⊕⊕	⊕	⊕
Bias due to missing data	⊖	⊖	⊕	⊕⊕	⊕⊕	⊕	⊕	⊖	⊕
Bias in measurement of outcome	⊖	⊖	⊕	⊖	⊕	⊖	⊖	⊖	⊕
Bias in selection of the reported result	⊖	⊖	⊖	⊖	⊕	⊖	⊕	⊖	⊖
Overall risk of Bias	⊕	⊕	⊕	⊕⊕	⊕⊕	⊕	⊕⊕	⊕	⊕
Key: ⊖ = low risk of bias; ⊕ = moderate risk of bias; ⊕⊕ = serious risk of bias; ⊕⊕⊕ = critical risk of bias									

Appendix 10 - Process of accounting for changes in NJR operating surgeon grade categories.

The current NJR MDS form (version 7.0) uses the following 'operating surgeon grade' categories: Consultant; SPR/ST3-ST8; F1-ST2; Specialty Doctor/SAS; and Other. However, since April 2003 there have been seven different versions of the MDS form, and the 'operating surgeon grade' categories have been updated with each iteration to reflect current training terminology. The following table details the process by which these variations have been accounted for, e.g. by recoding NJR records of 'operating surgeon grade' to correspond to the current MDS 7.0 categories.

<table>
<tr><th>NJR record of operating surgeon grade
(Based on numerous previous MDS versions)</th><th>Recoded to MDS 7.0 operating surgeon grade categories</th><th>Surgeon grade (binary variable)</th></tr>
<tr><td>Consultant</td><td>Consultant</td><td>Consultant</td></tr>
<tr><td>SpR/ST3-ST8</td><td rowspan="2">SpR/ST3-ST8</td><td rowspan="6">Trainee</td></tr>
<tr><td>SPR</td></tr>
<tr><td>F1-ST2</td><td rowspan="3">F1-ST2</td></tr>
<tr><td>House Officer (HO)</td></tr>
<tr><td>Senior House Officer (SHO)</td></tr>
<tr><td>Fellow</td><td rowspan="3">Other</td></tr>
<tr><td>Other</td><td rowspan="5">Not included in primary analyses</td></tr>
<tr><td>Visiting Overseas Specialist</td></tr>
<tr><td>Specialty Doctor/SAS</td><td rowspan="3">Specialty Doctor/SAS</td></tr>
<tr><td>Associate Specialist</td></tr>
<tr><td>Staff grade</td></tr>
</table>

Appendix 11 - Comparison of demographic characteristics for included and missing THR cases (Chapter 3).

Variable	Prior to exclusion of missing data (n=626,947)	Following exclusion of missing data (n=603,474)	Excluded due to missing data (n=23,473)
Surgeon grade (%)			
Consultant	566,295 (90.3)	545,337 (90.4)	20,958 (89.3)
Trainee	60,652 (9.7)	58,137 (9.6)	2,515 (10.7)
Mean age (SD)	69.7 (10.2)	69.7 (10.3)	70.8 (9.9)
Female (%)	382,649 (61.0)	367,746 (60.9)	14,903 (63.5)
Side (%right)	279,236 (55.5)	334,749 (55.5)	12,962 (55.2)
ASA grade (%)			
ASA I	95,483 (15.2)	90,782 (15.0)	4,701 (20.0)
ASA II	438,836 (70.0)	423,304 (70.1)	15,532 (66.2)
ASA ≥III	92,628 (14.8)	89,338 (14.8)	3,240 (13.8)
IMD decile (%)			
1	34,250 (5.5)	33,151 (5.5)	1,099 (4.7)
2	38,468 (6.1)	37,151 (6.2)	1,317 (5.6)
3	44,863 (7.2)	43,299 (7.2)	1,564 (6.7)
4	55,101 (8.8)	53,099 (8.8)	2,002 (8.5)
5	65,075 (10.4)	62,799 (10.4)	2,276 (9.7)
6	72,952 (11.6)	70,373 (11.7)	2,579 (11.0)
7	76,046 (12.1)	73,528 (12.2)	2,518 (10.7)
8	77,545 (12.4)	74,825 (12.4)	2,720 (11.6)
9	79,600 (12.7)	76,697 (12.7)	2,903 (12.4)
10	81,382 (13.0)	78,552 (13.0)	2,830 (12.1)
Missing*	1,665 (0.3)	-	1,665 (7.09)
Anaesthetic (%)			
Spinal	311,632 (49.7)	306,494 (50.8)	5,138 (21.9)
General	262,434 (41.9)	254,995 (42.3)	7,439 (31.7)
Epidural	24,144 (3.9)	23,010 (3.8)	1,134 (4.8)
Nerve block	18,580 (3.0)	18,329 (3.0)	251 (1.1)
Other	669 (<0.1)	646 (0.1)	23 (<0.1))
Missing*	9,488 (1.5)	-	9,488 (40.4)
Approach (%)			
Posterior	369,984 (59.0)	361,051 (59.8)	8,933 (38.06)
Other	256,963 (41.0)	242,423 (40.2)	14,540 (61.9)
Fixation (%)			
Cemented	230,727 (36.8)	218,918 (36.3)	11,809 (50.3)
Uncemented	251,336 (40.1)	244,389 (40.5)	6,947 (29.6)
Hybrid	126,857 (20.2)	122,609 (20.3)	4,248 (18.1)
Reverse hybrid	18,027 (2.9)	17,558 (2.9)	469 (2.0)
Bearing (%)			
Metal-on-polyethylene	404,988 (64.6)	388,375 (64.4)	16,613 (70.8)
Ceramic-on-polyethylene	112,830 (18.0)	109,637 (18.2)	3,193 (13.6)
Ceramic-on-ceramic	107,008 (17.1)	103,360 (17.1)	3,648 (15.5)
Other	2,121 (0.3)	2,102 (0.3)	19 (0.1)
Head size (%)			
<32mm	310,399 (49.5)	292,390 (48.5)	18,009 (76.7)
32mm	183,926 (29.3)	180,627 (29.9)	3,299 (14.05)
≥36mm	132,622 (21.2)	130,457 (21.6)	2,165 (9.2)
Funding source (%)			
NHS	520,288 (83.0)	513,585 (85.1)	6,703 (28.6)
Private	91,375 (14.6)	89,889 (14.9)	1,486 (6.3)
Missing*	15,284 (2.4)	-	15,284 (65.1)
Year of operation (%)			
2003-2012	358,462 (57.2)	335,692 (55.6)	22,770 (97.0)
2013-2016	268,485 (42.8)	267,782 (44.4)	703 (3.0)

ASA, American Society of Anaesthesiologists; NHS, National Health Service; IMD, Index of Multiple Deprivation. Data are n (%) or mean (SD); denoted where applicable. *Reason for exclusion for missing data not mutually exclusive. Some cases were excluded for missing data in more than one field.

Appendix 12 - Chapter 3 (THR) model specification summarising the exposures and confounding variables used in the analyses.

Model	Exposure (surgeon groups)	Method	Confounding variables included in model
1	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Unadjusted
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. SAS; v. Other (subcategorised according to consultant supervision)	Cox	
2	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Patient factors†
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. SAS; v. Other (subcategorised according to consultant supervision)	Cox	
3	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Patient factors† Operation factors‡
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. SAS; v. Other (subcategorised according to consultant supervision)	Cox	
4	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Patient factors† Operation factors‡ Healthcare setting factors§
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. SAS; v. Other (subcategorised according to consultant supervision)	Cox	
†**Patient factors:** age (continuous); sex (female); ASA (II); IMD decile (least deprived)			
‡**Operation factors:** bearing material (MoPe), mode of fixation (uncemented), head size (<32mm), approach (posterior), anaesthetic (spinal)			
§**Healthcare setting factors:** funding (NHS); year of operation (most recent). ***Baseline category in brackets was the most frequently occurring***			

Appendix 13 - Comparison of demographic characteristics for included and missing TKR cases (Chapter 4).

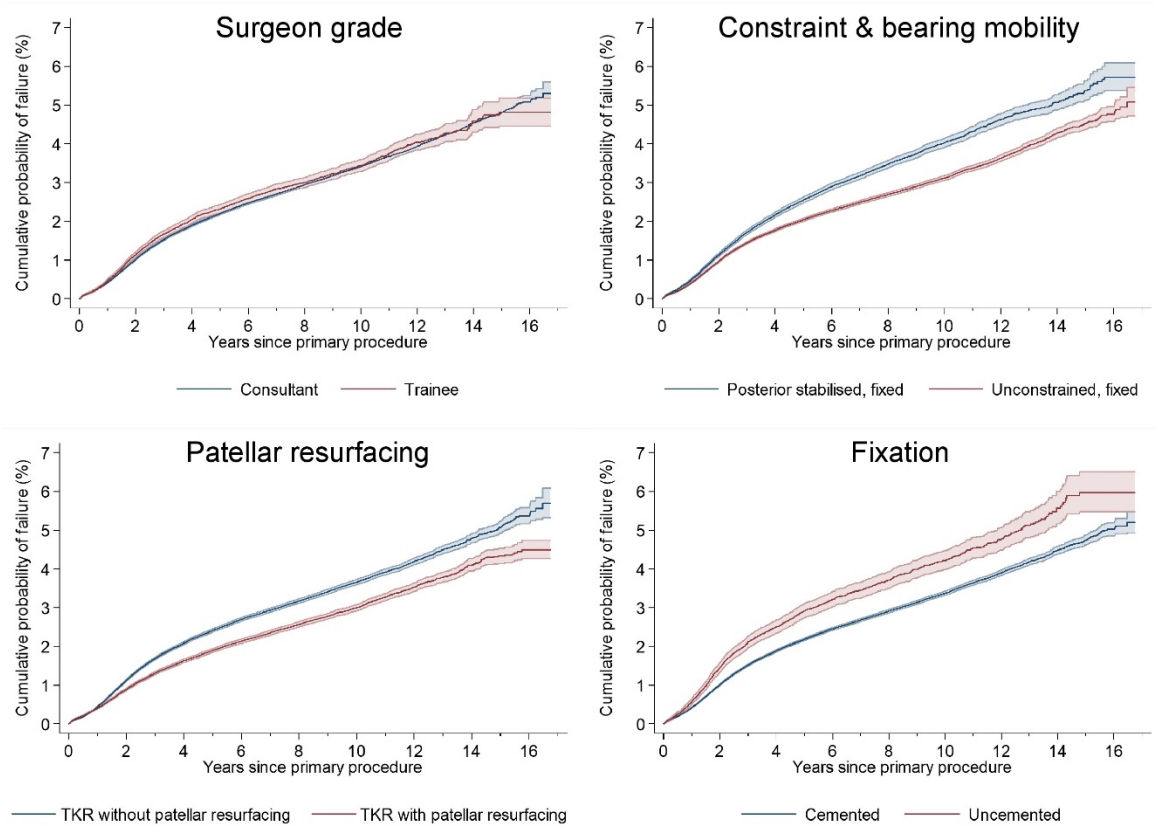
Variable	Prior to exclusion of missing data (n=998,753)	Following exclusion of missing data (n=953,081)	Excluded due to missing data (n=45,672)
Surgeon grade (%)			
Consultant	896,709 (89.8)	856,537 (89.9)	40,172 (88.0)
Trainee	102,044 (10.2)	96,544 (10.1)	5,500 (12.0)
Mean age (SD)	69.6 (9.2)	69.6 (9.2)	70.9 (9.1)
Female (%)	569,794 (57.1)	543,745 (57.1)	26,049 (57.0)
Side (% right)	526,323 (52.7)	502,320 (52.7)	24,003 (52.6)
ASA grade (%)			
ASA I	105,463 (10.6)	98,704 (10.4)	6,759 (14.8)
ASA II	725,306 (72.6)	693,851 (72.8)	31,455 (68.9)
ASA ≥III	167,984 (16.8)	160,526 (16.8)	7,458 (16.3)
IMD decile (%)			
1 (most deprived)	67,545 (6.8)	65,797 (6.9)	1,748 (3.8)
2	73,734 (7.4)	71,685 (7.5)	2,049 (4.5)
3	81,768 (8.2)	79,496 (8.3)	2,272 (5.0)
4	92,220 (9.2)	89,470 (9.4)	2,730 (6.0)
5	103,281 (10.3)	100,228 (10.5)	3,053 (6.7)
6	110,417 (11.1)	107,082 (11.2)	3,335 (7.3)
7	113,407 (11.4)	110,245 (11.6)	3,162 (6.9)
8	114,210 (11.4)	110,789 (11.6)	3,421 (7.5)
9	114,871 (11.5)	111,428 (11.7)	3,443 (7.5)
10 (least deprived)	109,964 (11.0)	106,861 (11.2)	3,103 (6.8)
Missing*	17,356 (1.7)	-	17,356 (38.0)
Anaesthetic (%)			
Spinal	662,750 (66.4)	642,616 (67.4)	20,134 (44.1)
General	362,662 (36.3)	347,900 (36.5)	14,762 (32.3)
Epidural	57,755 (5.8)	53,310 (5.6)	4,445 (9.7)
Nerve block	145,468 (14.6)	139,816 (14.7)	5,652 (12.4)
Missing*	11,335 (1.1)	-	11,335 (24.8)
Approach (%)			
Lateral parapatellar	8,482 (0.9)	7,990 (0.8)	492 (1.1)
Medial parapatellar	934,521 (93.6)	891,688 (93.6)	42,833 (93.8)
Mid-vastus	24,301 (2.4)	23,928 (2.5)	373 (0.8)
Sub-vastus	11,220 (1.1)	10,687 (1.1)	533 (1.2)
Other	20,216 (2.0)	18,788 (2.0)	1,428 (3.1)
Fixation (%)			
Cemented	949,518 (95.1)	907,750 (95.2)	41,768 (91.5)
Uncemented	41,134 (4.1)	37,819 (4.0)	3,315 (7.3)
Hybrid	8,101 (0.8)	7,512 (0.8)	589 (1.3)
Constraint (%)			
Constrained condylar	7,595 (0.8)	7,361 (0.8)	234 (0.5)
Monobloc poly tibia	15,965 (1.6)	15,566 (1.6)	399 (0.9)
Posterior stabilised, fixed	227,934 (22.8)	217,158 (22.8)	10,776 (23.6)
Posterior stabilised, mobile	11,879 (1.2)	10,950 (1.2)	929 (2.0)
Preassembled/hinged/linked	1,168 (0.1)	1,074 (0.1)	94 (0.2)
Unconstrained, fixed	675,096 (67.6)	647,270 (67.9)	27,826 (60.9)
Unconstrained, mobile	59,114 (5.9)	53,702 (5.6)	5,412 (11.9)
Missing*	2 (<0.01)	-	2 (<0.01)
Patellar resurfacing (%)			
Patella resurfaced	382,782 (38.3)	366,344 (38.4)	16,438 (36.0)
Funding source (%)			
NHS	867,465 (86.9)	846,385 (88.8)	21,080 (46.2)
Private	109,579 (11.0)	106,696 (11.2)	2,883 (6.3)
Missing*	21,709 (2.2)	-	21,709 (47.5)
Year of operation (%)			
2003-2011	379,041 (37.9)	342,217 (35.9)	36,824 (80.6)
2012-2019	619,712 (62.0)	610,864 (64.1)	8,848 (19.4)

ASA, American Society of Anaesthesiologists; NHS, National Health Service; IMD, Index of Multiple Deprivation. Data are n (%) or mean (SD); denoted where applicable. *Reason for exclusion for missing data not mutually exclusive. Some cases were excluded for missing data in more than one field.

Appendix 14 - Chapter 4 (TKR) model specification summarising the exposures and confounding variables used in the analyses.

Model	Exposure (surgeon groups)	Method	Confounding variables included in model
1	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Unadjusted
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Specific training grade: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	FPM	
2	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors†
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Specific training grade: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	FPM	
3	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors† Operation factors‡
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Specific training grade: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	FPM	
4	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors† Operation factors‡ Healthcare setting factors§
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Specific training grade: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	FPM	
†**Patient factors:** age (65-75); gender (female); ASA (II); IMD decile (least deprived)			
‡**Operation factors:** approach (medial parapatellar); mode of fixation (cemented); constraint (unconstrained fixed); anaesthetic (spinal); patellar resurfacing (not resurfaced)			
§**Healthcare setting factors:** funding (NHS); year of operation (most recent). ***Baseline category in brackets was the most frequently occurring***			
Time-dependent effects: Surgeon grade was specified as having a time-dependent effect and was modelled with 2 df. Remaining confounding variables were modelled with fixed effects. The baseline hazard was modelled with 8 df.			

Appendix 15 - Kaplan-Meier plots (one minus survival) demonstrating the cumulative probability of failure according to the surgeon grade (exposure A); constraint; patellar resurfacing; and fixation.



Appendix 16 - Comparison of demographic characteristics for included and missing UKR cases (Chapter 5).

	Prior to exclusion of missing data (n=111,326)	**Following exclusion of missing data (n=106,206)**	**Excluded due to missing data (n=5,120)**
Surgeon grade (%)			
Consultant	106,697 (95.8)	101,824 (95.9)	4,873 (95.2)
Trainee	4,629 (4.2)	4,382 (4.1)	247 (4.8)
Mean age (SD)	63.9 (9.7)	63.8 (9.7)	65.2 (9.7)
Female (%)	51,535 (46.3)	49,077 (46.2)	2,458 (48.0)
Side (%right)	55,672 (50.0)	53,127 (50.0)	2,545 (49.7)
IMD decile (%)			
1 (most deprived)	4,683 (4.2)	4,545 (4.3)	138 (2.7)
2	5,882 (5.3)	5,729 (5.4)	153 (3.0)
3	7,090 (6.4)	6,878 (6.5)	212 (4.1)
4	9,126 (8.2)	8,803 (8.3)	323 (6.3)
5	10,864 (9.8)	10,511 (9.9)	353 (6.9)
6	12,805 (11.5)	12,429 (11.7)	376 (7.3)
7	13,682 (12.3)	13,311 (12.5)	371 (7.3)
8	13,742 (12.3)	13,357 (12.6)	385 (7.5)
9	15,112 (13.6)	14,723 (13.9)	389 (7.6)
10 (least deprived)	16,339 (14.7)	15,920 (15.0)	419 (8.2)
Missing	2,001 (1.8)	-	2,001 (39.1)
ASA grade (%)			
ASA I	23,608 (21.2)	22,349 (21.0)	1,259 (24.6)
ASA II	78,067 (70.1)	74,669 (70.3)	3,398 (66.4)
ASA ≥III	9,651 (8.7)	9,188 (8.7)	463 (9.0)
Anaesthetic (%)			
Spinal	61,869 (55.6)	60,121 (56.6)	1,748 (34.1)
General	52,280 (47.0)	49,976 (47.1)	2,304 (45.0)
Epidural	4,912 (4.4)	4,621 (4.4)	291 (5.7)
Nerve block	18,374 (16.5)	17,795 (16.8)	579 (11.3)
Missing*	1,124 (1.0)	-	1,124 (22.0)
Approach (%)			
Lateral parapatellar	3,588 (3.2)	3,421 (3.2)	167 (3.3)
Medial parapatellar	98,790 (88.7)	94,131 (88.6)	4,659 (91.0)
Mid-vastus	4,175 (3.8)	4,099 (3.9)	76 (1.5)
Sub-vastus	1,687 (1.5)	1,639 (1.5)	169 (3.3)
Other	3,085 (2.8)	2,916 (2.8)	48 (0.9)
Missing*	1 (<0.1)	-	1 (<0.1)
Fixation (%)			
Cemented	86,909 (78.1)	82,414 (77.6)	4,495 (87.8)
Uncemented	21,729 (19.5)	21,245 (20.0)	484 (9.5)
Hybrid	2,688 (2.4)	2,547 (2.4)	141 (2.8)
Bearing mobility (%)			
Fixed	36,173 (32.5)	35,180 (33.1)	993 (19.4)
Mobile	68,892 (61.9)	65,164 (61.4)	3,728 (72.8)
Monobloc poly tibia	6,261 (5.6)	5,862 (5.5)	399 (7.8)
Funding source (%)			
NHS	84,064 (75.5)	81,965 (77.2)	2,099 (41.0)
Private	24,899 (22.4)	24,241 (22.8)	658 (12.9)
Missing*	2,363 (2.1)	-	2,363 (46.2)
Year of operation (%)			
2003-2011	41,050 (36.9)	37,134 (35.0)	3,916 (76.5)
2012-2019	70,276 (63.1)	69,072 (65.0)	1,204 (23.5)

ASA, American Society of Anaesthesiologists; NHS, National Health Service; IMD, Index of Multiple Deprivation; BMI, Body Mass Index. Data are n (%) or mean (SD); denoted where applicable. *Reason for exclusion for missing data not mutually exclusive. Some cases were excluded for missing data in more than one field.

Appendix 17 - Chapter 5 (UKR) model specification summarising the exposures and confounding variables used in the analyses.

Model	Exposure (surgeon groups)	Method	Confounding variables included in model
1	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Unadjusted
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow (subcategorised according to consultant supervision)	FPM	
2	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors†
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow (subcategorised according to consultant supervision)	FPM	
3	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors† Operation factors‡
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow (subcategorised according to consultant supervision)	FPM	
4	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	FPM	Patient factors† Operation factors‡ Healthcare setting factors§
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	FPM	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow (subcategorised according to consultant supervision)	FPM	
†**Patient factors:** age (55-64yrs); gender (male); ASA (II); IMD decile (least deprived)			
‡**Operation factors:** approach (medial parapatellar); mode of fixation (cemented); bearing mobility (mobile bearing); anaesthetic (spinal)			
§**Healthcare setting factors:** funding (NHS); year of operation (most recent). ***Baseline category in brackets was the most frequently occurring***			
Time-dependent effects: The following confounding variables were specified as having time-dependent effects: age (modelled with 1 degree of freedom (df)); sex (1 df); IMD decile (1 df); approach (1 df); fixation (3 df); bearing mobility (2 df); funder (3 df); year of operation (2 df). The baseline hazard was modelled with 6 df. Surgeon grade and the remaining confounding variables were modelled with fixed effects.			

Appendix 18 - Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Checklist.

Item No	Recommendation	Reported on Page No
Title and abstract		
1	a) Indicate the study's design with a commonly used term in the title or the abstract b) Provide in the abstract an informative and balanced summary of what was done and what was found	134
Introduction		
2	**Background/rationale:** Explain the scientific background and rationale for the investigation being reported	136-137
3	**Objectives:** State specific objectives, including any prespecified hypotheses	137
Methods		
4	**Study design:** Present key elements of study design early in the paper	137-138
5	**Setting:** Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	137-138
6	**Participants:** a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up b) For matched studies, give matching criteria and number of exposed and unexposed	137-138
7	**Variables:** Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	140-144
8	**Data sources/measurement:** For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	140-144
9	**Bias:** Describe any efforts to address potential sources of bias	144
10	**Study size:** Explain how the study size was arrived at	139
11	**Quantitative variables:** Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	144-145
12	**Statistical methods:** a) Describe all statistical methods, including those used to control for confounding b) Describe any methods used to examine subgroups and interactions c) Explain how missing data were addressed d) If applicable, explain how loss to follow-up was addressed e) Describe any sensitivity analyses	144-145
Results		
13	**Participants:** a) Report numbers of individuals at each stage of study—e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed b) Give reasons for non-participation at each stage c) Consider use of a flow diagram	139, 145
14	**Descriptive data:** a) Give characteristics of study participants (e.g. demographic, clinical, social) and information on exposures and potential confounders b) Indicate number of participants with missing data for each variable of interest c) Summarise follow-up time (e.g. average and total amount)	139, 145, 148, 151, 154
15	**Outcome data:** Report numbers of outcome events or summary measures over time	145-156
16	**Main results:** a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included	145-156

	b) Report category boundaries when continuous variables were categorized c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
17	**Other analyses:** Report other analyses done—e.g. analyses of subgroups and interactions, and sensitivity analyses	156
Discussion		
18	**Key results:** Summarise key results with reference to study objectives	157
19	**Limitations:** Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	157-158
20	**Interpretation:** Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	157, 158-160
21	**Generalisability:** Discuss the generalisability (external validity) of the study results	157
Other information		
22	**Funding:** Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	134

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

Appendix 19 - Process of accounting for variations in grade terminology during the course of the study (Chapter 6).

UK training grade terminology has evolved over the past 20 years. Our process of accounting for variations in grade terminology over the course of this study (Chapter 6) is documented in the following table.

Grade recorded on the operation note	**Current equivalent grade categories**	**Surgeon grade (binary variable)**
Consultant	Consultant	Consultant
House Officer (or F1)	F1-ST2	Trainee
Senior House Officer (SHO) (or F2-ST2)		
Registrar (or ST3-ST8)	ST3-ST8	
Fellow	Fellow	

Appendix 20 - Chapter 6 (THR) model specification showing the exposures and confounding variables used in the analysis.

Model	Exposure (surgeon groups)	Method	Confounding variables included in model
1	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Unadjusted
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	Cox	
2	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Patient factors†
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	Cox	
3	**A. Surgeon grade:** i. Consultant; ii. Trainee (overall)	Cox	Patient factors† Operation factors‡
	B. Supervision: i. Consultant; ii. Trainee supervised by a scrubbed consultant; iii. Trainee not supervised by a scrubbed consultant	Cox	
	C. Sensitivity analysis: i. Consultant; ii. F1-ST2; iii. ST3-ST8; iv. Fellow	Cox	
†**Patient factors:** age (continuous); sex (female); ASA (2); indication (OA)			
‡**Operation factors:** cup fixation (cemented); cup brand (Ogee); head size (28mm); bearing (metal on polyethylene). ***Baseline category in brackets was the most frequently occurring***			

Appendix 21 - The unadjusted cumulative probability of failure of THRs according to surgeon grade (exposure A) and supervision (exposure B).

Follow-up (years)	Consultant			Trainee (exposure A)			Trainee supervised by a scrubbed consultant (exposure B)			Trainee not supervised by a scrubbed consultant (exposure B)		
	Number at risk*	Number of revisions	% Failure, (95% CI)	Number at risk*	Number of revisions	% Failure, (95% CI)	Number at risk*	Number of revisions	% Failure, (95% CI)	Number at risk*	Number of revisions	% Failure, (95% CI)
0-1	119	0	0	400	3	0.8 (0.2 to 2.3)	204	2	1.0 (0.3 to 3.9)	196	1	0.5 (0.1 to 3.6)
1-2	116	0	0	385	1	1.0 (0.4 to 2.7)	195	0	1.0 (0.3 to 3.9)	190	1	1.0 (0.3 to 4.1)
2-3	112	0	0	373	1	1.3 (0.5 to 3.1)	190	0	1.0 (0.3 to 3.9)	183	1	1.6 (0.5 to 4.9)
3-4	109	0	0	359	0	1.3 (0.5 to 3.1)	187	0	1.0 (0.3 to 3.9)	172	0	1.6 (0.5 to 4.9)
4-5	100	0	0	343	1	1.6 (0.7 to 3.5)	178	0	1.0 (0.3 to 3.9)	165	1	2.2 (0.8 to 5.8)
5-6	94	0	0	319	0	1.6 (0.7 to 3.5)	164	0	1.0 (0.3 to 3.9)	155	0	2.2 (0.8 to 5.8)
6-7	92	2	2.2 (0.6 to 8.5)	304	0	1.6 (0.7 to 3.5)	155	0	1.0 (0.3 to 3.9)	149	0	2.2 (0.8 to 5.8)
7-8	86	1	3.4 (1.1 to 10.1)	280	1	1.9 (0.9 to 4.1)	140	0	1.0 (0.3 to 3.9)	140	1	2.9 (1.2 to 7.0)
8-9	76	0	3.4 (1.1 to 10.1)	262	1	2.3 (1.2 to 4.7)	131	0	1.0 (0.3 to 3.9)	131	1	3.7 (1.7 to 8.1)
9-10	73	0	3.4 (1.1 to 10.1)	247	1	2.7 (1.4 to 5.3)	125	0	1.0 (0.3 to 3.9)	122	1	4.5 (2.2 to 9.4)
10-11	69	1	4.9 (1.8 to 12.6)	229	0	2.7 (1.4 to 5.3)	119	0	1.0 (0.3 to 3.9)	110	0	4.5 (2.2 to 9.4)
11-12	59	0	4.9 (1.8 to 12.6)	187	0	2.7 (1.4 to 5.3)	90	0	1.0 (0.3 to 3.9)	97	0	4.5 (2.2 to 9.4)
12-13	41	1	7.7 (3.0 to 18.8)	146	0	2.7 (1.4 to 5.3)	60	0	1.0 (0.3 to 3.9)	86	0	4.5 (2.2 to 9.4)
13-14	26	0	7.7 (3.0 to 18.8)	72	0	2.7 (1.4 to 5.3)	31	0	1.0 (0.3 to 3.9)	41	0	4.5 (2.2 to 9.4)
14-15	8	0	7.7 (3.0 to 18.8)	14	0	2.7 (1.4 to 5.3)	3	0	1.0 (0.3 to 3.9)	11	0	4.5 (2.2 to 9.4)

Data are the number at risk, the number of revision events, the unadjusted cumulative probability of failure and the 95% CI. *Number at risk at the beginning of time period.

Appendix 22 - Demographic and operation-level data for PROMs responders and non-responders.

Variable	PROMs response		
	Responder (n=188)	**Non-responder (N=342)**	**p-value**
Mean (SD) age at operation (years)	71.1 (7.1)	76.0 (8.2)	<0.01
Sex (%)			
Female	132 (70.2)	223 (65.2)	0.24
Side (%)			
Right	100 (53.2)	190 (55.6)	0.60
ASA (%)			<0.01
1	10 (5.3)	13 (3.8)	
2	156 (83.0)	221 (64.6)	
≥3	22 (11.7)	108 (31.6)	
Indication (%)			0.90
Osteoarthritis (OA)	182 (96.8)	325 (95.0)	
Rheumatoid arthritis (RA)	2 (1.1)	7 (2.1)	
Avascular necrosis (AVN)	2 (1.1)	5 (1.5)	
Hip fracture (NOF)	1 (0.5)	2 (0.6)	
Other	1 (0.5)	3 (0.9)	
Bearing surface (%)			<0.01
Metal on poly (MoP)	141 (75.0)	294 (86.0)	
Metal on ceramic (MoC)	0 (0.0)	1 (0.3)	
Metal on metal (MoM)	19 (10.1)	28 (8.2)	
Ceramic on poly (CoP)	1 (0.5)	4 (1.2)	
Ceramic on ceramic (CoC)	26 (13.8)	9 (2.6)	
Ceramic on metal (CoM)	1 (0.5)	6 (1.8)	
Acetabular component (%)			<0.01
Pinnacle	113 (60.1)	134 (39.2)	
Ogee	69 (36.7)	198 (57.9)	
Duraloc	5 (2.7)	3 (0.9)	
Omnifit	0 (0.0)	3 (0.9)	
Trident	0 (0.0)	3 (0.9)	
ASR	1 (0.5)	1 (0.3)	
Fixation (%)			<0.01
Hybrid	119 (63.3)	144 (42.1)	
Cemented	69 (36.7)	198 (57.9)	
Head size (%)			0.01
28mm	140 (74.5)	282 (82.5)	
36mm	45 (23.9)	49 (14.3)	
Other	3 (1.6)	11 (3.22)	

Data are the frequency (n), proportion (%), mean, or standard deviation (SD); denoted where applicable. Statistical tests were used to assess equivalence between the responder and non-responder groups. For age, the Shapiro-Wilk test for normality was used. This suggested that age was not normally distributed, thus the Wilcoxon rank sum test was used for this variable. Chi-squared tests were used for the remaining categorical variables.