
Article:
Prats-Uribe, A., Kolovos, S., Berencsi, K. et al. (15 more authors) (2021) Unicompartmental compared with total knee replacement for patients with multimorbidities : a cohort study using propensity score stratification and inverse probability weighting. Health Technology Assessment, 25 (66). pp. 1-125. ISSN 1366-5278

https://doi.org/10.3310/hta25660

NIHR | National Institute for Health Research

Journals Library

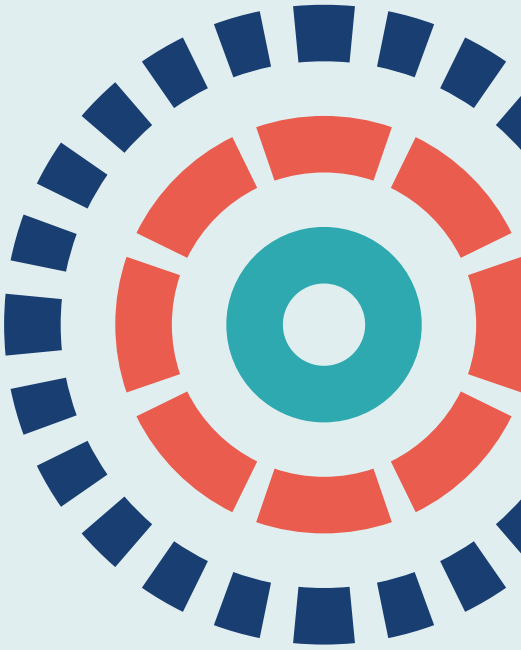
Health Technology Assessment

Volume 25 • Issue 66 • November 2021

ISSN 1366-5278

Unicompartmental compared with total knee replacement for patients with multimorbidities: a cohort study using propensity score stratification and inverse probability weighting

Albert Prats-Uribe, Spyros Kolovos, Klara Berencsi, Andrew Carr, Andrew Judge, Alan Silman, Nigel Arden, Irene Petersen, Ian J Douglas, J Mark Wilkinson, David Murray, Jose M Valderas, David J Beard, Sarah E Lamb, M Sanni Ali, Rafael Pinedo-Villanueva, Victoria Y Strauss and Daniel Prieto-Alhambra



DOI 10.3310/hta25660

Unicompartmental compared with total knee replacement for patients with multimorbidities: a cohort study using propensity score stratification and inverse probability weighting

Albert Prats-Uribe,[1] Spyros Kolovos,[1] Klara Berencsi,[1] Andrew Carr,[1] Andrew Judge,[1,2] Alan Silman,[1] Nigel Arden,[1,3,4] Irene Petersen,[5] Ian J Douglas,[6] J Mark Wilkinson,[7,8] David Murray,[1] Jose M Valderas,[9] David J Beard,[1] Sarah E Lamb,[1,10] M Sanni Ali,[1,6] Rafael Pinedo-Villanueva,[1] Victoria Y Strauss[1*] and Daniel Prieto-Alhambra[1]

[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK
[2]Musculoskeletal Research Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Southmead Hospital, Bristol, UK
[3]Centre for Sport, Exercise and Osteoarthritis Research Versus Arthritis, Botnar Research Centre, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK
[4]Medical Research Council Lifecourse Epidemiological Unit, University of Southampton, Southampton, UK
[5]Department of Primary Care and Population Health, University College London, London, UK
[6]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
[7]Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK
[8]Research Committee, National Joint Registry for England, Wales, Northern Ireland and the Isle of Man, Hemel Hempstead, UK
[9]College of Medicine and Health, University of Exeter, Exeter, UK
[10]University of Exeter Medical School, Institute of Health Research, College of Medicine and Health, Exeter, UK

*Corresponding author

Declared competing interests of authors: Albert Prats-Uribe reports grants from Fundación Alfonso Martin Escudero (Madrid, Spain) and the Medical Research Council (London, UK). Andrew Judge was a sub-panel member of the National Institute for Health Research (NIHR) Programme Grants for Applied Research (PGfAR) programme (September 2015–August 2020); has received personal fees from Freshfields Bruckhaus Deringer (London, UK); and was a member of the Data Safety and Monitoring Board (December 2012 to June 2016) (which involved receipt of fees) from Anthera Pharmaceuticals Inc. (Hayward, CA, USA). Nigel Arden reports grants from Merck & Co. Inc. (Darmstadt, Germany) and

personal fees from Merck & Co. Inc., Flexion Therapeutics (Burlington, MA, USA), Regeneron (Tarrytown, NY, USA) and Pfizer Inc. (New York, NY, USA)/Eli Lilly and Company (Indianapolis, IN, USA) outside the submitted work. David Murray reports research grants from NIHR Health Technology Assessment (HTA), specifically for TOPKAT (HTA 08/14/08), outside the submitted work; institutional grant and consultancy fees to the University of Oxford with Zimmer Biomet (Warsaw, IN, USA); and royalties from knee replacement-related patents. David J Beard reports grants from the NIHR HTA programme during the conduct of the study, specifically for TOPKAT. Sarah E Lamb was on the NIHR HTA Additional Capacity Funding Board (2012–15), HTA End of Life Care and Add-on Studies Board (2015), HTA Prioritisation Group Board (2010–15) and the HTA Trauma Board (2007–8). Rafael Pinedo-Villanueva reports personal fees from Mereo BioPharma Group plc (London, UK), Kyowa Kirin International (Galashiels, UK) and UCB Biopharma SPRL (Brussels, Belgium) outside the submitted work. Daniel Prieto-Alhambra reports NIHR HTA Funding Committee membership (November 2017–present); research grants from Amgen Inc. (Thousand Oaks, CA, USA), Johnson & Johnson (New Brunswick, NJ, USA), and UCB Biopharma SPRL; speaker services and advisory board membership fees paid to his department/research group from Amgen; and consultancy fees paid to his department/research group from UCB Biopharma SPRL. He also reports that Janssen Pharmaceutica (Beerse, Belgium), on behalf of Innovative Medicines Initiative-funded European Health Data Evidence Network and European Medical Information Framework consortiums, and Synapse Management Partners have supported internal training programmes organised by his department and open training programmes for external participants.

Published November 2021
DOI: 10.3310/hta25660

This report should be referenced as follows:

Prats-Uribe A, Kolovos S, Berencsi K, Carr A, Judge A, Silman A, *et al*. Unicompartmental compared with total knee replacement for patients with multimorbidities: a cohort study using propensity score stratification and inverse probability weighting. *Health Technol Assess* 2021;**25**(66).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (SciSearch®) and *Current Contents®*/ Clinical Medicine.

Health Technology Assessment

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 4.014

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, the Cochrane Library and Clarivate Analytics Science Citation Index.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 15/80/40. The contractual start date was in June 2017. The draft report began editorial review in January 2020 and was accepted for publication in July 2020. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health and Social Care.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

NIHR Journals Library Editor-in-Chief

NIHR Journals Library Editors

Please visit the website for a list of editors: www.journalslibrary.nihr.ac.uk/about/editors

Editorial contact: journals.library@nihr.ac.uk

Abstract

Unicompartmental compared with total knee replacement for patients with multimorbidities: a cohort study using propensity score stratification and inverse probability weighting

Albert Prats-Uribe,[1] Spyros Kolovos,[1] Klara Berencsi,[1] Andrew Carr,[1] Andrew Judge,[1,2] Alan Silman,[1] Nigel Arden,[1,3,4] Irene Petersen,[5] Ian J Douglas,[6] J Mark Wilkinson,[7,8] David Murray,[1] Jose M Valderas,[9] David J Beard,[1] Sarah E Lamb,[1,10] M Sanni Ali,[1,6] Rafael Pinedo-Villanueva,[1] Victoria Y Strauss[1*] and Daniel Prieto-Alhambra[1]

[1]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK
[2]Musculoskeletal Research Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Southmead Hospital, Bristol, UK
[3]Centre for Sport, Exercise and Osteoarthritis Research Versus Arthritis, Botnar Research Centre, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK
[4]Medical Research Council Lifecourse Epidemiological Unit, University of Southampton, Southampton, UK
[5]Department of Primary Care and Population Health, University College London, London, UK
[6]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
[7]Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK
[8]Research Committee, National Joint Registry for England, Wales, Northern Ireland and the Isle of Man, Hemel Hempstead, UK
[9]College of Medicine and Health, University of Exeter, Exeter, UK
[10]University of Exeter Medical School, Institute of Health Research, College of Medicine and Health, Exeter, UK

*Corresponding author Victoria.strauss@csm.ox.ac.uk

Background: Although routine NHS data potentially include all patients, confounding limits their use for causal inference. Methods to minimise confounding in observational studies of implantable devices are required to enable the evaluation of patients with severe systemic morbidity who are excluded from many randomised controlled trials.

Objectives: Stage 1 – replicate the Total or Partial Knee Arthroplasty Trial (TOPKAT), a surgical randomised controlled trial comparing unicompartmental knee replacement with total knee replacement using propensity score and instrumental variable methods. Stage 2 – compare the risk benefits and cost-effectiveness of unicompartmental knee replacement with total knee replacement surgery in patients with severe systemic morbidity who would have been ineligible for TOPKAT using the validated methods from stage 1.

Design: This was a cohort study.

Setting: Data were obtained from the National Joint Registry database and linked to hospital inpatient (Hospital Episode Statistics) and patient-reported outcome data.

Participants: Stage 1 – people undergoing unicompartmental knee replacement surgery or total knee replacement surgery who met the TOPKAT eligibility criteria. Stage 2 – participants with an American Society of Anesthesiologists grade of ≥ 3.

Intervention: The patients were exposed to either unicompartmental knee replacement surgery or total knee replacement surgery.

Main outcome measures: The primary outcome measure was the postoperative Oxford Knee Score. The secondary outcome measures were 90-day postoperative complications (venous thromboembolism, myocardial infarction and prosthetic joint infection) and 5-year revision risk and mortality. The main outcome measures for the health economic analysis were health-related quality of life (EuroQol-5 Dimensions) and NHS hospital costs.

Results: In stage 1, propensity score stratification and inverse probability weighting replicated the results of TOPKAT. Propensity score adjustment, propensity score matching and instrumental variables did not. Stage 2 included 2256 unicompartmental knee replacement patients and 57,682 total knee replacement patients who had severe comorbidities, of whom 145 and 23,344 had linked Oxford Knee Scores, respectively. A statistically significant but clinically irrelevant difference favouring unicompartmental knee replacement was observed, with a mean postoperative Oxford Knee Score difference of < 2 points using propensity score stratification; no significant difference was observed using inverse probability weighting. Unicompartmental knee replacement more than halved the risk of venous thromboembolism [relative risk 0.33 (95% confidence interval 0.15 to 0.74) using propensity score stratification; relative risk 0.39 (95% confidence interval 0.16 to 0.96) using inverse probability weighting]. Unicompartmental knee replacement was not associated with myocardial infarction or prosthetic joint infection using either method. In the long term, unicompartmental knee replacement had double the revision risk of total knee replacement [hazard ratio 2.70 (95% confidence interval 2.15 to 3.38) using propensity score stratification; hazard ratio 2.60 (95% confidence interval 1.94 to 3.47) using inverse probability weighting], but half of the mortality [hazard ratio 0.52 (95% confidence interval 0.36 to 0.74) using propensity score stratification; insignificant effect using inverse probability weighting]. Unicompartmental knee replacement had lower costs and higher quality-adjusted life-year gains than total knee replacement for stage 2 participants.

Limitations: Although some propensity score methods successfully replicated TOPKAT, unresolved confounding may have affected stage 2. Missing Oxford Knee Scores may have led to information bias.

Conclusions: Propensity score stratification and inverse probability weighting successfully replicated TOPKAT, implying that some (but not all) propensity score methods can be used to evaluate surgical innovations and implantable medical devices using routine NHS data. Unicompartmental knee replacement was safer and more cost-effective than total knee replacement for patients with severe comorbidity and should be considered the first option for suitable patients.

Future work: Further research is required to understand the performance of propensity score methods for evaluating surgical innovations and implantable devices.

Trial registration: This trial is registered as EUPAS17435.

Funding: This project was funded by the National Institute for Health Research (NIHR) Health Technology Assessment programme and will be published in full in *Health Technology Assessment*; Vol. 25, No. 66. See the NIHR Journals Library website for further project information.

Contents

List of tables

List of figures

List of abbreviations

APC	Admitted Patient Care
ASA	American Society of Anesthesiologists
ASMD	absolute standardised mean difference
ATE	average treatment effect
ATT	average treatment effect on treated
BMI	body mass index
CAG	Confidentiality Advisory Group
CI	confidence interval
EHDEN	European Health Data and Evidence Network
EQ-5D	EuroQol-5 Dimensions
EQ-5D-3L	EuroQol-5 Dimensions, three-level version
FDA	Food and Drug Administration
HES	Hospital Episode Statistics
HRG	Healthcare Resource Group
HRQoL	health-related quality of life
HTA	Health Technology Assessment
ICD-10	*International Classification of Diseases*, Tenth Revision
ICER	incremental cost-effectiveness ratio
IMD	Index of Multiple Deprivation
IPW	inverse probability weighting
IQR	interquartile range
IV	instrumental variable
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NJR	National Joint Registry
OHDSI	Observational Health Data Sciences and Informatics
OKS	Oxford Knee Score
OPCS-4	Office of Population Censuses & Surveys Version 4
PPI	patient and public involvement
PROM	patient-reported outcome measure
PS	propensity score
PSS_{exp}	propensity score stratification in the unicompartmental knee replacement cohort
PSS_{whole}	propensity score stratification in the whole cohort
QALY	quality-adjusted life-year
RCT	randomised controlled trial
SD	standard deviation
TKR	total knee replacement
TOPKAT	Total or Partial Knee Arthroplasty Trial
UKR	unicompartmental knee replacement
WTP	willingness to pay

Plain English summary

We compared the risks and benefits of partial and total knee replacements in NHS patients with a complex medical history who would normally be excluded from randomised trials on this topic. We used information that was collected during hospital appointments for people who had a knee replacement between 2009 and 2016. It is difficult to directly compare the two groups because each individual patient has a different medical history. We tested advanced statistical methods to account for these differences.

In stage 1, we showed that some of these advanced statistical methods could replicate the results of a recently published surgical trial using routine data from the NHS. We compared patients in the trial with similar patients who were operated on in the NHS. Three of the proposed methods showed results similar to those obtained from the Total or Partial Knee Arthroplasty Trial (TOPKAT).

In stage 2, we used the successful methods from stage 1 to study the risks, benefits and costs of partial and total knee replacement surgery in patients with complex medical histories. Two of the statistical methods found that patients who had a partial knee replacement had less self-reported pain and better function after surgery than patients who had a total knee replacement. All three methods found that partial knee replacement was safer, was associated with a lower risk of blood clots (a known complication of knee surgery) and had lower mortality over 5 years. However, patients who had a partial knee replacement were twice as likely as those with a total knee replacement to need a second surgery within 5 years.

We found that partial knee replacements were less costly to the NHS and were associated with better overall quality of life for patients than total knee replacement.

Scientific summary

Background

Routinely collected NHS clinical data and national registries offer new opportunities for the comparative assessment of health technologies in actual practice conditions. This is particularly interesting for elderly and complex patients with multiple comorbidities, who are excluded from many randomised controlled trials. Surgical randomised controlled trials are particularly challenging owing to ethics difficulties, scarce surgeon equipoise and the need for specialised and experienced treatment centres and teams.

Two procedures for knee arthroplasty that are offered in the NHS (unicompartmental and total knee replacement) were compared in a National Institute for Health Research Health Technology Assessment programme-funded surgical randomised controlled trial [08/14/08; Total or Partial Knee Arthroplasty Trial (TOPKAT)]. Although TOPKAT offered top-quality information on the comparative effects of these surgeries for relatively healthy (American Society of Anesthesiologists grade of 1 or 2) patients, data from the National Joint Registry suggest that almost one in six patients undergoing unicompartmental or total knee replacement surgery in the UK have an American Society of Anesthesiologists grade of ≥ 3. The TOPKAT findings are, thus, hard to interpret for a substantial proportion of NHS patients.

Routinely collected data contain information on, potentially, all NHS patients, regardless of their medical history. These data sets offer an opportunity for research that includes elderly and multimorbid participants. However, the lack of random allocation of treatments in such databases does pose challenges, including confounding by indication. If confounding is not accounted for and minimised, it can lead to bias.

Objectives

In stage 1 of the Unicompartmental (vs. Total) knee replacement for patients with Multimorbidity Study (UTMoSt), we assessed whether or not the available analytical methods could obtain comparable findings to those from TOPKAT, using participants in the National Joint Registry who would have been eligible for TOPKAT (American Society of Anesthesiologists grade of 1 or 2). The proposed propensity score and instrumental variable methods were each applied to the data set. Those offering results comparable to TOPKAT were deemed valid and were used in stage 2.

In stage 2 of UTMoSt, the validated methods from stage 1 were used to compare the benefits (postoperative patient-reported outcome measures), risks (revision, complications and mortality), hospital costs and cost-effectiveness of unicompartmental and total knee replacement among National Joint Registry participants who would not have been eligible for TOPKAT (American Society of Anesthesiologists grade of ≥ 3).

Methods

For data sources, National Joint Registry participants undergoing total or unicompartmental knee replacement with linked, routinely collected data from the NHS hospital inpatient records were included in safety analyses. Those with linked patient-reported outcome measure data were included in the primary outcome analyses.

The participants in stage 1 were total and unicompartmental knee replacement recipients recorded in the National Joint Registry with linked data who would have been eligible for TOPKAT. In stage 2, participants were recruited who had an American Society of Anesthesiologists grade of 3 or 4, indicating severe systemic comorbidities that would have excluded them from TOPKAT. The comparison was unicompartmental versus total knee replacement.

The primary outcome was postoperative Oxford Knee Score (patient-reported outcome measure). The secondary outcomes were safety outcomes, including 90-day risks of venous thromboembolism, myocardial infarction and prosthetic joint infection (stage 2 only), and 5-year risks of revision and mortality. The health economic analysis outcomes were health-related quality of life (EuroQol-5 Dimensions) and NHS hospital costs (stage 2 only).

Statistics

In stage 1, four propensity score-based approaches and inverse probability weighting were used to account for measured confounding: (1) propensity score matching (1 : 5), (2) stratification based on the distribution of the propensity score in the whole cohort, (3) stratification based on the unicompartmental knee replacement cohort and (4) propensity score adjustment (linear and non-linear models). For each outcome, a logistic regression model was used to calculate the propensity score for unicompartmental knee replacement using patient-level characteristics, including demographics, preoperative patient-reported outcome measures, comorbidities and procedures recorded within the 3 years before surgery. Missing body mass index data and preoperative patient-reported outcome measures were imputed using multiple imputation by chained equations. Covariate balance was assessed using absolute standardised mean difference, with a predefined cut-off point of 0.1.

We also explored four potential instrumental variables: surgeon preference, hospital preference, geographical location and calendar time. When certain assumptions are fulfilled, instrumental variable analyses can account for measured and unmeasured confounders. Key instrumental variable assumptions were checked with *F*-statistics, odds ratios (strength of the instrument) and absolute standardised mean differences (lack of an association between the instrument and the confounders).

We compared the results obtained for each method with the TOPKAT findings using the TOPKAT outcome analysis methods: multilevel linear regression for postoperative Oxford Knee Score and a multilevel Poisson model for 5-year revision or death. Two-stage analyses were used for instrumental variables. We predefined three criteria by which an analytical method would be considered unable to replicate the TOPKAT findings and, therefore, be invalid for stage 2: chi-squared test p-value < 0.05, a relatively large tau^2 or an $I^2 > 40\%$. We also used two newly proposed methods to assess the methods' validity: whether or not the obtained treatment effect estimates fall within the trial's 95% confidence interval and statistical significance agreement. We performed sensitivity analyses on the valid methods, including restricting the analysis to surgeries performed by lead surgeons with ≥ 10, ≥ 30 and ≥ 50 index surgeries in the previous year.

In stage 2, for each valid method and each outcome, patient-level characteristics overall and for unicompartmental knee replacement patients were compared using absolute standardised mean difference with a cut-off point of 0.1. Differences in postoperative Oxford Knee Score between unicompartmental knee replacement patients and total knee replacement patients were estimated using multilevel linear regression. For each of the 90-day postoperative complications, the relative risk and 95% confidence interval were estimated using Poisson models with robust standard errors. Cause-specific hazard models were fitted to estimate the risk of 5-year revision or mortality, censoring patients when they had revision or mortality (a competing event). Prespecified interactions between surgery types and sex, age or American Society of Anesthesiologists grade were assessed with a

p-value of < 0.1. Long-term complications were also assessed when restricting the analysis to patients operated on by experienced surgeons.

For the health economic evaluation, multilevel regression analyses were performed to estimate the differences in costs and quality-adjusted life-years between unicompartmental knee replacement and total knee replacement patients. The regression models for quality-adjusted life-years also included the preoperative utility score as a covariate. The incremental cost-effectiveness ratio was calculated by dividing the difference in costs by the difference in quality-adjusted life-years between unicompartmental knee replacement and total knee replacement patients. The uncertainty surrounding the incremental cost-effectiveness ratio was estimated using non-parametric bootstrapping with 1000 replications.

Results

In stage 1, 21,026 National Joint Registry participants undergoing unicompartmental knee replacement and 273,530 participants undergoing total knee replacement would have been eligible for TOPKAT. Of these participants, 1197 unicompartmental knee replacement and 125,834 total knee replacement patients had postoperative Oxford Knee Score data and could be included in the Oxford Knee Score analysis.

In the Oxford Knee Score analysis, inverse probability weighting and propensity score stratification based on the whole cohort resulted in unresolved imbalances, whereas propensity score matching and propensity score stratification based on the unicompartmental knee replacement cohort achieved good balance. All of the propensity score-based methods resulted in an average treatment effect estimate favouring unicompartmental knee replacement, but with at least 1 point less than the effect seen in the trial, ranging from 0.10 (propensity score non-linear adjustment) to 0.76 (propensity score stratification based on the unicompartmental knee replacement cohort), compared with 1.91 in TOPKAT.

Propensity score stratification based on the unicompartmental knee replacement cohort was the preferred method ($I^2 = 35\%$, chi-squared test $p = 0.21$ and $\tau^2 = 0.23$), followed by inverse probability weighting ($I^2 = 48\%$, chi-squared test $p = 0.17$ and $\tau^2 = 0.43$) and propensity score stratification based on the whole cohort ($I^2 = 53\%$, chi-squared test $p = 0.14$ and $\tau^2 = 0.48$).

A surgeon-level eligibility criterion was then applied to mimic surgeon eligibility in TOPKAT, including only participants operated on by surgeons who had performed ≥ 10 surgeries of the same type in the previous year. The treatment estimates from all three methods moved closer to the TOPKAT findings, with average treatment effects of 1.37 (95% confidence interval 0.54 to 2.20) for propensity score stratification based on the unicompartmental knee replacement cohort, 1.37 (95% confidence interval 0.54 to 2.20) for propensity score stratification based on the whole cohort and 1.32 (95% confidence interval 0.32 to 2.33) for inverse probability weighting, compared with 1.91 (95% confidence interval 0.20 to 3.62) in TOPKAT. All three methods had an I^2 of 0% and small τ^2, indicating that they were able to replicate TOPKAT findings.

Only five of the potential instrumental variables passed both testable assumptions: the three lead surgeon-based preference instruments (based on 20, 30 and 50 previous surgeries) and two of the consultant surgeon-based preference instruments (based on 30 and 50 previous surgeries). The other tested instrumental variables violated either one or both of the testable assumptions. The five selected instrumental variables then all failed to produce a comparable treatment effect estimate with TOPKAT, with a chi-squared test *p*-value < 0.001 and $I^2 > 90\%$. All of the instrumental variable analyses passed the statistical significance agreement tests and showed a significant improvement in postoperative Oxford Knee Score favouring unicompartmental knee replacement, as in TOPKAT.

In stage 2, the comparative safety analyses included 57,682 total knee replacement patients and 2256 unicompartmental knee replacement patients. Of these patients, only 145 unicompartmental knee replacement and 23,344 total knee replacement patients were included in the Oxford Knee Score analysis. Propensity score stratification based on the unicompartmental knee replacement cohort yielded excellent covariate balance both between and within strata. Propensity score stratification based on the whole cohort had excellent average covariate balance between the 10 strata. Four covariates remained imbalanced after inverse probability weighting. Propensity score stratification based on the unicompartmental knee replacement cohort and on the whole cohort resulted in statistically significant positive effects for unicompartmental knee replacement, with an estimated mean postoperative Oxford Knee Score difference of 1.83 (95% confidence interval 0.10 to 3.56) points and 1.82 (95% confidence interval 0.10 to 3.56) points in favour of unicompartmental knee replacement, respectively, which is close to the effect seen in TOPKAT. Inverse probability weighting analysis found an insignificant effect in postoperative Oxford Knee Score.

Unicompartmental knee replacement patients had a lower relative risk of developing venous thromboembolism in the 90 days after surgery than total knee replacement patients, with relative risks of 0.33 (95% confidence interval 0.15 to 0.74) based on propensity score stratification and 0.39 (95% confidence interval 0.16 to 0.96) based on inverse probability weighting. No significant differences in myocardial infarction or prosthetic joint infection risks were found between unicompartmental knee replacement and total knee replacement patients. Unicompartmental knee replacement patients experienced a higher risk of revision over 5 years than total knee replacement patients, with hazard ratios of 2.70 (95% confidence interval 2.15 to 3.38) in propensity score stratification analyses and 2.60 (95% confidence interval 1.94 to 3.97) in inverse probability weighting. They also had reduced all-cause mortality in propensity score stratification analyses, with a hazard ratio of 0.52 (95% confidence interval 0.36 to 0.74). However, this difference was attenuated when using inverse probability weighting.

American Society of Anesthesiologists grade and sex had significant interactions with total knee replacement and unicompartmental knee replacement: women had a higher risk of revision than men, and people with an American Society of Anesthesiologists grade of 4 had a much higher revision risk than patients with an American Society of Anesthesiologists grade of 3, although statistical power was a concern.

The crude mean cost of a primary knee replacement was £6246 (standard deviation £779) for unicompartmental knee replacement patients and £6627 (standard deviation £1402) for total knee replacement patients. The mean costs for complications were £3560 (standard deviation £6) for unicompartmental knee replacement patients and £3986 (standard deviation £3853) for total knee replacement patients. The mean differences in quality-adjusted life-years gained were 0.147 (95% confidence interval −0.507 to 0.803) and 0.330 (95% confidence interval −0.305 to 0.967) in favour of unicompartmental knee replacement when using inverse probability weighting and propensity score stratification, respectively. Unicompartmental knee replacement costs were £334 (95% confidence interval £306 to £362) and £359 (95% confidence interval £339 to £378) lower than total knee replacement costs, using inverse probability weighting and propensity score stratification, respectively.

Conclusions

Propensity score-based stratification and inverse probability weighting successfully replicated the TOPKAT findings in the primary outcome (postoperative Oxford Knee Score) analyses, indicating that these methods can be used to minimise confounding in observational studies on the comparative effectiveness of implantable medical devices. Propensity score adjustment, propensity score matching and instrumental variable methods led to results that departed from those observed in TOPKAT. More research is required on the best use of analytical methods and design of observational post-marketing research of medical devices.

In stage 2, unicompartmental knee replacement had similar effectiveness for patients with multimorbidity as for the healthier (stage 1 and TOPKAT) population. There was little or no clinically relevant difference in postoperative Oxford Knee Score between unicompartmental knee replacement and total knee replacement patients. A strongly protective effect against postoperative venous thromboembolism for patients undergoing unicompartmental knee replacement was identified. In the long term, unicompartmental knee replacement was associated with an almost threefold higher revision risk than total knee replacement, but also with a reduction in all-cause mortality of almost 50%. Cost-effectiveness analyses showed that unicompartmental knee replacement dominated in patients with substantial comorbidity (American Society of Anesthesiologists grade of ≥ 3), as it was both more beneficial and less expensive than the alternative (total knee replacement) in this patient subgroup. These findings should guide future clinical guidelines on knee replacement for patients with severe multimorbidity.

Trial registration

This trial is registered as EUPAS17435.

Funding

This project was funded by the National Institute for Health Research (NIHR) Health Technology Assessment programme and will be published in full in *Health Technology Assessment*; Vol. 25, No. 66. See the NIHR Journals Library website for further project information.

Chapter 1 Introduction

Background

Surgical randomised controlled trials (RCTs) generate gold standard evidence on the causal effects of surgery. Recent evidence suggests that they are both safe and useful for informing clinical practice in surgical specialties.[1] However, such studies remain uncommon owing to, for example, resource intensity, time required from design to completion, ethics considerations, need for surgeon equipoise and other feasibility issues.[2,3]

Non-randomised studies that rely on routinely collected data offer an efficient alternative for the comparative assessment of established surgical interventions and/or implantable medical devices available in the NHS. When conducted well, so-called 'real-world' evidence studies offer results that are potentially generalisable to the whole population of NHS patients, regardless of comorbidities, socioeconomic status, sex or age, including patients who would have been excluded from RCTs. However, observational studies are limited by confounding indication and related channelling bias owing to non-random allocation of treatment alternatives. Although analytical and study design methods have been used in drug safety studies to minimise confounding, there are few data on their performance in comparative effectiveness and safety research of surgery and medical devices.

Large-scale analyses of medicines are used to inform regulatory and clinical decision-making, and both the US Food and Drug Administration (FDA) and the European Medicines Agency[1] have recently published guidelines on the use of routinely collected data for regulatory purposes. Collaborations such as the Observational Health Data Sciences and Informatics (OHDSI) [www.ohdsi.org (accessed 20 December 2019)] and the European Health Data and Evidence Network (EHDEN) [www.ehden.eu (accessed 20 December 2019)] are accelerating the creation of multinational networks and tools for curating and analysing real-world data at scale. The combined existence of data and best practices for analyses are leading to high-impact publications that will influence clinical guidelines by replacing 'expert opinion' for which RCT evidence is lacking.[2]

Well-designed randomisation in clinical trials eliminates systematic bias. In surgical RCTs evaluating implantable devices or alternative surgical procedures, randomisation can account for patient characteristics and surgeon characteristics and expertise. High levels of adherence in RCTs reduce performance bias and attrition bias. Well-designed, well-conducted RCTs, thus, have excellent internal validity.

Randomised controlled trials give detailed evidence about the potential effects of new interventions under ideal circumstances, and are considered the gold standard for casual inference and health technology assessment. However, the main criticism of RCTs is that their rigid eligibility criteria can mean that trial participants are not representative of the full target population. The more restrictive the trial, the more limited the trial's external validity.

An example of ongoing debate in surgery is the choice of total knee replacement (TKR) or unicompartmental knee replacement (UKR), also known as partial knee replacement, for severe knee osteoarthritis. In response, the National Institute for Health Research (NIHR) funded a surgical RCT called TOPKAT (Total or Partial Knee Arthroplasty Trial) [Health Technology Assessment (HTA) 08/14/08].[4] This recently concluded multicentre RCT successfully recruited, randomised and followed up participants for 5 years. The trial results were reported in *The Lancet*[5] and the full report has now been published in the NIHR HTA journal.[4] In brief, TOPKAT demonstrated that UKR had a small benefit in a patient-reported outcome over TKR of < 2 points in the Oxford Knee Score (OKS) in the short term (1 year), but no difference in the longer term (5 years). UKR was more cost-effective than TKR over the 5 years of follow-up.

TOPKAT was a relatively pragmatic trial that excluded only patients with an unusually high American Society of Anesthesiologists (ASA) grade of ≥ 3, owing to severe comorbidity and potentially limited lifespan. The National Joint Registry (NJR) report suggests that only about 17% of people receiving knee replacement surgery have an ASA grade of ≥ 3[3] and, therefore, would have been ineligible for TOPKAT. TOPKAT has, alongside some observational studies, been identified by *NIHR Signals* as potentially relevant for informing future National Institute for Health and Care Excellence (NICE) guidelines and NHS practices.[6]

There is an opportunity to complement the results from TOPKAT with good-quality data on the performance of these two surgical approaches for multimorbid patients requiring knee surgery, which TOPKAT cannot provide. Observational data from the NJR can potentially provide insights into the impact of different types of knee replacement for all NHS patients. A recent *Lancet* paper used one of the most widely extended methods [propensity score (PS)] to minimise bias.[7–10] The authors acknowledged that unmeasured confounders (such as unrecorded conditions, disease severity or drug use) could at least partially explain the study findings, as PS can account only for measured confounders. Such unresolved bias can sometimes be minimised with alternative pharmacoepidemiological analytical methods, such as instrumental variables (IVs)[11] or high-dimensional PSs.[12]

The FDA and colleagues from a number of academic institutions are replicating previous drug RCTs using observational methods to demonstrate their usefulness for drug and vaccine safety and comparative effectiveness research.[6,13] However, to our knowledge, these methods have not yet been used to replicate the results of surgical or implantable device RCTs. There is a need for a better understanding of the performance of these methods in comparative effectiveness and safety studies to evaluate surgical and implantable medical device innovations using routinely collected data. The existence of a national, multicentre surgical RCT comparing two common surgical techniques in TOPKAT, and the availability of good-quality national data on these treatments and the primary study outcome from the NJR, offers a unique opportunity to study the validity of analytical methods for researching surgical and medical device innovations using observational data.

Evidence explaining why this research is needed now

The recent multicentre RCT TOPKAT provided high-quality evidence on the clinical effectiveness and cost-effectiveness of UKR compared with TKR for medial compartmental knee osteoarthritis. However, the results might not be generalisable to patients with an ASA grade of ≥ 3, equivalent to severe or very severe systemic disease, as they were ineligible for participation in the trial.[3] Recent NJR documents have reported[3] that differences in patient-reported outcome measures (PROMs) exist according to ASA grade, and that there are known associations between comorbidities and postoperative complications and mortality.[14]

Patients with an ASA grade of ≥ 3 currently represent only 17% of those undergoing knee replacement surgery. However, this proportion will probably increase as our population ages, and will probably account for a high proportion of the NHS expenditure on knee replacement surgery and related hospital admissions, given their baseline medical history and risk factors. The difficulties in recruiting older people and patients with severe and/or multiple comorbidities for surgical RCTs are well known. Alternative solutions are needed to generate evidence for this group of people.[3] This study follows previously published NIHR themed calls on evaluating interventions and services for older people with multimorbidity or complex health needs.

TOPKAT's finding that UKR was not associated with an excess risk of revision contradicted all previous observational research, including previous publications using NJR data,[9,10] a *BMJ* meta-analysis[15] and a recent multinational analysis led by OHDSI and EHDEN collaborators.[16] This discrepancy could have been driven by residual confounding in the observational studies or heterogeneity in the population of

patients, surgeons or hospitals in the RCT and population-based cohort analyses. Evidence is required on the mechanisms behind this difference in results to inform future related research and health-care delivery in the NHS.

A new European regulation for medical devices will be implemented in May 2021, which will require a more comprehensive evaluation of implantable devices, including orthopaedic prostheses.[17] There is, therefore, an urgent need for methodological guidelines on using real-world data to inform the post-marketing use, effectiveness and safety of medical devices.

Research aims and objectives

We undertook the Unicompartmental (vs. Total) knee replacement for patients with Multimorbidity Study (UTMoSt) in two stages:

- Stage 1 – we studied the validity of different methods previously used in drug and vaccine studies to minimise confounding for assessing the comparative effectiveness of alternative surgical procedures and implantable devices. We used knee replacement (UKR and TKR) in patients with an ASA grade of 1 or 2 (eligible for TOPKAT) as an example, and TOPKAT as a gold standard for comparison. Methods that gave results comparable to TOPKAT were deemed valid and were used in stage 2.
- Stage 2 – we used the methods that were able to replicate the RCT findings in stage 1 to compare the benefits (OKS), risks (revision surgery, complications and mortality), hospital costs and cost-effectiveness of UKR versus TKR among NJR participants with multiple and severe comorbidities (ASA grade of ≥ 3).

Structure of this report

Chapter 2 describes the data sources, defines the exposures (UKR and TKR) and potential confounders, and summarises the statistical methods used to minimise confounding.

Chapters 3–5 report the main findings of stage 1. *Chapter 3* reports results from stage 1 based on PS analyses. *Chapter 4* reports results from stage 1 based on IVs. Stage 1 conclusions and implications for stage 2 are discussed in *Chapter 5*.

Chapters 6–8 report stage 2. *Chapter 6* describes the population, *Chapter 7* reports the results of the safety-effectiveness analyses and *Chapter 8* reports the health economics analysis.

We synthesise and discuss the results, study strengths and limitations, future research and implications of the study findings in *Chapter 9*.

Chapter 2 Data sources and analytical methods

Data sources

National Joint Registry

The NJR for England, Wales, Northern Ireland and the Isle of Man collects information on hip, knee, ankle, elbow and shoulder joint replacement surgery carried out in both NHS and private hospitals, and monitors the performance of joint replacement implants in primary and revision operations. The NJR started collecting data for knee and hip replacements in April 2003. The collection of these data became mandatory in April 2011. Although compliance was initially low, at 43% in 2004, it had risen to 95% by 2015.

Based on the NJR's 15th annual report,[18] it contains over 2 million procedure records, including over 1 million verifiable primary knee replacements recorded up to 31 December 2017, with a maximum follow-up of over 14 years. The most common reason for knee replacement is osteoarthritis. In total, 85% of primary knee surgeries are a TKR. Revision operations data are matched to the primary operation data in NJR using unique patient identifiers. There are 33,292 first revisions linked to the NJR primary knee operation.[19] The NJR also links to the Office for National Statistics for mortality information. TKR and UKR are reported to have a 90-day cumulative mortality rate of 0.31 [95% confidence interval (CI) 0.30 to 0.33] and 0.08 (95% CI 0.06 to 0.10), respectively.[19]

A data request for NJR knee replacements taking place until the end of 2016 was granted in October 2017 (NJR Application Reference RSC2016/13).

Hospital Episode Statistics

The NJR is linked to the Hospital Episode Statistics (HES), which covers NHS hospitals and independent sectors that provide NHS services in England. At the time of writing, HES included admitted patient care data from 1997, outpatient data from 2003, accident and emergency data from 2007 and diagnostic imaging data from 2012. All data recorded in HES are submitted by contributing hospitals for reimbursement purposes. HES is an administrative data set, but inpatient records have been used extensively for research purposes, including for previous NIHR-funded work,[12] resulting in high-impact publications.[15] Previous reports have demonstrated that HES records musculoskeletal procedures and outcomes accurately and completely, when compared with primary care electronic medical records.[14]

For each hospital admission recorded in HES, information is available on hospital diagnoses and procedures, administrative details (e.g. date of admission and discharge) and basic sociodemographic data (e.g. the region and ethnic background). Diagnoses are coded using the *International Classification of Diseases*, Tenth Revision (ICD-10).[20] The procedure undertaken is coded using the Office of Population Censuses and Surveys Version 4 (OPCS-4) codes.[21]

Patient-reported outcome measures database

In 2009, NHS England introduced the routine collection of PROMs associated with a short list of elective surgeries, including knee replacement.[22,23] The PROMs database contains patients' perspectives of their knee and hip operations, which are collected with self-completed questionnaires before surgery and 6 months post surgery. The initial focus was on four procedures: hip replacement, knee replacement, hernia repair and varicose vein surgery. The postoperative questionnaires were sent to patients by post 6 months after surgery and were returned by post.

For knee replacement, the PROMs database uses the OKS to measure patients' perspective of their knee pain and function. The OKS has 12 questions, each with five possible responses, and results in a score ranging from 0 to 48.[24,25] The PROMs database also collects quality-of-life data [EuroQol-5 Dimensions, three-level version (EQ-5D-3L)] for all four surgeries. EQ-5D-3L is a quality-of-life measure that is made up of the EuroQol-5 Dimensions (EQ-5D) index and a health visual analogue scale. The EQ-5D index contains five questions for each of the five subscales: mobility, self-care, daily activities, pain or discomfort, and depression or anxiety. Each question has three possible responses.[26] The raw score has been weighted in accordance with UK preferences to represent the whole of UK society, resulting in a score (EQ-5D utility index) ranging from –0.59 (worst state) to 1.00 (best state). The EuroQol visual analogue scale is a patient self-assessment of their health in general, with a score from 0 (worst imaginable) to 100 (best imaginable).[22]

Data linkage

The PROMs data were matched to HES by NHS Digital (DARS-NIC-172121-G0Z1H-v0.11) using the probabilistic linkage methods. A rank score was created based on patient-identifiable fields, provider codes and operation codes and dates, for which the highest score of 1 was given if identical information in patient-identifiable fields was recorded in the databases.[27] NJR patients were matched to HES/PROMs in a deterministic fashion that required the same information in patient-identifiable fields. The data linkage was approved and conducted by NHS Digital (DARS-NIC-172121-G0Z1H-v0.11). *Figure 1* shows a flow chart of data sources used in this study.

Methods

Target population

The target population in stage 1 was NJR patients who fulfilled the eligibility criteria for TOPKAT[4] and had a record of TKR/UKR in the primary procedure field of the NJR from 2009 to 2016.

TOPKAT eligibility criteria were applied as closely as possible based on the information recorded in the linked data set described in *Data sources*.[28] *Table 1* shows our operationalisation of the TOPKAT eligibility criteria based on HES inpatient data within 3 years before the operation date in the NJR.

The operation date recorded in the NJR was considered the index date. For NJR patients with two primary knee replacements, one on each side, only information related to the earliest operation was used and the index date was the operation date for the first knee replacement.

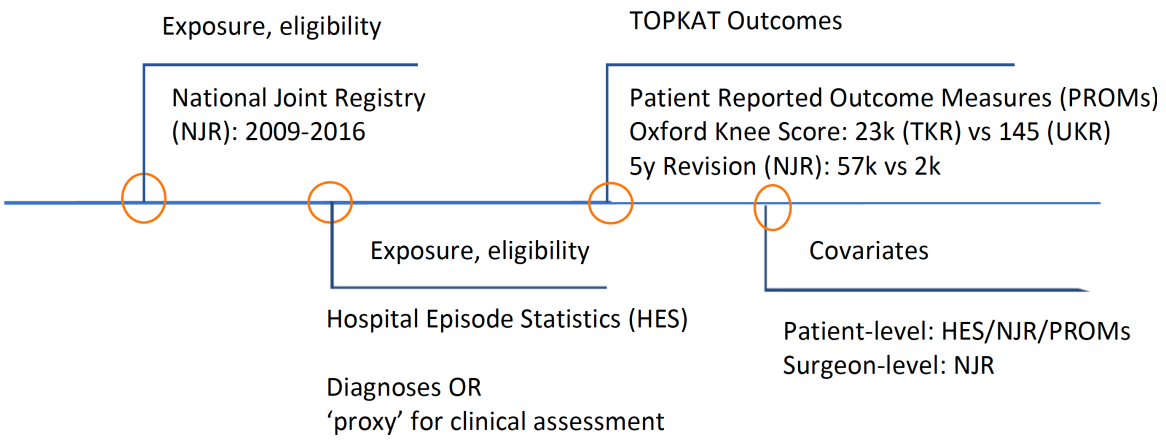


FIGURE 1 Data source flow chart.

TABLE 1 Patients' eligibility criteria used in TOPKAT and this study (UTMoSt)

TOPKAT criteria	UTMoSt general criteria	UTMoSt stage 1 criteria
TOPKAT surgery type		
TKR or UKR	Only TKR/UKR recorded in the surgery type were included	
Trial participants		
Patients participated in the trial once. They were not randomised twice if they had a knee replacement on the other knee after they had become a trial participant	Only the first record of TKR/UKR was included if there were multiple knee replacement surgeries in the NJR	
–	Patients received their surgery before 31 December 2016 to allow their postoperative OKS to be collected	
Consented to trial participation	Patients who had opted out from the use of their data for research were excluded	
–	Patients without IMD data were excluded	
	Patients without their postoperative OKS collected were excluded from the OKS cohort (primary analysis)	
Inclusion criteria		
Medial compartment osteoarthritis with exposed bone on both femur and tibia	Data unavailable as clinical assessment was not recorded in the NJR	
Functionally intact anterior cruciate ligament	Patients with a record of previous cruciate ligament injury (see *Table 31*) in HES were excluded	
Full-thickness and good-quality lateral cartilage present	Data unavailable as clinical assessment was not recorded in the NJR	
Correctable intra-articular varus deformity	Data unavailable as clinical assessment was not recorded in the NJR	
Medically fit showing an ASA grade of 1 or 2		Patients with ASA grade of 1 or 2 in NJR were included
Clinical exclusion criteria		
Require revision knee replacement surgery		Not applicable as only primary TKR/UKR procedures were included
Have rheumatoid arthritis or other inflammatory disorders		Patients with a record of rheumatoid arthritis or other inflammatory disorders (see *Table 32*) were excluded
Are unlikely to be able to perform required clinical assessment tasks		Clinical assessment was not recorded in NJR
Have symptomatic foot, hip or spinal pathology		Patients with a record of foot, hip or spinal pain (see *Tables 33* and *34*) in the 1 year before surgery were excluded
Previous knee surgery other than diagnostic arthroscopy and medial meniscectomy		Patients with a record of prior knee surgery (see *Table 35*) were excluded
Previously had septic arthritis		Patients with a record of septic arthritis (see *Table 36*) were excluded
Have significant damage to the patellofemoral joint especially on the lateral facet		Patients with a record of patellofemoral damage (see *Table 37*) were excluded

IMD, Index of Multiple Deprivation.

Patients who did not have Index of Multiple Deprivation (IMD) data were excluded. It was impossible to impute IMD data because predictors of missing IMD are likely to be unknown in the HES, NJR and PROMs databases.

An opt-out rule was applied and an updated list of patients was obtained from NHS Digital in October 2019. People who were not on this list of patient identifiers were excluded from the final analytical data set. The code lists used can be found in *Appendix 2*.

Methods to minimise confounding

Randomised controlled trials provide gold standard evidence to evaluate the casual effect of an intervention/treatment. Well-implemented randomisation within a RCT ensures exchangeability; the probability of being exposed or not being exposed to a given treatment is independent of the patient characteristics and, therefore, not conditional on measured and unmeasured confounders.[29,30] The casual effect can, therefore, be estimated in RCT data as equivalent to the differences in outcome risk or probability between trial participants who are assigned to each intervention group, without further adjustment.

However, randomisation is not always feasible owing to time, costs and ethical and practical constraints. Unfortunately, RCTs are not the default for testing medical devices and surgical interventions.[31] There is a growing interest in observational studies to evaluate the casual effect of medical and surgical interventions, after minimising confounding.

As stated in *Chapter 1*, *Research aims and objectives*, the aim of stage 1 was to prospectively emulate a surgical target trial to evaluate the risk–benefit of UKR versus TKR using real-world data. We analysed the same primary outcome that was used in TOPKAT: patient-reported postoperative OKS. Secondary outcomes included those from TOPKAT (e.g. 5-year revision risk) and safety events (postoperative complications and mortality). Methods that replicated the primary outcome results from TOPKAT were considered to sufficiently minimise confounding and were taken forward to stage 2 of the project.

We tested the following methods:

- propensity score methods –
    - PS matching
    - PS stratification
    - PS adjustment
    - inverse probability (of treatment) weighting.
- instrumental variable method, considering as the IV –
    - surgeon preference
    - hospital preference
    - geographical location
    - calendar time.

Although initially proposed in our grant application, we did not test high-dimensional PSs. Covariate prioritisation and selection algorithms in high-dimensional PSs have been developed based on binary covariates and a binary outcome. Unfortunately, high-dimensional PSs have not been implemented and/or validated for continuous outcome data and could, therefore, not be used to analyse the OKS. Categorising the OKS was considered as a potential solution, but would have resulted in an unacceptable loss of statistical power. The results would also not have been comparable to those from TOPKAT or the other tested methods.[32]

Propensity score methods

Propensity score methods have been widely used in observational studies to summarise measured covariate information and minimise systematic differences between exposed and unexposed participants when randomisation is not possible.[33] PSs can be used to adjust such differences via study design or when estimating an exposure's casual effect.

A PS is defined as the probability that a participant receives a treatment (UKR in our example) based on their characteristics. Multivariable logistic regression equations are often used to estimate PSs using all potential confounders available in a data set. By definition, PS methods can account only for observed confounders, although some have speculated that proxies might be available in the data when granular information on specific confounders is missing.

We considered 18 patient-level baseline sociodemographic and clinical characteristics from the HES/PROMs/NJR linked data set for inclusion in the PS equation (*Table 2*). The limitations of the proposed methods for estimating PSs based on the available data are discussed in *Chapter 5*. Missing data on body mass index (BMI) and preoperative PROMs characteristics, including EQ-5D, general health and OKS, were imputed using multiple imputation by chained equations with an assumption of missing at random and 10 imputed data sets. In each imputed data set, multiple logistic regression equations were used to calculate one PS.

Propensity score matching

Once generated using the algorithm described in *Propensity score methods*, every exposed (UKR) patient was PS matched with up to five unexposed (TKR) patients without replacement on a calliper width of 0.2 standard deviations (SDs) of the logit of the PS.[34] The UKR (exposed) and TKR (unexposed) matched participants were, on average, comparable regarding the available confounders. This concept is analogous to a RCT, in which participants in the treatment and control/placebo arms have similar characteristics but the groups might still be unbalanced with respect to unmeasured characteristics with PS methods.

This matching method has been shown to efficiently minimise confounding by indication in pharmaco-epidemiological (drug safety and comparative effectiveness) studies.[35,36] PS matching using calliper widths excludes the small proportion of patients with an extremely high or extremely low probability of treatment who are not present in both groups. As a result, the obtained treatment estimate in such analyses is represented by the average treatment effect in the exposed group, also called the average treatment effect on treated (ATT).[37] The method assumes that the PS-matched exposed cohort are similar to the unexposed cohort in all respects except for the treatment received. Any differences in outcome between the PS-matched exposed and the unexposed patients can be interpreted as the effect of the treatment.

Covariate balance was assessed using absolute standardised mean differences (ASMDs), with a cut-off point of 0.1[34] in each imputed data set. An ASMD of > 0.1 indicated that the covariate was distributed differently in the UKR and TKR groups. These unbalanced covariates were included as covariates in the outcome analyses and, therefore, double adjusted for.

Propensity score stratification

In PS stratification, all participants in the data set were ranked according to their estimated PSs and were stratified into equal subgroups (i.e. deciles) called strata. Within each stratum, exposed and unexposed patients had roughly similar PSs, implying that the groups had similar distributions of available confounders. The mean PS stratification is commonly used to estimate a treatment's effect in the target population, otherwise called the average treatment effect (ATE). Treatment effects were first calculated separately in each stratum, and were then averaged with a weight of the proportion of all participants within that stratum.[5] The standard error of the pooled treatment effect was estimated using the jack-knife approach.[38,39]

TABLE 2 A description of patient-level covariates included in the PS models

Covariate	Data source	Description
Sociodemographic and clinical factors		
Age	NJR	Age at operation
Sex	NJR	Sex
Rural/urban	HES	The official statistic classifying locations as rural or urban areas: urban, town and fringe, village or isolated
IMD	HES	Index of Multiple Deprivation. Patients' deprivation status in percentile
BMI	NJR	Calculated from height and weight
PROMs preoperative OKS	PROMs	Self-reported preoperative OKS score, ranging from 0 to 44
PROMs EQ-5D	PROMs	Self-reported preoperative EQ-5D visual analogue scale, ranging from 0 to 100
PROMs general health	PROMs	Self-reported preoperative general health, ranging from 0 (excellent) to 5 (poor)
Charlson Comorbidity Index	HES	The Charlson Comorbidity Index score recorded in HES (the code list is shown in *Tables 38–54*): 0, 1, 2, 3 and 4
Gastrointestinal disease	HES	An ICD-10 code starting with 'K2', 'K3', 'K4', 'K5', 'K6', 'K7', 'K8' or 'K9' (gastrointestinal disease) recorded in HES in the 3 years before the operation
Osteoarthritis and other joint problems	HES	An ICD-10 code for other joint problems in HES in the 3 years before the operation (code list is shown in *Table 55*)
Mental health	HES	An ICD-10 code starting with 'H' (mental health) in HES in the 3 years before the operation
Respiratory disease	HES	An ICD-10 code starting with 'J4', 'J5', 'J6', 'J7', 'J8' or 'J9' (respiratory disease) in HES in the 3 years before the operation
Cardiovascular disease	HES	An ICD-10 code starting with 'I' (cardiovascular disease) in HES in the 3 years before the operation
Thyroid problems	HES	An ICD-10 code starting with 'E0' (thyroid problems) in HES in the 3 years before the operation
Foot, hip and spinal pain	HES	An ICD-10 code for foot, hip or spinal pain problems in HES in the 3 years before the operation (code list shown in *Table 33*)
Coxarthrosis	HES	An ICD-10 code starting with 'M16' (hip osteoarthritis) in HES in the 3 years before the operation
Neurological disorders	HES	An ICD-10 code starting with 'G1', 'G2', 'G3', 'G4', 'G5', 'G6', 'G7', 'G8' or 'G9' (neurological disorders) in HES in the 3 years before the operation
Other arthrosis	HES	An ICD-10 code starting with 'M19' (other arthrosis) in HES in the 3 years before the operation
Polyarthrosis	HES	An ICD-10 code starting with 'M15' (polyarthrosis) in HES in the 3 years before the operation
Spondylosis	HES	An ICD-10 code starting with 'M47' (spondylosis) in HES in the 3 years before the operation

BMI, body mass index.

Some evidence suggested that stratification into more PS strata results in further bias reductions. However, when exposure is infrequent,[40] many strata mean that extreme strata are dominated by exposed or unexposed participants. Desai and colleagues[41] suggested basing the boundaries between strata on the PS distribution of the exposed group. This solution yielded better bias reduction than traditional PS stratification in their simulated examples. We used 10 strata based on the distribution of PSs in the whole data set (PS deciles) and the exposed group (UKR recipients). Both results were compared with the gold standard (TOPKAT estimates of OKS).

The key advantage of PS stratification over PS matching is that it does not exclude any participants, which preserves the sample size and improves precision.

The PS distributions of TKR and UKR in each stratum were compared to check whether or not they were roughly similar, which is an important assumption of this method. We assessed covariate balance using ASMD with a cut-off value of 0.1 per stratum per imputed data set. We report any covariate with a mean ASMD of > 0.1 across the 10 strata in any of the imputed data sets. The covariates with a mean ASMD of > 0.1 across strata were included in the outcome analyses for double covariate adjustment.

Propensity score adjustment

For PS adjustment, the estimated PS was included as a covariate to estimate the exposure's casual effect: the outcome variable was regressed on the exposure and created the PS.[33] This is probably the simplest PS approach. When PS is treated as a continuous variable in a regression analysis, the underlying model assumes a linear association between the PS and the outcome, and no interaction between the PS, exposure and study outcome.

We explored non-linear PS adjustments using fractional polynomial regression. The statistical significance of the interaction term was assessed using likelihood ratio tests between the model with and the model without the interaction term.[42] ATE can be estimated in the regression model without the interaction term. Both ATT and ATE can be estimated in the regression model with the interaction term.

Propensity score adjustment is less preferable than the other PS methods because it does not facilitate transparent reporting of covariate imbalance and its findings are difficult to interpret if there is a non-linear relationship between the PS and the outcome.[43]

Inverse probability weighting

Inverse probability weighting (IPW) creates a pseudo-population in which exposed and unexposed participants are assigned to weights equal to the inverse of the PS and the inverse of 1 minus the PS, respectively.[37,40,44] The weights are used like survey sampling weights in the estimation of treatment effects. ATE is the typical focus of IPW, similar to PS stratification and PS adjustment (without the interaction term).

One of the limitations of IPW is that rare/infrequent exposure leads to large weights that have an exaggerated influence on the obtained treatment effect estimates. To address this problem, we used the weight stabilisation method to create the weight.[40,44]

Covariate imbalance was evaluated using ASMD with a cut-off value of 0.1. Any covariate with an ASMD of > 0.1 in any of the 10 imputed data sets was included as a covariate in the outcome analyses.

Instrumental variable analyses

All PS-based methods are prone to residual confounding because they can account for measured confounders only.[8] By contrast, under certain assumptions, IV methods can account for both observed and unobserved variables.[36]

The IV methods rely on the existence of an 'instrument', an observed variable that is related to the exposure or treatment under study, that is independent of all known (and potentially unknown) confounders and is associated with the outcomes of interest through the treatment effects only. This situation resembles a RCT, in which treatment allocation typically almost perfectly coincides with the actual treatment received. In the case of a double-blinded RCT, treatment assignment affects only the outcome through the allocated treatment. IV methods are, thus, called pseudo-randomisation.

We constructed the following instruments and tested them against the underlying IV assumptions:

1. preference-based instruments – physician (here, the surgeon) preference for a treatment (here, UKR), surgical experience (with UKR) and hospital volume (of UKR)
2. geographical location
3. calendar time (i.e. date of surgery).

Construction of instrumental variables

To calculate surgeon preference for UKR, we sorted the whole NJR data set, regardless of eligibility for our study, in increasing order of operation dates and applied one of three approaches:

1. surgeon preference based on the last 20 consecutive procedures (UKR/TKR)
2. surgeon preference based on the last 30 consecutive procedures (UKR/TKR)
3. surgeon preference based on the last 50 consecutive procedures (UKR/TKR).

For each patient, we observed the surgeon's previous 20, 30 or 50 knee replacement surgeries and calculated their preference as the proportion of UKR. This proportion was used as an IV at the patient level to account for changes in preference over time.

Surgeon experience and hospital volume were estimated based on the number of knee replacement procedures undertaken by each of the surgeons or in each of the centres identified in the NJR in the previous year, and in total.

Patient region of residence and the proportion of UKR surgeries carried out in each region were used to construct the geographical location instrument. Regional instruments have previously been used to evaluate surgical techniques using observational data.[45]

Calendar time was constructed based on the recorded surgery date. We determined secular trends in UKR surgery in the NJR data, and established whether or not there was an inflexion point showing when UKR uptake increased. This method has been used in pharmacoepidemiology in situations where uptake of a medication changes after launch or when marketing or production of a drug or drug class stops.[46]

Instrumental variable assumptions and diagnostics

Instrumental variables rely on three strong assumptions:[47,48]

1. There is a strong association between the IV and the exposure of interest.
2. The IV must not have direct effects on the outcome, except through its association with the exposure.
3. The IV is independent of confounders.

The first assumption can be tested with the *F*-statistic value from the first-stage linear regression. The assumption is said to hold when the odds ratio is > 2.[49,50] The second and third assumptions are not verifiable or directly testable because they involve unobservable variables;[51] we used circumstantial evidence to support them. For the second assumption, we assumed that surgeon and hospital allocation, region of residence, and date of surgery were random and not associated with any potential confounders. We used a falsification test based on the standardised difference to

test for the third assumption. If the IV was associated with measured confounders, then it might also have been associated with unmeasured confounders. A cut-off point of 10% for the standardised difference in means or proportions of confounders between IV groups has been proposed to formally test this assumption.[52,53] If any of the proposed instruments violated this assumption, it was deemed not valid and not used in the IV analyses.

Stage 1 outcomes

As stated in *Chapter 1*, *Research aims and objectives*, stage 1 aimed to emulate TOPKAT's results using observational data to identify the best methods for minimising confounding. Many of the TOPKAT outcomes could not be obtained from routinely collected data. However, TOPKAT collected OKS data 1 year after randomisation. For most TOPKAT participants, their postoperative OKS was collected 9–12 months after surgery. The postoperative OKS recorded in the PROMs database was requested from patients by 6 months after surgery. For many patients in the PROMs databases, the postoperative OKS was, therefore, collected 6–12 months after surgery, similar to when the postoperative OKS was collected in TOPKAT. Another TOPKAT end point, revision, was mandatory in the NJR data collection. We used the 6-month to 12-month postoperative OKS and 5-year revision as stage 1 outcomes.

Patients were followed up from the start date of their surgery to the earliest of:

- end of enrolment in the database, for example owing to emigration, or 31 December 2016
- date of revision surgery (for the revision outcome)
- death
- end of 5-year observation period.

Outcome analyses

Propensity score-based methods

The same statistical approaches that were used in TOPKAT were applied for the PS-based methods: linear regression for postoperative OKS and Poisson regression for 5-year revisions.[5] Like TOPKAT, the dependency of different patients who were operated on by the same lead surgeons was implemented as a cluster level in the linear and Poisson regressions.

Instrumental variable analyses

The two-stage least-squares method was used.[54] The first model estimated the effect of an IV of interest on the exposure (UKR vs. TKR). The predicted exposure based on the IV was used in the second model to compare outcomes between exposed (UKR) and unexposed (TKR) recipients.

Evaluating the stage 1 methods

For each outcome (OKS and revision), we report the difference in effect estimates and the overlap in OKS 95% CIs.

For each method, as prespecified in the UTMoSt stage 1 protocol, we conducted a random-effect meta-analysis of the estimates derived from TOPKAT. A method was considered invalid and, therefore, excluded from stage 2 if any of the following were true:

1. The chi-squared test had a p-value of < 0.05, which suggests statistical heterogeneity between the estimates.
2. The I^2 was $> 40\%$, which suggests a considerably important difference between the TOPKAT and the method estimates.[55,56]
3. The between-method variance, tau^2, was large. There was no predefined cut-off point for the variance.

Only the observational methods that passed all of these tests in the OKS cohort were considered valid approaches for minimising confounding and were used in stage 2.

We used three other methods to test the validity of using these analytical approaches to deal with confounding. An analysis was deemed valid if the results yielded an OKS estimate that fell within the 95% CI of the TOPKAT estimate.[57] An analysis was considered to have successfully mimicked TOPKAT if the statistical significance of the treatment estimate agreed with that seen in the trial (statistical significance agreement test).[58] As suggested by our co-investigators, an analytical approach was considered unable to replicate TOPKAT if the OKS had a minimally clinically significant difference of < 4.[55]

Sensitivity and subgroup analyses

UTMoST's eligibility criteria did not take into account the fact that surgeons' experience was used as an inclusion criterion for participating surgeons in TOPKAT. To explore the impact of surgeons' experience, we planned several ad hoc analyses after a co-applicant meeting in February 2019. We conducted a sensitivity analysis restricted to participants operated on by surgeons who had carried out ≥ 10 knee replacements of the same type as the index in the previous year, as this was the inclusion criterion for surgeons in TOPKAT.

To explore the impact of surgeon experience on the observed effects, we performed sensitivity analyses restricted to surgeries performed by lead surgeons with ≥ 30 and ≥ 50 surgeries of the same type as the index surgery in the previous year. Owing to limited power, these additional analyses could be carried out for secondary outcomes only, as the number of patients with a linked OKS was limited.

No sensitivity, subgroup or interaction analyses between age, sex or ASA and TKR/UKR were conducted. The aim of stage 1 was to compare TOPKAT's main results with the results obtained after using each observational method for accounting for confounding, not to evaluate the treatment effect of UKR in the observational data or in different population strata.

Ethics and scientific approval

No additional ethics approval was required as this study used pseudo-anonymised, routinely collected data from HES, NJR and PROMs. A NJR data request was approved by the NJR research subcommittee (reference number RSC2016/13). The HES PROMs and linkage to NJR data request was approved by NHS Digital (reference number DARS-NIC-172121-G0Z1H). The Confidentiality Advisory Group (CAG) approved the data linkage (reference 17/CAG/0174).

Chapter 3 Stage 1 patients' characteristics and propensity score-based analyses

Study population and participant flow

The NJR database contained 868,785 records of TKR or UKR. Of these, 553,567 records had unique HES linkage data, which were required for inclusion. HES covers treatment centres and hospitals in England only, whereas the NJR contains patient records for the whole of the UK.

After removing 2099 duplicate records, 514,435 patients were reported to have had TKR and 39,132 patients were reported to have had UKR in the linked database. Additional exclusion criteria were applied, as follows:

- We included only the first and unilateral knee replacement procedures. We excluded 5141 patients undergoing coded 'bilateral knee replacement surgery' and 74,601 patients undergoing knee replacements on both knees on the same date (suggestive of bilateral knee replacement surgery).
- We excluded patients whose surgeries were carried out after 2016 or who died before postoperative OKS data collection.
- We excluded 5402 patients who received patellofemoral or lateral knee replacements.
- We excluded 4567 patients without IMD data and 52 patients with inconsistent age data in HES and NJR, with a difference of > 3 years.

The final data set included 425,284 TKR and 32,293 UKR patients for further analyses. *Figure 2* summarises patient flow through the study.

To replicate TOPKAT's findings in stage 1, we also applied the trial's eligibility criteria:

- We excluded 75,074 patients (TKR, $n = 72{,}183$; UKR, $n = 2891$) who had a preoperative ASA score of > 2.
- We excluded another 79,571 TKR patients and 8376 UKR patients using TOPKAT's clinical eligibility criteria listed in *Table 1*.

After we applied the TOPKAT criteria, 273,530 TKR patients and 21,026 UKR patients were included in the stage 1 revision analysis. They formed the revision cohort. Of these, 1197 UKR and 125,834 TKR patients had postoperative OKS data and could be used to analyse the primary outcome: postoperative OKS. They formed the OKS cohort. *Figure 3* summarises patient flow from the full cohort to the stage 1 revision and OKS cohorts.

Table 3 shows the unadjusted patient-level characteristics in the revision and OKS cohorts before matching, stratification or other strategies were used to minimise confounding. The two cohorts were generally comparable. However, UKR patients in the OKS cohort appeared healthier than those in the revision cohort.

Overall, in the revision cohort there were noticeable differences between patients who received TKR and patients who received UKR in terms of sex (43% vs. 52% men, respectively), health status (11% vs. 21% rated as fit and healthy, respectively), comorbidity levels (69% vs. 73% with no reported comorbidity, respectively) and age [mean (SD): 70.2 (8.9) years vs. 64.3 (9.5) years, respectively]. TKR patients had a lower mean preoperative OKS [mean (SD): 19.3 (6.8)] than UKR patients [mean (SD): 21.3 (6.2)]. UKR patients were more likely to have comorbid osteoarthritis and other joint problems (18% vs. 13% in TKR) and cardiovascular disease (58% vs. 46%) than TKR patients.

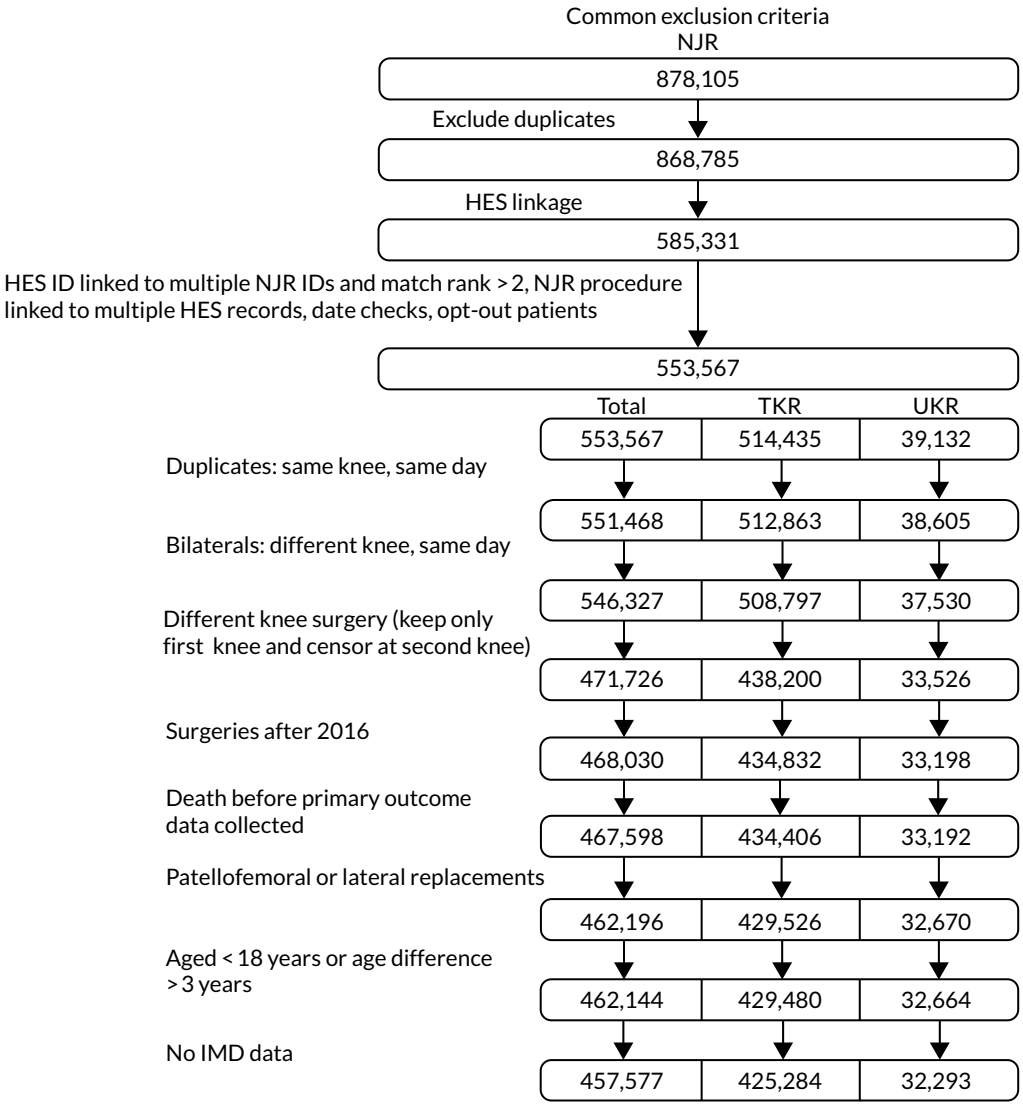


FIGURE 2 Patient flow showing the common exclusion criteria used for the whole study.

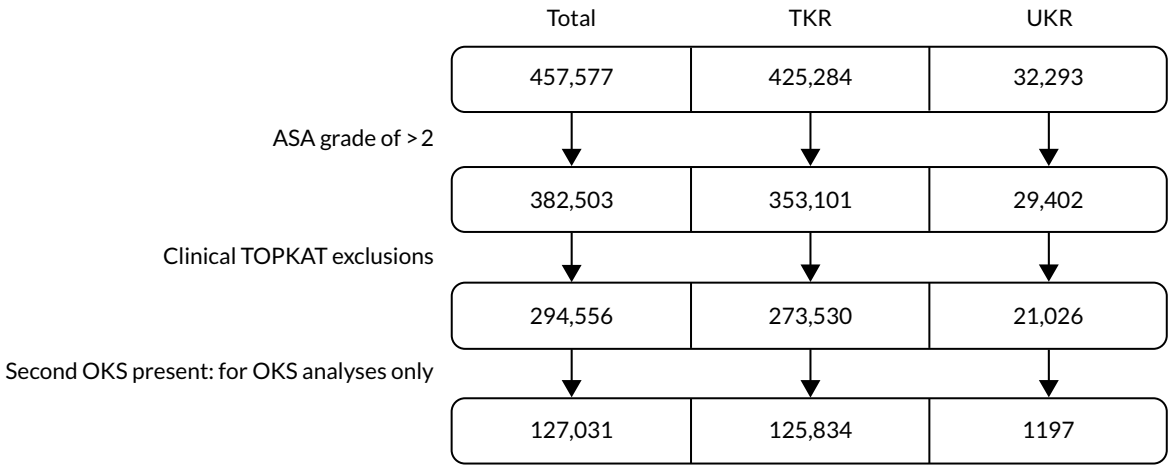


FIGURE 3 Patient flow showing the selection of patients from the full cohort for the stage 1 revision and OKS cohorts.

TABLE 3 Baseline patient-level characteristics for patients who received TKR or UKR surgeries

	Revision cohort		OKS cohort	
Stage 1	**TKR (*N* = 273,530)**	**UKR (*N* = 21,026)**	**TKR (*N* = 125,834)**	**UKR (*N* = 1197)**
Sex, *n* (%)				
Female	155,267 (57)	10,016 (48)	70,671 (56)	576 (48)
Male	118,263 (43)	11,010 (52)	55,163 (44)	621 (52)
Rural Index, *n* (%)				
Urban	203,938 (74)	14,607 (70)	92,052 (73)	844 (71)
Town and fringe	32,573 (12)	2698 (13)	15,730 (13)	164 (14)
Village	26,012 (10)	2596 (12)	12,637 (10)	138 (12)
Isolated	11,007 (4)	1125 (5)	5415 (4)	51 (4)
IMD, *n* (%)				
Least deprived 10%	29,339 (11)	2917 (14)	14,168 (11)	149 (12)
Less deprived				
10–19%	31,518 (12)	2871 (14)	15,194 (12)	137 (11)
20–29%	31,946 (12)	2669 (13)	15,435 (12)	142 (12)
30–39%	32,593 (12)	2480 (12)	15,405 (12)	138 (12)
40–49%	31,209 (11)	2456 (12)	14,611 (12)	164 (14)
More deprived				
10–19%	20,502 (7)	1224 (6)	8628 (7)	102 (9)
20–29%	23,357 (9)	1415 (7)	10,110 (8)	84 (7)
30–39%	26,174 (10)	1917 (9)	11,621 (9)	123 (10)
40–49%	29,479 (11)	2156 (10)	13,557 (11)	106 (9)
Most deprived 10%	17,413 (6)	921 (4)	7105 (6)	52 (4)
ASA, *n* (%)				
P1: fit and healthy	30,224 (11)	4394 (21)	13,849 (11)	242 (20)
P2: mild disease not incapacitating	243,306 (89)	16,632 (79)	111,985 (89)	955 (80)
Charlson Comorbidity Index score, *n* (%)				
0	187,509 (69)	15,408 (73)	86,474 (69)	915 (76)
1	58,781 (21)	4134 (20)	26,733 (21)	224 (19)
2	17,834 (7)	996 (5)	8357 (7)	41 (3)
3	6172 (3)	308 (1)	2846 (2)	13 (1)
4	3234 (1)	180 (1)	1424 (1)	4 (0)
Age (years), mean (SD)	70.2 (8.9)	64.3 (9.5)	70.4 (8.6)	64.9 (9.4)
BMI (kg/m^2), mean (SD)	30.5 (5.1)	30.0 (4.9)	30.4 (5.0)	29.6 (4.7)
PROMs				
Preoperative OKS, mean (SD)	19.3 (6.8)	21.3 (6.2)	19.7 (7.6)	21.9 (7.5)
EQ-5D, mean (SD)	69.2 (19.4)	69.7 (19.2)	70.0 (19.2)	71.1 (19.0)
General health, *n* (%)				
Excellent	161,904 (59)	6546 (31)	88,778 (71)	604 (50)
1	43,913 (16)	6643 (32)	1433 (1)	33 (3)
2	30,058 (11)	4400 (21)	10,398 (8)	181 (15)
3	26,008 (9)	2217 (10)	17,504 (14)	271 (23)
4	10,024 (4)	834 (4)	6886 (5)	94 (8)
Poor	1623 (1)	386 (2)	835 (1)	14 (1)

continued

TABLE 3 Baseline patient-level characteristics for patients who received TKR or UKR surgeries (*continued*)

	Revision cohort		OKS cohort	
Stage 1	**TKR (*N* = 273,530)**	**UKR (*N* = 21,026)**	**TKR (*N* = 125,834)**	**UKR (*N* = 1197)**
Medical history*, n *(%)				
Gastrointestinal disease	52,029 (19)	3621 (17)	25,142 (20)	174 (15)
Osteoarthritis and other joint problems	49,941 (18)	2696 (13)	23,578 (19)	149 (12)
Mental health	25,823 (9)	2380 (11)	11,421 (9)	101 (8)
Respiratory diseases	37,754 (14)	2827 (13)	17,078 (14)	147 (12)
Cardiovascular diseases	157,504 (58)	9592 (46)	73,382 (58)	515 (43)
Thyroid problems	20,724 (8)	1249 (6)	9742 (8)	80 (7)
Foot, hip or spinal pain	3096 (1)	205 (1)	1519 (1)	15 (1)
Coxarthrosis	8966 (3)	381 (2)	4395 (3)	25 (2)
Neurological disorders	16,435 (6)	1208 (6)	7491 (6)	67 (6)
Other arthrosis	12,818 (5)	708 (3)	5930 (5)	41 (3)
Polyarthrosis	15,935 (6)	675 (3)	7520 (6)	29 (2)
Spondylosis	7378 (3)	349 (2)	3501 (3)	17 (1)

Similar differences in sex, mean age, ASA grade and the PROMs for general health and mean preoperative OKS were observed between TKR and UKR patients in the OKS cohort. However, UKR patients were generally healthier than TKR patients in the OKS cohort. In addition, more UKR patients had no Charlson Comorbidity Index scores than TKR patients (76% vs. 69%, respectively). TKR patients were more likely than UKR patients in the OKS cohort to have a history of gastrointestinal disease (20% vs. 15%), osteoarthritis and other joint problems (19% vs. 12%), or cardiovascular disease (58% vs. 43%).

Covariate balance assessment

Covariate balance assessments were conducted for PS matching, PS stratification and IPW methods.

Propensity score matching

We PS matched 1197 UKR patients to 5652 TKR patients in the PS-matched postoperative OKS cohort. Before PS matching, TKR patients had much lower PS values than UKR patients. The wide range of ASMD values for the different characteristics shows the degree of mismatch (*Figure 4a*). Such differences disappeared in the PS-matched cohort, with estimated ASMD values for all baseline characteristics dropping to below 0.1 after matching (see *Figure 4a*). The TKR and UKR groups were, therefore, well-balanced after matching.

Baseline characteristics for the whole OKS cohort (before PS matching) and the matched cohort (after PS matching) are detailed in *Appendix 1*, *Table 23*. The characteristics of TKR patients were different in the matched and unmatched cohorts. After matching, TKR patients became more like the UKR patients: they were healthier and younger and a greater proportion were men.

We PS matched 21,026 UKR patients and 92,071 TKR patients from the revision cohort to form the PS-matched revision cohort, excluding 181,459 TKR patients in the process. The UKR and TKR patients in the matched revision cohort had generally similar baseline characteristics (see *Appendix 1*, *Table 24*). All patient-level covariates were well below the prespecified threshold of an ASMD of ≤ 0.1 after matching, suggesting that PS matching produced excellently balanced matched samples of TKR and UKR patients (see *Figure 4b*).

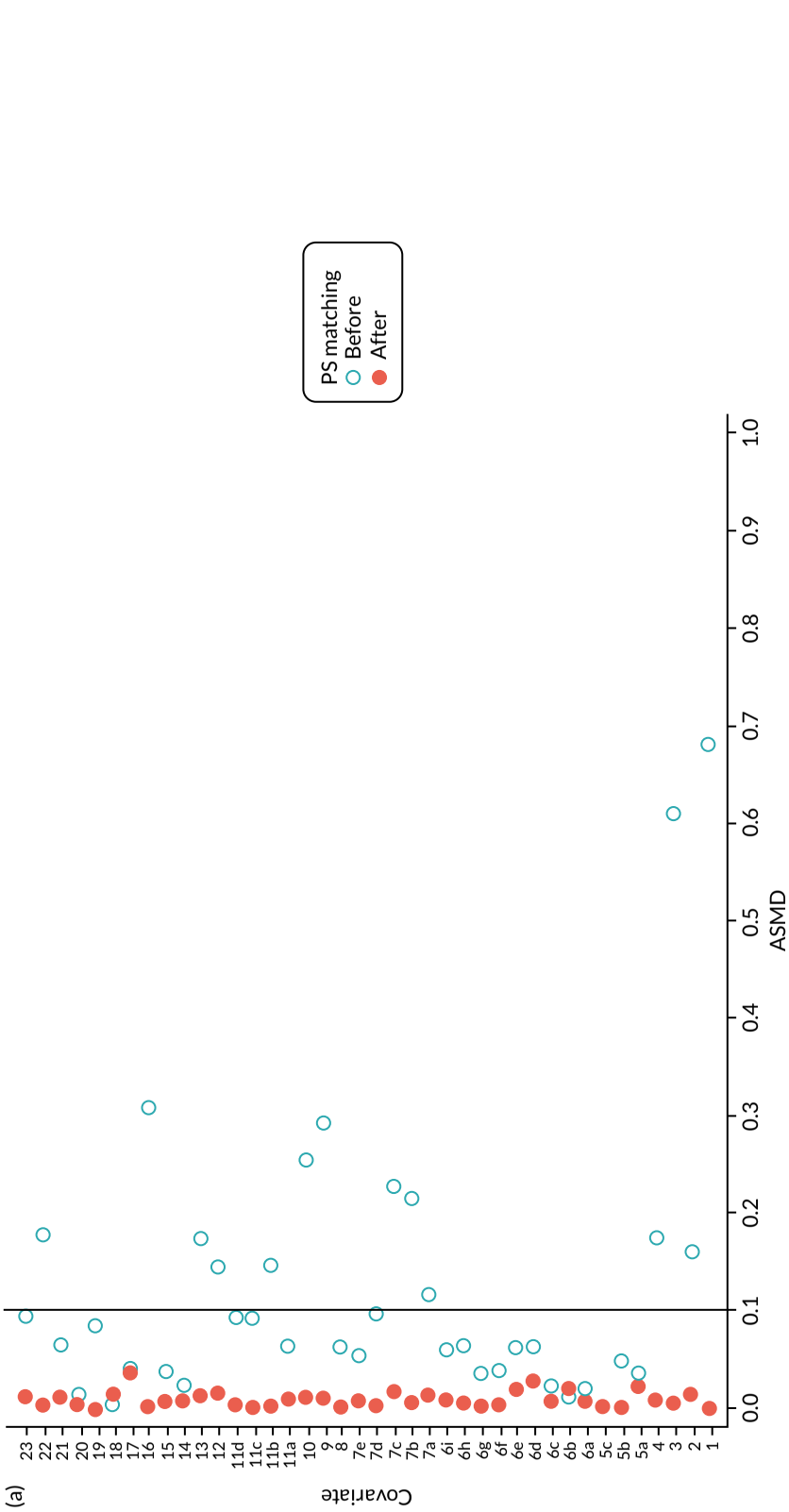


FIGURE 4 The ASMD of each covariate included in the PS matching for the (a) postoperative OKS and (b) revision cohorts, before and after PS matching. 1, overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

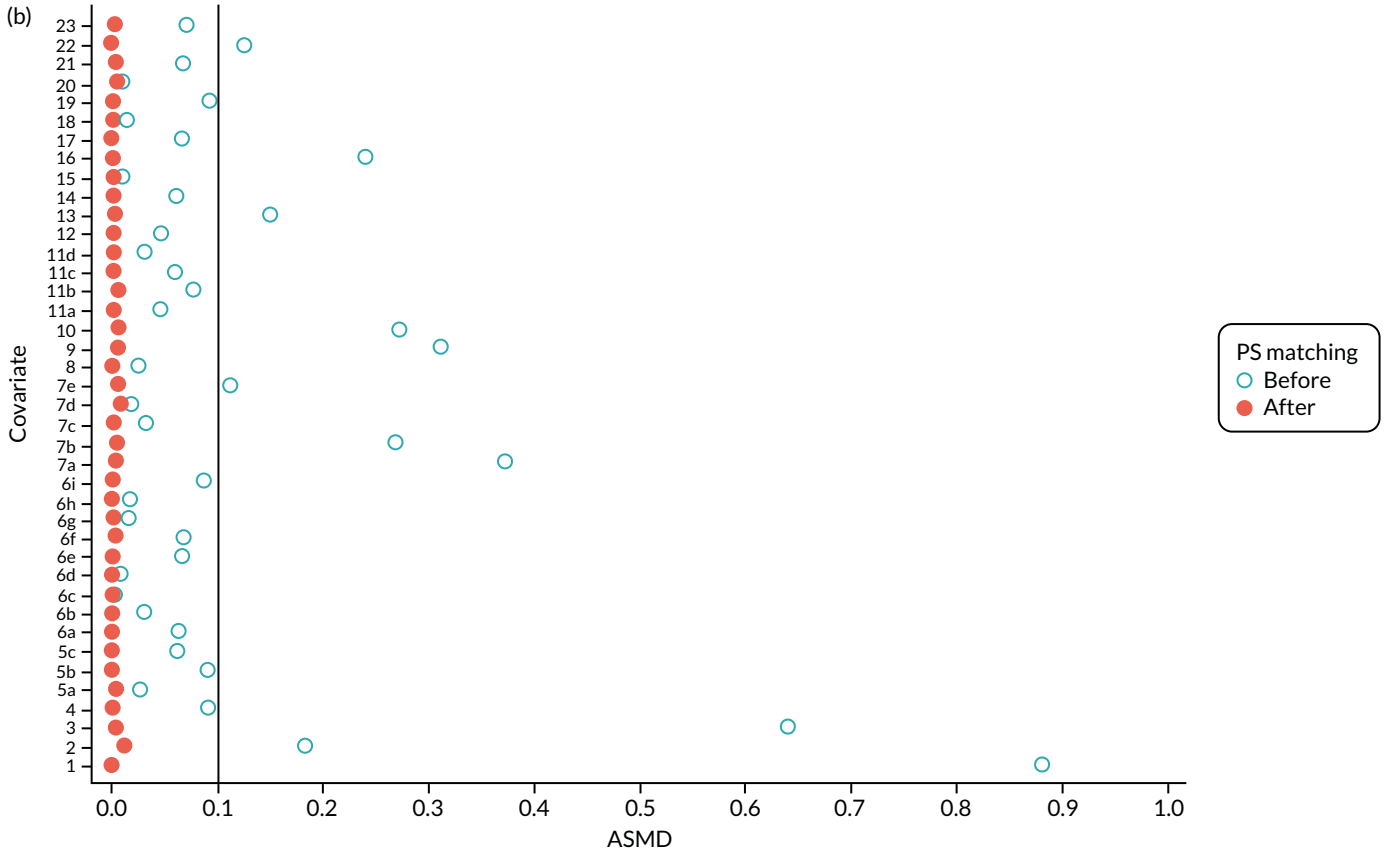


FIGURE 4 The ASMD of each covariate included in the PS matching for the (a) postoperative OKS and (b) revision cohorts, before and after PS matching. 1, overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis.

Propensity score stratification

As defined in *Chapter 2*, *Propensity score stratification*, two sets of 10 strata were created in the OKS and revision cohorts. One set was created by splitting the distribution of the estimated PS stratification in the whole cohort (PSS_{whole}) into 10, and the other set was created by splitting the distribution of the PS stratification in the UKR cohort (PSS_{exp}) into 10.

In the OKS cohort, PSS_{exp} stratification based on the PS distribution in the UKR cohort resulted in similar PS distributions for TKR and UKR patients in each stratum (see *Figure 20a*) and equal proportions of UKR and TKR patients between and within strata (see *Figure 20b*), suggesting good overall covariate balance. By contrast, PSS_{whole} stratification based on the PS distribution in the whole study population resulted in covariate imbalances in 6 out of the 10 strata: strata 1–6 were dominated by TKR patients and had < 1% UKR patients.

Figure 5a shows the covariate balance (mean ASMD) for each confounder across strata in the OKS cohort when using the PSS_{whole} method. Overall, the PS remained imbalanced, with an ASMD of > 0.1. In particular, BMI remained imbalanced between TKR and UKR patients, with an ASMD of 0.11. Covariate balance within strata was not always achieved, especially in strata 1–6. This is not surprising, as there were < 1% UKR patients included in these strata.

By contrast, PSS_{exp} stratification balanced all covariates, with an average ASMD of ≤ 0.1 across strata (see *Figure 5b*). This method also had better covariate balance within strata in most strata. In conclusion, in the OKS cohort, PSS_{exp} resulted in a balanced distribution of baseline characteristics between TKR and UKR patients. BMI remained imbalanced when using the PSS_{whole} method and was, therefore, included as a covariate adjustment when estimating the exposure effect [see *Primary outcome (postoperative Oxford Knee Score) results and comparison with the TOPKAT findings*].

In the revision cohort, in both methods, TKR and UKR patients have similar PSs (see *Figure 21*). The PSS_{whole} method performed better for the revision cohort than for the OKS cohort. Only stratum 1 in the revision cohort had < 1% UKR patients. The PSS_{whole} method achieved within-stratum covariate balance, except for sex; age; BMI; IMD; preoperative general health; EQ-5D; OKS; Charlson Comorbidity Index; mental health diseases; cardiovascular diseases; thyroid problems; foot, hip and spinal pain; and coxarthrosis in some strata. On average across the 10 strata, only the preoperative OKS had an imbalanced distribution after stratification, with a mean ASMD of 0.14 (see *Figure 5c*). It was, therefore, included in the exposure effect estimation.

By contrast, the PSS_{exp} method resulted in a mean ASMD of ≤ 0.1 across strata for all covariates, which indicates good average covariate balance (see *Figure 5d*). Within-stratum covariate balance was also achieved for all covariates except sex, BMI, general health, EQ-5D and OKS in some strata. In conclusion, the PSS_{exp} method resulted in better covariate balance than the PSS_{whole} method in the revision cohort, as was found with the OKS cohort.

Inverse probability weighting

In the OKS pseudo-population, the 1197 UKR patients had a stabilised weight ranging from 0.04 to 7.90 [interquartile range (IQR) 0.37–1.30], with a mean of 1. The TKR patients had a stabilised weight ranging from 0.99 to 1.35 (IQR 0.99–1.00), with a mean of 1.

The UKR and TKR patients in the OKS cohort had similar distributions in all covariates included in the PS except BMI, which had an ASMD just above 0.1 (*Figure 6a*). This imbalance was of limited clinical relevance: UKR patients had a mean BMI of 29.87 kg/m^2 and TKR patients had a mean BMI of 30.43 kg/m^2.

In the revision cohort, the 21,026 UKR patients had a weight ranging from 0.09 to 27.73, and the 273,530 TKR patients had a weight ranging from 0.93 to 12.30; both had a mean weight of 1. Both groups had a balanced distribution for all of the covariates, with an ASMD of ≤ 0.1 (see *Figure 6b*).

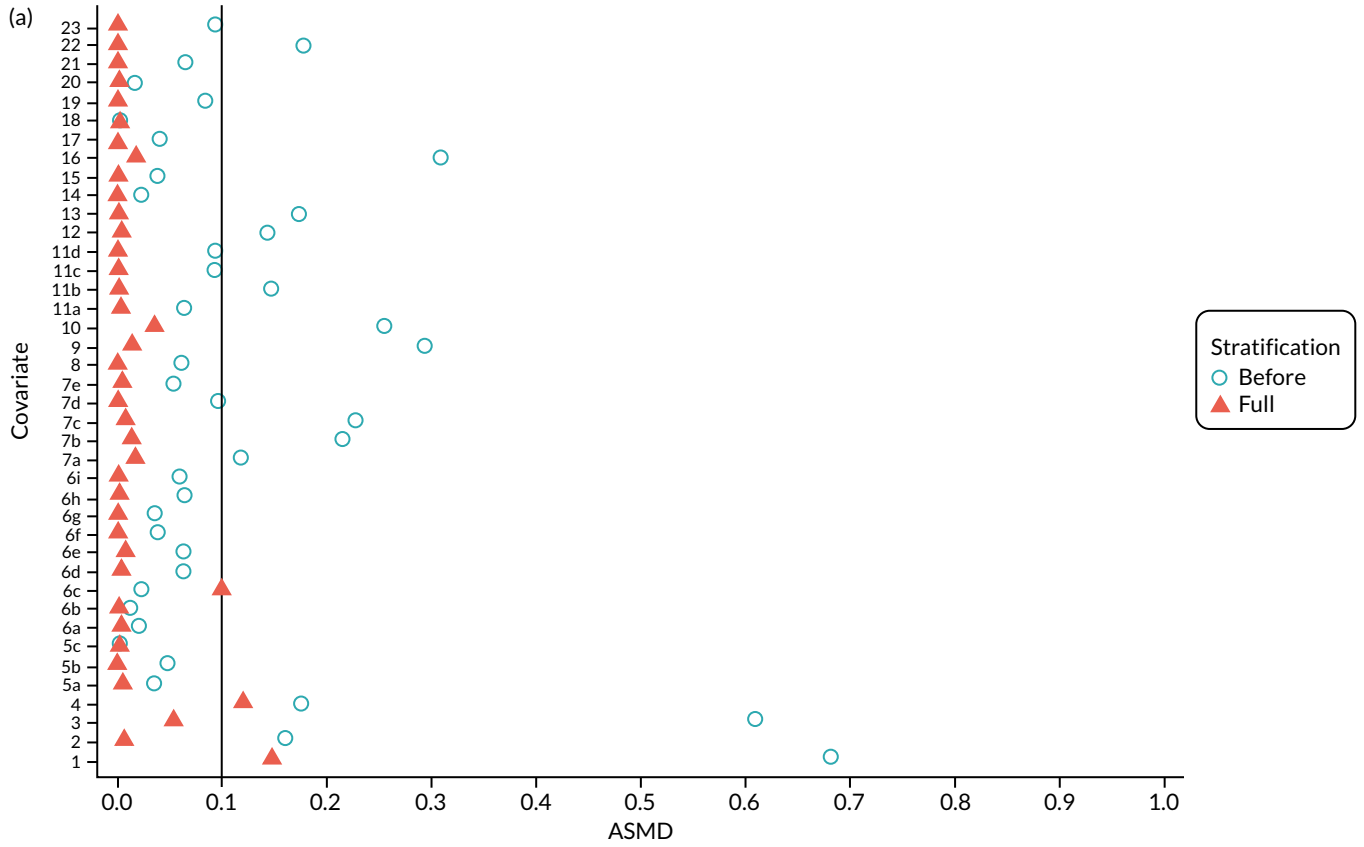


FIGURE 5 The ASMD for each covariate included in the PS stratification for the (a) ASMD for the OKS cohort with full PSS; (b) ASMD for the OKS cohort with exposure PSS; (c) safety cohort with full PSS; and (d) safety cohort with exposure PSS. 1, Overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

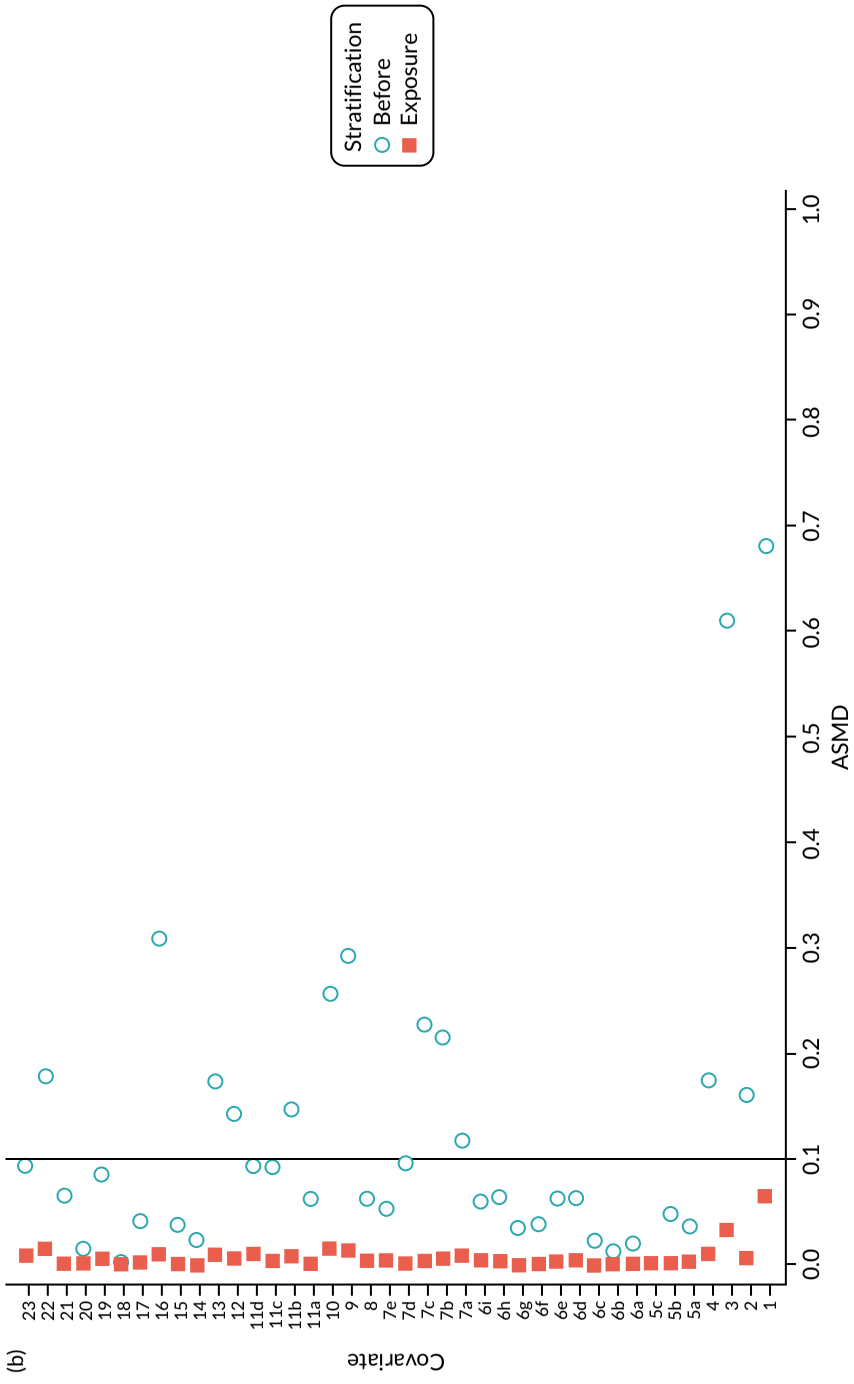


FIGURE 5 The ASMD for each covariate included in the PS stratification for the (a) ASMD for the OKS cohort with full PSS; (b) ASMD for the OKS cohort with exposure PSS; (c) safety cohort with full PSS; and (d) safety cohort with exposure PSS. 1, Overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

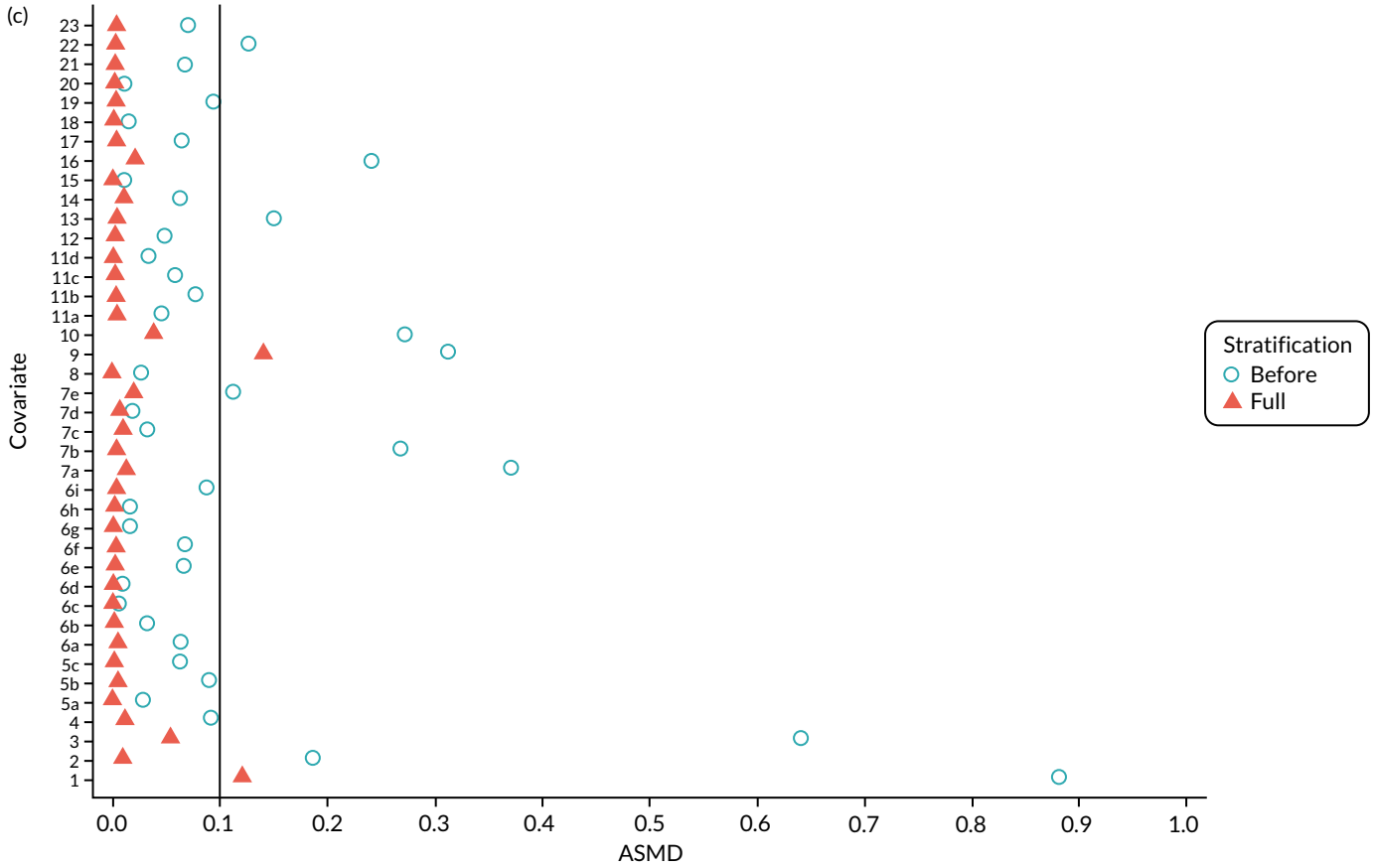


FIGURE 5 The ASMD for each covariate included in the PS stratification for the (a) ASMD for the OKS cohort with full PSS; (b) ASMD for the OKS cohort with exposure PSS; (c) safety cohort with full PSS; and (d) safety cohort with exposure PSS. 1, Overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

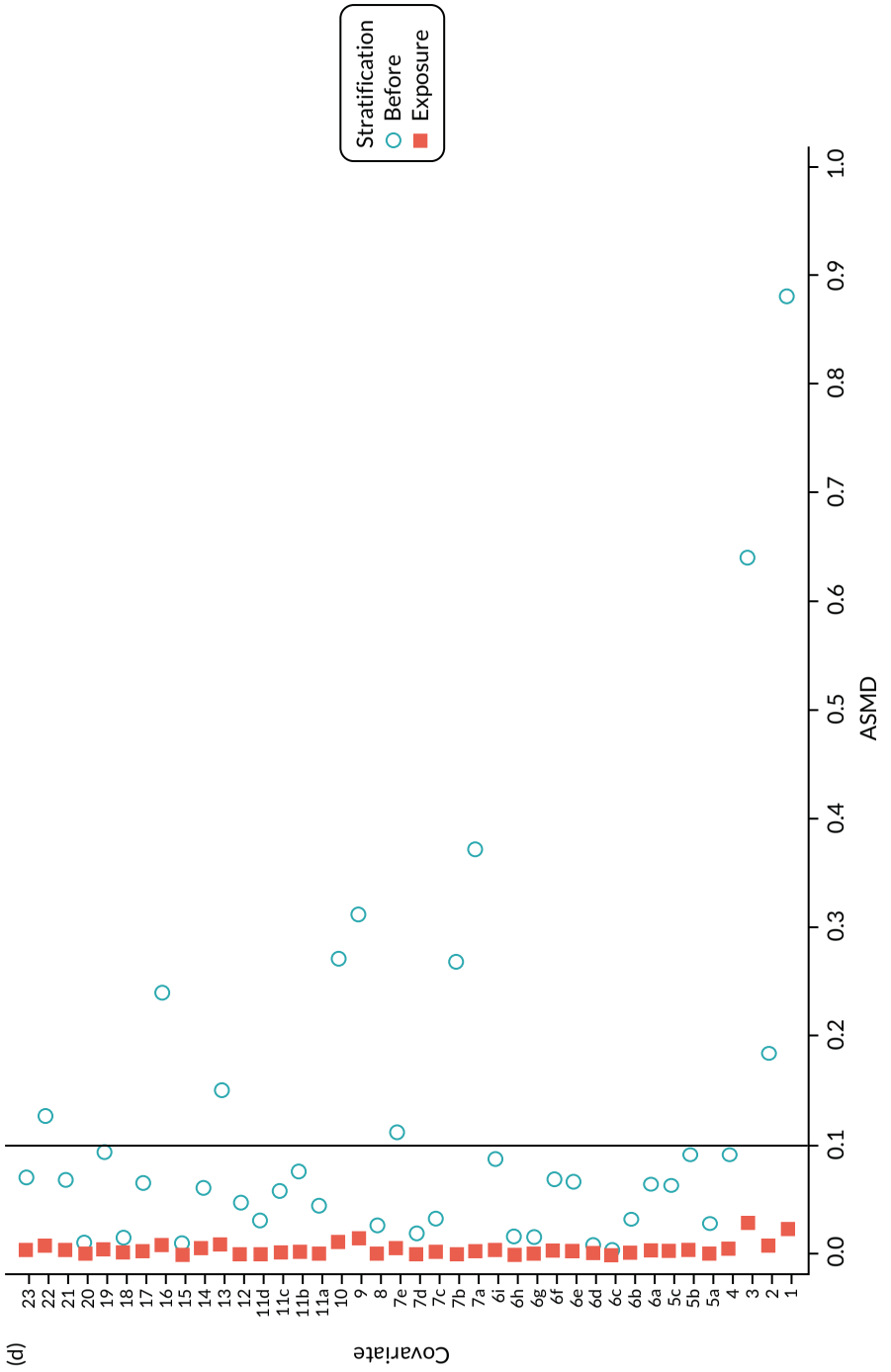


FIGURE 5 The ASMD for each covariate included in the PS stratification for the (a) ASMD for the OKS cohort with full PSS; (b) ASMD for the OKS cohort with exposure PSS; (c) safety cohort with full PSS; and (d) safety cohort with exposure PSS. 1, Overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis.

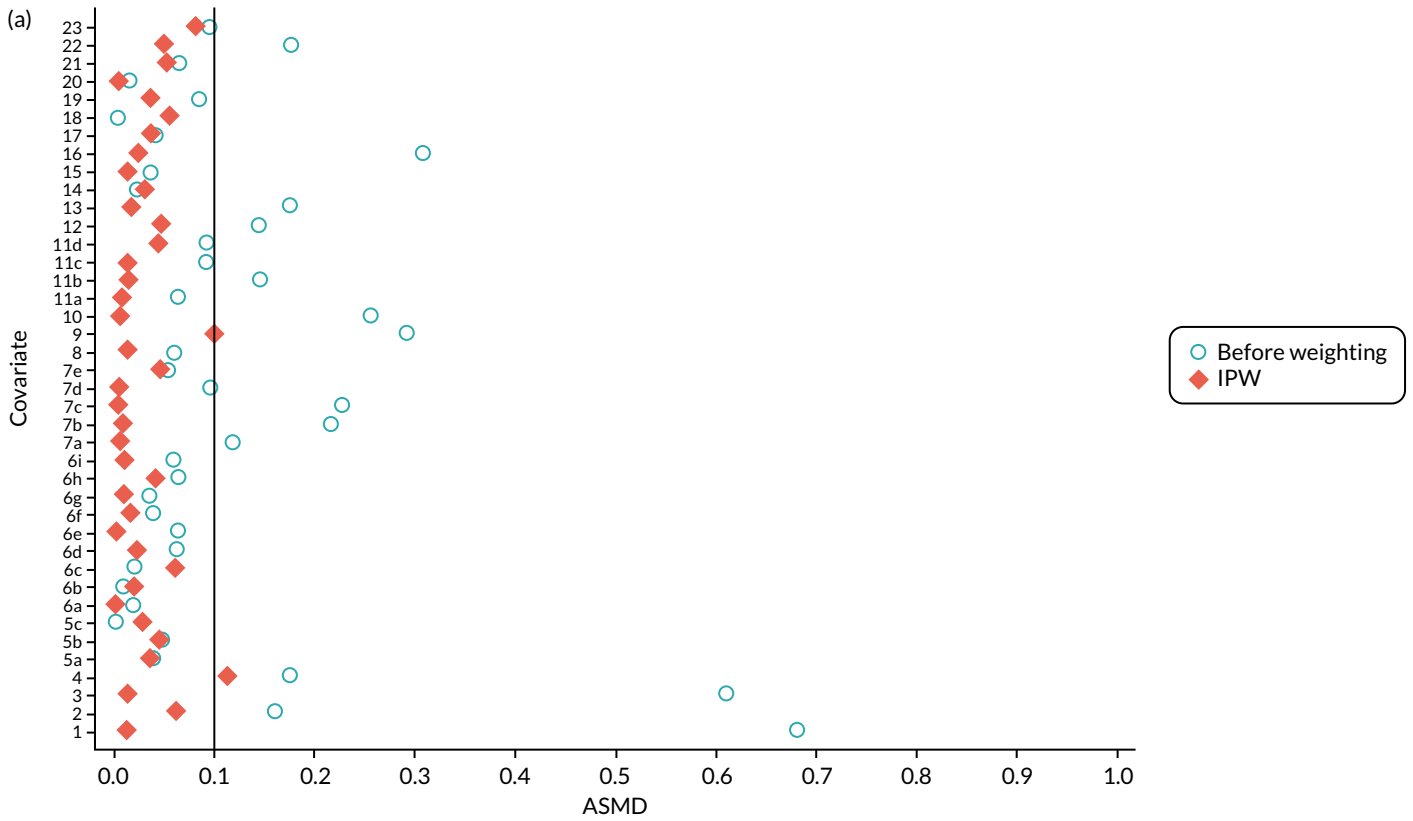


FIGURE 6 The ASMD for each covariate included in the PS matching for the (a) postoperative OKS cohort and (b) the revision cohort, before and after IPW. 1, overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

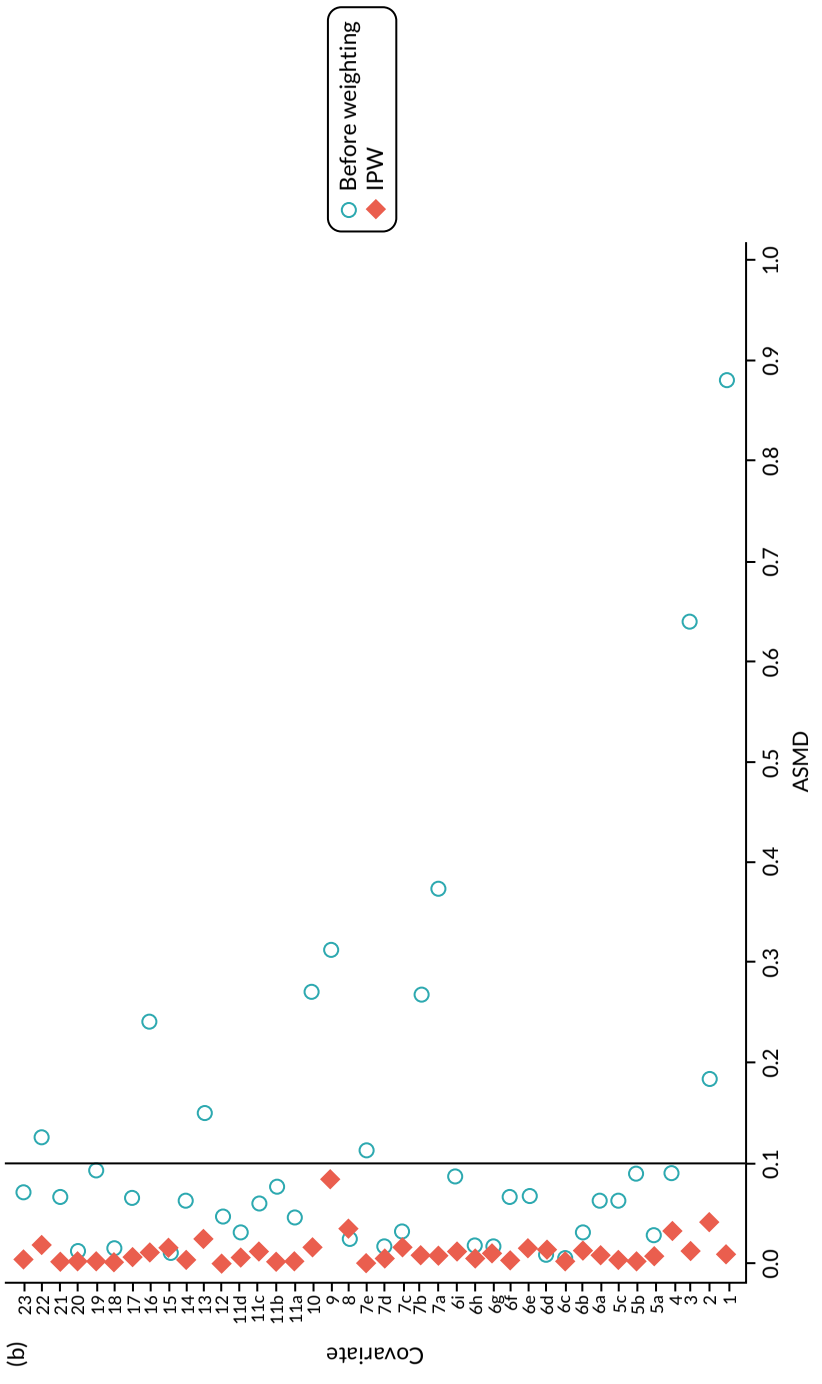


FIGURE 6 The ASMD for each covariate included in the PS matching for the (a) postoperative OKS cohort and (b) the revision cohort, before and after IPW. 1, overall PS; 2, males; 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis.

Inverse probability weighting minimised confounding to an acceptable degree based on the prespecified threshold (ASMD of ≤ 0.1) in both cohorts. Only BMI in the OKS cohort remained unbalanced, but this imbalance was of little clinical significance (< 0.6 kg/m^2 difference in means between UKR and TKR recipients) and was adjusted for in the final analyses.

Primary outcome (postoperative Oxford Knee Score) results and comparison with the TOPKAT findings

One of the key limiting factors for replicating RCTs with observational data is identifying a population with similar characteristics to the trial participants. *Table 4* shows preoperative and postoperative OKSs collected in TOPKAT and estimated using each of the tested methods: crude scores (no adjustment for confounding), PS matching, IPW, PSS_{whole}, PSS_{exp}, and linear and non-linear PS adjustment.

TABLE 4 The preoperative and postoperative OKSs collected in TOPKAT and estimated from the crude analysis and with each PS method

Preoperative and postoperative OKSs	Treatment group, mean (SD)		Mean difference/ effect size (95% CI)
	TKR	UKR	
TOPKAT			
Preoperative OKS	19.00 (7.2)	18.80 (7.0)	–
Postoperative OKS	35.10 (10.3)	36.90 (9.9)	1.91 (0.20 to 3.62)
Crude			
Preoperative OKS	19.68 (7.56)	21.88 (7.52)	–
Postoperative OKS	35.80 (9.35)	36.74 (9.77)	0.76 (0.22 to 1.29)
PSM			
Preoperative OKS	21.96 (7.76)	21.88 (7.52)	–
Postoperative OKS	36.71 (9.14)	36.74 (9.77)	0.27 (–0.38 to 0.92)
IPW			
Preoperative OKS	19.70 (7.57)	20.41 (7.42)	–
Postoperative OKS	35.80 (9.35)	36.64 (9.50)	0.58 (–0.19 to 1.35)
PSS_{whole}			
Preoperative OKS	19.68 (11.64)	21.88 (7.94)	–
Postoperative OKS	35.80 (11.35)	36.74 (10.13)	0.56 (–0.03 to 1.16)
PSS_{exp}			
Preoperative OKS	19.68 (13.30)	21.88 (7.77)	–
Postoperative OKS	35.80 (12.31)	36.74 (9.87)	0.76 (0.15 to 1.36)
PS_{Alin}			
Preoperative OKS	19.68 (7.56)	21.88 (7.52)	–
Postoperative OKS	35.80 (9.35)	36.74 (9.77)	0.14 (–0.39 to 0.68)
$PS_{Anonlin}$			
Preoperative OKS	19.68 (7.56)	21.88 (7.52)	–
Postoperative OKS	35.80 (9.35)	36.74 (9.77)	0.10 (–0.44 to 0.63)

PSM, propensity score matching; PS_{Alin}, propensity score linear adjustment; $PS_{Anonlin}$, propensity score non-linear adjustment.

Most of the tested methods produced similar mean baseline OKSs for patients receiving TKR and UKR to those reported in TOPKAT:

- The crude/unadjusted mean preoperative OKS for patients receiving TKR (19.68) and UKR (21.88) differed by about 2 points at baseline. The mean preoperative OKS for TKR participants was similar to that in TOPKAT (mean preoperative OKS of 19.0). However, the mean preoperative OKS for UKR participants was about 3 points higher than that in TOPKAT (mean preoperative OKS of 18.8).
- PS stratification (both PSS_{whole} and PSS_{exp}) and adjustment (both linear and non-linear) included the whole population and resulted in the same mean preoperative OKS as the crude/unadjusted analysis.
- PS matching produced more similar mean preoperative OKSs for the TKR and the UKR patients than the crude/unadjusted analysis (TKR: 21.96; UKR: 21.88). Both groups were more different from the trial participants than in the crude analysis, with a mean preoperative OKS more than 2 points higher in both patient groups in the PS-matched cohort than in TOPKAT.
- The pseudo-population created by IPW had similar baseline preoperative OKSs for TKR (average 19.70) and UKR (average 20.41). The means differed from those in TOPKAT by < 1.5 points.

At the postoperative time point, approximately 6–8 months after the operation, TOPKAT and all of the tested methods showed a large improvement in OKS from baseline, in line with previous literature on knee replacement surgery.[23]

All of the applied analytical methods obtained a treatment effect estimate that favoured UKR surgery over TKR surgery, as TOPKAT did. However, PS matching, IPW, PSS_{whole} and PS adjustment all obtained estimates with 95% CIs that included the null effect (0). Only PSS_{exp} found a statistically significant difference between UKR and TKR, with a point estimate (95% CI) of 0.76 (0.15 to 1.36). All of the tested methods yielded a treatment effect estimate at least 1 point lower than the 1-year effect observed in TOPKAT (*Figure 7*). Although none of the obtained estimates was completely covered by the 95% CI from TOPKAT, all of the estimates overlapped partially with it.

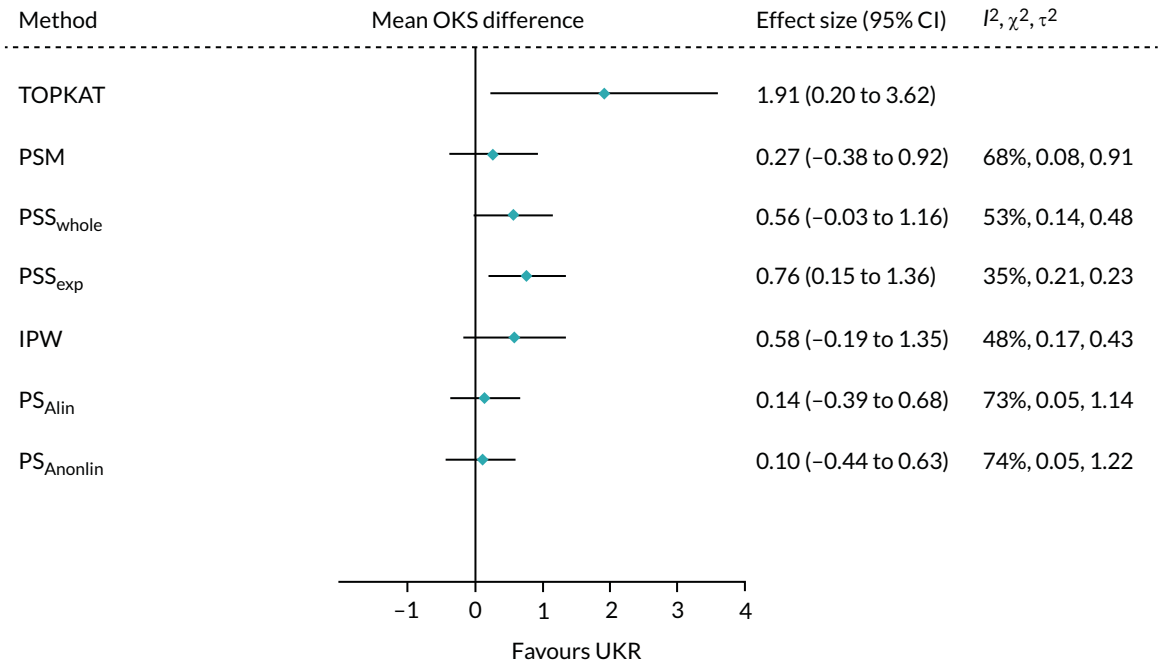


FIGURE 7 Forest plot of the postoperative OKS effect size for TOPKAT and each of the tested PS methods, with heterogeneity measures (I^2, χ^2 and τ^2). PSM, propensity score matching; PS_{Alin}, propensity score linear adjustment; $PS_{Anonlin}$, propensity score non-linear adjustment with PS^0 and $ln(PS)^0$.

All of the methods except PS adjustment had a chi-squared *p*-value of > 0.05, implying that any differences in the treatment effects collected in TOPKAT and calculated using the tested methods were likely attributable to chance.

Propensity score stratification yielded the treatment effect estimates closest to TOPKAT, with the smallest τ^2 value (PSS_{whole}: 0.23; PSS_{exp}: 0.48). PSS_{exp} was the only method with a small heterogeneity: $I^2 < 40\%$. IPW showed moderate heterogeneity ($I^2 = 48\%$; $\tau^2 = 0.43$).

Propensity score matching resulted in a point estimate of 0.27, which was close to the lower 95% CI of the TOPKAT estimate, 0.20. It also had high heterogeneity ($I^2 = 68\%$; $\tau^2 = 0.91$). PS linear and non-linear adjustments resulted in OKS effect sizes that were even more different from TOPKAT. They also had the largest I^2 and τ^2 values, suggesting that PS matching and adjustment could not replicate the TOPKAT findings.

Five-year revision risks for unicompartmental knee replacement

Overall, 852 out of 21,026 UKR participants (4.1%) and 4090 out of 273,530 TKR participants (1.5%) in UTMoSt stage 1 underwent revision surgery within 5 years of the index procedure. In the PS-matched sample, 852 out of 21,026 (4.1%) UKR and 1383 out of 71,045 (1.5%) TKR patients underwent revision surgery.

All of the tested methods yielded a greater than twofold (statistically significant) increase in the risk of 5-year revision for UKR participants compared with TKR participants. By contrast, TOPKAT found no significant difference in risk between UKR and TKR participants. Potential reasons underlying these differences are discussed in detail in *Chapter 5*. However, as observed in *Figure 8*, all of the methods in UTMoSt yielded treatment effect estimates fully covered by the 95% CI observed in TOPKAT. PSS_{whole} and PSS_{exp} were the only methods with moderate to high heterogeneity ($I^2 = 71\%$ and $I^2 = 63\%$, respectively). The other methods had no heterogeneity in their estimates of revision risk compared with TOPKAT, with $I^2 = 0\%$. None of the tested methods had a significant chi-squared result when compared with TOPKAT, suggesting that any differences in treatment estimates were probably a result of chance. Even the smallest recorded chi-squared results (PSS_{whole}: $p = 0.07$; PSS_{exp}: $p = 0.10$) were still insignificant.

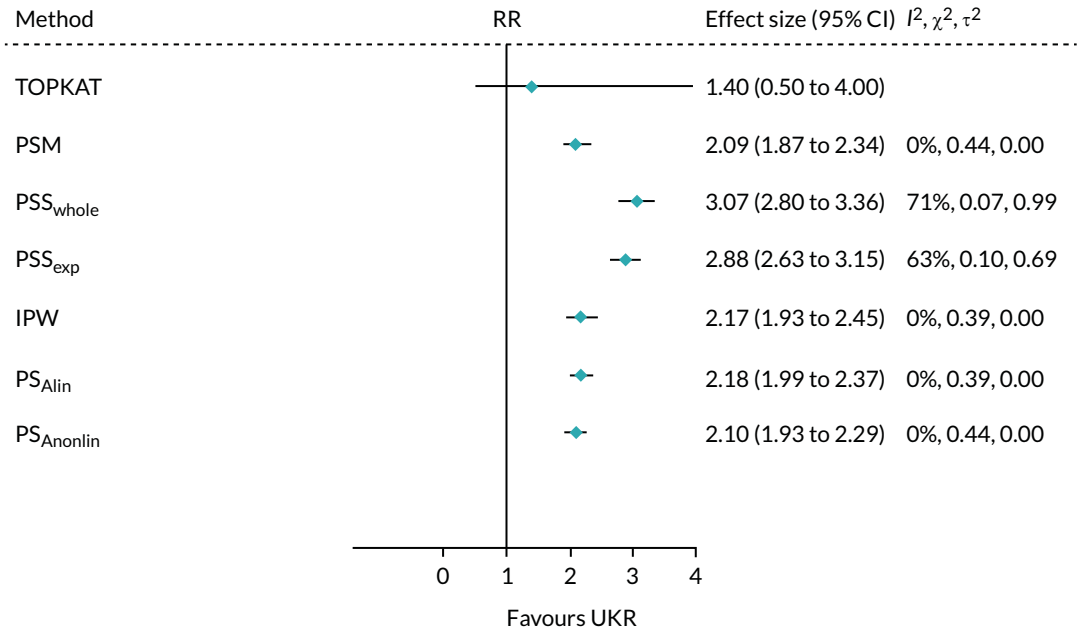


FIGURE 8 Forest plot of the 5-year relative risk of revision for TOPKAT and each of the PS methods, with heterogeneity measures (I^2, χ^2 and τ^2). PSM, propensity score matching; PS_{Alin}, propensity score linear adjustment; $PS_{Anonlin}$, propensity score non-linear adjustment with $PS^{0.5}$ and $\ln(PS)^0$; RR, relative risk.

Within this 5-year window, 496 out of 21,026 (2.4%) UKR participants and 14,004 out of 273,530 (5.1%) TKR participants died. As UKR appeared to be associated with a consistent reduction in mortality in all of our analytical methods (*Table 5*), this result suggests that further modelling to account for risk of death as a competing event may be warranted. However, such modelling was not carried out in TOPKAT. We, therefore, did not account for risk of death, so that we could compare our findings with those in TOPKAT as planned.

Sensitivity analyses

Oxford Knee Score cohort

We conducted a sensitivity analysis of patients whose surgery was performed by an 'experienced' lead surgeon who had performed at least 10 surgeries of the same type in the previous year (*Table 6*). We used the same volume-based definition of 'experienced' (i.e. number of surgeries performed) as that used to recruit participating surgeons in TOPKAT, although arguably volume does not accurately represent a surgeon's true experience.

TABLE 5 Five-year death rates and relative risk (95% CI) for TOPKAT and each of the PS methods

	Treatment group, number of patients who died within 5 years/total patients (%)		
	UKR	TKR	Relative risk (95% CI)
TOPKAT	11/264 (4.2%)	6/264 (2.3%)	N/A
PSM	496/21,026 (2.4%)	2969/71,045 (4.2%)	0.64 (0.57 to 0.71)
PSS_{whole}	496/21,026 (2.4%)	14,004/273,530 (5.1%)	0.48 (0.44 to 0.53)
PSS_{exp}			0.46 (0.42 to 0.51)
IPW			0.62 (0.55 to 0.71)
PS_{Alin}			0.64 (0.58 to 0.70)
$PS_{Anonlin}$			0.64 (0.58 to 0.70)

N/A, not applicable; PS_{Alin}, propensity score linear adjustment; $PS_{Anonlin}$, propensity score non-linear adjustment; PSM, propensity score matching.

TABLE 6 Number of participants and surgeons in the OKS and full cohorts, according to surgeon expertise in performing the index procedure

	OKS cohort				Full cohort			
	Patients		Surgeons		Patients		Surgeons	
Surgeon expertise	TKR	UKR	TKR	UKR	TKR	UKR	TKR	UKR
All, *n*	125,834	1197	3895	452	273,530	21,026	4597	1462
≥ 10 surgeries in the previous year, *n* (%)	114,871 (91.3)	602 (50.3)	2625 (67.4)	164 (36.3)	248,785 (91.0)	13,334 (63.4)	3001 (65.3)	474 (32.4)
≥ 30 surgeries in the previous year, *n* (%)	91,504 (72.7)	217 (18.1)	1556 (39.9)	43 (9.5)	195,898 (71.6)	5555 (26.4)	1730 (37.6)	128 (8.8)
≥ 50 surgeries in the previous year, *n* (%)	66,166 (52.6)	83 (6.9)	996 (25.6)	17 (3.8)	139,396 (51.0)	2550 (12.1)	1109 (24.1)	51 (3.5)

We found 2625 out of 3895 (67.4%) patients were operated on by TKR lead surgeons and 164 out of 452 (36.3%) patients were operated on by UKR lead surgeons in our cohort. The proposed sensitivity analysis that was restricted to patients operated on by experienced surgeons included 602 out of 1197 (50.2%) UKR patients and 114,871 out of 125,834 (91.3%) TKR patients from our OKS cohort. Their baseline characteristics are reported in *Appendix 1*, *Table 25*.

We applied IPW, PSS_{whole} and PSS_{exp} to the subcohort of patients in the OKS cohort who had been operated on by experienced surgeons. The resulting treatment effects were closer to that seen in TOPKAT than when using the full OKS cohort. The results from this sensitivity analysis (*Figure 9*) suggested that restricting the analysis to surgeons eligible for the trial would result in a treatment effect closer to that seen in TOPKAT than the treatment effect obtained for the full cohort.

The treatment effect estimates obtained for the experienced surgeon cohort lay fully within the 95% CI of the TOPKAT estimate. Heterogeneity was dramatically lower in the experienced surgeon cohort than in the full OKS cohort, with the I^2 for all three methods dropping to 0% and $\tau^2 = 0$. These results implied that surgeon experience contributed to the differences in treatment effect observed between the main OKS cohort analysis and TOPKAT.

Revision cohort

We also examined the association between UKR (vs. TKR) and 5-year revision and death risks stratified by surgeon experience. We defined three subcohorts of the revision cohort, based on whether the surgeon had performed ≥ 10, ≥ 30 or ≥ 50 surgeries of the same type as the index surgery in the previous year. This restricted the analysis to 248,785 out of 273,530 (91.0%), 195,898 out of 273,530 (71.6%) and 139,396 out of 273,530 (51.0%) TKR participants, and to 13,334 out of 21,026 (63.4%), 5555 out of 21,026 (26.4%) and 2550 out of 21,026 (12.1%) UKR participants, respectively (see *Table 6*). These cohorts included patients operated on by 3001 out of 4597 (65.3%) surgeons who had performed ≥ 10 TKR surgeries in the previous year, 1730 out of 4597 (37.6%) surgeons who had performed ≥ 30 TKR surgeries in the previous year and 1109 out of 4597 (24.1%) surgeons who had performed ≥ 50 TKR surgeries in the previous year. These cohorts also included 474 out of 1462 (32.4%) surgeons who had performed ≥ 10 UKR surgeries in the previous year, 128 out of 1462 (8.8%) surgeons who had

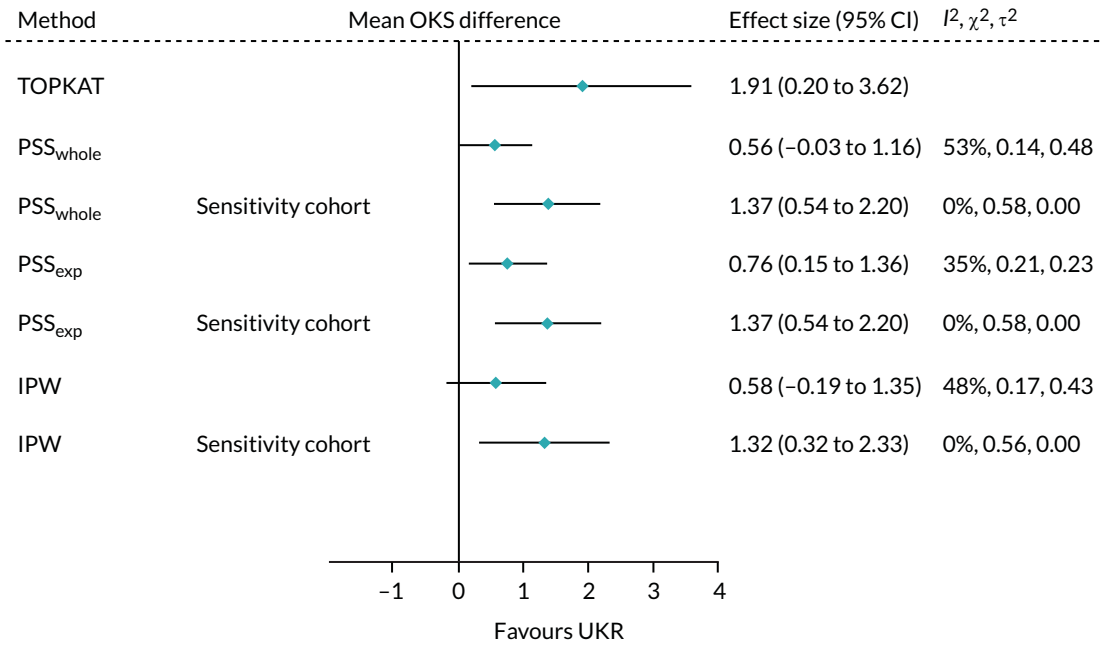


FIGURE 9 Forest plot of the postoperative OKS effect size for TOPKAT and each of the validated methods in the whole OKS cohort and in the sensitivity cohort of patients operated on by surgeons who had performed ≥ 10 surgeries of the same type in the previous year, with heterogeneity measures (I^2, χ^2 and τ^2).

performed ≥ 30 UKR surgeries in the previous year and 51 out of 1462 (3.5%) surgeons who had performed ≥ 50 UKR surgeries in the previous year. Baseline characteristics of the full revision cohort and the three subcohorts are reported in *Appendix 1*, *Table 26*.

Table 7 shows the number and percentage of UKR/TKR patients who underwent revision surgery or died within 5 years of their index operation in TOPKAT, the full revision cohort and the three experienced surgeon subcohorts. The proportion of TKR patients undergoing revision decreased with surgeon experience from 1.5% in the full cohort to 1.3% among patients operated on by the most experienced surgeons. The decrease in the proportion of patients undergoing revision was more striking for UKR patients, dropping from 4.1% in the full cohort to 3.3% among patients operated on by surgeons who had performed ≥ 10 UKR surgeries in the previous year, 2.5% of those operated on by surgeons who had performed ≥ 30 UKR surgeries in the previous year and 1.9% of those operated on by surgeons who had performed ≥ 50 UKR surgeries in the previous year.

Mortality did not change substantially with surgeon volume in the TKR cohorts (4.7% in the full cohort vs. 4.6% for those operated on by the highest-volume surgeons). However, a monotonic decrease in mortality was seen among UKR patients, with mortality dropping from 2.4% in the full cohort to 2.3%, 2.2% and 1.7% in those operated on by surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 UKR surgeries in the previous year, respectively.

When using PSS_{whole} to adjust for covariates, the 5-year relative risk of revision decreased from 3.07 (95% CI 2.80 to 3.36) in the main cohort to 1.49 (95% CI 1.05 to 2.10) in the highest-volume surgeon cohort. The effect of UKR (vs. TKR) on 5-year revision risk in the highest-volume surgeon cohort was much closer to that seen in TOPKAT (*Figure 10*) than in the other two surgeon groups. There was no heterogeneity between TOPKAT and the highest-volume surgeon group, with $I^2 = 0\%$, $\chi^2 > 0.9$ and $\tau^2 = 0$. Similar trends were observed when using PSS_{exp}.

When using IPW, the risk, again, decreased with an increase in surgeon experience, but smaller differences were observed than those observed when using either PS stratification method. IPW yielded almost identical findings for the highest-volume surgeon cohort as those seen in TOPKAT, with relative risks of 1.39 (95% CI 0.93 to 2.07) for IPW and 1.40 (95% CI 0.50 to 4.00) for TOPKAT. Restriction to high-volume surgeons did not have a striking effect on the observed association between UKR (vs. TKR) and 5-year mortality following surgery (*Figure 11*).

TABLE 7 Number (%) of participants undergoing revision surgery and dying in the 5 years after index surgery in TOPKAT, the full UTMoSt cohort (main) and the three subcohorts of participants operated on by experienced surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 surgeries of the same type as the index surgery in the year before the index surgery

	5-year revision, *n/N* (%)		5-year mortality, *n/N* (%)	
	UKR	TKR	UKR	TKR
TOPKAT	10/264 (3.8)	8/264 (3.0)	11/264 (4.2)	6/264 (2.3)
Main	852/21,026 (4.1)	4090/273,530 (1.5)	496/21,026 (2.4)	14,004/273,530 (5.1)
≥ 10 surgeries	435/13,334 (3.3)	3633/248,785 (1.5)	313/13,334 (2.3)	12,452/248,785 (5.0)
≥ 30 surgeries	137/5555 (2.5)	2670/195,898 (1.4)	122/5555 (2.2)	9472/195,898 (4.8)
≥ 50 surgeries	48/2550 (1.9)	1791/139,396 (1.3)	43/2550 (1.7)	6403/139,396 (4.6)

n/N refers to number of patients undergoing surgery/dying over the total number of patients for that group.

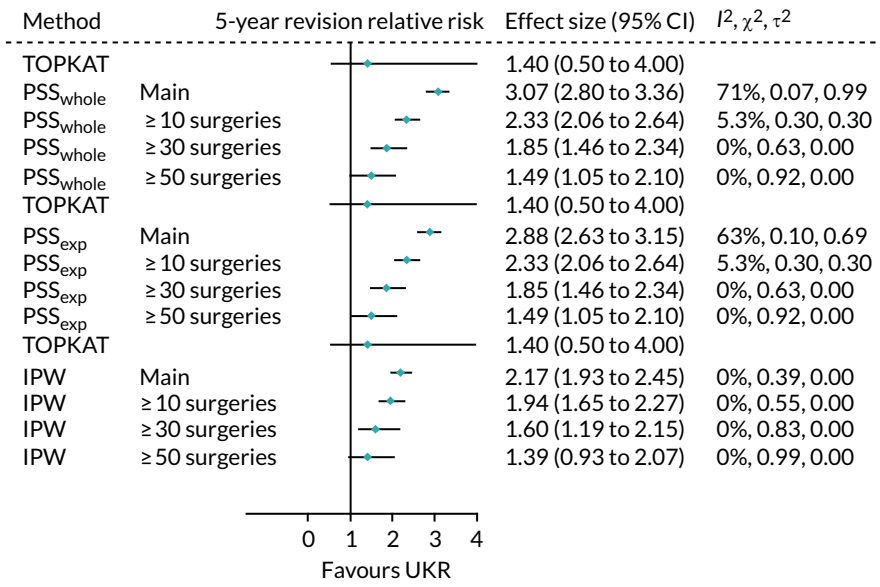


FIGURE 10 Forest plot of the relative risk of revision surgery within 5 years of initial surgery for TOPKAT and each of the validated methods in the full revision cohort (main) and the sensitivity cohorts of patients operated on by surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 surgeries of the same type in the previous year, with heterogeneity measures (I^2, χ^2 and τ^2).

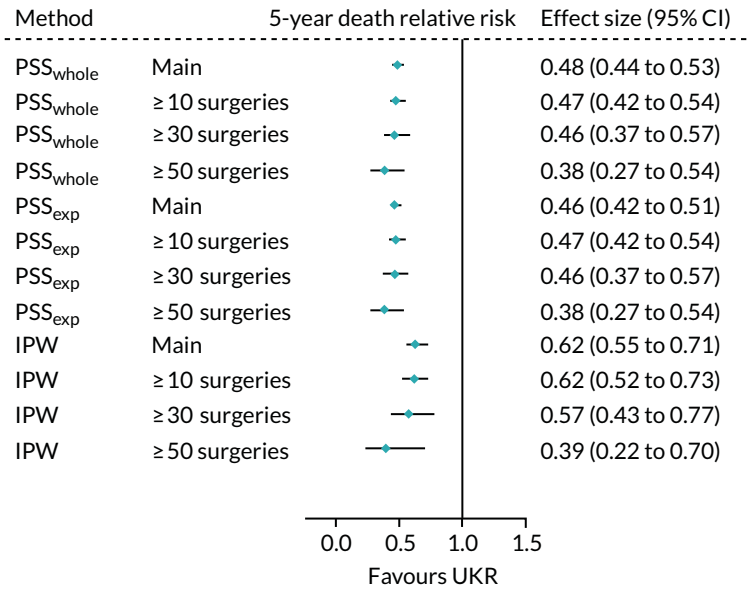


FIGURE 11 Forest plot of the estimated relative risk of death within 5 years of surgery, by index surgery type. Estimates were made using each of the validated methods with the full revision cohort (main) and the sensitivity subcohorts of patients operated on by surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 surgeries of the same type in the previous year, with heterogeneity measures (I^2, χ^2 and τ^2).

Chapter 4 Testing instrumental variable analyses

Patient characteristics

Eligible patient cohort

As mentioned in *Chapter 3*, 127,031 patients (TKR, $n = 125{,}834$; UKR, $n = 1197$ recipients) were eligible for inclusion in our primary analysis of OKS, and 294,556 patients were eligible for inclusion in our secondary analysis of revision surgery and death (TKR, $n = 273{,}530$; UKR, $n = 21{,}026$ recipients). We reported the baseline characteristics for these cohorts in *Chapter 3*, *Study population and participant flow*.

The revision cohort was used to construct IVs, with additional patients excluded as needed for the preference-based instruments. For example, to estimate surgeon-based preference for UKR based on the previous 10 surgeries, we excluded the first 10 patients for each surgeon in the data set, as the surgeon would not yet have an estimated preference. *Table 8* illustrates this example.

The construction of the three proposed surgeon preference instruments detailed in *Chapter 2*, *Propensity score methods*, resulted in the exclusion of 20, 30 and 50 previous surgeries per surgeon, respectively. In practical terms, a higher proportion of participants were excluded for instruments that required more surgeries for their estimation. For example, 17,857 patients were excluded when estimating based on the lead surgeon having performed 20 previous surgeries, 25,141 were excluded when based on 30 surgeries and 39,243 were excluded when based on 50 surgeries. As expected, these additional exclusions predominantly affected patients receiving TKR surgery, who accounted for 17,696 out of 17,857 (99.1%), 24,908 out of 25,141 (99.1%) and 38,860 out of 39,243 (99.0%) of the excluded patients, respectively.

Exclusions for each IV are reported later in this chapter.

TABLE 8 Illustrative example of the construction of preference-based IVs

Surgeon ID	Patient ID	Date of surgery	Treatment	Preference for UKR
12345	1	January 2010	TKR	N/A
12345	2	January 2010	UKR	N/A
12345	3	February 2010	TKR	N/A
12345	4	February 2010	TKR	N/A
12345	5	February 2010	UKR	N/A
12345	6	February 2010	TKR	N/A
12345	7	March 2010	TKR	N/A
12345	8	March 2010	TKR	N/A
12345	9	March 2010	TKR	N/A
12345	10	March 2010	TKR	N/A
12345	11	March 2010	TKR	0.20
12345	12	April 2010	TKR	0.20
12345	13	April 2010	TKR	0.10

N/A, not applicable.

Note
All data in the table are fake and not true patient data. The preference for UKR is calculated using the first 10 treatments of the surgeon.

Instrumental variable creation

Possible instruments for analysis were generated and tested against the IV assumptions. We tested three types of preference-based instruments (surgeon-, hospital- and region-based preference) and volume-, area- and calendar-time-based instruments.

Surgeon preference for unicompartmental knee replacement

Surgeon preference for UKR was an obvious IV option. It is equivalent to a widely used IV in drug safety research: physician prescription preference. Given the richness of the data available to us, we considered different units for calculating surgeon preference based on different surgeon categories in the NJR: lead surgeon, consultant surgeon and surgical unit. For each unit, we followed these steps:

1. sort patients by surgeon (lead, consultant or surgical unit) pseudonymised identifiers provided by the NJR and NHS Digital
2. sort patients by date of operation within each surgeon ID/cluster
3. exclude the number of (20, 30 or 50) first surgeries performed by each of the surgeons
4. calculate the preference for UKR surgery at the patient level as the proportion of patients within the previous number of (20, 30 or 50) surgeries who had received a UKR
5. categorise the estimated surgeon preference into two groups, high and low preference for UKR, using the instrument-specific median-estimated preference as a cut-off value.

This method accounted for time-varying preference, as a preference was assigned to each patient based on the previous number of (20, 30 or 50) patients operated on by the same surgeon, rather than all available data. *Table 9* shows the resulting data set for further analysis and testing.

TABLE 9 Illustrative example of the construction of the analytical data set for IV analyses. All data in the table are fake and not true patient data

Surgeon ID	Patient ID	Date of surgery	Treatment	Preference for UKR	Binary IV
12345	1	January 2010	TKR	N/A	N/A
12345	2	January 2010	UKR	N/A	N/A
12345	3	February 2010	TKR	N/A	N/A
12345	4	February 2010	TKR	N/A	N/A
12345	–				
12345	20	March 2010	TKR	N/A	N/A
12345	21	March 2010	TKR	0.20	High
12345	22	April 2010	TKR	0.20	High
12345	23	April 2010	TKR	0.10	Low
12345	–				
12345	56	November 2010	UKR	0.11	Low
12345	57	November 2010	TKR	0.14	High
12345	58	January 2011	TKR	0.15	High

N/A, not applicable.

Note
These are fake data produced for illustrative purposes. The fake data set's median preference is set to 0.12, which is then used to categorise the instrument into two groups.

Other preference-based instrumental variables

We estimated hospital preference for UKR surgery in a similar way to surgeon-level preference, but instead based it on the first 20, 30 or 50 surgeries performed in each hospital. We also estimated regional preference based on the first 20, 30 or 50 surgeries performed in each region. The same general principles and steps that were used for the surgeon-level instruments were used to estimate hospital- and region-based preferences.

Volume-based instrumental variables

We calculated and tested two kinds of volume-based variables: total number of UKR and TKR surgeries performed by a surgeon (lead surgeon, consultant surgeon or surgical unit) in the whole cohort study period and the total number of surgeries performed by a surgeon in the previous year. We dichotomised these variables ($\geq$ or $<$ the median) into high- and low-volume surgeons.

Area-based instrumental variables

We tested two area-based instruments: area of treatment and area of residence. We used the government office region (of treatment or of residence) recorded in the HES data set. We calculated the median prevalence of UKR surgery using the data available for the full cohort. We then dichotomised regions as high or low uptake of UKR based on the median prevalence.

Calendar time

Calendar time is used in drug safety research as an instrument when there are clear changes in a medicine's secular trends of use over time, such as when a new product is approved for use or when the conditions of use change dramatically over time. We attempted to identify such a change in the use/uptake of UKR surgery in our analytical data set. However, although the use of UKR surgery increased from 2011, the change was minimal, from an average prevalence of 6.64% before 2011 to 7.33% after 2011 (*Figure 12*). We, therefore, halted our attempt to build and test a calendar-time-based instrument.

Instrumental variable selection

Once generated, each of the proposed IVs was shortlisted for analysis based on instrument diagnostics. The proposed diagnostics tested the first and third assumptions for IVs, detailed in *Chapter 2*, *Instrumental variable analyses*:

- The IV must be strongly associated with the exposure of interest.
- The IV must be independent of confounders.

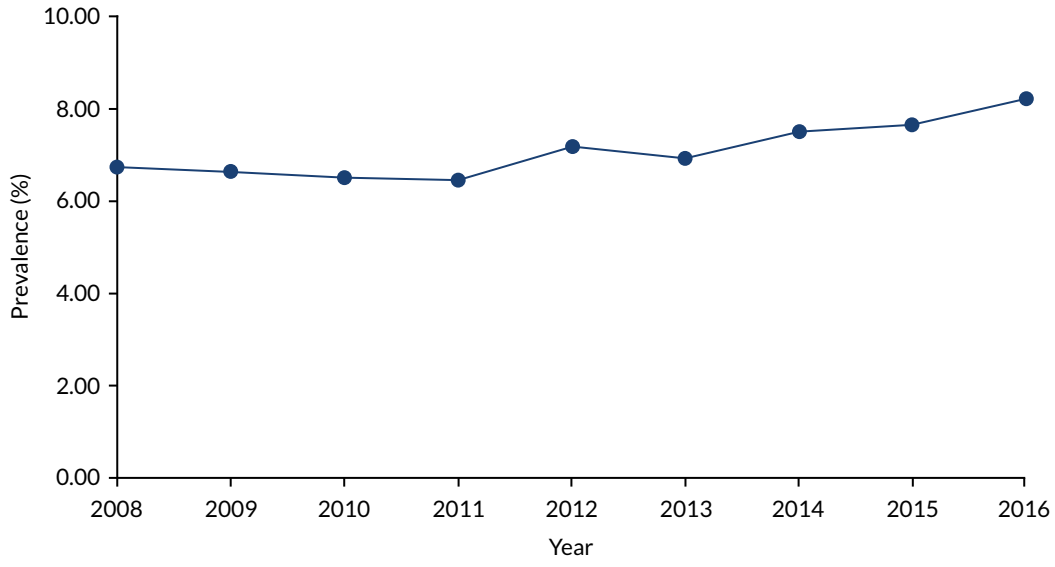


FIGURE 12 Secular trends in the prevalence (%) of UKR (vs. TKR) in the analytical data set per calendar year.

Although the latter assumption can be tested only for known or recorded confounders, violation of this assumption for known confounders would immediately rule out that instrument. We tested this assumption using the estimated ASMD in the known confounders between high-preference and low-preference groups. Instruments with an ASMD of ≤ 0.10 were assumed to have balanced the known confounders across the groups, as proposed by Ali *et al.*,[52] indicating that the instrument was independent of that confounder. Instruments with an ASMD of > 0.10 for any of the known confounders, indicating imbalance, were not taken forward for the final analysis.

Many approaches have been proposed to characterise the strength of the association between an IV and an exposure of interest. One of the more intuitive approaches is to use the *F*-statistic: the odds ratio from a logistic regression model based on the instrument as an independent variable and the outcome as a dependent variable. Previous simulation studies[18,46] have demonstrated that an instrument strength equivalent to an odds ratio of > 2 gives unbiased results. We followed this approach and took forward instruments with an estimated odds ratio of > 2.0 only.

The summary estimates obtained from these two tests and the decision of whether or not to shortlist each tested instrument for use in the final analysis are reported in *Table 10*.

Surgeon-based preference instrumental variables

Three surgeon-level preferences were estimated: lead surgeon, consultant surgeon and surgical-unit preferences. Instruments were built using the 20, 30 and 50 previous surgeries for each preference.

TABLE 10 Summary of diagnostics for each of the tested instruments

Unit	IV	Per cent additionally excluded	Odds ratio (95% CI)	*F*-statistic	Maximum ASMD	Short-listed
Lead surgeon	Last 20 preference	14.10%	12.34 (10.25 to 14.88)	1202.9	0.097	✗
	Last 30 preference	19.80%	16.96 (13.38 to 21.77)	1035.85	0.089	✗
	Last 50 preference	30.90%	25.15 (17.84 to 36.59)	754.51	0.083	✗
Consultant surgeon	Last 20 preference	7.20%	10.34 (8.71 to 12.26)	1155.47	0.108	
	Last 30 preference	11.40%	13.81 (11.17 to 17.23)	1023.41	0.098	✗
	Last 50 preference	20.50%	21.52 (15.78 to 30.08)	782.35	0.091	✗
Surgical unit	Last 20 preference	0.60%	2.58 (2.30 to 2.90)	279.43	0.136	
	Last 30 preference	1.00%	2.72 (2.40 to 3.08)	273.31	0.114	
	Last 50 preference	2.10%	2.80 (2.46 to 3.18)	277.42	0.126	
Lead surgeon	Total experience	0%	1.20 (1.07 to 1.35)	9.65	0.059	
	Yearly experience	0%	1.04 (0.93 to 1.17)	0.42	0.059	
Consultant surgeon	Total experience	0%	0.99 (0.88 to 1.11)	0.05	0.065	
	Yearly experience	0%	0.87 (0.77 to 0.98)	5.86	0.063	
Surgical unit	Total experience	0%	0.79 (0.70 to 0.88)	17.09	0.092	
	Yearly experience	0%	0.73 (0.65 to 0.82)	28.55	0.08	
	Area of residence	0%	1.37 (1.22 to 1.53)	29.05	0.158	
	Area of treatment	0%	1.67 (1.48 to 1.88)	75.44	0.144	

All of the estimated surgeon-based preference instruments were associated with the exposure (see *Table 10*), with odds ratios ranging from 3.82 (95% CI 3.71 to 3.93) for surgical unit preference based on the previous 20 surgeries to 29.54 (95% CI 2.50 to 31.80) for lead surgeon preference based on the previous 30 surgeries.

However, all of the surgical unit-based and one of the consultant surgeon-based (20 surgeries) preference instruments resulted in unacceptable imbalance (ASMD of > 0.1) for at least one known confounder. Socioeconomic deprivation was the most commonly imbalanced confounder. Baseline characteristics for each of the prespecified confounders stratified by instrument status are reported in *Tables 11–13*. These four instruments were rejected and the remaining five instruments (lead surgeon preference based on 20, 30 and 50 surgeries, and consultant surgeon preference based on 30 and 50 surgeries) were shortlisted for further testing.

TABLE 11 Covariate balance for a selected list of confounders stratified by lead surgeon preference for UKR surgery, estimated based on the previous 20, 30 and 50 surgeries

	ASMD based on		
Covariate	**20 previous surgeries**	**30 previous surgeries**	**50 previous surgeries**
Sex	0.033	0.027	0.026
Age at primary surgery	0.037	0.041	0.042
BMI	0.012	0.019	0.017
IMD socioeconomic status	0.097	0.089	0.083
Preoperative OKS	0.038	0.031	0.017
Myocardial infarction	0.020	0.019	0.022
Heart failure	0.002	0.009	0.004
Peripheral artery disease	0.008	0.008	0.004
Cerebrovascular disease	0.006	0.007	0.008
Dementia	0.007	0.008	0.009
Respiratory/pulmonary disease	0.006	0.010	0.006
Peptic ulcer	0.000	0.001	0.003
Mild liver disease	0.002	0.000	0.002
Severe liver disease	0.005	0.001	0.006
Diabetes	0.026	0.021	0.019
Diabetes with complications	0.016	0.012	0.012
Hemi/paraplegia	0.011	0.012	0.006
Chronic kidney disease	0.003	0.004	0.009
Solid tumours/malignancies	0.001	0.002	0.001
Metastatic cancer	0.008	0.011	0.016
Foot, hip or spinal pain	0.006	0.006	0.009
Previous arthroscopy	0.021	0.034	0.040
Hip osteoarthritis	0.010	0.014	0.019
Previous knee washout	0.020	0.014	0.012
Hip replacement	0.015	0.016	0.022
Previous knee injections	0.015	0.002	0.001

TABLE 12 Covariate balance for a selected list of confounders stratified by consultant surgeon preference for UKR surgery, estimated based on the previous 20, 30 and 50 surgeries

	ASMD based on		
Confounder	20 previous surgeries	30 previous surgeries	50 previous surgeries
Sex	0.031	0.029	0.026
Age at primary surgery	0.012	0.019	0.021
BMI	0.012	0.018	0.010
IMD socioeconomic status	0.108	0.098	0.091
Preoperative OKS	0.043	0.038	0.030
Myocardial infarction	0.016	0.018	0.022
Heart failure	0.007	0.011	0.008
Peripheral artery disease	0.005	0.010	0.011
Cerebrovascular disease	0.003	0.004	0.006
Dementia	0.011	0.009	0.010
Respiratory/pulmonary disease	0.004	0.010	0.004
Peptic ulcer	0.006	0.008	0.005
Mild liver disease	0.008	0.003	0.001
Severe liver disease	0.012	0.012	0.001
Diabetes	0.025	0.022	0.013
Diabetes with complications	0.016	0.012	0.012
Hemi/paraplegia	0.013	0.012	0.010
Chronic kidney disease	0.000	0.002	0.007
Solid tumours/malignancies	0.003	0.002	0.000
Metastatic cancer	0.008	−0.011	0.009
Foot, hip or spinal pain	0.005	0.006	0.010
Previous arthroscopy	0.016	0.026	0.037
Hip osteoarthritis	0.009	0.012	0.015
Previous knee washout	0.029	0.024	0.016
Hip replacement	0.012	0.013	0.017
Previous knee injections	0.015	0.006	0.016

TABLE 13 Covariate balance for a selected list of confounders stratified by surgical unit preference for UKR surgery, estimated based on the previous 20, 30 and 50 surgeries

	ASMD based on		
Confounder	20 previous surgeries	30 previous surgeries	50 previous surgeries
Sex	0.016	0.030	0.032
Age at primary surgery	0.038	0.038	0.048
BMI	0.000	0.011	0.010
IMD socioeconomic status	0.136	0.114	0.126
Preoperative OKS	0.061	0.049	0.056
Myocardial infarction	0.004	0.009	0.005
Heart failure	0.007	0.002	0.006
Peripheral artery disease	0.013	0.006	0.011

TABLE 13 Covariate balance for a selected list of confounders stratified by surgical unit preference for UKR surgery, estimated based on the previous 20, 30 and 50 surgeries (*continued*)

Confounder	ASMD based on		
	20 previous surgeries	30 previous surgeries	50 previous surgeries
Cerebrovascular disease	0.001	0.007	0.006
Dementia	0.001	0.005	0.008
Respiratory/pulmonary disease	0.006	0.001	0.000
Peptic ulcer	0.000	0.004	0.008
Mild liver disease	0.012	0.014	0.016
Severe liver disease	0.018	0.013	0.016
Diabetes	0.019	0.019	0.024
Diabetes with complications	0.001	0.007	0.005
Hemi/paraplegia	0.008	0.006	0.010
Chronic kidney disease	0.004	0.006	0.009
Solid tumours/malignancies	0.019	0.019	0.021
Metastatic cancer	0.003	0.005	0.003
Foot, hip or spinal pain	0.000	0.003	0.005
Previous arthroscopy	0.023	0.024	0.026
Hip osteoarthritis	0.008	0.001	0.004
Previous knee washout	0.041	0.038	0.042
Hip replacement	0.008	0.006	0.008
Previous knee injections	0.033	0.031	0.025

Volume-based instrumental variables

Similar to preference, surgeon experience with UKR surgery was estimated by lead surgeon, consultant surgeon and surgical unit. None of these instruments had a strong enough association with the exposure for further analysis, as based on the prespecified threshold of an odds ratio of > 2.0 (see *Table 10*). Odds ratios for these instruments ranged from 0.73 (95% CI 0.65 to 0.82) for surgical unit based on the previous year to 1.20 (95% CI 1.07 to 1.35) for lead surgeon overall experience. However, all of these instruments resulted in acceptable confounder imbalances, with ASMDs well below the prespecified threshold. The highest ASMDs ranged from 0.059 for lead surgeon yearly and total experience to 0.092 for surgical unit total experience.

As the first assumption was violated, none of the volume-based instruments was taken forward for further analysis.

Area-based instrumental variables

Two area-based instruments were estimated based on the patient's area of residence and the hospital/treatment centre in which the knee replacement operation took place. Both instruments were too weakly associated with the exposure, with an odds ratio of 1.37 (95% CI 1.22 to 1.53) for residence and 1.67 (95% CI 1.48 to 1.88) for hospital/treatment centre (see *Table 10*). Neither area-based instrument reduced confounding for known variables, with maximum imbalances recorded in preoperative OKS (ASMD of 0.16) and socioeconomic status measured with IMD (ASMD of 0.14), respectively.

Neither area-based IV was selected for further analysis.

Calendar time

As explained in *Calendar time*, no calendar time could be identified for use as an IV, as no strong changes in secular trends of UKR surgery uptake were identified in the study period.

Instrumental variables selected for further analysis

After applying the prespecified criteria, five IVs were taken forward for further analysis:

- lead surgeon preference for UKR –
  - based on the previous 20 surgeries
  - based on the previous 30 surgeries
  - based on the previous 50 surgeries
- consultant surgeon preference for UKR –
  - based on the previous 30 surgeries
  - based on the previous 50 surgeries.

Results from the selected instrumental variables

Figure 13 shows the two-stage regression results for the association between UKR and postoperative OKS (primary outcome) for the five selected instruments compared with the TOPKAT results. All of the proposed instruments gave different results to TOPKAT. None of the estimates or their CIs overlapped with the main estimate obtained from TOPKAT or its upper or lower CI limits.

Quantitative estimates suggested that these results departed significantly from those obtained from TOPKAT, with tau^2 estimates ranging from 85.3 (consultant surgeon preference based on the previous 50 surgeries) to 190.88 (lead surgeon preference based on the previous 20 surgeries), I^2 ranging from 92.7% to 97.7% for the same instruments and all chi-squared test p-values < 0.001 (*Table 14*).

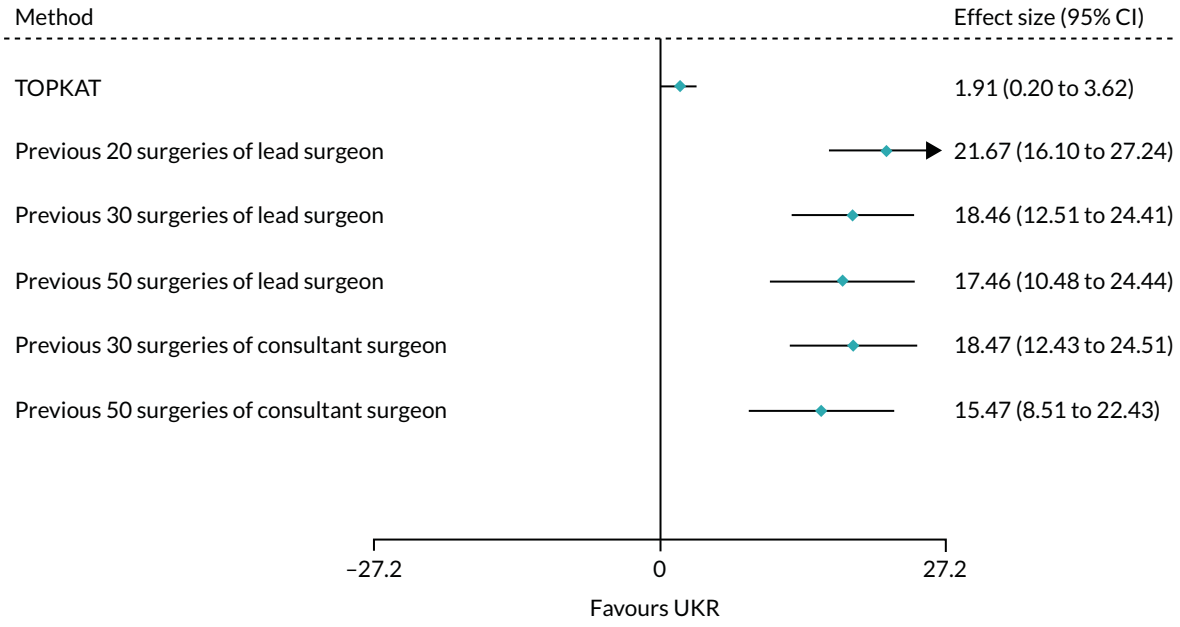


FIGURE 13 Association between UKR (vs. TKR) and postoperative OKS recorded in TOPKAT and estimated with IV analysis using the five shortlisted IVs.

TABLE 14 Consistency of results obtained from IV analyses compared with TOPKAT findings

Instrument	τ^2	I^2	Chi-squared test *p*-value
Lead surgeon			
Preference from last 20 surgeries	190.88	97.70%	< 0.001
Preference from last 30 surgeries	132.02	96.40%	< 0.001
Preference from last 50 surgeries	114.17	94.40%	< 0.001
Consultant surgeon			
Preference from last 30 surgeries	131.94	96.30%	< 0.001
Preference from last 50 surgeries	85.27	92.70%	< 0.001

Conclusions from instrumental variable analysis

Only 5 of the 17 tested potential instruments passed our diagnostic tests and were eligible for two-stage regression analysis. Of the instruments that failed, four failed owing to residual confounding (one or more variables with an imbalance of > 0.1), six owing to instrument weakness (odds ratio of > 2 in the association between instrument and UKR exposure) and two (both area-based instruments) owing to both measures.

The surgeon preference-based instruments that failed owing to unresolved confounding did so because of an imbalance in socioeconomic status. Surgeon preference for UKR surgery may be geographically determined and associated to some degree with socioeconomic status. Previous studies have reported on the heterogeneity of UKR surgery use nationally and on the determinants of surgeon[47] and patient[48] choice. There is evidence for inequality in access to knee replacement generally, assessed as provision versus need.[49] However, to our knowledge, there are no data available on the potential heterogeneity in access to UKR nationally and/or globally and on its effect on patient outcomes. A recent study by Garriga *et al.*[50] reported geographical variation in outcomes of primary knee replacement, with greater surgical volume (by surgeon and hospital) associated with better patient outcomes, and UKR surgery associated with a lower risk of complications than TKR surgery.

All of the volume-based and regional/area-based instruments tested were not sufficiently strongly associated with the use of UKR. The area-based instruments also resulted in residual imbalances in preoperative OKS and socioeconomic status.

In conclusion, none of the tested IVs yielded results comparable with those obtained from TOPKAT. We, therefore, did not use IV analysis in UTMoSt stage 2. The reasons underlying this result could form the basis of future methodological research. They could include the violation of untested assumptions (e.g. a direct association between surgeon preference for UKR and postoperative OKS), the presence of residual confounding for unobserved variables or the non-normal distribution of OKSs.

Chapter 5 Conclusions from UTMoSt stage 1

Study participants identified from NHS routine practice and their eligibility for surgical randomised controlled trials

Applying the TOPKAT ASA grade inclusion criterion (no patients with an ASA grade of > 2) to our real-world set of knee replacement patients (NJR linked to HES and PROMs) excluded 75,074 out of 457,577 (16%) patients, which was expected from NJR annual reports. Another 87,947 (23% of the remaining 382,503) patients were excluded because of other TOPKAT exclusion criteria.

Although TOPKAT was a relatively pragmatic trial, about one-third of the patients who receive a knee replacement in the NHS would not have been eligible for the trial, in some cases because of indication for UKR, as about 50% of TKR patients were not eligible for UKR and, therefore, would not have been eligible for TOPKAT. Of particular concern is the use of RCT-based efficacy and safety data for patients with multiple comorbidities and/or complex health needs. Overall, the proportion excluded was lower than what has been observed when users of widely used medicines are compared with participants in pivotal RCTs.[53] However, it still raises concerns about the external validity of RCT findings when applied to the general population and patients with severe comorbidity.

Surgical RCTs are potentially limited by the participation of more academic and specialised surgeons, hospitals and treatment centres. Including only patients operated on in the NHS by surgeons who had performed ≥ 10 UKR surgeries in the previous year (following the TOPKAT published protocol)[28] excluded almost half of the patients undergoing UKR surgery in the NHS from our data set (see *Table 6*). This surgeon-based exclusion preferentially affected the treatment under study: UKR. One-third of surgeons would not have been eligible for inclusion in TOPKAT based on their lack of previous experience with TKR, but up to two-thirds would have been excluded because of their lack of previous experience with UKR.

Results from propensity score analyses

Covariate balance

Propensity score methods are one of the recommended approaches for minimising confounding in drug safety and comparative effectiveness research using observational data. Although in principle these methods are also useful for studying medical devices like the two types of knee replacement here, uncertainty remains in how PSs should be used for surgical epidemiology.

We have demonstrated that some PS-based methods were able to replicate the efficacy results of TOPKAT, which is considered the gold standard when studying comparative efficacy. We tested methods that estimated ATE and ATT. Although all of the tested methods improved the balance between UKR and TKR patients, some failed to achieve sufficient balance (ASMD of < 0.1) for all of the prespecified confounders. When analysing the primary outcome (postoperative OKS), PS stratification based on the full cohort and IPW resulted in unresolved imbalances (ASMD of ≥ 0.1) for at least one confounder. PS matching and stratification based on the UKR cohort reached good balance for all measured confounders. Unfortunately, balance for measured confounders is not equivalent to comparability or exchangeability because unmeasured/unresolved confounding is always possible in observational analyses, regardless of the use of PSs and IPW.

Concordance between propensity score analyses and TOPKAT results

In accordance with our protocol, and applying patient-level eligibility criteria, our main primary outcome analyses (postoperative OKS) found a small, but not clinically relevant, additional benefit for UKR over TKR. This effect ranged from an ATE estimate of 0.10 (PS non-linear adjustment) to 0.76 (PS stratification based on the UKR cohort) points in postoperative OKS, compared with 1.91 points in TOPKAT.

Most of the tested methods did not have a significant chi-squared test result ($p < 0.05$), suggesting that any differences between the estimated results and TOPKAT results were due to chance. The exceptions were linear and non-linear PS adjustment, which both had a chi-squared test *p*-value of 0.05. As detailed in *Chapter 2*, such binary testing is dependent on power. We also used I^2 and τ^2 to quantitatively assess the methods' ability to mimic TOPKAT. I^2 values ranged from 35% (PSS_{exp}) to 74% (non-linear PS adjustment). Using a prespecified but poorly justified threshold of $I^2 < 40\%$, PSS_{exp} was the preferred method ($I^2 = 35\%$, chi-squared test $p = 0.21$, $\tau^2 = 0.23$), followed by IPW ($I^2 = 48\%$, chi-squared test $p = 0.17$, $\tau^2 = 0.43$) and PSS_{whole} ($I^2 = 53\%$, chi-squared test $p = 0.14$, $\tau^2 = 0.48$).

We conducted a prespecified sensitivity analysis that included only participants who had been operated on by surgeons who had performed ≥ 10 surgeries of the same type in the previous year. This inclusion restriction drove the treatment estimates from all three methods closer to the TOPKAT findings: ATE of 1.37 (95% CI 0.54 to 2.20) for PSS_{exp}, 1.37 (95% CI 0.54 to 2.20) for PSS_{whole} and 1.32 (95% CI 0.32 to 2.33) for IPW, compared with 1.91 (95% CI 0.20 to 3.62) in TOPKAT.

PSS_{whole}, PSS_{exp} and IPW were considered valid methods following our prespecified criteria, with $I^2 = 0\%$ and $\tau^2 = 0.0$ for all three and high chi-squared test *p*-values of 0.58, 0.58 and 0.56, respectively.

Our results suggest that replication of surgical trials requires identification not only of 'eligible' patients/participants but also of potentially eligible surgeons to preclude learning curves and surgical volume-related performance. The results also suggest that PSS_{whole}, PSS_{exp} and IPW are valid methods for replicating the TOPKAT findings, when applied to the trial-eligible population of patients and surgeons.

Additional methods have recently been proposed for replicating RCTs using observational data, including ensuring that the obtained treatment effect estimates lie within the 95% CI of the trial's estimate[57] and ensuring statistical significance agreement.[58] All of the treatment effect estimates obtained from the full OKS cohort fell within the 95% CIs obtained from TOPKAT (95% CI 0.20 to 3.62), except the estimates from linear and non-linear PS adjustment. However, none of the estimates' 95% CIs was totally covered by the TOPKAT 95% CI. The results from PSS_{exp} were the closest, with 95% CI 0.15 to 1.36 versus TOPKAT's 95% CI 0.20 to 3.62. The sensitivity analyses that were restricted to surgeons eligible for the trial resulted in 95% CIs that were fully covered by the TOPKAT 95% CI for all three tested methods (PSS_{whole}: 95% CI 0.54 to 2.20; PSS_{exp}: 95% CI 0.54 to 2.20; IPW: 95% CI 0.32 to 2.33; versus TOPKAT: 95% CI 0.20 to 3.62). This result reiterates the importance of considering surgeon-related eligibility criteria for any future research that replicates surgical or medical device RCTs using real-world data. The outcome of UKR observed in the overall cohort could also be potentially improved by centralising such procedures in the hands of treatment centres and/or surgeons who have performed a larger number of such surgeries, as was undertaken in TOPKAT.

If we consider agreement in statistical significance, the only valid method was PSS_{exp}, as it yielded a statistically significant effect in favour of UKR similar to the TOPKAT findings. When including only patients operated on by surgeons who would have been eligible for the trial, PSS_{whole}, PSS_{exp} and IPW all obtained a statistically significant positive effect for UKR on postoperative OKS, similar to TOPKAT. All three methods would be considered valid under this additional criterion when restricting the analysis to the population of surgeons and patients eligible for TOPKAT.

Table 15 summarises the tests used to assess the validity of the proposed analytical methods in the overall OKS cohort and the subset of patients operated on by surgeons considered eligible for TOPKAT.

In summary, only PS stratification based on the exposed (UKR) cohort replicated the TOPKAT findings according to all known criteria when the whole cohort of patients eligible for the trial was analysed. However, three of the proposed PS methods (stratification based on the whole cohort or the exposed/UKR participants and IPW) successfully replicated TOPKAT for the primary outcome analysis when the analysis was restricted to patients operated on in the NHS by surgeons with sufficient experience to have been eligible for TOPKAT. We, therefore, selected these three methods for UTMoSt stage 2, in which we focused on the patients who were not eligible for TOPKAT.

Results from instrumental variable analysis

Assumptions and diagnostics

To determine whether or not IV analysis was a valid method, we tested two of the assumptions underlying IV analysis:

1. the association between the exposure (UKR) and the instrument, as a measure of instrument strength
2. covariate balance as a proxy for the lack of association between an instrument and known/recorded confounders.

Instrument strength

Logistic regression analysis between the built instruments and the exposure (UKR surgery) suggested that all of the surgeon-based preference variables were strong instruments, with odds ratios ranging from > 2.5 to > 25. Experience-based and area-based (areas of residence and treatment) variables were weak instruments, with an odds ratio of < 2 or, in some cases, no association with the exposure (odds ratio close to 1). These weak instruments were rejected for stage 2 analyses as they violated the first of the tested assumptions.

TABLE 15 Summary of the validity of the proposed methods for replicating the surgical RCT, TOPKAT, in the whole OKS cohort and in the sensitivity analysis restricted to patients operated on by surgeons with sufficient experience to participate in the RCT

	Whole OKS cohort				Sensitivity analysis (eligible surgeons)			
Proposed method	Chi-squared test	I^2	Coverage	SSA	Chi-squared test	I^2	Coverage	SSA
PSM	✓	✗	✓	✗	–	–	–	–
PS_{whole}	✓	✗	✓	✗	✓	✓	✓	✓
PSS_{exp}	✓	✓	✓	✓	✓	✓	✓	✓
IPW	✓	✗	✓	✗	✓	✓	✓	✓
PS_{Alin}	✓	✗	✗	✗	–	–	–	–
$PS_{Anonlin}$	✓	✗	✗	✗	–	–	–	–

PS_{Alin}, propensity score linear adjustment; $PS_{Anonlin}$, propensity score non-linear adjustment; PSM, propensity score matching; SSA, statistical significance agreement; ✗, failed; ✓, passed.

Notes
Chi-squared test: $p < 0.05$; I^2: I^2 below the prespecified threshold of 40%; coverage: treatment effect estimate included in the 95% CI of the treatment effect obtained from the RCT.

Among the preference-based variables, instrument strength (based on the estimated odds ratio) increased with the number of surgeries used to estimate preference. Surgeon-level preference was a stronger instrument than surgical unit-level preference for UKR, with an odds ratio of > 10 versus 2–3, respectively. Similar effects were seen when *F*-statistics were used instead of odds ratios. However, instruments estimated based on a larger number of surgeries resulted in a higher proportion of patient exclusions. For example, lead surgeon preference estimated based on 20 surgeries excluded just below 15% of the eligible patients, whereas preference based on 50 surgeries excluded almost 31% of the eligible patients. The potential effect of these exclusions on selection bias and/or external validity needs investigation, and should be considered in future research.

Covariate balance

Most of the proposed IVs achieved satisfactory covariate balance for the known confounders of interest, as defined by a prespecified threshold of an ASMD of < 0.10. The area-based instruments (areas of residence and treatment), surgical unit preference and consultant surgeon preference based on 20 surgeries (but not based on 30 or 50 surgeries) consistently failed to achieve covariate balance. Socioeconomic status was the most commonly imbalanced confounder in all of these analyses.

Combining the results of the instrument strength and confounder balance falsification tests resulted in the selection of five IVs for analysis: the three lead surgeon-based preference instruments (based on 20, 30 and 50 previous surgeries) and two of the consultant surgeon-based preference instruments (based on 30 and 50 previous surgeries). All other tested IVs violated either one or both of the IV assumptions.

Concordance between instrumental variable analysis and TOPKAT findings

The five selected IV analyses failed to fulfil the two prespecified criteria for selection for UTMoSt stage 2. As shown in *Table 16*, all five selected IV analyses resulted in chi-squared test *p*-values of < 0.001, suggesting that the observed differences between their results and TOPKAT were unlikely to be because of chance. All five gave $I^2 > 90\%$, which was well above the predefined threshold of 40%, suggesting heterogeneity in the obtained findings compared with the trial results.

All five also failed the coverage criterion, as none of the five treatment effect estimates (ranging from 15.47 to 21.67 points in OKS) was covered by the 95% CI obtained from TOPKAT (95% CI 0.20 to 3.62). All five IV analyses passed the statistical significance agreement test, as they all found a statistically significant improvement in postoperative OKS favouring UKR, just like in TOPKAT.

TABLE 16 Summary of the validity of each of the shortlisted IV analyses for replicating TOPKAT

IV analyses	Chi-squared test	I^2	Coverage	SSA
Lead surgeon preference				
Last 20 surgeries	✗	✗	✗	✓
Last 30 surgeries	✗	✗	✗	✓
Last 50 surgeries	✗	✗	✗	✓
Consultant surgeon preference				
Last 30 surgeries	✗	✗	✗	✓
Last 50 surgeries	✗	✗	✗	✓

SSA, statistical significance agreement; ✗, failed; ✓, passed.

Notes
Chi-squared test: *p*-value of < 0.05; I^2: I^2 below the prespecified threshold of 40%; coverage: treatment effect estimate included in the 95% CI of the treatment effect obtained from the RCT.

However, this criterion does not take into account the magnitude of the observed effect; therefore, it failed to detect the difference in results from the conducted IV analyses and TOPKAT.

In summary, although five instruments were shortlisted, none passed the two prespecified criteria. None of the IV analyses was used in UTMoSt stage 2.

Strengths and limitations

To our knowledge, this study is the first attempt to investigate different statistical methods that account for confounding in trial replications. A trial duplication study[59] run by the FDA and colleagues from a number of academic institutions focused on a framework of trial replication rather than method comparisons.

Our study team was blinded to TOPKAT findings during the analyses of PS-based methods and IVs, guarding against bias. We also used the same outcomes and outcome analysis methods as TOPKAT to ensure a practical comparison.

Our stage 1 analyses have a number of limitations. First, one of the key TOPKAT inclusion criteria was that patients had medial compartment osteoarthritis with exposed bone on both the femur and the tibia. However, the NJR database does not record the indication for a surgery. We unsuccessfully attempted to emulate this criterion by exploring osteoarthritis records in the HES data, but very few UKR or TKR patients had osteoarthritis recorded in the HES database. Such information is more likely to be recorded in the primary care consultation database, but linkage to primary care data is not routinely available. Despite our best efforts, our stage 1 populations (both revision and OKS cohorts) might, therefore, have differed from the TOPKAT participants. Compared with TOPKAT participants, TKR patients included in our stage 1 cohorts were generally older (mean age of 70.2 years and 70.4 years in the revision and OKS cohorts, respectively, vs. 65.2 years in TOPKAT) and had noticeably higher preoperative EQ-5D scores. These participant differences might have led to differences in the results obtained from UTMoSt stage 1 and TOPKAT. It is, therefore, reassuring that some of the proposed methods replicated TOPKAT successfully, but it is possible that some of the methods deemed invalid (e.g. PS matching) could have obtained more similar findings to TOPKAT in a more ideal scenario.

Second, clinical covariates included in the PS were based on HES inpatient data, implying that residual confounding might have been an issue owing to unmeasured variables. For example, osteoarthritis consulted in primary care would not have been identified in the HES inpatient database. Nevertheless, our linked observational data with IPW, PSS_{exp} and PSS_{cohort} yielded similar results to the gold standard TOPKAT estimates for the postoperative OKS. PS-based methods, therefore, have potential for trial replication or generalisation, and some of the included covariates might act as proxies for the unavailable confounders.

Third, there was a small difference in the timing of postoperative OKS collection in TOPKAT (1 year after randomisation) and UTMoSt (6–12 months after surgery). This difference might also have contributed to the differences in results in some of the analyses.

Finally, we estimated PSs using the most commonly used technique, logistic regression. Further research could explore alternative methods, such as machine learning, large-scale PSs or multilevel PSs. In PS matching, we used only calliper matching to form matched cohorts. Although this method has performed well when treatment effects are examined,[35,36] PS matching with different algorithms may have led to results closer to the TOPKAT estimates.

We used only the IPW estimator to calculate treatment effect. Further studies could explore the augmented IPW estimator method in large samples. This approach requires PS estimates and separate outcome models for the exposed and unexposed cohorts. However, simulation studies have found that this method might not be suitable in small samples, such as our UKR OKS cohort.[60] Further studies could explore this method in large samples.

Conclusions and implications for UTMoSt stage 2

UTMoSt stage 1 demonstrated that surgical RCTs can be replicated using routine data recorded in actual NHS practice conditions. This is a pragmatic validation of methods routinely used in post-marketing drug safety observational research for surgical and medical device epidemiology, but does not imply that surgical RCTs are no longer needed. In fact, UTMoSt stage 1 was possible only because of the existence of a surgical RCT (TOPKAT) and good-quality routine data from the NJR linked to patient-reported outcomes and hospital inpatient records.

Some, but not all, of the methods tested obtained treatment effect estimates comparable to those obtained from TOPKAT for the trial's primary outcome (postoperative OKS). When focused on the target population, only PS stratification based on the distribution of the PS in the exposed (UKR) arm (PSS_{exp}) successfully replicated TOPKAT according to all of the criteria used (chi-squared test, I^2 for heterogeneity, coverage, and statistical significance agreement). Two other methods (PSS_{whole} and IPW) replicated the trial according to two of the four criteria (chi-squared test and coverage), but did not pass the heterogeneity (prespecified, $I^2 < 40\%$) or statistical significance agreement tests.

Our findings demonstrate challenges when replicating surgical trials, compared with ongoing international efforts to replicate RCT findings on the effects of medicines. The effects of surgeon expertise on outcome can affect the replicability of surgical RCTs, which are typically conducted in specialised treatment centres by experienced surgeons who can deliver both procedures and are in equipoise. Restricting the UTMoSt stage 1 analyses to patients operated on by surgeons with the volume required to participate in TOPKAT drove the obtained treatment estimates much closer to those seen in the trial, compared with using all eligible patients. Three tested methods (PSS_{exp}, PSS_{whole} and IPW) were able to replicate the TOPKAT findings in accordance with all four criteria when using the restricted population (chi-squared test p-values > 0.5, $I^2 = 0\%$, $\tau^2 = 0.0$, and statistical significance agreement and coverage). These three methods were selected for the UTMoSt stage 2 analysis, which is reported in *Chapters 6–8*.

Unfortunately, none of the proposed IVs passed the two prespecified tests for concordance with TOPKAT. The reasons underlying this failure warrant further investigation, for which we will seek funding from other streams focused on methodological research. Shortlisted instruments passed falsification tests for two of three key assumptions. We, therefore, speculate that the direct effect of surgeon expertise on the primary outcome (postoperative OKS) observed in the sensitivity analysis could explain to some degree the failure of preference-based instruments to replicate TOPKAT. None of the built instruments was selected for UTMoSt stage 2.

Chapter 6 Stage 2 methods

UTMoSt stage 2 aimed to evaluate the risks, benefits, costs and cost-effectiveness of UKR (compared with TKR) for patients who would not have been eligible for the TOPKAT surgical RCT.

This chapter details the methods used in stage 2, focusing on where the methods differed from stage 1. We used the analysis methods that had been shown to replicate TOPKAT's findings in UTMoSt stage 1. These validated methods, PS stratification and IPW, were discussed in detail in *Chapter 2*, *Propensity score stratification*; *Chapter 2*, *Inverse probability weighting*; and *Chapter 3*. Stage 2 used the same observational data sources as stage 1, which were presented in detail in *Chapter 2*; however, the source population, study population, outcomes and statistical analyses differed from stage 1.

Target population

The target population for UTMoSt stage 2 was NJR participants undergoing primary UKR or TKR surgery who had severe comorbidities, defined as an ASA grade of 3 or 4 at the time of surgery.

Besides the general inclusion and exclusion criteria listed in *Table 1*, we also excluded patients who met any of the following exclusion criteria:

- NJR participants with no possible linkage to an episode in HES, as this was needed to study complications and costs
- NJR participants with no linked data available on pre or postoperative PROMs, as these were required for comparative effectiveness analyses
- NJR participants with previous cruciate ligament injury or inflammatory arthritis, as they would not have been eligible for UKR
- participants for whom revision costs could not be estimated because the linked HES episode did not provide valid/adequate information on a Healthcare Resource Group (HRG), as this was needed for the health economic analysis.

In stage 1, patients with a record of foot, hip or spinal pain in the 1 year before surgery were excluded. These patients were not excluded in stage 2, as this information was instead included as a PS covariate. We did not include the stage 1 exclusion criteria of prior knee surgery, patella dislocation or septic arthritis as PS covariates because too few patients had positive records.

Outcomes

UTMoSt stage 2 used the same primary outcome (postoperative OKS) as TOPKAT and UTMoSt stage 1. The secondary outcomes of interest were:

- 5-year risk of revision identified in the NJR or mortality identified in the HES data set
- 90-day risk of postoperative complications, including myocardial infarction, venous thromboembolism and prosthetic joint infection. These complications were identified using the primary diagnosis ICD-10 code in the HES data set. Code lists for these outcomes were prespecified based on previous research and are shown in *Tables 27–29*.[61]

Participants were followed up from the index surgery date to the earliest of:

- end of enrolment in the database or 31 December 2016
- date of revision surgery (for secondary outcome analyses)
- date of a surgery for the other knee (for secondary outcome analyses)
- death
- end of 5 years of observation after the index surgery date.

We censored people at the surgery date in case of contralateral knee replacement, which would have made it difficult to attribute any surgical complications and costs to a specific knee.

Statistical analyses

Patient-level characteristics of the included TKR and UKR patients were compared using ASMDs with a cut-off value of 0.1. Any remaining imbalance (ASMD of > 0.1) in patient-level characteristics was accounted for by including the non-balanced covariate in the subsequent outcome analyses.

The following analyses were conducted for each of the proposed outcomes:

- Primary outcome – differences in postoperative OKS between UKR and TKR patients were estimated using multilevel linear regression (cluster 1: lead surgeons; cluster 2: patients).
- Secondary outcomes – postoperative complications. For each 90-day risk of an adverse event, we compared the cumulative incidence of all adverse events of interest between UKR and TKR patients. Relative risk and 95% CIs were estimated using Poisson models with robust standard errors. Mortality was not considered a competing risk in the 90-day risk because of low mortality rates over this period.
- Secondary outcomes – mortality and revision risk. Incidence rates and 95% CIs of revision and mortality for UKR and TKR patients were estimated using Poisson models, with the jack-knife method for CI calculations, and reported per 1000 person-years. Cause-specific hazard models were fitted to estimate risk of revision or mortality, censoring patients when they had a competing event (revision or mortality).

All outcome analyses were conducted in each of the imputed data sets and combined using Rubin's rules.

The methods used to analyse hospital costs, health-related quality of life (HRQoL) and derived cost-effectiveness analyses are shown in *Chapter 8*.

Sensitivity analyses

No sensitivity analyses were conducted for postoperative OKS or any of the 90-day postoperative complications. As few patients had these outcomes, there was a lack of statistical power to detect a significant difference.

Sensitivity analyses were conducted for 5-year revision risk. Three predefined interactions were tested for using multiplicative terms in the above models. Stratified analyses by sex, age (younger or older than the median age in the study data sets) and ASA grade were reported if the p-value was < 0.1. To explore the impact of learning curves, revision analyses were restricted to surgeries undertaken by lead surgeons who had performed at least 10, 30 or 50 surgeries of the same type in the previous year.

Chapter 7 Stage 2 patient characteristics

Study population and participant flow

Of the 457,577 patients (UKR, *n* = 32,293; TKR, *n* = 425,284) available in the source-linked data, 383,522 patients (UKR, *n* = 29,403; TKR, *n* = 353,119) had an ASA grade of 1 or 2 and were, therefore, eligible for TOPKAT and UTMoSt stage 1, but excluded from stage 2 (*Figure 14*). The stage 2 exclusion criteria (see *Chapter 6*, *Target population*) excluded a further 15,117 patients. The resulting cohort for analysing safety outcomes (safety cohort) comprised 57,682 TKR patients and 2256 UKR patients. Of these patients, OKS postoperative data were available for 145 UKR and 23,344 TKR patients (OKS cohort).

Table 17 shows the unadjusted baseline characteristics for the safety and OKS cohorts. TKR patients in the OKS cohort had similar baseline characteristics to those in the safety cohort. UKR patients in the OKS cohort were healthier (34% vs. 38% had a Charlson Comorbidity Index score of 0) and more likely to live in the countryside (22% vs. 16% with a Rural Index of 3 or 4) than UKR patients in the safety cohort.

In the safety cohort, UKR patients were younger than TKR patients [mean age (SD): 69 (10) years vs. 73.5 (8.9) years, respectively] and were more likely to be men (57% vs. 44%, respectively) and live in the countryside (16% vs. 11%, respectively, with a Rural Index of 3 or 4) or the least deprived areas (26% vs. 18%, respectively). UKR patients were less likely than TKR patients to have a history of osteoarthritis and other joint problems (19% vs. 26%, respectively).

In the OKS cohort, there are similar noticeable differences in sex, Rural Index and socioeconomic status between UKR and TKR patients. UKR patients were also more likely than TKR patients to be healthy (Charlson Comorbidity Index score of 0) (34% vs. 39%, respectively). There was a vast difference in the number and proportion of UKR and TKR patients who responded to the postoperative OKS: 145 out of 2256 (6.4%) UKR patients versus 23,344 out of 57,682 (40.5%) TKR patients.

Covariate balance assessment

Oxford Knee Score cohort

We applied the three validated methods (IPW, PSS_{whole} and PSS_{exp}) to the OKS cohort to compare the treatment effect, as measured by the OKS, for UKR and TKR recipients.

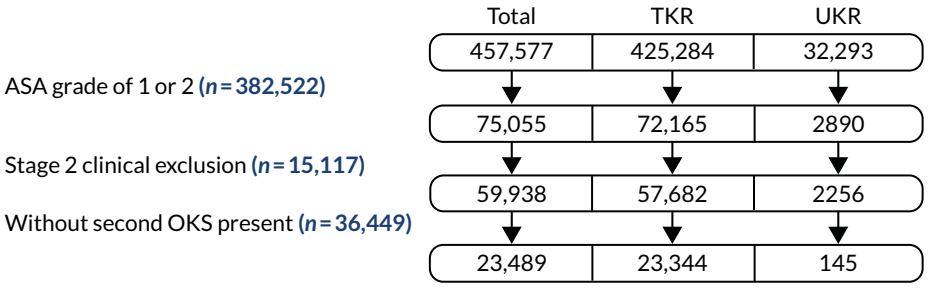


FIGURE 14 Stage 2-specific eligibility criteria and resulting patient selection.

TABLE 17 Baseline patient-level characteristics for patients who received TKR or UKR

Characteristic	Safety cohort		OKS cohort	
	TKR	UKR	TKR	UKR
Total number of patients, *N*	57,682	2256	23,344	145
Sex, *n* (%)				
Female	32,086 (56)	978 (43)	12,683 (54)	68 (47)
Male	25,596 (44)	1278 (57)	10,661 (46)	77 (53)
Rural Index, *n* (%)				
1	44,296 (77)	1629 (72)	17,626 (76)	97 (67)
2	6803 (12)	271 (12)	2926 (13)	16 (11)
3	4853 (8)	252 (11)	2067 (9)	21 (14)
4	1730 (3)	104 (5)	725 (3)	11 (8)
IMD, *n* (%)				
Least deprived 10%	4784 (8)	309 (14)	2026 (9)	16 (11)
Less deprived				
10–19%	5756 (10)	274 (12)	2464 (11)	20 (14)
20–29%	6281 (11)	246 (11)	2634 (11)	10 (7)
30–39%	6298 (11)	230 (10)	2683 (11)	20 (14)
40–49%	6391 (11)	268 (12)	2617 (11)	18 (12)
More deprived				
10–19%	5400 (9)	163 (7)	2011 (9)	11 (8)
20–29%	5570 (10)	166 (7)	2143 (9)	8 (6)
30–39%	5857 (10)	231 (10)	2307 (10)	18 (12)
40–49%	6205 (11)	230 (10)	2616 (11)	11 (8)
Most deprived 10%	5140 (9)	139 (6)	1843 (8)	13 (9)
ASA grade, *n* (%)				
P3: incapacitating systemic disease	56,625 (98)	2232 (99)	22,973 (98)	142 (98)
P4: life-threatening disease	1057 (2)	24 (1)	371 (2)	3 (2)
Charlson Comorbidity Index score, *n* (%)				
0	22,672 (39)	863 (38)	9162 (39)	50 (34)
1	18,369 (32)	750 (33)	7511 (32)	58 (40)
2	8665 (15)	349 (15)	3486 (15)	21 (14)
3	4476 (8)	172 (8)	1823 (8)	10 (7)
4	3500 (6)	122 (5)	1362 (6)	6 (4)
Age (years), mean (SD)	73.5 (8.9)	69.0 (10.0)	73.5 (8.6)	69.8 (10.2)
BMI (kg/m^2), mean (SD)	32.6 (6.4)	32.6 (6.1)	32.6 (6.3)	32.6 (6.1)
PROMs				
Preoperative OKS, mean (SD)	16.4 (7.6)	19.2 (8.0)	17.0 (7.6)	19.4 (8.6)
EQ-5D Health Scale, mean (SD)	61.8 (20.5)	63.7 (20.5)	62.7 (20.1)	63.7 (22.2)
EQ-5D, mean (SD)	0.3 (0.3)	0.4 (0.3)	0.3 (0.3)	0.4 (0.3)

TABLE 17 Baseline patient-level characteristics for patients who received TKR or UKR (*continued*)

Characteristic	Safety cohort		OKS cohort	
	TKR	UKR	TKR	UKR
General health, *n* (%)				
0	40,968 (71)	1399 (62)	16,522 (71)	78 (54)
1	9563 (17)	430 (19)	4052 (17)	35 (24)
≥ 2	7151 (12)	427 (19)	2770 (12)	32 (22)
Medical history, *n* (%)				
Gastrointestinal disease	16,270 (28)	584 (26)	6741 (29)	36 (25)
Osteoarthritis and other joint problems	15,064 (26)	420 (19)	6196 (27)	35 (24)
Mental health	7503 (13)	326 (14)	2819 (12)	14 (10)
Respiratory diseases	15,186 (26)	622 (28)	6024 (26)	37 (26)
Cardiovascular diseases	47,105 (82)	1745 (77)	19,269 (83)	110 (76)
Thyroid problems	6354 (11)	204 (9)	2630 (11)	8 (6)
Foot, hip or spinal pain	2220 (4)	76 (3)	897 (4)	4 (3)
Coxarthrosis	2354 (4)	58 (3)	969 (4)	6 (4)
Neurological disorders	7495 (13)	322 (14)	3014 (13)	18 (12)
Other arthrosis	4904 (9)	116 (5)	2005 (9)	13 (9)
Polyarthrosis	4390 (8)	95 (4)	1762 (8)	4 (3)
Spondylosis	2531 (4)	68 (3)	1039 (4)	2 (1)

Propensity score stratification based on the distribution of PSs in the whole cohort (PSS_{whole}) resulted in few UKR patients in some groups, leading to imbalanced PS distributions in some strata (see *Figure 22a*). Within-stratum covariate balance was not always achieved in all strata in each of the 10 imputed data sets. For example, across the 10 imputed data sets, all covariates had an ASMD of > 0.1 in stratum 1. This is not surprising because there were only two (0.09%) UKR patients in this stratum, which was defined by a low PS and, therefore, a low probability of UKR treatment. Strata 2–5 had between six and 17 UKR patients each, and the distribution of some covariates remained imbalanced when UKR and TKR patients were compared in some of the 10 imputed data sets. Within-stratum covariate balance improved in strata 6–10 because they were defined by a higher PS and probability of treatment; therefore, strata 6–10 included a larger number and higher proportion of UKR patients. However, the average ASMD across strata for each of the covariates was ≤ 0.1 (*Figure 15a*), indicating that good balance was achieved for all individual covariates across strata using this widely accepted, pre-defined threshold.

The PSS_{exp} stratified based on the distribution of PSs in the exposure (UKR) cohort and most accurately replicated the TOPKAT findings in UTMoSt stage 1. It resulted in equal numbers of UKR patients in each stratum, and obtained better balance than PSS_{whole} in the PS distribution between UKR and TKR patients (see *Figure 22b*). It also led to better within-stratum covariate balance for each of the identified confounders, although most covariates had an ASMD of > 0.1 in at least one stratum. Overall, average covariate balance across the 10 strata was achieved (*Figure 15b*).

The IPW pseudo-population included 145 UKR patients with a stabilised weight ranging from 0.08 to 4.45 (IQR 0.38–1.28) and 23,344 TKR patients with a stabilised weight close to 1 (minimum, 25th percentile, 75th percentile, maximum: 0.99, 1.00, 1.00, 1.10). Four covariates remained imbalanced

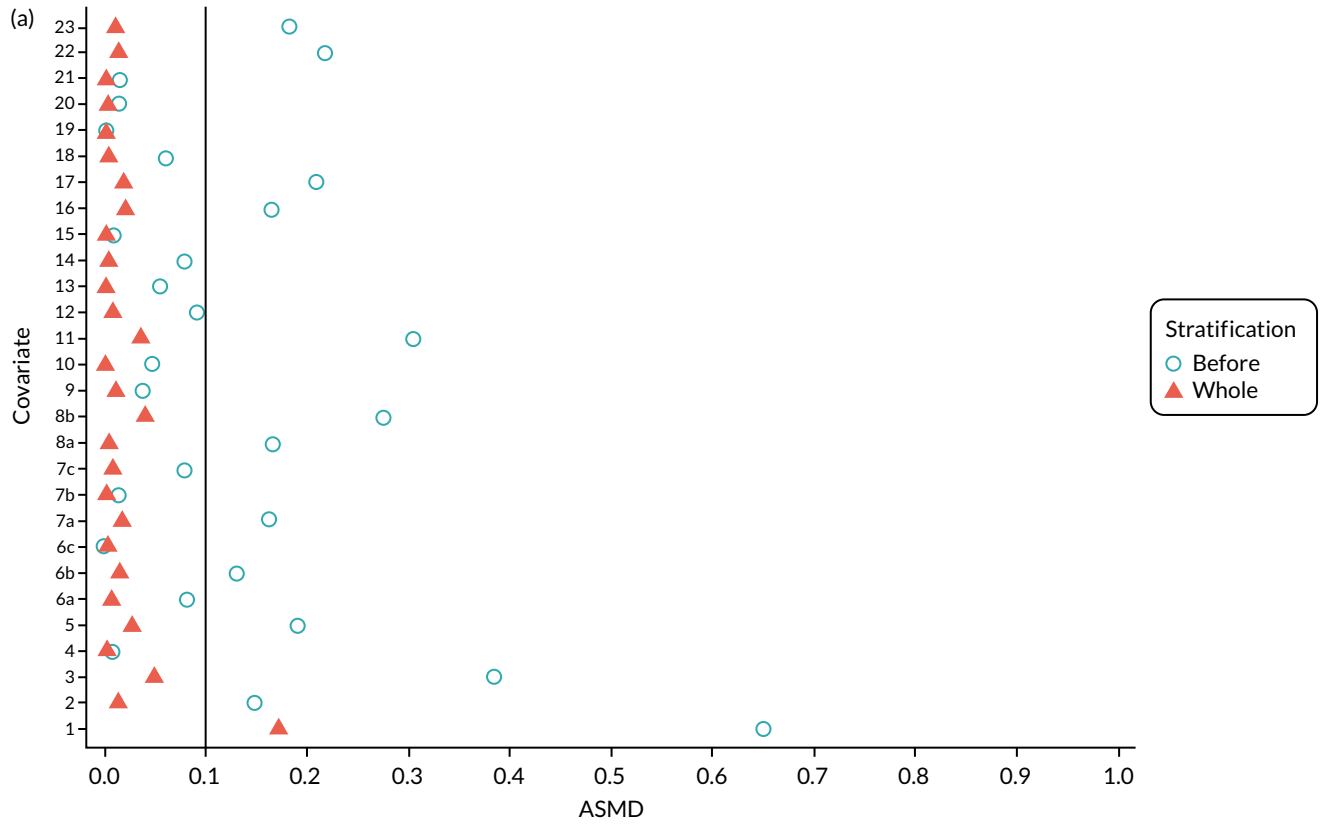


FIGURE 15 The ASMD of each covariate included in the PS for the postoperative OKS cohort before and after balancing covariates by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, IMD – general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

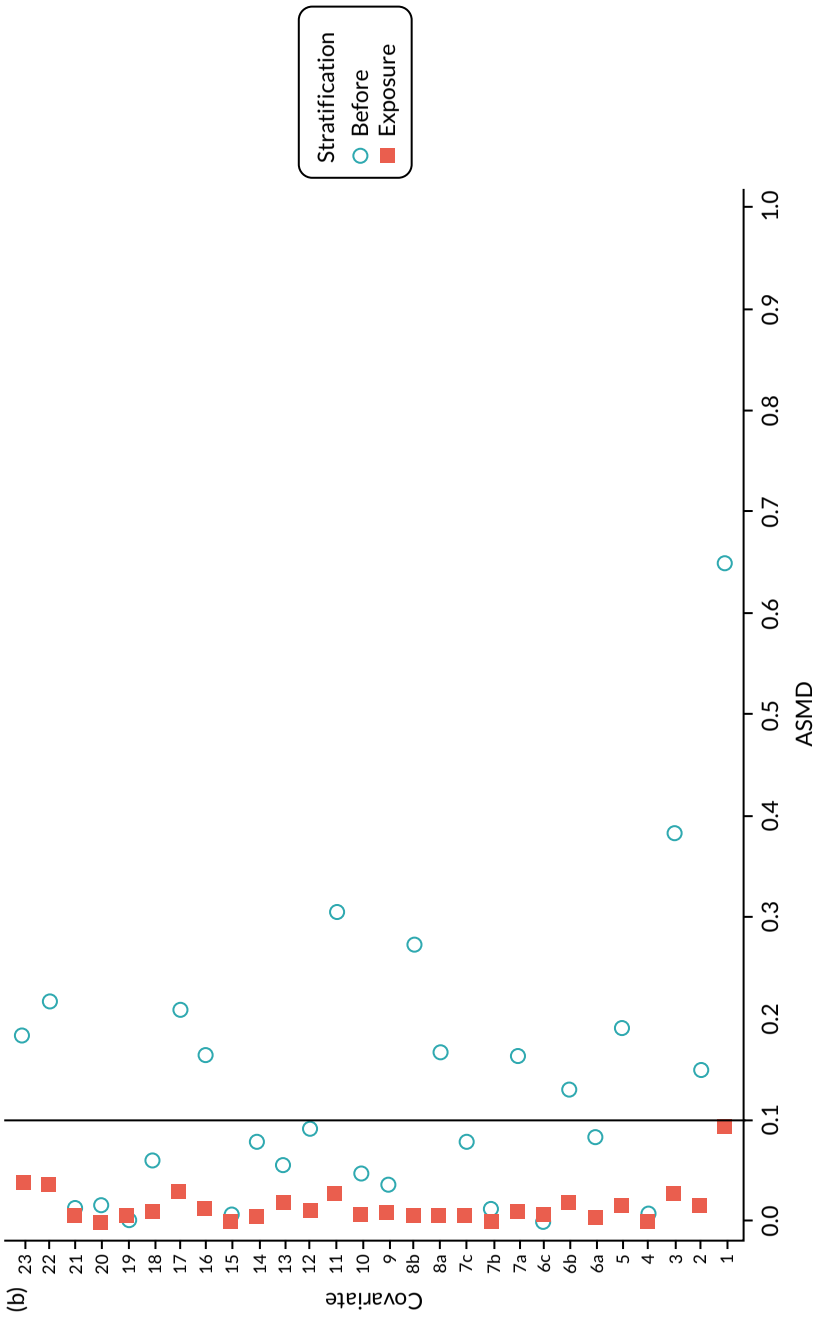


FIGURE 15 The ASMD of each covariate included in the PS for the postoperative OKS cohort before and after balancing covariates by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, IMD – general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

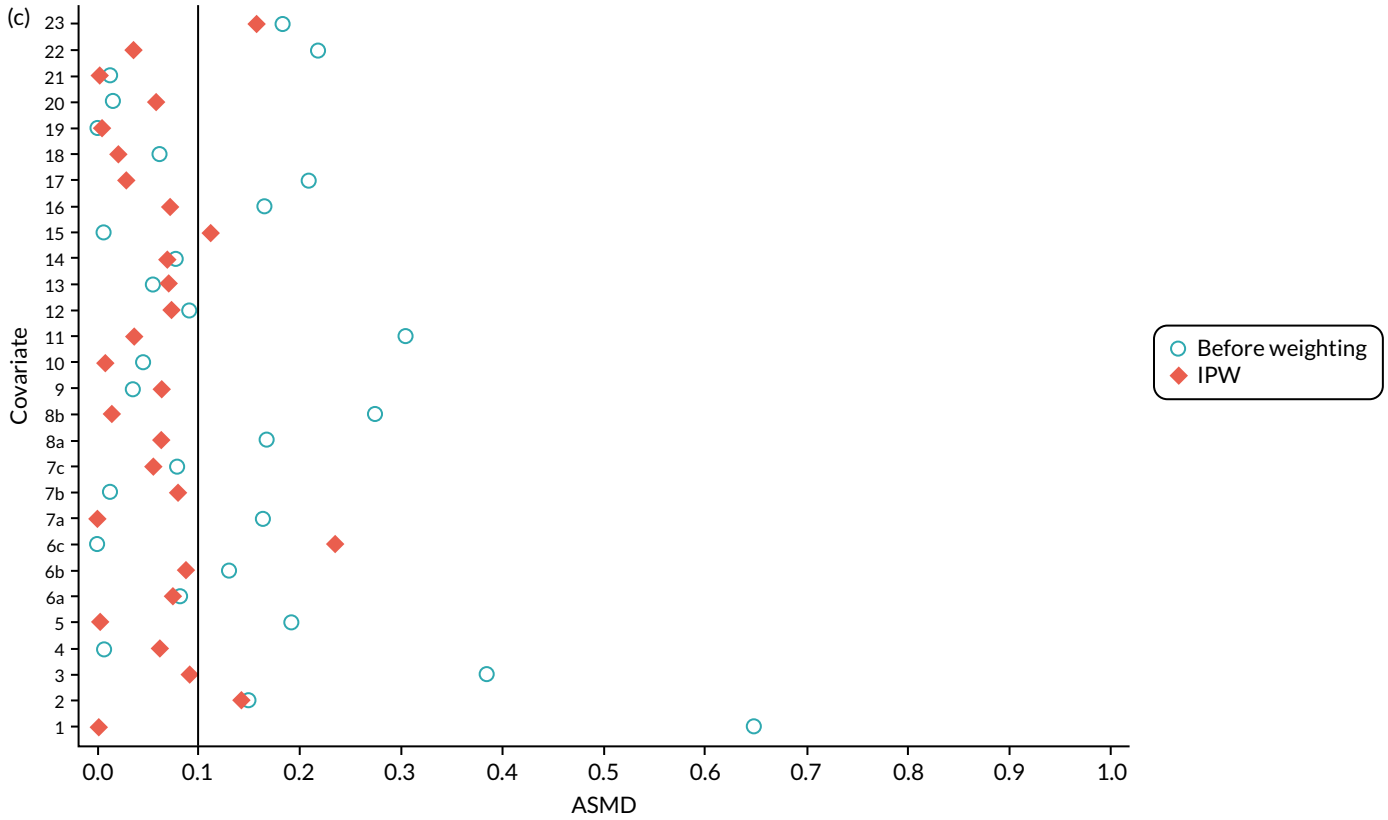


FIGURE 15 The ASMD of each covariate included in the PS for the postoperative OKS cohort before and after balancing covariates by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, IMD – general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis.

(ASMD of > 0.1) after IPW: respiratory disease, sex, socioeconomic deprivation and history of spondylosis (*Figure 15c*). UKR patients had a higher prevalence than TKR patients of respiratory disease (31% vs. 26%, respectively), male sex (53% vs. 46%, respectively) and residence in more deprived areas (40% vs. 29%, respectively). They were also less likely than TKR patients to have spondylosis (2% vs. 4%, respectively). These covariates were further (double) adjusted in the outcome analyses in *Primary outcome analyses: postoperative Oxford Knee Score* and *Comparative safety analyses*.

Safety cohort

We also applied the three validated methods to the safety cohort. PSS_{whole} resulted in similar overall PS distributions for UKR and TKR patients in each stratum (see *Figure 23a*). UKR patients had a higher predicted probability than TKR patients of receiving UKR based on their baseline characteristics, which was indicated through a higher PS. As a result, the lower PS quintiles (strata 1–3) each included < 1% of the UKR patients. Within-stratum covariate balance between UKR and TKR was much better than that in the OKS cohort, probably because the safety cohort had higher power. Overall, PSS_{whole} stratification controlled confounding to an acceptable degree based on ASMD (*Figure 16a*).

The PSS_{exp} stratification yielded a more equally distributed PS than PSS_{whole} (see *Figure 23b*) between UKR and TKR patients. The within-stratum covariate balance was also better with PSS_{exp} than with PSS_{whole}, with fewer variables with an ASMD of > 0.1. Overall, good covariate balance was achieved for all of the observed confounders, with average ASMDs of < 0.1 across all strata (see *Figure 16b*).

In the IPW pseudo-population, 2256 UKR patients were given a stabilised weight ranging from 0.09 to 9.45 and TKR patients were given weights of around 1. There were balanced distributions in all of the covariates (ASMD of ≤ 0.1) (see *Figure 16c*).

Primary outcome analyses: postoperative Oxford Knee Score

Table 18 shows the pre and postoperative OKS estimates from the stage 1 (ASA grade of 1 or 2) and stage 2 (ASA grade of 3 or 4) OKS cohort for the validated analyses using IPW, PSS_{whole} and PSS_{exp}:

- Stage 2 patients had a baseline (preoperative) OKS, on average, about 3 points lower than stage 1 patients for both UKR and TKR recipients and for any of the three methods. This difference was probably a result of stage 2 participants having higher comorbidity than stage 1 participants, or having surgery delayed because of high ASA grade.
- Stage 2 TKR and UKR patients in the IPW pseudo-population had similar mean (SD) preoperative OKSs [16.99 (7.56) versus 17.28 (8.37), respectively].
- PSS_{whole} and PSS_{exp} both included the whole cohort and, therefore, found a better mean (SD) preoperative OKS for stage 2 UKR patients [19.44 (8.55)] than TKR patients [16.97 (7.55)]. However, the variance in the mean preoperative OKS for UKR patients was larger than that for TKR patients, resulting in an agreeable degree of balance, as demonstrated in *Figure 15a* and *Figure 15b*.

Approximately 6–8 months after the operation, the mean OKS was more than twice the preoperative OKS in all participant groups, indicating a dramatic improvement because of surgery in both stage 1 and stage 2 patients. PSS_{exp}, the preferred method from stage 1, resulted in a statistically significant positive effect for UKR, with an estimated mean postoperative OKS difference of 1.83 (95% CI 0.10 to 3.56) points in favour of UKR. The other validated PS stratification method (PSS_{whole}) found very similar results, with an estimated mean difference in postoperative OKS of 1.82 (95% CI 0.10 to 3.56) points, again favouring UKR. IPW analyses found a non-significant difference in postoperative OKS between TKR and UKR, with a mean difference between groups of 1.00 (95% CI −1.28 to 3.27) points.

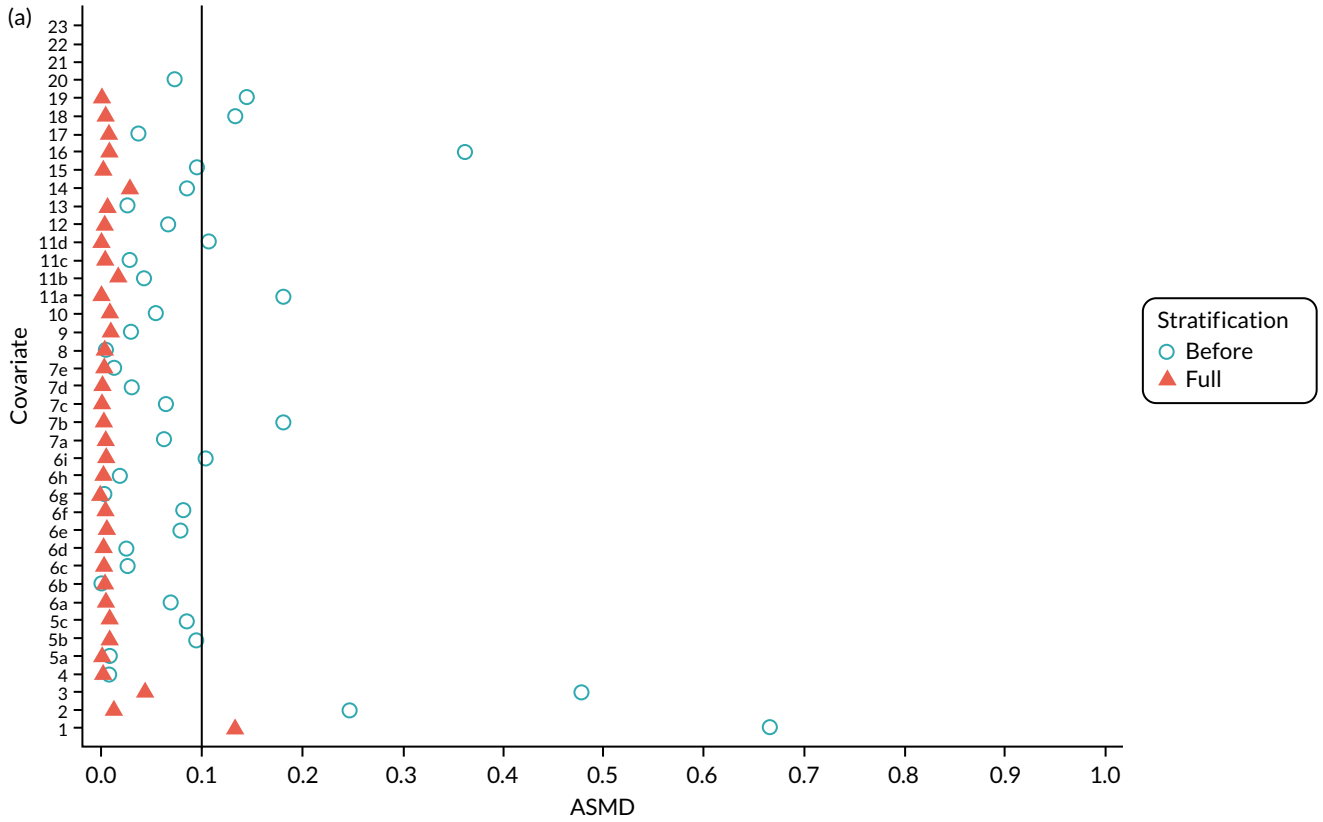


FIGURE 16 The ASMD of each covariate included in the PS for the postoperative safety cohort before and after covariate balancing by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

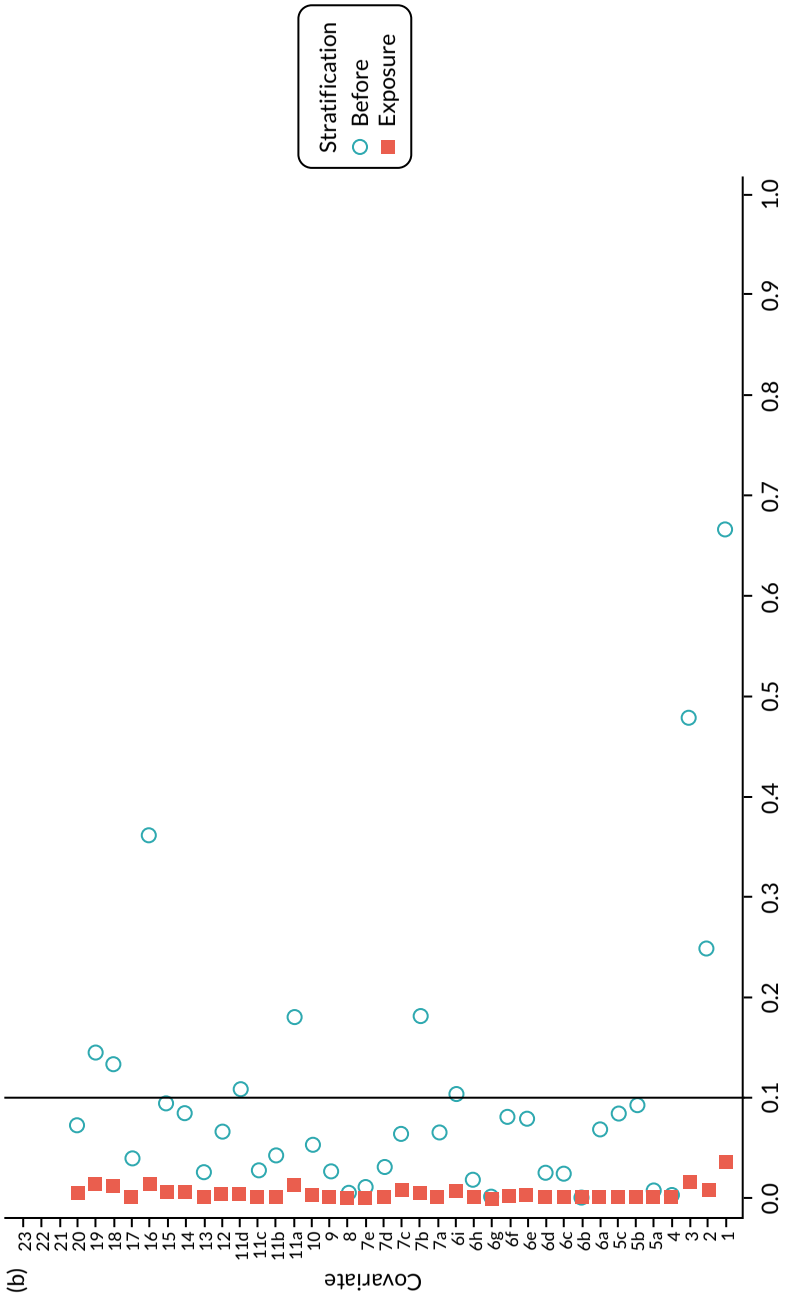


FIGURE 16 The ASMD of each covariate included in the PS for the postoperative safety cohort before and after covariate balancing by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis. (*continued*)

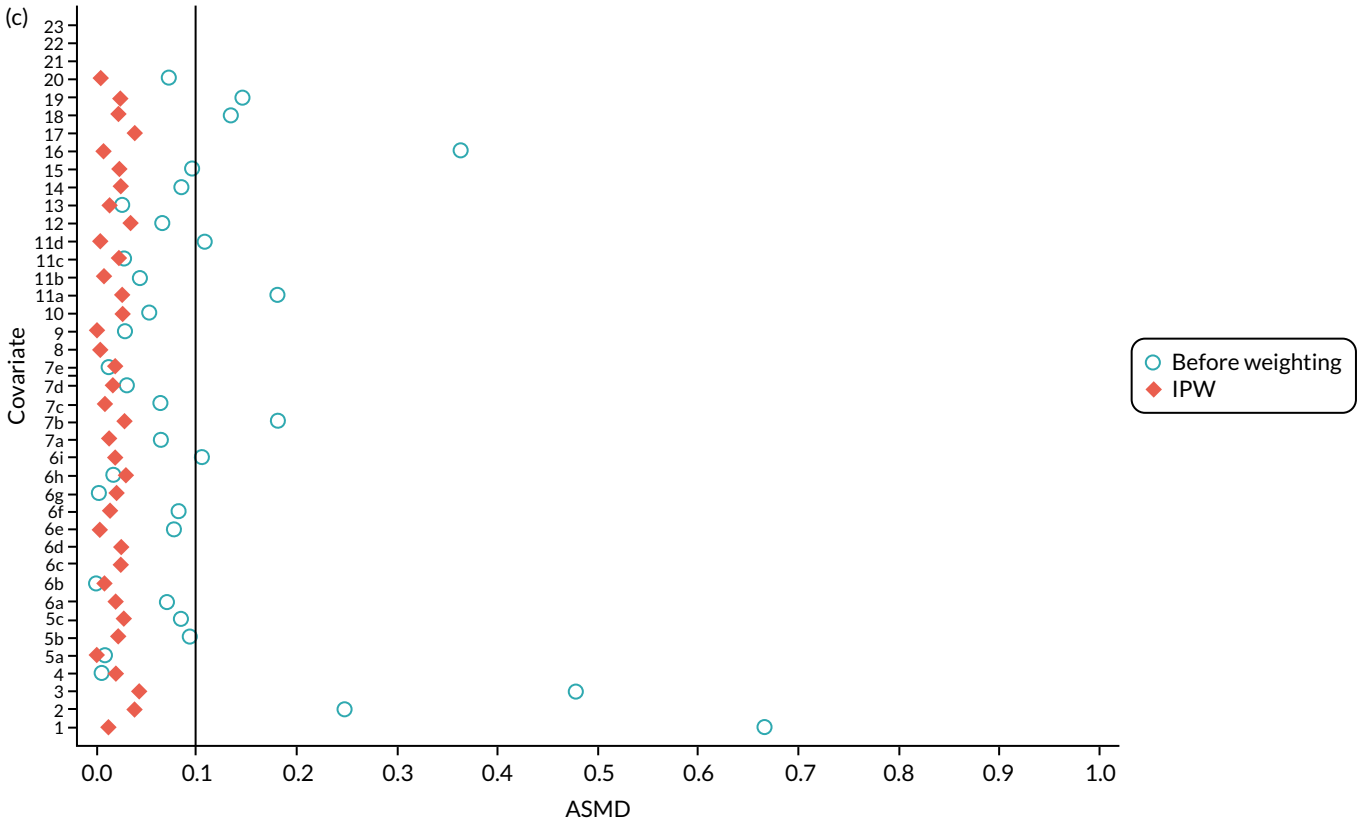


FIGURE 16 The ASMD of each covariate included in the PS for the postoperative safety cohort before and after covariate balancing by (a) PSS_{whole}, (b) PSS_{exp} and (c) IPW. 1, overall PS; 2, males; 3a, Rural Index – urban (≥ 10,000); 3, age; 4, BMI; 5a, Rural Index – town and fringe; 5b, Rural Index – village; 5c, Rural Index – isolated; 6a, IMD – less deprived 10–20%; 6b, IMD – less deprived 21–30%; 6c, IMD – less deprived 31–40%; 6d, IMD – less deprived 41%–50%; 6e, IMD – more deprived 10–20%; 6f, IMD – more deprived 21–30%; 6g, IMD – more deprived 31–40%; 6h, IMD – more deprived 41–50%; 6i, IMD – most deprived; 7a, general health = 1; 7b, general health = 2; 7c, general health = 3; 7d, general health = 4; 7e, general health = 5; 8, preoperative quality-of-life measure (EQ-5D); 9, preoperative OKS; 10, ASA grade of 2, mild diseases; 11a, Charlson Comorbidity Index score = 1; 11b, Charlson Comorbidity Index score = 2; 11c, Charlson Comorbidity Index score = 3; 11d, Charlson Comorbidity Index score = 4; 12, gastrointestinal diseases; 13, osteoarthritis and other joint problems; 14, mental health; 15, respiratory diseases; 16, cardiovascular diseases; 17, thyroid problems; 18, foot, hip or spinal pain; 19, coxarthrosis; 20, neurological disorders; 21, other arthrosis; 22, polyarthrosis; and 23, spondylosis.

TABLE 18 Pre and postoperative OKS in the stage 1 and 2 cohorts, calculated by PSS_{whole}, PSS_{exp} and IPW

Cohort	Treatment group, mean (SD)		Treatment effect UKR, mean difference (95% CI)
	TKR	UKR	
Stage 1 IPW			
Preoperative OKS	19.70 (7.57)	20.41 (7.42)	–
Postoperative OKS	35.80 (9.35)	36.64 (9.50)	0.58 (–0.19 to 1.35)
Stage 2 IPW			
Preoperative OKS	16.99 (7.56)	17.28 (8.37)	–
Postoperative OKS	32.60 (10.24)	33.65 (10.87)	1.00 (–1.28 to 3.27)
Stage 1 PSS_{whole}			
Preoperative OKS	19.68 (11.64)	21.88 (7.94)	–
Postoperative OKS	35.80 (11.35)	36.74 (10.13)	0.56 (–0.03 to 1.16)
Stage 2 PSS_{whole}			
Preoperative OKS	16.97 (7.55)	19.43 (8.55)	–
Postoperative OKS	32.59 (10.24)	34.56 (10.53)	1.82 (0.10 to 3.56)
Stage 1 PSS_{exp}			
Preoperative OKS	19.68 (13.30)	21.88 (7.77)	–
Postoperative OKS	35.80 (12.31)	36.74 (9.87)	0.76 (0.15 to 1.36)
Stage 2 PSS_{exp}			
Preoperative OKS	16.97 (7.55)	19.44 (8.55)	–
Postoperative OKS	32.59 (10.24)	34.57 (10.53)	1.83 (0.10 to 3.56)

Comparative safety analyses

Short-term (90-day postoperative) complications

The 90-day cumulative incidence of postoperative venous thromboembolism observed was lower for UKR participants (relative risk 2.66, 95% CI 1.20 to 5.91, per 1000 people) than for TKR participants (relative risk 7.96, 95% CI 7.26 to 8.71, per 1000 people), resulting in a crude relative risk of 0.33 (95% CI 0.15 to 0.75) in favour of UKR patients. The differences were not attenuated and persisted after adjusting for confounding using the validated methods. Adjustment with PSS_{whole} or PSS_{exp} resulted in a relative risk of 0.33 (95% CI 0.15 to 0.74), and with IPW resulted in a relative risk of 0.39 (95% CI 0.16 to 0.96).

By contrast, UKR and TKR patients had similar 90-day cumulative incidences of myocardial infarction and prosthetic joint infection. No significant differences in the risk of myocardial infarction or prosthetic joint infection were noted after adjustment with any of the three methods (*Table 19*).

Long-term (5-year) complications

The cumulative risk of revision increased faster for UKR patients than for TKR patients over 5 years of follow-up (*Figure 17*). The incidence rates of revision were 13.09 (95% CI 10.64 to 16.09) after UKR and 4.88 (95% CI 4.56 to 5.22) after TKR, an almost threefold increase in revision risk for UKR compared with TKR (crude hazard ratio 2.70, 95% CI 2.16 to 3.37). Adjustment for confounding using the validated methods did not attenuate this risk, with a resulting cause-specific hazard ratio of 2.70 (95% CI 2.15 to 3.38) for PSS_{whole} and PSS_{exp}, and 2.60 (95% CI 1.94 to 3.47) for IPW (*Table 20*).

TABLE 19 Short-term (90-day) complications after UKR or TKR

Complication	Cumulative incidence (95% CI)	Crude RR (95% CI)	Method, RR (95% CI)		
			PSS_{whole}	PSS_{exp}	IPW
Venous thromboembolism					
TKR ($n = 459$)	7.96 (7.26 to 8.71)	REF	REF	REF	REF
UKR ($n = 6$)	2.66 (1.20 to 5.91)	0.33 (0.15 to 0.75)	0.33 (0.15 to 0.74)	0.33 (0.15 to 0.74)	0.39 (0.16 to 0.96)
Myocardial infarction					
TKR ($n = 281$)	4.87 (4.34 to 5.47)	REF	REF	REF	REF
UKR ($n = 8$)	3.55 (1.77 to 7.07)	0.73 (0.36 to 1.47)	0.73 (0.36 to 1.45)	0.73 (0.36 to 1.45)	0.64 (0.29 to 1.45)
Prosthetic joint infection					
TKR ($n = 111$)	1.92 (1.60 to 2.32)	REF	REF	REF	REF
UKR ($n = 4$)	1.77 (0.67 to 4.71)	0.92 (0.34 to 2.50)	0.85 (0.33 to 2.19)	0.85 (0.33 to 2.19)	0.55 (0.18 to 1.71)

REF, reference; RR, relative risk.

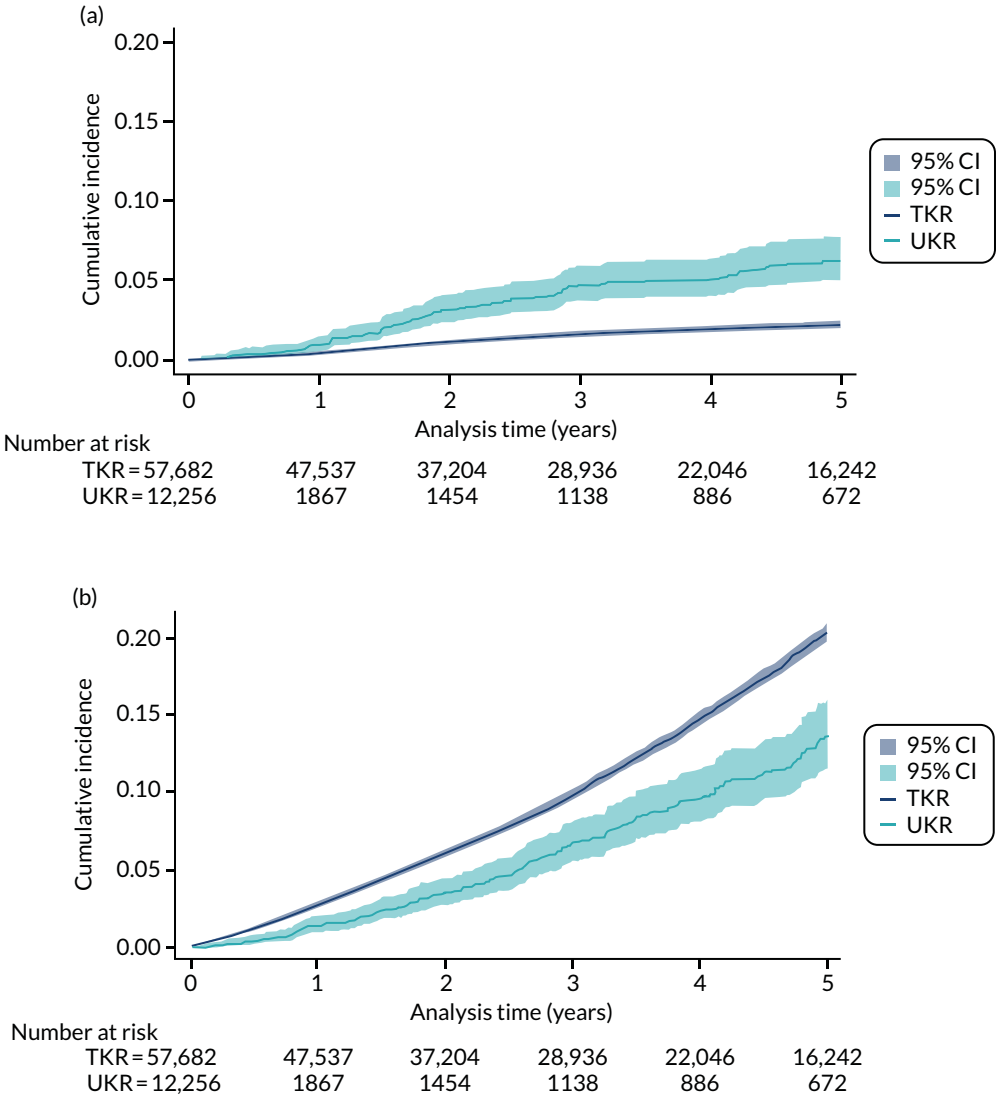


FIGURE 17 Cumulative incidence functions of (a) risk of revision and (b) mortality, for UKR (UKR = 1) and TKR (UKR = 0) over 5 years of follow-up. Number at risk: number of patients with a particular surgery who did not experience any of the outcomes.

TABLE 20 Long-term (5-year) complications after UKR or TKR

Complication	Cumulative incidence (95% CI)	Crude HR (95% CI)	Method, HR (95% CI)		
			PSS_{whole}	PSS_{exp}	IPW
Revision surgery					
TKR ($n = 847$)	4.88 (4.56 to 5.22)	REF	REF	REF	REF
UKR ($n = 90$)	13.09 (10.64 to 16.09)	2.70 (2.16 to 3.37)	2.70 (2.15 to 3.38)	2.70 (2.15 to 3.38)	2.60 (1.94 to 3.47)
All-cause mortality					
TKR ($n = 6401$)	36.89 (36.00 to 37.81)	REF	REF	REF	REF
UKR ($n = 164$)	23.85 (20.46 to 27.79)	0.64 (0.55 to 0.75)	0.64 (0.55 to 0.75)	0.64 (0.55 to 0.75)	0.83 (0.67 to 1.03)

HR, hazard ratio; REF, reference.

Participants receiving UKR surgery had lower 5-year mortality than TKR patients (see *Figure 17b*). UKR surgery, therefore, appeared to be associated with reduced all-cause mortality in the unadjusted analysis, with a crude hazard ratio of 0.64 (95% CI 0.55 to 0.75).

The observed decrease in mortality associated with UKR (vs. TKR) remained after adjustment for confounding using PSS_{whole} or PSS_{exp}, both resulting in a cause-specific hazard ratio of 0.64 (95% CI 0.55 to 0.75) (see *Table 20*). However, the observed effect on mortality was attenuated and became non-significant when using IPW, with a cause-specific hazard ratio of 0.83 (95% CI 0.67 to 1.03).

Sensitivity analyses

Prespecified interactions and stratified analyses

Significant interactions, predefined as having a p-value of < 0.1, were identified with ASA grade ($p = 0.07$ for PS stratification methods) and sex ($p = 0.05$ for IPW and $p = 0.02$ for PS stratification methods), but not with age ($p = 0.48$ for IPW and $p = 0.68$ for PS stratification methods).

When we stratified the analysis by sex, female UKR patients had a higher excess risk of revision (cause-specific hazard ratios around 3.5) than male UKR patients (hazard ratios around 2.0). When we stratified the analysis by ASA grade, the increase in revision risk associated with UKR was higher in patients with an ASA grade of 4 than patients with an ASA grade of 3. The hazard ratio estimates for patients with an ASA grade of 4 were around 8.0, but the CI could not be calculated owing to limited power for this analysis. *Table 21* gives the full results of the stratified analyses.

Analysis restricted to high-volume surgeons

We restricted the three validated analyses to surgeries performed by experienced surgeons to examine whether or not patients' risk of long-term complications changed. As in *Chapter 3*, *Revision cohort*, we defined three subcohorts of the safety cohort based on the number of surgeries of the same type performed by the lead surgeon in the previous year: ≥ 10, ≥ 30 and ≥ 50 surgeries. Of the 57,682 TKR patients included in the total cohort, 51,118 (89%), 38,321 (66%) and 25,944 (45%) were included in these three lead surgeon subcohorts, respectively. Of the 2256 UKR patients included in the total cohort, a smaller proportion were included [1449 (64%), 610 (27%) and 242 (11%), respectively]. See *Appendix 1*, *Table 30* for the baseline characteristics for these subcohorts.

TABLE 21 Sex-specific and ASA grade-specific cause-specific hazard ratios for UKR (vs. TKR) revision and mortality over 5-year follow-up

	Method, HR (95% CI)		
	PSS_{whole}	PSS_{exp}	IPW
Women (UKR, n = 978; TKR, n = 32,086)			
Revision	3.52 (2.59 to 4.78)	3.52 (2.59 to 4.78)	3.46 (2.33 to 5.14)
Death	0.53 (0.39 to 0.71)	0.53 (0.39 to 0.71)	0.77 (0.52 to 1.13)
Men (UKR, n = 1278; TKR, n = 25,596)			
Revision	2.07 (1.46 to 2.92)	2.07 (1.46 to 2.92)	1.86 (1.16 to 2.99)
Death	0.66 (0.53 to 0.81)	0.66 (0.53 to 0.81)	0.86 (0.66 to 1.13)
ASA grade of 3 (UKR, n = 2232; TKR, n = 56,625)			
Revision	2.63 (2.10 to 3.30)	2.63 (2.10 to 3.30)	2.51 (1.88 to 3.36)
Death	0.65 (0.55 to 0.77)	0.65 (0.55 to 0.77)	0.83 (0.67 to 1.04)
ASA grade of 4[a] (UKR, n = 24; TKR, n = 1057)			
Revision	8.77 (–)	8.77 (–)	7.93 (–)
Death	0.53 (–)	0.53 (–)	0.71 (–)

HR, hazard ratio.
a 95% CIs cannot be provided for ASA grade 4 owing to the small number of UKR cases.

Cause-specific risks of revision and mortality for UKR (vs. TKR) patients calculated with the three analytical methods are reported in *Figure 18*. The observed excess risk of revision seen in UKR patients was somewhat reduced when the analyses were restricted to those operated on by high-volume surgeons. The cause-specific hazard ratio of 2.60 (95% CI 1.94 to 3.47) in the main cohort decreased to 2.05 (95% CI 1.03 to 4.09) when restricting to high-volume surgeons with ≥ 30 surgeries in the past year and using IPW, and to 1.65 (95% CI 1.01 to 2.69) when using PS stratification. However, the CIs of these estimates overlapped, indicating no significant difference (see *Figure 18a*).

Excess revision risks increased again when restricting to the highest-volume surgeons (≥ 50 surgeries of the same type in the previous year). However, this subanalysis was limited by low statistical power, resulting in wide CIs.

Restricting to high-volume surgeons did not have striking effects on the observed association with 5-year mortality following surgery (see *Figure 18b*). The overlapping CIs of these estimates suggested no clear trend in differential mortality between UKR and TKR with increasing surgeon volume.

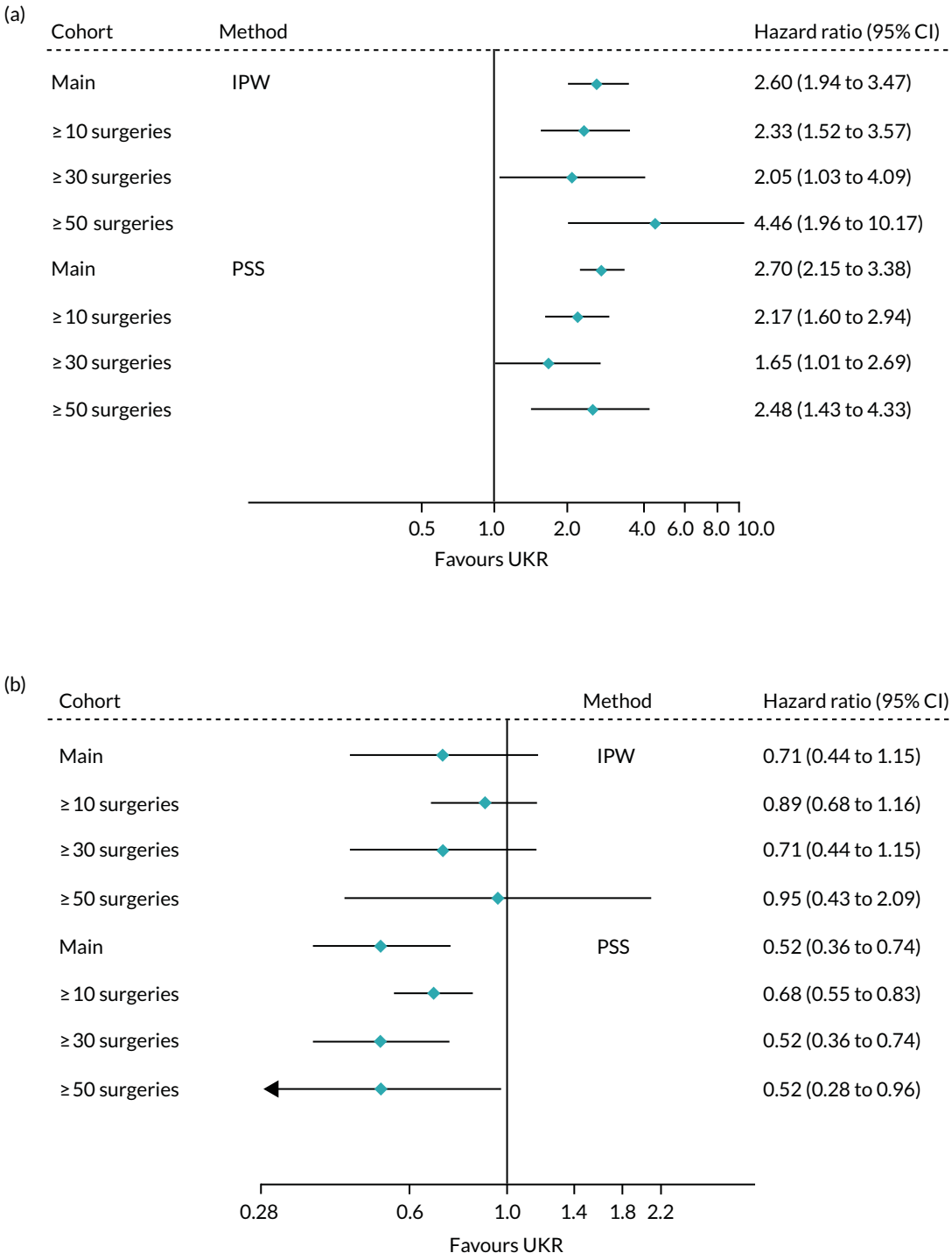


FIGURE 18 Cause-specific hazard ratios for risk of (a) 5-year revision and (b) mortality for patients undergoing UKR (vs. TKR) in sensitivity analyses restricted to lead surgeons with ≥ 10, ≥ 30 or ≥ 50 surgeries of a particular type in the previous year.

Chapter 8 Economic evaluation

This chapter reports an analysis of the cost-effectiveness of UKR, compared with TKR, for complex patients with an ASA grade of 3 or 4 identified in routinely collected data in England.

Introduction

The economic burden to the health-care system associated with knee replacements is substantial, owing to the large number of replacements performed each year and the risk of complications and revision surgery. The relative burden, and, more specifically, the cost-effectiveness, of UKR compared with TKR is not known. The length of hospital stay is shorter for UKR than for TKR, which may lead to less intense health-care resource use.[9] A previous modelling study[62] using data from the NJR for England, Wales, Northern Ireland and the Isle of Man found that UKR was expected to generate lower lifetime costs and better quality-of-life gains than TKR. Most of the evidence for the cost-effectiveness of UKR over TKR comes from modelling studies that extrapolate outcomes observed for a specific period over a lifetime based on assumptions.[63] Economic evaluations alongside RCTs typically require fewer assumptions, but examine cost-effectiveness over shorter time periods.

An economic evaluation was carried out alongside TOPKAT that examined the clinical effectiveness and cost-effectiveness of UKR compared with TKR for selected participants over 5 years.[5] TOPKAT included 232 patients who underwent a UKR surgery and 238 who underwent a TKR surgery; these patients were recruited from 27 sites across the UK. The authors concluded that UKR and TKR were both effective and had similar clinical outcomes, complication rates and revision rates, but that UKR resulted in lower health-care costs and better cost-effectiveness 5 years after surgery than TKR.[5,62]

TOPKAT and its economic evaluation may have low external validity owing to the strict patient inclusion criteria, a limitation that affects most RCTs. Patients were eligible for TOPKAT if they were medically fit and had an ASA grade of 1 or 2.[28] Although patients suffering from multiple comorbidities (i.e. ASA grade of 3 or 4) routinely receive knee replacements, they are often ineligible for clinical trials. It remains unknown whether or not UKR is cost-effective over TKR for these patients with multiple comorbidities. We used routinely collected data and the methods assessed in the previous chapters to evaluate the cost-effectiveness of UKR compared with TKR for complex patients with an ASA grade of 3 or 4.

Methods

Study design and setting

The population of this study comprised patients who received a UKR or a TKR for any indication. Routinely collected data from the NJR were linked to inpatient hospital data from the English NHS HES and PROMs databases.[18,64,65] The NJR collects data on knee and other joint replacements, with compliance of approximately 95% for primary knee replacements and 90% for revisions.[66]

The HES database contains data on all hospital admissions to the NHS and to independent sector providers paid for by the NHS.[20] Data are collected primarily for administrative purposes; however, these data have also been used extensively for research purposes.[20] Each record in HES Admitted Patient Care (APC) contains information on diagnoses (ICD-10) and the procedures undertaken, using OPCS-4. The HES PROMs database collects patients' perceptions, for example on their HRQoL and the quality of care delivered to NHS patients undergoing primary or revision knee replacement surgery. Linked data from the NJR, HES APC and HES PROMs were extracted from 2009 to 2016. *Chapter 2* gives more details about the study design, data linkage and data sources.

Study population

Patients who received a UKR and TKR for any indication were identified in the NJR data set using procedures from the OPCS-4 classification system, as described by the NJR.[21] From the cohort of UKR and TKR patients, we included patients with an ASA grade (collected by the NJR) of 3 or 4 to build a cohort of patients with multiple comorbidities.

The primary outcome of this economic evaluation was the incremental cost-effectiveness ratio (ICER), which generates a measure of additional units of cost per, in this case, extra unit of quality-adjusted life-years (QALYs) if a UKR was carried out instead of a TKR. As QALYs are derived from a combination of time and responses to the EQ-5D-3L questionnaire (included in HES PROMs), we included only patients who had completed one or both preoperative and postoperative PROM questionnaires. The analysis time horizon was 5 years, matching the TOPKAT exposure time.

The inclusion criteria for the final cohort are listed in detail in *Table 1*.

Outcome measures

The main outcomes for this economic evaluation were QALYs and costs, which are discussed in the following sections. Detailed information about other study variables can be found in *Chapter 2*.

Quality of life

Health-related quality of life was derived from the PROMs database, and was measured before and 6 months after surgery (primary knee replacement or revision). We linked records in the PROMs database to records for a primary or revision knee replacement for the corresponding patient in the NJR if the preoperative PROMs were completed between 1 year before and 31 days after the surgery date. When more than one PROM record satisfying these conditions was available, we chose the record completed closest to the surgery date.

Health-related quality of life was measured using the EQ-5D-3L questionnaire.[67] EQ-5D was used to estimate utility scores by applying the UK value set tariff, which incorporates the preferences of the general population.[68] Estimated utility scores are a preference-based measure of HRQoL ranging between −0.59 (worst state) and 1.00 (perfect health), with death anchored at zero. QALYs were estimated using the area-under-the-curve method: utility scores were multiplied by the length of time a patient spent in a particular health state.[69] Transitions between health states were linearly interpolated.

Costs

This economic evaluation included hospital costs for primary knee replacement, complications after primary knee surgery (myocardial infarction, venous thromboembolism and prosthetic joint infection) and revision surgery. Hospital costs were based on patients' HRGs. HRGs were assigned to spells, which are defined as uninterrupted inpatient stays at one hospital that may include several finished consultant episodes if a patient is seen by various consultants during the same stay. Spells with clinically similar treatments and comparable levels of resource consumption were classified into the same HRG groups.[70] For spells with stay lengths beyond the trim point of their HRG code, additional cost was added for the number of excess bed-days. Unbundled costs were also considered, which refer to significant elements of activity and costs (e.g. diagnostic imaging) that are not included in the core HRGs.[71] The HRG4 Reference Costs Grouper was used to derive the core and unbundled HRG codes, and the *NHS Reference Costs 2017/18*[72] were used to derive a cost from each HRG code.

Hospital spells related to primary and revision knee replacements were identified using the recommended NJR OPCS-4 procedures as described above.[21] Spells associated with myocardial infarction and venous thromboembolism were identified using the primary ICD-10 diagnosis code for related hospital episodes that occurred the day of the primary knee replacement or within the following 90 days.[61] Spells related to prosthetic joint infection were identified by the primary ICD-10 diagnosis code of a hospital episode and a concurrent or subsequent operation (OPCS-4) code of debridement, antibiotics, irrigation, retention or revision of the prosthesis within 1 year of the diagnosis.[73]

When HES data were added into the HRG4 Reference Costs Grouper to derive the HRG codes for primary knee replacement, 3% of the admissions generated invalid codes because of missing or invalid information. Approximately 30% of these errors occurred because of the diagnostic ICD-10 code 'I48.9, Unspecified atrial fibrillation and atrial flutter', which is a non-billable code. To reduce the number of invalid HRG codes, we deleted this code when it was recorded as a secondary diagnosis. This increased the number of valid HRG codes and did not affect the actual HRG codes produced or costs of the primary operations. All remaining spells without a valid HRG code for primary knee replacement (approximately 2% of the sample) were excluded from the analyses. To maintain consistency, the secondary diagnostic code 'I48.9' was deleted from all further HRG estimations (i.e. revision and complications).

Economic evaluation

This economic evaluation was carried out from a health-care payer perspective, including costs incurred by the health-care system. The health-care payer perspective is recommended by NICE for health technology appraisals.[74] Costs and QALYs were discounted at a rate of 3.5% per year, in accordance with NICE guidelines.[74]

Methods to minimise confounding

The choice of treatment in an observational study may be influenced by outcome predictors, creating a risk of confounding by indication. Various methods to account for potential confounding have been developed.[8] UTMoSt stage 1 aimed to identify the methods that best minimised confounding by testing which methods gave results that were similar to those in TOPKAT.

This economic evaluation was conducted using the methods found to be valid in UTMoSt stage 1, as described in *Chapters 3* and *4*. The tested methods were described in detail in *Chapter 2*, *Propensity score stratification* and their assessment in *Chapter 2*, *Inverse probability weighting*.

Propensity score stratification

Propensity scores represent the probability that a patient will receive the treatment of interest (i.e. UKR) according to their sociodemographic and clinical characteristics.[44] Multivariable logistic regression equations were used to calculate a PS for each outcome of interest. In PS stratification, different strata were calculated using the PSs. Strata were analysed separately and the results averaged by the proportion of total participants within that stratum.[44] Two PS stratification approaches were used, based on the PS distribution in the whole sample (PSS_{whole}) and in the exposed group who received a UKR (PSS_{exp}).[40]

Inverse probability weighting

In IPW, the outcomes in the analyses were weighted by the inverse probability of patients receiving the treatment of interest (in this case UKR) based on measured confounders (clinical and demographic patient characteristics). Confounding was removed by creating a pseudo-population in which the treatment was not dependent on the included confounders.[75] Any covariate showing imbalance (ASMD of > 0.1) in any of our 10 imputed data sets was included as a covariate in the regression models.

Missing data

Missing data on BMI, PROMs (including preoperative and postoperative EQ-5D for primary and revision knee surgery), general health variables and costs for revision or complications were imputed using multiple imputation by chained equations.[76] We assumed that data were missing at random.[77] Predictive mean matching was used to impute HRQoL data because utility scores (derived from EQ-5D) are not normally distributed and are bounded at 1.00.[78,79] Predictive mean matching is a semiparametric approach in which the missing value is imputed with an observed value from another individual whose predicted value is similar to the predicted value of the individual with the missing observation.[80] We used the five nearest neighbours (closest observations) to draw imputed values.[81]

Ten imputed data sets were created and each was analysed separately before pooling the estimates and standard errors from the imputed data set according to Rubin's rules.[82]

Statistical analysis

Using IPW, linear mixed-regression analyses were performed to estimate the differences in costs and QALYs between UKR and TKR while accounting for the hierarchical structure of the data. A two-level structure was used in which patients were nested within lead surgeons who carried out the knee replacement. Four variables (i.e. sex, deprivation index, and having respiratory disease or spondylosis up to 3 years before the index date) showed imbalance based on the ASMD and were included as covariates in the regression model. In the linear mixed-regression analysis, we assumed that the outcome (i.e. cost and QALYs) for each patient was predicted by the intercept, which varied across lead surgeons. We also assumed that the slopes were the same across the lead surgeons, so that the outcome would not change if a surgeon-level variable, such as surgeon's experience, increased.

Using the PS stratification methods, separate stratum estimates were obtained using multilevel linear mixed regression analysis and pooled together. The preoperative utility score was included as a covariate in the regression models for QALYs.[82,83]

The ICER was calculated by dividing the difference in costs by the difference in QALYs between UKR and TKR. The uncertainty surrounding the ICER was estimated using non-parametric bootstrapping with 1000 replications. To illustrate this uncertainty, the bootstrapped cost and QALY pairs were plotted on a cost-effectiveness plane,[84] with the incremental costs between UKR and TKR on the *y*-axis and incremental QALYs on the *x*-axis. The cost-effectiveness plane is divided into four quadrants, with the north-east quadrant indicating when UKR is more expensive and more effective than TKR, and the south-east quadrant indicating when UKR is less expensive and more effective than TKR (i.e. UKR dominates TKR). The south-west quadrant indicates when UKR is less expensive and less effective than TKR, and the north-west quadrant indicates when UKR is more expensive and less effective than TKR.

Results

Patient characteristics

Figure 1 shows a flow chart describing the inclusion of patients in the study. We identified 868,785 patients who received a UKR or TKR in the NJR, of whom 553,567 could be linked to HES. We excluded patients with an ASA grade of 1 or 2, who opted out of HES or who had missing PROMs, leaving 23,489 patients for inclusion in the analysis. Of these patients, 145 had received a UKR and 23,344 had received a TKR. A detailed description of the patient inclusion criteria can be found in *Chapters 3* and *7*.

Before the cohorts were balanced, the mean age of UKR patients was 69.8 (SD 10.0) years and the mean age of TKR patients was 73.5 (SD 8.6) years. Sixty-eight (47%) UKR patients and 2683 (11%) TKR patients were female. Within 3 months after their knee replacement, two (1.4%) UKR patients and 98 (0.4%) TKR patients had a myocardial infarction, 170 (0.7%) TKR patients had a venous thromboembolism, and 41 (0.2%) TKR patients had a prosthetic joint infection. During the 5 years of follow-up, four (2.7%) UKR patients and 327 (1.4%) TKR patients had a revision, and 18 (12%) UKR patients and 2084 (8.9%) TKR patients died. Additional information on the patients' clinical and demographic characteristics can be found in *Chapter 7*.

Nine (6%) primary UKR patients and 1363 (6%) primary TKR patients had a missing preoperative EQ-5D, while 13 (9%) UKR patients and 1404 (6%) TKR patients did not report a postoperative measure. For revision procedures following a UKR, three (75%) patients were missing a preoperative EQ-5D and three (75%) patients were missing a postoperative EQ-5D. For revision procedures following a TKR, 256 (78%) patients were missing a preoperative EQ-5D and 280 (86%) patients were missing a postoperative EQ-5D.

Costs

Before using any method to balance the two cohorts, the mean cost of a primary knee replacement was £6246 (SD £779) for UKR patients and £6627 (SD £1402) for TKR patients. The mean cost for patients who experienced complications was £3560 (SD £6) for UKR and £3986 (SD £3853) for TKR. The mean cost for patients undergoing a revision surgery was £5103 (SD £3953) for primary UKR patients and £9161 (SD £4303) for primary TKR patients. The mean total costs without discounting were £6436 (SD £1247) for UKR patients and £6808 (SD £2017) for TKR patients. The mean discounted total costs were £6206 (SD £1177) for UKR patients and £6565 (SD £1920) for TKR patients.

Quality-adjusted life-years

Before using any method to balance the two cohorts, the mean preoperative estimated health utility score for primary knee replacement was 0.389 (SD 0.318) for UKR patients and 0.342 (SD 0.317) for TKR patients. The postoperative mean health utility scores were 0.708 (SD 0.288) for UKR patients and 0.667 (SD 0.278) for TKR patients. For revision surgery, the mean preoperative utility score was 0.248 (SD 0.319) for UKR and 0.223 (SD 0.312) for TKR patients, and the mean postoperative score was 0.553 (SD 0.312) for UKR and 0.383 (SD 0.348) for TKR patients. Without discounting, UKR patients gained 2.47 (SD 1.51) and TKR patients gained 2.05 (SD 1.38) mean QALYs over 5 years. After discounting, UKR patients gained 2.24 (SD 1.34) and TKR patients gained 1.87 (SD 1.23) mean QALYs over 5 years.

Cost-effectiveness analysis

The results from the cost-effectiveness analysis using each of the three adjustment methods are presented in *Table 22*, and the uncertainty surrounding the ICER shown in *Figure 19*. As we used regression equations with difference as the outcome, the results give differences between UKR and TKR with no separate adjusted means for costs and QALYs.

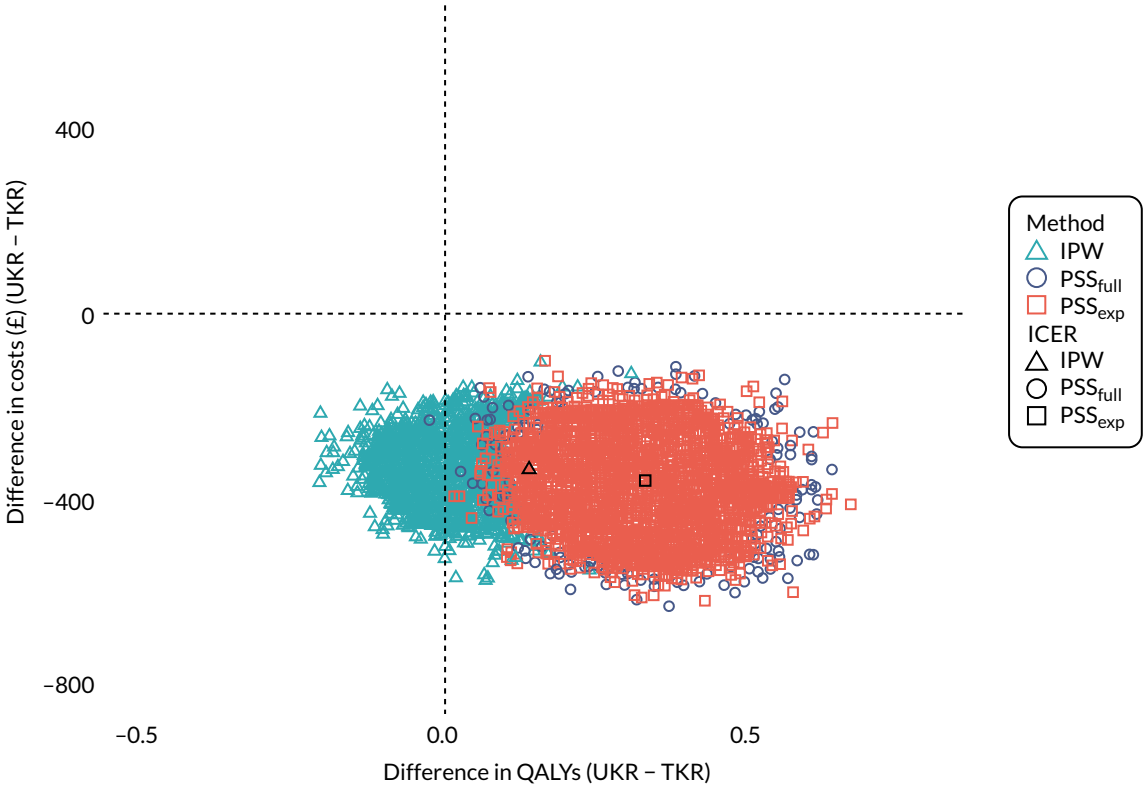


FIGURE 19 Cost-effectiveness plane for UKR compared with TKR in patients with an ASA grade of 3 or 4.

TABLE 22 Cost-effectiveness analysis results, stage 2 UTMoSt

Method	Difference in costs (£) (95% CI)	Difference in QALYs (95% CI)	ICER	Distribution of bootstrapped cost and QALYs on cost-effectiveness plane by quadrant (%)			
				NW	NE	SW	SE
IPW	−334 (−362 to −306)	0.135 (−0.482 to 0.753)	UKR dominant	–	–	10	90
PSS_{whole}	−359 (−378 to −339)	0.330 (−0.305 to 0.967)	UKR dominant	–	–	–	100
PSS_{exp}	−359 (−378 to −340)	0.330 (−0.309 to 0.970)	UKR dominant	–	–	–	100

NE, north-east; NW, north-west; SE, south-east; SW, south-west.

Conducting the cost-effectiveness analysis using the IPW method to minimise the risk of confounding led to UKR costs that were £334 (95% CI £306 to £362) lower than TKR costs. The mean difference in QALYs gained was 0.147 (95% CI −0.507 to 0.803) favouring UKR; however, this difference was not statistically significant. UKR dominated TKR as it was less expensive and more effective. Using the IPW method, the probability of UKR being more cost-effective than TKR was 100% at a willingness-to-pay (WTP) threshold of £0 (see *Figure 19*). From the bootstrapped analysis, 90% of resulting cost-and-effect pairs were located in the south-east quadrant (UKR less expensive and more effective than TKR) and 10% were located in the south-west quadrant (UKR less expensive and less effective than TKR) of the cost-effectiveness plane (see *Table 22*).

Using PSS_{whole} to minimise the risk of confounding produced a mean difference in costs of £359 (95% CI £339 to £378), again favouring UKR over TKR. The improvement in QALYs was 0.330 (95% CI −0.305 to 0.967) greater for UKR patients than for TKR patients; however, again, this difference was not statistically significant. UKR, therefore, dominated TKR because it was less expensive and more effective. The probability of UKR being more cost-effective than TKR was 100% at any positive WTP threshold (see *Figure 19*). *Table 22* shows that all of the bootstrapped cost-and-effect pairs were located in the south-east quadrant (UKR less expensive and more effective than TKR) of the cost-effectiveness plane.

Using PSS_{exp} to minimise the risk of confounding produced a mean difference in costs of £359 (95% CI £340 to £378) favouring UKR over TKR. The mean difference in QALYs gained was 0.330 (95% CI −0.309 to 0.970), also favouring UKR over TKR. Once again, this difference was not statistically significant. UKR, therefore, dominated TKR. The probability of UKR being more cost-effective than TKR was 100% at any positive WTP threshold (see *Figure 19*). All of the bootstrapped cost-and-effect pairs were located in the south-east quadrant (UKR less expensive and more effective than TKR) of the cost-effectiveness plane (see *Table 22*).

Discussion

The aim of this study was to assess the cost-effectiveness of UKR compared with TKR for complex patients with comorbidities (ASA grade of 3 or 4), who are generally ineligible for clinical trials, using routinely collected data. The three methods validated in stage 1 to replicate the TOPKAT results in routinely collected data were used to assess cost-effectiveness for this patient group during the 5 years after primary knee replacement. Under all three methods, UKR was associated with lower mean hospital costs and higher mean QALYs gained than TKR. Lower costs for patients with UKR were mainly a result of the lower costs of primary knee replacement. The mean difference in QALYs between the two groups favoured UKR by between 0.1 and 0.3 discounted QALYs over the 5 years and was not statistically significant.

We found UKR to be more effective and less costly than TKR at 5-year follow-up for patients with an ASA grade of 3 or 4. This finding agrees with previous findings using trial and routinely collected data.[5,60,62,63,71,72] Burn *et al.*[62] compared UKR and TKR lifetime QALYs gained and health-care costs in a modelling study using data from the UK. They found that UKR was related to more QALYs gained and lower costs than TKR for different sex and age groups; however, no separate analysis was undertaken for complex patients with multiple comorbidities. Similarly, TOPKAT found that UKR was more effective (mean QALY difference 0.240, 95% CI 0.046 to 0.434) and less expensive (mean cost difference −£910, 95% CI −£1503 to −£317) than TKR during 5-year follow-up. However, the trial was restricted to patients with an ASA grade of 1 or 2.[5] UKR, therefore, appears to be consistently associated with more QALYs gained and lower costs than TKR regardless of patient comorbidity. To our knowledge, this is the first study assessing cost-effectiveness specifically for patients with multiple comorbidities and applying previously assessed methods that minimise the risks of confounding by indication that generally affect observational studies.

The difference in costs between UKR and TKR was probably driven by the lower cost of primary UKR surgery to the health-care system. The difference in the cost of primary knee replacement may be because UKR is a less invasive surgery than TKR.[62] For instance, TOPKAT found that UKR had lower primary surgery costs owing to shorter hospital stays, fewer perioperative complications and lower implant device costs than TKR.[5] Although we also found that the costs for revision and complications were lower for UKR than for TKR, these events were rare.

The mean difference in QALYs gained favoured UKR over TKR, but this difference was not marked and not statistically significant. The uncertainty around the mean difference in QALYs between the two groups was also reflected in the cost-effectiveness plane. When using the IPW method, 10% of the cost and QALY pairs were located in the south-west quadrant, indicating that UKR is less expensive and less effective than TKR. TOPKAT found that the 0.2 difference in QALYs between the two groups was statistically significant, whereas Burn *et al.*[62] found that the mean difference was statistically significant for some patient subgroups but not others.[5] As we identified only 145 UKR patients meeting the inclusion criteria, it is possible that the lack of statistical significance in our findings was because of the sample size. However, we found a difference in QALYs that varied between 0.1 and 0.3 depending on the confounding minimisation method used, which was consistent with the TOPKAT findings. These findings suggest that patients' complexity and comorbidities did not prevent them from experiencing the quality-of-life benefits of a UKR rather than a TKR. We found that UKR and TKR both resulted in substantial improvements in utility scores from preoperative measurements to measurements 6 months after surgery, indicating that both procedures were highly beneficial for most patients.

We used the three PS methods validated in UTMoSt stage 1 (IPW, PSS_{whole} and PSS_{exp}) as being able to account for confounding bias and mimic TOPKAT. All three methods found similar mean differences in costs between UKR and TKR and overlapping 95% CIs. PSS_{whole} and PSS_{exp} resulted in wider 95% CIs than IPW for mean difference in QALYs gained between the two groups. Overall, the choice of method to account for potential confounding did not substantially alter the study's conclusions about the cost-effectiveness of the two surgeries.

Strengths and limitations

The conduct of this study faced some limitations. Although IPW and PS stratification were used to achieve balance in a wide range of observed characteristics, we cannot rule out remaining imbalances in unobserved factors that could bias the study findings. A common concern when using routinely collected data is that the data were not primarily collected for research purposes. The accuracy of the information extracted from the linked databases has to be further explored. However, to increase our confidence that the correct cohort of patients was included in the study, we used validated OPCS-4 codes to identify primary knee replacements and revisions and ICD-10 codes for complications.[21] Another important limitation of working with routinely collected data is missing values for some variables. We addressed this problem by applying recommended, commonly used methods to impute the missing data[76] so that we did not have to exclude patients with missing data.

One of the key strengths of our study is that it was conducted using real-world, routinely collected data. Trial data alone could not compare the cost-effectiveness of UKR and TKR for patients with multiple comorbidities. By using real-world data, we were able to address this cost-effectiveness question for the first time for this group of patients. To minimise the risk of confounding by indication, we used three methods that had been shown to replicate the findings of TOPKAT (stage 1). We used the same methods for this economic evaluation as TOPKAT[5] and estimated the costs associated with HRGs for reimbursement using the same data source (HES APC). As recommended by NICE, we derived QALYs from HRQoL outcomes obtained from patient-completed EQ-5D questionnaires (HES PROMs).[85]

Conclusion

Our findings showed that UKR was more cost-effective and led to lower additional costs and higher additional QALYs than TKR for patients with multiple comorbidities who were eligible for either surgery. The costs were lower mainly because UKR primary replacements cost less than TKR primary replacements. UKR and TKR were both associated with a substantial improvement in health-utility estimates from before surgery to 6 months after primary surgery, indicating that both procedures were beneficial for complex, highly comorbid patients.

Chapter 9 Conclusions and discussion of study findings

Study conclusions: UTMoSt stage 1

To the best of our knowledge, UTMoSt stage 1 is one of the first studies in the world to attempt to mimic a surgical RCT using real-world data and different analytical methods to minimise confounding. A number of recent initiatives aim to prove the value of real-world data for regulatory and clinical decision-making. Funded and sponsored by the public [FDA (Silver Spring, MD, USA); US National Heart, Lung, and Blood Institute (Bethesda, MD, USA); and Harvard University (Cambridge, MA, USA)] and private institutions [Aetion Inc. (New York, NY, USA)], RCT DUPLICATE[73] is perhaps the best known of these initiatives. Another initiative, the LEGEND study,[86] conducted by the OHDSI collaboration, has reported preliminary results in the area of hypertension. In a multidatabase, multidrug, PS-based analysis, the authors found that well-performed pharmacoepidemiological analyses provided findings that were highly consistent with previously performed head-to-head trials (slides available for review[86]). To our knowledge, all of these initiatives attempt to replicate the findings of RCTs for the study of medicines, but none has attempted to mimic surgical or medical device trials.

Our stage 1 findings show that the replication of surgical RCTs shares some challenges with studies replicating drug or medicinal product RCTs, but also has its own challenges. In common with the replication of drug RCTs, we found that even in a relatively pragmatic, post-marketing RCT, such as TOPKAT, a good proportion of actual NHS patients would not have been eligible. About 37% of patients undergoing TKR and 35% of patients undergoing UKR in the NHS would not have been eligible for TOPKAT based on their clinical characteristics. There is clearly a space for real-world data to contribute useful information on the risk–benefit and cost-effectiveness of medical devices and alternative surgical approaches and procedures for the many patients (here around one in three) for whom no RCT-derived evidence is available. UTMoSt stage 2 delivered the best available information in the absence of RCT data for NHS patients who were not eligible for TOPKAT. The results are summarised in *Study conclusions: UTMoSt stage 2*.

More worryingly, the generalisability of surgical and medical device evaluation RCTs can be limited by inclusion criteria that dictate which surgeons and treatment centres are eligible to participate. Surgical RCTs tend to require surgeons with a certain level of expertise with both treatments under study and seek clinicians who are in equipoise when deciding what treatment is best for the eligible patients. TOPKAT in particular used an experience-based design, including surgeons who had carried out $\geq$ 10 procedures (UKR or TKR) in the previous year of the same type as the allocated treatment.[28]

Restricting UTMoSt analyses to participants operated on by surgeons with such levels of experience excluded another 9% of possible TKR participants and one in three TKR surgeons from UTMoSt stage 1. Probably because of the lower uptake of UKR, the same criterion excluded almost half of the possible UKR participants and 64% of UKR surgeons from UTMoSt stage 1. TOPKAT surgeons had completed many surgeries in their careers before TOPKAT: a median of 100 (IQR 50–200) UKR procedures and 300 (IQR 260–400) TKR procedures.[87] Although these numbers are not directly equivalent to the previous-year surgery counts that we used, they do indicate that the surgeons participating in TOPKAT probably had greater expertise than strictly required for participation. For illustrative purposes, we reported the impact of restricting the analysis to operations performed by surgeons who had performed $\geq$ 30 and $\geq$ 50 surgeries of the same type in the previous year. These criteria excluded, respectively, > 80% and > 90% of the UKR patients operated on in the NHS who were initially eligible for UTMoSt stage 1. We discuss the impact of restricting to expert surgeons below.

Our stage 1 findings demonstrated that some (but not all) of the previously approved methods for the study of drug safety and post-marketing comparative effectiveness research can also be applied to the study of implantable devices and surgical procedures. Our results showed that some PS methods could reliably approximate the TOPKAT primary outcome results (postoperative OKS). Methods that estimated the ATEs (PS stratification and IPW) more closely approximated the TOPKAT findings than other tested methods. In the full cohort analysis, only PS stratification based on the distribution of the PS in the UKR (exposed) cohort, PSS_{exp}, passed all of the proposed diagnostics and was classified as able to replicate the TOPKAT findings. PSS_{whole} and IPW came close to passing these diagnostics.

The sensitivity analysis was restricted to patients operated on by surgeons who would have been eligible for TOPKAT, based on their experience with the index surgery, resulting in findings much closer to those seen in the RCT, with ATEs of 1.32 (95% CI 0.32 to 2.33) (IPW) and 1.37 (95% CI 0.54 to 2.20) (PS stratification) compared with 1.91 (95% CI 0.20, 3.62) in TOPKAT, all in favour of UKR over TKR. PSS_{whole}, PSS_{exp} and IPW were, therefore, deemed valid and taken forward into UTMoSt stage 2.

As well as the effect of surgeon experience on the PROM OKS, we found strong evidence of an interaction between surgeon experience and the association between type of surgery (UKR vs. TKR) and 5-year revision risk. Using the three validated methods, we demonstrated that the excess risk observed among patients undergoing UKR in the whole cohort decreased dramatically when UKR was performed by surgeons with more UKR experience. UKR patients in the whole cohort had more than double the risk of 5-year revision surgery than TKR patients. However, this increased risk dropped to around 40–50%, and was no longer significantly different from TKR patients' risk when the analysis was restricted to patients operated on by surgeons who had performed ≥ 50 surgeries of the same type in the previous year. This finding came closest to replicating the TOPKAT findings, for which the estimated odds ratio for revision was 1.40 (95% CI 0.50 to 4.00) for UKR versus TKR, but was limited by low statistical power and, therefore, needs further research in other settings.

These findings have both methodological consequences for future research in the field and clinical relevance. Given the observed interaction between the surgeons' experience and the obtained health outcomes (both OKS and revision risk), we suggest that UKR surgery be centralised in specialised treatment centres and provided by specialist surgeons. Our results suggest that patients will have the best possible results when their UKR surgeries are performed by surgeons who perform ≥ 50 UKR surgeries per year. This is equivalent to around one UKR surgery per working week. Conversely, UKR surgeries performed by surgeons who have performed ≤ 10 UKRs in the previous year may have suboptimal patient benefit and could even lead to higher risks of revision than TKR surgery.

Study conclusions: UTMoSt stage 2

UTMoSt stage 2 provided evidence on the comparative effectiveness, safety and cost-effectiveness of UKR compared with TKR for patients with multiple comorbidities, as measured by an ASA grade of 3 or 4. These patients would not have been eligible for TOPKAT, and it is unlikely that there will be a follow-up trial to include this subpopulation. To our knowledge, the results summarised here are the best-quality data available to date for the approximately 15–20% of patients who undergo knee replacement surgery in the NHS while having relatively poor health status.

The stage 2 analyses included a much smaller number of participants than stage 1: 2256 UKR patients and 57,682 TKR patients, of whom only 145 UKR patients and 23,344 TKR patients contributed primary outcome data. However, the safety analyses included almost 10 times more UKR patients and more than 200 times more TKR patients than TOPKAT.

Given that no RCT has included people comparable to the participants of UTMoSt stage 2, there are no gold standard data available for comparison. We, therefore, judged the performance of our analytical

methods by their ability to minimise confounding for the variables available in our analytical data set. These analyses were still limited by the effect of potential unobserved confounders that were not recorded in any of the three linked data sources used for UTMoSt (NJR, HES and NHS PROMs database).

Of the tested methods, PS stratification based on the whole cohort most efficiently minimised confounding for the primary outcome analysis. PS stratification based on the UKR cohort and IPW led to unacceptable imbalances in some confounders and required double adjustment for the imbalanced variables. The safety analysis included a much larger population than the primary outcome analysis. The three methods successfully achieved balance for all of the known confounders available in the UTMoSt data set for this larger analysis.

When using the preferred method of PS stratification based on the UKR cohort, UKR had similar effectiveness (compared with TKR) to that observed in TOPKAT and UTMoSt stage 1, with an ATE of 1.83 (95% CI 0.10 to 3.56) OKS points in favour of UKR. Although statistically significant, this increased patient-reported benefit for UKR over TKR is not likely to be clinically relevant. After double adjustment for unresolved imbalances, PSS_{whole} yielded an estimate of 1.82 (95% CI 0.10 to 3.56) and an IPW estimate of 1.00 (95% CI −1.28 to 3.27). In summary, UKR had similar comparative effectiveness in patients with severe systemic disease and/or substantial functional limitations, as defined by an ASA grade of ≥ 3, as in the overall population. There were small, clinically irrelevant differences in postoperative OKS between UKR and TKR in these patients. Sensitivity analyses of effectiveness restricted to more experienced surgeons could not be performed because of limited statistical power.

Safety considerations, particularly short-term complications, are particularly important for patients with multimorbidity. As the two surgical approaches had no appreciable difference in benefit, any differences in risks would be highly relevant for decision-making. All three analyses suggested a strongly protective effect against postoperative venous thromboembolism for UKR patients, with a 60–67% relative reduction in risk compared with TKR patients. Venous thromboembolism is the most common postoperative complication of knee replacement and affected up to 8% of TKR patients and just below 3% of UKR patients in UTMoSt. Acute myocardial infarction and prosthetic joint infection were also analysed. In the first 90 days after surgery, almost 5% of TKR patients and just over 3.5% of UKR patients experienced myocardial infarction, and 1.9% of TKR patients and 1.8% of UKR patients experienced a prosthetic joint infection. However, no clear statistical difference for either complication could be demonstrated.

TOPKAT's sample size did not give sufficient power to reliably study postoperative complications. However, our findings are consistent with a recent multinational collaboration study led by EHDEN (of which we are members) and OHDSI.[16] For example, Burn *et al.*[16] analysed over 32,000 UKR participants and more than 250,000 TKR participants after PS matching and found a 50% reduction in the risk of 90-day postoperative venous thromboembolism, but no significant reduction in the risk of infection. These results are also consistent with a meta-analysis published in *BMJ* in 2019,[88] which found that UKR was associated with a 60% reduction in the risk of postoperative venous thromboembolism compared with TKR.

The UTMoSt stage 2 results suggested that the relative effects of UKR and TKR on short-term postoperative complications in patients with severe systemic disease were consistent with those seen in previous literature for the wider population. UKR seemed to be safer in the short term and resulted in a lower (about 40–50% reduced) risk of postoperative venous thromboembolism in the 90 days after knee replacement surgery. This is highly relevant for patients and clinicians, as such complications can have deleterious effects on patients with baseline comorbidity.

We also assessed the effects of UKR (vs. TKR) on long-term consequences, 5-year revision surgery and mortality. UTMoSt stage 2 showed that, as expected, patients with complex health needs were more

likely to die than have revision surgery. The 5-year cumulative mortality rate was 24% for UKR patients and 37% for TKR patients, compared with cumulative revision rates of 13% and 5%, respectively. Survival analyses found that UKR was associated with an almost threefold higher revision risk than TKR in these patients with systemic comorbidity, but with > 30% reduction in all-cause mortality. However, these analyses were hampered by a lack of reliable information on post-revision mortality, as the analyses were censored at the earliest of both events, and on cause of death. More data are, therefore, required to elucidate how UKR (vs. TKR) reduced long-term mortality and how the observed excess revision risk for UKR patients affects subsequent (post-revision) risk of death.

TOPKAT found that UKR had a lower health-care cost (to the NHS) and better cost-effectiveness at 5 years post surgery,[5] with an additional benefit of 0.24 QALYs and a cost-saving of £910 per procedure over TKR. In UTMoSt stage 2, we used an economic evaluation to compare the cost-effectiveness of UKR and TKR for patients with severe systemic disease, as defined by an ASA grade of ≥ 3, who would have been excluded from TOPKAT. We used the three validated PS methods to minimise confounding, then analysed cost-effectiveness using similar health economics methods to those used in TOPKAT. These analyses demonstrated that UKR had an average cost of £6246 and TKR had a slightly higher average cost of £6627. Although UKR was associated with an increased revision risk, the cost of a revision after UKR was just over £5100, which was substantially lower than the cost of over £9100 associated with a revision after TKR. After discounting, UKR had a mean gain in quality of life within 5 years of 2.24 QALYs, higher than the 1.87 QALYs for TKR. The UTMoSt cost-effectiveness analysis suggested that UKR dominated TKR for patients with substantial comorbidity (ASA grade of 3 or 4), as it was more beneficial and less expensive.

The analyses performed in UTMoSt stage 2 were limited by the potential for residual confounding and information bias related to the use of data routinely collected for clinical purposes rather than for research purposes. However, it is unlikely that better data will be obtained from randomised studies any time soon. The striking finding that UKR was dominant over TKR for patients with multiple comorbidity should guide future provision of care in the NHS.

Public and patient involvement

A patient representative for the National Rheumatoid Arthritis Society was a co-applicant for the grant application. She assessed the study's topic and relevance. She was involved in co-investigator meetings for the study, during which the study progress and results were assessed and evaluated. She has not raised any concerns about the work. We also had a plan to include a patient and public involvement (PPI) representative in the Study Steering Committee. However, despite a substantial effort by Versus Arthritis, no PPI representative was recruited.

Implications for future research and clinical practice

The results of UTMoSt stage 1 have clinical and methodological implications. Despite challenges inherent in the nature of the data used for these analyses, our findings suggest that real-world evidence and some PS methods can reliably mimic surgical RCTs. This is of fundamental importance and is very timely, as coming changes in the regulation of medical devices will probably require comprehensive observational post-marketing surveillance.

Key recommendations arising from stage 1 for future research include:

- PS stratification and IPW are useful for minimising confounding when evaluating the comparative effectiveness and risks of alternative medical devices and surgical procedures. PS matching, PS adjustment and IV analyses should be used with caution as they failed to replicate the RCT results in our study.

- More methodological research is needed to produce guidance on which analytical methods are preferable in different scenarios and circumstances for the post-marketing surveillance of medical devices and surgical epidemiology. Real-world evidence studies will continue to emerge in the absence of equivalent RCTs. It is likely that observational post-marketing safety studies will grow in number with upcoming European and global regulations, creating an urgent need for better guidance on the best use of analytical methods when evaluating surgery and implantable medical devices.
- Future attempts to mimic surgical trials should take into account the eligibility criteria of both patients and surgeons contributing to RCTs, as the patients and clinicians participating in RCTs are not representative of routine NHS care.

The main clinical implication of UTMoSt stage 1 arises from the finding that the potential additional benefits of UKR in terms of reduced risk and increased benefit are achieved only when performed by surgeons with high volume in this surgical technique. Health-care services, including the NHS, should consider centralising the delivery of UKR in specialised centres or by specialised surgeons to maximise the potential cost-effectiveness gains described in TOPKAT and consistently demonstrated in a sub-analysis of UTMoSt stage 1 restricted to surgeons with higher volume of knee replacement surgery.[89]

UTMoSt stage 2 also has implications for clinical care. Although most surgeons support the use of UKR for fit young patients, our findings suggest that UKR has similar benefits over TKR for patients with severe systemic disease. We found that patients with severe systemic disease had better patient-reported outcomes (probably not clinically relevant) and dramatically fewer safety events, particularly thromboembolic events, after UKR than after TKR, as seen in fit young patients. Despite an excess revision risk, mortality was also lower among UKR patients. This information should be clearly communicated to patients. Patients with a limited lifespan might prefer a procedure that provides similar benefits but that is potentially safer in the short term, despite the fact that it may lead to higher risk of revision surgery in the long term.

From a NHS perspective, UKR is the preferable option for this patient subgroup where suitable, as it provides better and less expensive care than TKR. From NJR data, we estimate that about 50% of patients undergoing knee replacement would be suitable for UKR, but that $< 10\%$ receive it. More strikingly, less than 4% (2890/75,055) of patients with an ASA grade of ≥ 3 underwent UKR in our data set. There is a clear need for NICE guidelines on the use of UKR for patients with multimorbidity in need of knee replacement surgery.

Acknowledgements

We would like to thank Miss Susan Thwaite for her involvement in the study from the design stage as a PPI representative, and Professor Nicholas C Harvey, Professor Elaine Dennison and Mr Terry Lock for their role in the Steering Committee of the study. We acknowledge English-language editing by Dr Jennifer A de Beyer of the Centre for Statistics in Medicine, University of Oxford, Oxford, UK.

This study was approved by the Confidentiality Advisory Group (17/CAG/0174), National Joint Registry Research Sub-committee (RSC2016/13) and NHS Digital (DARS-NIC-172121-G0Z1H).

Contributions of authors

Albert Prats-Uribe (https://orcid.org/0000-0003-1202-9153) contributed to data management, cleaning and analysis, and led the drafting of the study report.

Spyros Kolovos (https://orcid.org/0000-0003-3201-1743) contributed to data management, cleaning and analysis, and led the drafting of the study report.

Klara Berencsi (https://orcid.org/0000-0002-2109-6369) contributed to data management, cleaning and analysis.

Andrew Carr (https://orcid.org/0000-0001-5940-1464) contributed to study design.

Andrew Judge (https://orcid.org/0000-0003-3015-0432) contributed to study design.

Alan Silman (https://orcid.org/0000-0001-8426-8925) contributed to study design.

Nigel Arden (https://orcid.org/0000-0002-3452-3382) contributed to study design.

Irene Petersen (https://orcid.org/0000-0002-0037-7524) contributed to study design.

Ian J Douglas (https://orcid.org/0000-0002-8970-1406) contributed to study design.

J Mark Wilkinson (https://orcid.org/0000-0001-5577-3674) contributed to study design.

David Murray (https://orcid.org/0000-0002-0839-3166) contributed to study design.

Jose M Valderas (https://orcid.org/0000-0002-9299-1555) contributed to study design.

David J Beard (https://orcid.org/0000-0001-7884-6389) contributed to study design.

Sarah E Lamb (https://orcid.org/0000-0003-4349-7195) contributed to study design.

M Sanni Ali (https://orcid.org/0000-0002-4192-7908) contributed to study design.

Rafael Pinedo-Villanueva (https://orcid.org/0000-0002-4723-5128) contributed to study design and data management, cleaning, and analysis, and led the drafting of the study report.

Victoria Y Strauss (https://orcid.org/0000-0002-5172-512X) contributed to data management, cleaning, and analysis and led the drafting of the study report.

Daniel Prieto-Alhambra (https://orcid.org/0000-0002-3950-6346) contributed to study design and data management, cleaning, and analysis, and led the drafting of the study report.

All authors provided feedback and critically reviewed the report, and approved the final version of the report for submission.

Data-sharing statement

All data requests should be submitted to the corresponding author for consideration. Access to anonymised data may be granted following review.

Patient data

This work uses data provided by patients and collected by the NHS as part of their care and support. Using patient data is vital to improve health and care for everyone. There is huge potential to make better use of information from people's patient records, to understand more about disease, develop new treatments, monitor safety, and plan NHS services. Patient data should be kept safe and secure, to protect everyone's privacy, and it's important that there are safeguards to make sure that it is stored and used responsibly. Everyone should be able to find out about how patient data are used. #datasaveslives You can find out more about the background to this citation here: https://understandingpatientdata.org.uk/data-citation.

References

1. Heads of Medicine Agencies, European Medicines Agency. *HMA-EMA Joint Big Data Taskforce: Summary Report*. Amsterdam: European Medicines Agency; 2019.
2. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019;**394**:1816–26. https://doi.org/10.1016/S0140-6736(19)32317-7
3. National Joint Registry (NJR). *12th Annual Report 2015. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Surgical Data to 31 December 2014*. Hempstead: NJR; 2015.
4. Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, *et al.* Total versus partial knee replacement in patients with medial compartment knee osteoarthritis: the TOPKAT RCT. *Health Technol Assess* 2020;**24**(20). https://doi.org/10.3310/hta24200
5. Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, *et al.* 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* 2019;**394**:746–56. https://doi.org/10.1016/S0140-6736(19)31281-4
6. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform* 2017;**66**:72–81. https://doi.org/10.1016/j.jbi.2016.12.005
7. Kendal AR, Prieto-Alhambra D, Arden NK, Carr A, Judge A. Mortality rates at 10 years after metal-on-metal hip resurfacing compared with total hip replacement in England: retrospective cohort analysis of hospital episode statistics. *BMJ* 2013;**347**:f6549. https://doi.org/10.1136/bmj.f6549
8. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6409. https://doi.org/10.1136/bmj.f6409
9. Liddle AD, Judge A, Pandit H, Murray DW. 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* 2014;**384**:1437–45. https://doi.org/10.1016/S0140-6736(14)60419-0
10. Liddle AD, Pandit H, Judge A, Murray DW. 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* 2015:**97-B**:793–801. https://doi.org/10.1302/0301-620X.97B6.35155
11. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;**19**:537–54. https://doi.org/10.1002/pds.1908
12. Judge A, Javaid MK, Leal J, Hawley S, Drew S, Sheard S, *et al.* Models of care for the delivery of secondary fracture prevention after hip fracture: a health service cost, clinical outcomes and cost-effectiveness study within a region of England. *Health Serv Deliv Res* 2016;**4**(28). https://doi.org/10.3310/hsdr04280
13. Patorno E, Schneeweiss S, Gopalakrishnan C, Martin D, Franklin JM. Using real-world data to predict findings of an ongoing phase IV cardiovascular outcome trial – cardiovascular safety of linagliptin vs. glimepiride. *Diabetes care* 2019;**42**:dc190069. https://doi.org/10.2337/dc19-0069

14. Hawley S, Cordtz R, Dreyer L, Edwards CJ, Arden NK, Delmestri A, *et al.* Association between NICE guidance on biologic therapies with rates of hip and knee replacement among rheumatoid arthritis patients in England and Wales: an interrupted time-series analysis. *Semin Arthritis Rheum* 2018;**47**:605–10. https://doi.org/10.1016/j.semarthrit.2017.09.006

15. Kynaston-Pearson F, Ashmore AM, Malak TT, Rombach I, Taylor A, Beard D, *et al.* Primary hip replacement prostheses and their evidence base: systematic review of literature. *BMJ* 2013;**347**:f6956. https://doi.org/10.1136/bmj.f6956

16. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, *et al.* Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study. *Lancet Rheumatol* 2019;**1**:E229–36. https://doi.org/10.1016/S2665-9913(19)30075-X

17. European Medicines Agency. *First Guidance on New Rules for Certain Medical Devices*. 2019. URL: www.ema.europa.eu/en/news/first-guidance-new-rules-certain-medical-devices (accessed November 2019).

18. National Joint Registry (NJR). *15th Annual Report 2018. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Surgical Data to 31st December 2017*. Hempstead: NJR; 2018.

19. National Joint Registry (NJR). *Patient Characteristics for Primary Knee Replacement Procedures*. Hempstead: NJR; 2017. https://reports.njrcentre.org.uk/knees-primary-procedures-patient-characteristics (accessed 1 December 2019).

20. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017;**46**:1093–1093i. https://doi.org/10.1093/ije/dyx015

21. National Joint Registry (NJR). *OPCS Codes Relevant to Procedures Recorded on the NJR*. Hempstead: NJR; 2016.

22. Devlin N, Appleby J. *Getting the Most Out of PROMS: Putting Health Outcomes at the Heart of NHS Decision-making*. London: The King's Fund; 2010.

23. Partridge T, Carluke I, Emmerson K, Partington P, Reed M. Improving patient reported outcome measures (PROMs) in total knee replacement by changing implant and preserving the infrapatella fatpad: a quality improvement project. *BMJ Qual Improv Rep* 2016;**5**:u204088.w3767. https://doi.org/10.1136/bmjquality.u204088.w3767

24. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996;**78**:185–90. https://doi.org/10.1302/0301-620X.78B2.0780185

25. Murray D, Fitzpatrick R, Rogers K, Pandit H, Beard D, Carr A, *et al.* The use of the Oxford hip and knee scores. *J Bone Joint Surg Br* 2007;**89**:1010–4. https://doi.org/10.1302/0301-620X.89B8.19424

26. Brooks R. EuroQol: the current state of play. *Health Policy* 1996;**37**:53–72. https://doi.org/10.1016/0168-8510(96)00822-6

27. Secondary Care Analysis (PROMs) ND. *Patient Reported Outcome Measures (PROMs) in England: A Guide to PROMs Methodology*. 2017. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms#proms-special-topics (accessed 1 December 2019).

28. Beard D, Price A, Cook J, Fitzpatrick R, Carr A, Campbell M, *et al.* Total or partial knee arthroplasty trial-TOPKAT: study protocol for a randomised controlled trial. *Trials* 2013;**14**:292. https://doi.org/10.1186/1745-6215-14-292

29. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;**155**:176–84. https://doi.org/10.1093/aje/155.2.176

30. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;**60**:578–86. https://doi.org/10.1136/jech.2004.029496

31. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials* 2009;**10**:9. https://doi.org/10.1186/1745-6215-10-9

32. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;**332**:1080. https://doi.org/10.1136/bmj.332.7549.1080

33. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55. https://doi.org/10.1093/biomet/70.1.41

34. Nguyen TL, Collins GS, Spence J, Daurès JP, Devereaux PJ, Landais P, Le Manach Y. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Med Res Methodol* 2017;**17**:78. https://doi.org/10.1186/s12874-017-0338-0

35. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009;**20**:512–22. https://doi.org/10.1097/EDE.0b013e3181a663cc

36. Uddin MJ, Groenwold RH, Ali MS, de Boer A, Roes KC, Chowdhury MA, Klungel OH. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *Int J Clin Pharm* 2016;**38**:714–23. https://doi.org/10.1007/s11096-016-0299-0

37. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res* 2012;**21**:273–93. https://doi.org/10.1177/0962280210394483

38. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004;**23**:2937–60. https://doi.org/10.1002/sim.1903

39. Pitblado J. *Survey Data Analysis in Stata*. Canadian Stata Users Group Meeting, 22 October 2009. Toronto, Canada, *abstract no. 48*.

40. Ali MS, Prieto-Alhambra D, Lopes LC, Ramos D, Bispo N, Ichihara MY, *et al.* Propensity score methods in health technology assessment: principles, extended applications, and recent advances. *Front Pharmacol* 2019;**10**:973. https://doi.org/10.3389/fphar.2019.00973

41. Desai RJ, Rothman KJ, Bateman BT, Hernandez–Diaz S, Huybrechts KF. A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology* 2017;**28**:249–257. https://doi.org/10.1097/EDE.0000000000000595

42. Sauerbrei W, Royston P. *Fractional Polynomials*. URL: http://mfp.imbi.uni-freiburg.de/fp (accessed November 2019).

43. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf* 2004;**13**:855–7. https://doi.org/10.1002/pds.968

44. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;**46**:399–424. https://doi.org/10.1080/00273171.2011.568786

45. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;**297**:278–85. https://doi.org/10.1001/jama.297.3.278

46. Uddin MJ, Groenwold RH, de Boer A, Belitser SV, Roes KC, Hoes AW, Klungel OH. Performance of instrumental variable methods in cohort and nested case-control studies: a simulation study. *Pharmacoepidemiol Drug Saf* 2014;**23**:165–77. https://doi.org/10.1002/pds.3555

47. Beard DJ, Holt MD, Mullins MM, Malek S, Massa E, Price AJ. Decision making for knee replacement: variation in treatment choice for late stage medial compartment osteoarthritis. *Knee* 2012;**19**:886–9. https://doi.org/10.1016/j.knee.2012.05.005

48. Khatri PJ, O'Connor AM, Dervin GF. Decision support needs of patients choosing between unicompartmental and total knee arthroplasty for advanced medial compartment osteoarthritis of the knee. *J Arthroplasty* 2011;**26**:1343–9. https://doi.org/10.1016/j.arth.2010.12.016

49. Judge A, Welton NJ, Sandhu J, Ben-Shlomo Y. Equity in access to total joint replacement of the hip and knee in England: cross sectional study. *BMJ* 2010;**341**:c4092. https://doi.org/10.1136/bmj.c4092

50. Garriga C, Leal J, Sánchez-Santos MT, Arden N, Price A, Prieto-Alhambra D, *et al.* Geographical variation in outcomes of primary hip and knee replacement. *JAMA Netw Open* 2019;**2**:e1914325. https://doi.org/10.1001/jamanetworkopen.2019.14325

51. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology* 2006;**17**:260–7. https://doi.org/10.1097/01.ede.0000215160.88317.cb

52. Ali MS, Uddin MJ, Groenwold RH, Pestman WR, Belitser SV, Hoes AW, *et al.* Quantitative falsification of instrumental variables assumption using balance measures. *Epidemiology* 2014;**25**:770–2. https://doi.org/10.1097/EDE.0000000000000152

53. Reyes C, Pottegård A, Schwarz P, Javaid MK, Van Staa TP, Cooper C, *et al.* Real-life and RCT participants: alendronate users versus FITs' trial eligibility criterion. *Calcif Tissue Int* 2016;**99**:243–9. https://doi.org/10.1007/s00223-016-0141-7

54. Groenwold RHH, Uddin MJ, Roes KCB, de Boer A, Rivero-Ferrer E, Martin E, *et al.* Instrumental variable analysis in randomized trials with non-compliance and observational pharmacoepidemiologic studies. *OA Epidemiology* 2014;**2**:9.

55. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, Price AJ. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol* 2015;**68**:73–9. https://doi.org/10.1016/j.jclinepi.2014.08.009

56. Sedgwick P. Meta-analyses: what is heterogeneity? *BMJ* 2015;**350**:h1435. https://doi.org/10.1136/bmj.h1435

57. Rahal R, Collaboration OS. Estimating the reproducibility of psychological science. *Science* 2015;**349**:aac4716. https://doi.org/10.1126/science.aac4716

58. Franklin J. *Comparing Real World Evidence with Randomized Trial Results to Assess Validity. Anticipated Learnings from the RCT DUPLICATE Initiative.* Paper presented at the 35th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, 27 August 2019, Philadelphia, PA, USA.

59. RCT Duplicate. *Effectiveness Research with Real-world Data to Support FRA's Regulatory Decision Making: A Real World Evidence Demonstration Project*. URL: www.rctduplicate.org/fda-demonstration-project.html (accessed 1 December 2019).

60. Seaman SR, Vansteelandt S. Introduction to double robust methods for incomplete data. *Stat Sci* 2018;**33**:184–97. https://doi.org/10.1214/18-STS647

61. Burn E, Edwards CJ, Murray DW, Silman A, Cooper C, Arden NK, *et al.* The impact of rheumatoid arthritis on the risk of adverse events following joint replacement: a real-world cohort study. *Clin Epidemiol* 2018;**10**:697–704. https://doi.org/10.2147/CLEP.S160347

62. Burn E, Liddle AD, Hamilton TW, Judge A, Pandit HG, Murray DW, Pinedo-Villanueva R. Cost-effectiveness of unicompartmental compared with total knee replacement: a population-based study using data from the National Joint Registry for England and Wales. *BMJ Open* 2018;**8**:e020977. https://doi.org/10.1136/bmjopen-2017-020977

63. Peersman G, Jak W, Vandenlangenbergh T, Jans C, Cartier P, Fennema P. Cost-effectiveness of unicondylar versus total knee arthroplasty: a Markov model analysis. *Knee* 2014;**21**(Suppl. 1):37–42. https://doi.org/10.1016/S0968-0160(14)50008-7

64. NHS Digital. *Hospital Admitted Patient Care Activity, 2016–17*. Leeds: NHS Digital; 2017.

65. NHS Digital. *Finalised Patient Reported Outcome Measures (PROMs) in England for Hip and Knee Replacement Procedures (April 2017 to March 2018)*. Leeds: NHS Digital; 2019.

66. Porter M, Armstrong R, Howard P, Porteous M, Wilkinson JM. Orthopaedic registries – the UK view (National Joint Registry): impact on practice. *EFORT Open Rev* 2019;**4**:377–90. https://doi.org/10.1302/2058-5241.4.180084

67. The EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208. https://doi.org/10.1016/0168-8510(90)90421-9

68. Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997;**35**:1095–108. https://doi.org/10.1097/00005650-199711000-00002

69. Sassi F. Calculating QALYs, comparing QALY and DALY calculations. *Health Policy Plan* 2006;**21**:402–8. https://doi.org/10.1093/heapol/czl018

70. Healthcare Resource Groups 4 (HRG4). *Full Operational Information Standard* 2018. https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb-0070-healthcare-resource-groups-hrgs (accessed 1 December 2019).

71. National Casemix Classifications Service. *Guide to Unbundling*. 2019.

72. NHS Improvement. *Reference Costs 2017/18: Highlights, Analysis and Introduction to the Data*. London: NHS Improvement; 2018.

73. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther* 2019;**105**:867–77. https://doi.org/10.1002/cpt.1351

74. National Institute for Clinical Excellence (NICE). *Guide to the Processes of Technology Appraisal*. London: NICE; 2018.

75. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;**352**:i189. https://doi.org/10.1136/bmj.i189

76. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;**20**:40–9. https://doi.org/10.1002/mpr.329

77. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? *Int J Epidemiol* 2014;**43**:1336–9. https://doi.org/10.1093/ije/dyu080

78. Vroomen JM, Eekhout I, Dijkgraaf MG, van Hout H, de Rooij SE, Heymans MW, *et al.* Multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best? *Eur J Health Econ* 2016;**17**:939–50. https://doi.org/10.1007/s10198-015-0734-5

79. Faria R, Gomes M, Epstein D, White IR. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. *PharmacoEconomics* 2014;**32**:1157–70. https://doi.org/10.1007/s40273-014-0193-3

80. Little RJ. Missing-data adjustments in large surveys. *J Bus Econ Stat* 1988;**6**:287–96. https://doi.org/10.1080/07350015.1988.10509663

81. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014;**14**:75. https://doi.org/10.1186/1471-2288-14-75

82. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.

83. Hunter RM, Baio G, Butt T, Morris S, Round J, Freemantle N. An educational review of the statistical issues in analysing utility data for cost-utility analysis. *PharmacoEconomics* 2015;**33**:355–66. https://doi.org/10.1007/s40273-014-0247-6

84. Black WC. The CE plane: a graphic representation of cost-effectiveness. *Med Decis Making* 1990;**10**:212–14. https://doi.org/10.1177/0272989X9001000308

85. National Institute for Clinical Excellence (NICE). *Guide to the Methods of Technology Appraisal*. London: NICE; 2013.

86. Ryan P, Hripcsak G. *A Journey Toward Real-world Evidence for Regulatory Decision-making*. www.ohdsi.org/wp-content/uploads/2019/09/4-Plenary-3-Replicating-LEGEND.pdf (accessed 1 December 2019).

87. Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, *et al.* A multicentre randomised controlled trial assessing clinical and cost effectiveness of total versus partial knee replacement (TOPKAT). *Health Technol Assess* 2021; in press.

88. Wilson HA, Middleton R, Abram SGF, Smith S, Alvand A, Jackson WF, *et al.* Patient relevant outcomes of unicompartmental versus total knee replacement: systematic review and meta-analysis. *BMJ* 2019;**364**:l352. https://doi.org/10.1136/bmj.l352

89. National Institute for Health Research. *Partial Knee Replacement 'Could Be First Choice' for Suitable Patients with Osteoarthritis*. 2019. URL: https://discover.dc.nihr.ac.uk/content/signal-000824/partial-knee-replacement-could-be-first-choice-in-some-patients (accessed November 2019).

Appendix 1 Supplementary figures and tables

TABLE 23 Baseline patient-level characteristics before and after PS matching in the OKS cohort

	Before PS matching		After PS matching	
Characteristic	**TKR (*N* = 125,834)**	**UKR (*N* = 1197)**	**TKR (*N* = 5652)**	**UKR (*N* = 1197)**
Sex, *n* (%)				
Female	70,671 (56)	576 (48)	2691 (48)	576 (48)
Male	55,163 (44)	621 (52)	2961 (52)	621 (52)
Rural Index, *n* (%)				
1	92,052 (73)	844 (71)	4038 (71)	844 (71)
2	15,730 (13)	164 (14)	737 (13)	164 (14)
3	12,637 (10)	138 (12)	638 (11)	138 (12)
4	5415 (4)	51 (4)	239 (4)	51 (4)
IMD, *n* (%)				
Least deprived 10%	14,168 (11)	149 (12)	693 (12)	149 (12)
Less deprived				
10–19%	15,194 (12)	137 (11)	664 (12)	137 (11)
20–29%	15,435 (12)	142 (12)	702 (12)	142 (12)
30–39%	15,405 (12)	138 (12)	672 (12)	138 (12)
40–49%	14,611 (12)	164 (14)	714 (13)	164 (14)
More deprived				
10–19%	8628 (7)	102 (9)	453 (8)	102 (9)
20–29%	10,110 (8)	84 (7)	407 (7)	84 (7)
30–39%	11,621 (9)	123 (10)	578 (10)	123 (10)
40–49%	13,557 (11)	106 (9)	515 (9)	106 (9)
Most deprived 10%	7105 (6)	52 (4)	254 (4)	52 (4)
ASA grade, *n* (%)				
P1: fit and healthy	13,849 (11)	242 (20)	1100 (19)	242 (20)
P2: mild disease not incapacitating	111,985 (89)	955 (80)	4552 (81)	955 (80)
Charlson Comorbidity Index score, *n* (%)				
0	86,474 (69)	915 (76)	4319 (76)	915 (76)
1	26,733 (21)	224 (19)	1045 (18)	224 (19)
2	8357 (7)	41 (3)	204 (4)	41 (3)
3	6172 (3)	308 (1)	63 (1)	13 (1)
4	3234 (1)	180 (1)	21 (0)	4 (0)
Age (years), mean (SD)	70.4 (8.6)	64.9 (9.4)	65.3 (9.0)	64.9 (9.4)
BMI (kg/m^2), mean (SD)	30.4 (5.0)	29.6 (4.7)	29.6 (4.9)	29.6 (4.7)
PROMs				
Preoperative OKS, mean (SD)	19.7 (7.6)	21.9 (7.5)	21.82 (7.73)	21.88 (7.52)
EQ-5D, mean (SD)	70.0 (19.2)	71.1 (19.0)	71.09 (19.22)	71.13 (18.97)

continued

TABLE 23 Baseline patient-level characteristics before and after PS matching in the OKS cohort (*continued*)

Characteristic	Before PS matching		After PS matching	
	TKR (*N* = 125,834)	UKR (*N* = 1197)	TKR (*N* = 5652)	UKR (*N* = 1197)
General health, *n* (%)				
0	88,778 (71)	604 (50)	2918 (52)	604 (50)
1	1433 (1)	33 (3)	138 (2)	33 (3)
2	10,398 (8)	181 (15)	815 (14)	181 (15)
3	17,504 (14)	271 (23)	1273 (23)	271 (23)
4	6886 (5)	94 (8)	442 (8)	94 (8)
5	835 (1)	14 (1)	66 (1)	14 (1)
Medical history, *n* (%)				
Gastrointestinal disease	25,142 (20)	174 (15)	798 (14)	174 (15)
Other joint problems	23,578 (19)	149 (12)	751 (13)	149 (12)
Mental health	11,421 (9)	101 (8)	469 (8)	101 (8)
Respiratory diseases	17,078 (14)	147 (12)	686 (12)	147 (12)
Cardiovascular diseases	73,382 (58)	515 (43)	2465 (44)	515 (43)
Thyroid problems	9742 (8)	80 (7)	327 (6)	80 (7)
Foot, hip or spinal pain	1519 (1)	15 (1)	79 (1)	15 (1)
Coxarthrosis	4395 (3)	25 (2)	121 (2)	25 (2)
Neurological disorders	7491 (6)	67 (6)	306 (5)	67 (6)
Other arthrosis	5930 (5)	41 (3)	210 (4)	41 (3)
Polyarthrosis	7520 (6)	29 (2)	140 (2)	29 (2)
Spondylosis	3501 (3)	17 (1)	93 (2)	17 (1)

TABLE 24 Baseline patient-level characteristics before and after PS matching in the revision cohort

Characteristic	Before PS matching		After PS matching	
	TKR (*N* = 273,530)	UKR (*N* = 21,026)	TKR (*N* = 92,071)	UKR (*N* = 21,026)
Sex, *n* (%)				
Female	155,267 (57)	10,016 (48)	35,300 (50)	10,016 (48)
Male	118,263 (43)	11,010 (52)	35,745 (50)	11,010 (52)
Rural Index, *n* (%)				
1	203,938 (74)	14,607 (70)	50,141 (71)	14,607 (69)
2	32,573 (12)	2698 (13)	9035 (13)	2698 (13)
3	26,012 (10)	2596 (12)	8288 (12)	2596 (12)
4	11,007 (4)	1125 (5)	3581 (5)	1125 (5)
IMD, *n* (%)				
Least deprived 10%	29,339 (11)	2917 (14)	9315 (13)	2917 (14)
Less deprived				
10–19%	31,518 (12)	2871 (14)	9325 (13)	2871 (14)
20–29%	31,946 (12)	2669 (13)	8899 (13)	2669 (13)
30–39%	32,593 (12)	2480 (12)	8422 (12)	2480 (12)
40–49%	31,209 (11)	2456 (12)	8276 (12)	2456 (12)

TABLE 24 Baseline patient-level characteristics before and after PS matching in the revision cohort (*continued*)

Characteristic	Before PS matching		After PS matching	
	TKR (*N* = 273,530)	UKR (*N* = 21,026)	TKR (*N* = 92,071)	UKR (*N* = 21,026)
More deprived				
10–19%	20,502 (7)	1224 (6)	4394 (6)	1224 (6)
20–29%	23,357 (9)	1415 (7)	5087 (7)	1415 (7)
30–39%	26,174 (10)	1917 (9)	6570 (9)	1917 (9)
40–49%	29,479 (11)	2156 (10)	7339 (10)	2156 (10)
Most deprived 10%	17,413 (6)	921 (4)	3418 (5)	921 (4)
ASA grade, *n* (%)				
P1: fit and healthy	30,224 (11)	4394 (21)	12,213 (17)	4394 (21)
P2: mild disease not incapacitating	243,306 (89)	16,632 (79)	58,832 (83)	16,632 (79)
Charlson Comorbidity Index score, *n* (%)				
0	187,509 (69)	15,408 (73)	51,019 (72)	15,408 (73)
1	58,781 (21)	4134 (20)	14,467 (20)	4134 (20)
2	17,834 (7)	996 (5)	3759 (5)	996 (5)
3	2846 (2)	13 (1)	1153 (2)	308 (1)
4	1424 (1)	4 (0)	647 (1)	180 (1)
Age (years), mean (SD)	70.2 (8.9)	64.3 (9.5)	66.1 (9.1)	64.3 (9.5)
BMI (kg/m^2), mean (SD)	30.5 (5.1)	30.0 (4.9)	30.2 (5.1)	30.0 (4.9)
PROMs				
Preoperative OKS, mean (SD)	19.3 (6.8)	21.3 (6.2)	1.20 (1.25)	1.30 (1.23)
EQ-5D, mean (SD)	69.2 (19.4)	69.7 (19.2)	69.66 (19.43)	69.70 (19.17)
General health, *n* (%)				
0	161,904 (59)	6546 (31)	26,651 (38)	6546 (31)
1	43,913 (16)	6643 (32)	19,224 (28)	6643 (32)
2	30,058 (11)	4400 (21)	13,233 (18)	4400 (21)
3	26,008 (9)	2217 (10)	7838 (11)	2217 (10)
4	10,024 (4)	834 (4)	2926 (4)	834 (4)
5	1623 (1)	386 (2)	852 (1)	386 (2)
Medical history, *n* (%)				
Gastrointestinal disease	52,029 (19)	3621 (17)	12,701 (18)	3621 (17)
Other joint problems	49,941 (18)	2696 (13)	9998 (14)	2696 (13)
Mental health	25,823 (9)	2380 (11)	7645 (11)	2380 (11)
Respiratory diseases	37,754 (14)	2827 (13)	9636 (14)	2827 (13)
Cardiovascular diseases	157,504 (58)	9592 (46)	35,015 (49)	9592 (46)
Thyroid problems	20,724 (8)	1249 (6)	4568 (6)	1249 (6)
Foot, hip or spinal pain	3096 (1)	205 (1)	731 (1)	205 (1)
Coxarthrosis	8966 (3)	381 (2)	1461 (2)	381 (2)
Neurological disorders	16,435 (6)	1208 (6)	4201 (6)	1208 (6)
Other arthrosis	12,818 (5)	708 (3)	2511 (4)	708 (3)
Polyarthrosis	15,935 (6)	675 (3)	2665 (4)	675 (3)
Spondylosis	7378 (3)	349 (2)	1349 (2)	349 (2)

TABLE 25 Baseline characteristics of study participants receiving UKR vs. TKR in a sensitivity analysis of patients with OKS data and who were operated on by surgeons who had performed ≥ 10 surgeries of the same type in the previous year

Characteristic	Treatment group	
	TKR (*N* = 114,871)	UKR (*N* = 602)
Sex, *n* (%)		
Female	64,468 (56)	287 (48)
Male	50,403 (44)	315 (52)
Rural Index, *n* (%)		
1	83,810 (73)	396 (66)
2	14,446 (13)	97 (16)
3	11,587 (10)	79 (13)
4	5028 (4)	30 (5)
IMD, *n* (%)		
Least deprived 10%	12,981 (11)	75 (12)
Less deprived		
10–19%	13,992 (12)	72 (12)
20–29%	14,159 (12)	82 (14)
30–39%	14,140 (12)	65 (11)
40–49%	13,371 (12)	94 (16)
More deprived		
10–19%	7731 (7)	42 (7)
20–29%	9178 (8)	39 (6)
30–39%	10,551 (9)	59 (10)
40–49%	12,333 (11)	52 (9)
Most deprived 10%	6435 (6)	22 (4)
ASA grade, *n* (%)		
P1: fit and healthy	12,748 (11)	118 (20)
P2: mild disease not incapacitating	102,123 (89)	484 (80)
Charlson Comorbidity Index score, *n* (%)		
0	79,157 (69)	447 (74)
1	24,269 (21)	121 (20)
2	7582 (7)	23 (4)
3	2579 (2)	8 (1)
4	1284 (1)	3 (0)
Age (years), mean (SD)	70.3 (8.6)	65.6 (9.3)
BMI (kg/m^2), mean (SD)	30.4 (5.0)	29.5 (4.6)
PROMs		
Preoperative OKS, mean (SD)	19.7 (7.6)	22.1 (7.6)
EQ-5D, mean (SD)	70.0 (19.2)	71.3 (18.8)
General health, *n* (%)		
0	81,617 (71)	306 (51)
1	1311 (1)	14 (2)
2	9395 (8)	100 (17)
3	15,652 (14)	128 (21)
4	6148 (5)	48 (8)
5	748 (1)	6 (1)

TABLE 25 Baseline characteristics of study participants receiving UKR vs. TKR in a sensitivity analysis of patients with OKS data and who were operated on by surgeons who had performed ≥ 10 surgeries of the same type in the previous year (*continued*)

Characteristic	Treatment group	
	TKR (*N* = 114,871)	UKR (*N* = 602)
Medical history, *n* (%)		
Gastrointestinal disease	22,766 (20)	93 (15)
Osteoarthritis and other joint problems	21,434 (19)	71 (12)
Mental health	10,528 (9)	46 (8)
Respiratory diseases	15,503 (13)	79 (13)
Cardiovascular diseases	66,546 (58)	272 (45)
Thyroid problems	8868 (8)	39 (6)
Foot, hip or spinal pain	1408 (1)	7 (1)
Coxarthrosis	4000 (3)	14 (2)
Neurological disorders	6794 (6)	38 (6)
Other arthrosis	5340 (5)	15 (2)
Polyarthrosis	6877 (6)	9 (1)
Spondylosis	3196 (3)	7 (1)

TABLE 26 Baseline characteristics of study participants receiving UKR vs. TKR in a sensitivity analysis of patients operated on by surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 surgeries of the same type in the previous year

Characteristic	Full cohort ≥ 10 surgeries		Full cohort ≥ 30 surgeries		Full cohort ≥ 50 surgeries	
	TKR (*N* = 248,785)	UKR (*N* = 13,334)	TKR (*N* = 195,898)	UKR (*N* = 5555)	TKR (*N* = 139,396)	UKR (*N* = 2550)
Sex, *n* (%)						
Female	141,124 (57)	6401 (48)	110,807 (57)	2636 (47)	78,641 (56)	1242 (49)
Male	107,661 (43)	6933 (52)	85,091 (43)	2919 (53)	60,755 (44)	1308 (51)
Rural Index, *n* (%)						
1	185,028 (74)	8984 (67)	144,874 (74)	3513 (63)	102,350 (73)	1550 (61)
2	29,793 (12)	1810 (14)	23,913 (12)	815 (15)	17,349 (12)	379 (15)
3	23,784 (10)	1790 (13)	18,964 (10)	881 (16)	13,711 (10)	443 (17)
4	10,180 (4)	750 (6)	8147 (4)	346 (6)	5986 (4)	178 (7)
IMD, *n* (%)						
Least deprived 10%	26,808 (11)	1936 (15)	21,504 (11)	823 (15)	15,215 (11)	386 (15)
Less deprived						
10–19%	28,936 (12)	1908 (14)	23,211 (12)	827 (15)	16,656 (12)	366 (14)
20–29%	29,178 (12)	1711 (13)	23,177 (12)	775 (14)	16,663 (12)	383 (15)
30–39%	29,751 (12)	1572 (12)	23,744 (12)	674 (12)	17,103 (12)	292 (11)
40–49%	28,532 (11)	1605 (12)	22,448 (11)	716 (13)	16,100 (12)	347 (14)
More deprived						
10–19%	18,313 (7)	678 (5)	14,045 (7)	212 (4)	9735 (7)	91 (4)
20–29%	21,123 (8)	873 (7)	16,385 (8)	323 (6)	11,642 (8)	135 (5)
30–39%	23,669 (10)	1169 (9)	18,462 (9)	445 (8)	13,084 (9)	200 (8)
40–49%	26,781 (11)	1371 (10)	20,990 (11)	582 (10)	14,945 (11)	287 (11)
Most deprived 10%	15,694 (6)	511 (4)	11,932 (6)	178 (3)	8253 (6)	63 (2)

continued

TABLE 26 Baseline characteristics of study participants receiving UKR vs. TKR in a sensitivity analysis of patients operated on by surgeons who had performed ≥ 10, ≥ 30 and ≥ 50 surgeries of the same type in the previous year (*continued*)

Characteristic	Full cohort ≥ 10 surgeries		Full cohort ≥ 30 surgeries		Full cohort ≥ 50 surgeries	
	TKR (*N* = 248,785)	UKR (*N* = 13,334)	TKR (*N* = 195,898)	UKR (*N* = 5555)	TKR (*N* = 139,396)	UKR (*N* = 2550)
ASA grade, *n* (%)						
P1: fit and healthy	27,829 (11)	2707 (20)	22,227 (11)	1104 (20)	15,725 (11)	539 (21)
P2: mild disease not incapacitating	220,956 (89)	10,627 (80)	173,671 (89)	4451 (80)	123,671 (89)	2011 (79)
Charlson Comorbidity Index score, *n* (%)						
0	170,990 (69)	9694 (73)	134,945 (69)	3980 (72)	95,884 (69)	1862 (73)
1	53,212 (21)	2645 (20)	41,686 (21)	1124 (20)	29,628 (21)	481 (19)
2	16,101 (6)	652 (5)	12,654 (6)	304 (5)	9108 (7)	151 (6)
3	5586 (2)	222 (2)	4374 (2)	99 (2)	3166 (2)	38 (1)
4	2896 (1)	121 (1)	2239 (1)	48 (1)	1610 (1)	18 (1)
Age (years), mean (SD)	70.2 (9.0)	64.8 (9.5)	70.1 (9.0)	65.5 (9.6)	70.0 (9.0)	65.7 (9.6)
BMI (kg/m^2), mean (SD)	30.5 (5.1)	30.0 (4.9)	30.4 (5.1)	29.8 (5.0)	30.4 (5.1)	29.8 (4.8)
PROMs						
Preoperative OKS, mean (SD)	19.3 (6.8)	21.4 (6.2)	19.4 (6.8)	21.6 (6.2)	19.4 (6.9)	21.7 (6.1)
EQ-5D, mean (SD)	69.3 (19.4)	69.9 (19.2)	69.4 (19.4)	70.3 (19.1)	69.5 (19.4)	70.3 (19.2)
General health, *n* (%)						
0	147,872 (59)	4099 (31)	118,240 (60)	1743 (31)	85,617 (61)	817 (32)
1	40,068 (16)	4324 (32)	31,166 (16)	1846 (33)	21,766 (16)	860 (34)
2	27,233 (11)	2819 (21)	21,043 (11)	1121 (20)	14,603 (10)	502 (20)
3	23,188 (9)	1357 (10)	17,617 (9)	559 (10)	12,116 (9)	239 (9)
4	8944 (4)	489 (4)	6709 (3)	202 (4)	4548 (3)	94 (4)
5	1480 (1)	246 (2)	1123 (1)	84 (2)	746 (1)	38 (1)
Medical history, *n* (%)						
Gastrointestinal disease	46,976 (19)	2346 (18)	36,986 (19)	1025 (18)	26,435 (19)	449 (18)
Osteoarthritis and other joint problems	45,193 (18)	1655 (12)	35,589 (18)	677 (12)	25,192 (18)	296 (12)
Mental health	23,773 (10)	1487 (11)	19,113 (10)	630 (11)	13,867 (10)	286 (11)
Respiratory diseases	34,160 (14)	1793 (13)	26,882 (14)	750 (14)	19,222 (14)	306 (12)
Cardiovascular diseases	142,322 (57)	6275 (47)	111,485 (57)	2604 (47)	79,174 (57)	1167 (46)
Thyroid problems	18,786 (8)	794 (6)	14,744 (8)	322 (6)	10,479 (8)	151 (6)
Foot, hip or spinal pain	2831 (1)	127 (1)	2220 (1)	50 (1)	1574 (1)	28 (1)
Coxarthrosis	8158 (3)	245 (2)	6454 (3)	106 (2)	4518 (3)	42 (2)
Neurological disorders	14,848 (6)	796 (6)	11,684 (6)	335 (6)	8409 (6)	144 (6)
Other arthrosis	11,518 (5)	449 (3)	9029 (5)	204 (4)	6399 (5)	94 (4)
Polyarthrosis	14,466 (6)	371 (3)	11,306 (6)	140 (3)	7912 (6)	64 (3)
Spondylosis	6677 (3)	215 (2)	5344 (3)	76 (1)	3812 (3)	25 (1)

TABLE 27 Myocardial infarction ICD-10 codes

ICD-10 code	Description
I200	Unstable angina
I208	Other forms of angina pectoris
I209	Angina pectoris, unspecified
I210	Acute transmural myocardial infarction of anterior wall
I211	Acute transmural myocardial infarction of inferior wall
I212	Acute transmural myocardial infarction of other sites
I213	Acute transmural myocardial infarction of unspecified site
I214	Acute subendocardial myocardial infarction
I219	Acute myocardial infarction, unspecified
I220	Subsequent myocardial infarction of anterior wall
I221	Subsequent myocardial infarction of inferior wall
I228	Subsequent myocardial infarction of other sites
I229	Subsequent myocardial infarction of unspecified site
I241	Dressler syndrome
I248	Other forms of acute ischaemic heart disease
I249	Acute ischaemic heart disease, unspecified
I251	Atherosclerotic heart disease
I255	Ischaemic cardiomyopathy
I256	Silent myocardial ischaemia
I258	Other forms of chronic ischaemic heart disease
I259	Chronic ischaemic heart disease, unspecified

TABLE 28 Venous thromboembolism ICD-10 codes

ICD-10 code	Description
I801	Phlebitis and thrombophlebitis of femoral vein
I802	Phlebitis and thrombophlebitis of other deep vessels of lower extremities
I803	Phlebitis and thrombophlebitis of lower extremities, unspecified
I260	Pulmonary embolism with mention of acute cor pulmonale
I269	Pulmonary embolism without mention of acute cor pulmonale

TABLE 29 Prosthetic joint infection ICD-10 codes

ICD-10 code	Description
T845	Infection and inflammatory reaction due to internal joint prosthesis
T846	Infection and inflammatory reaction due to internal fixation device of unspecified site
T847	Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts
T857	Infection and inflammatory reaction due to other internal prosthetic devices, implants and grafts
T814	Infection following a procedure
T813	Disruption of wound, not elsewhere classified
AND	
Debridement and implant retention (up to 1 year from PJI diagnosis)	
OPCS-4	***Description***
W801	Open debridement and irrigation of joint
W802	Open debridement of joint NEC
W808	Other specified debridement and irrigation of joint
W809	Unspecified debridement and irrigation of joint
OR	
Revision (up to 1 year from PJI diagnosis)	

NEC, not elsewhere classified; PJI, prosthetic joint infection.

TABLE 30 Baseline characteristics of participants in the safety cohorts included in the sensitivity analysis of experienced surgeons

	Subcohort ≥ 10 surgeries		Subcohort ≥ 30 surgeries		Subcohort ≥ 50 surgeries	
Characteristic	TKR (*N* = 51,118)	UKR (*N* = 1449)	TKR (*N* = 38,321)	UKR (*N* = 610)	TKR (*N* = 25,944)	UKR (*N* = 242)
Sex, *n* (%)						
Female	28,470 (56)	627 (43)	21,296 (56)	280 (46)	14,310 (55)	117 (48)
Male	22,648 (44)	822 (57)	17,025 (44)	330 (54)	11,634 (45)	125 (52)
Rural Index, *n* (%)						
1	39,185 (77)	1026 (71)	29,239 (76)	404 (66)	19,669 (76)	159 (66)
2	6069 (12)	172 (12)	4661 (12)	78 (13)	3192 (12)	28 (12)
3	4319 (8)	181 (12)	3266 (9)	92 (15)	2259 (9)	42 (17)
4	1545 (3)	70 (5)	1155 (3)	36 (6)	824 (3)	13 (5)
IMD, *n* (%)						
Least deprived 10%	4275 (8)	220 (15)	3287 (9)	101 (17)	2277 (9)	44 (18)
Less deprived						
10–19%	5146 (10)	194 (13)	3925 (10)	99 (16)	2708 (10)	43 (18)
20–29%	5606 (11)	145 (10)	4246 (11)	49 (8)	2938 (11)	23 (10)
30–39%	5613 (11)	136 (9)	4237 (11)	56 (9)	2951 (11)	20 (8)
40–49%	5705 (11)	178 (12)	4252 (11)	74 (12)	2890 (11)	34 (14)

TABLE 30 Baseline characteristics of participants in the safety cohorts included in the sensitivity analysis of experienced surgeons (*continued*)

Characteristic	Subcohort ≥ 10 surgeries		Subcohort ≥ 30 surgeries		Subcohort ≥ 50 surgeries	
	TKR (*N* = 51,118)	UKR (*N* = 1449)	TKR (*N* = 38,321)	UKR (*N* = 610)	TKR (*N* = 25,944)	UKR (*N* = 242)
More deprived						
10–19%	4752 (9)	97 (7)	3534 (9)	41 (7)	2300 (9)	13 (5)
20–29%	4883 (10)	112 (8)	3624 (9)	49 (8)	2436 (9)	17 (7)
30–39%	5137 (10)	147 (10)	3782 (10)	49 (8)	2517 (10)	18 (7)
40–49%	5505 (11)	151 (10)	4118 (11)	64 (10)	2786 (11)	23 (10)
Most deprived 10%	4496 (9)	69 (5)	3316 (9)	28 (5)	2141 (8)	7 (3)
ASA grade, *n* (%)						
P3: incapacitating systemic disease	50,171 (98)	1432 (99)	37,637 (98)	608 (100)	25,508 (98)	242 (100)
P4: life-threatening disease	947 (2)	17 (1)	684 (2)	2 (0)	436 (2)	0 (0)
Charlson Comorbidity Index score, *n* (%)						
0	20,126 (39)	538 (37)	14,972 (39)	235 (39)	10,082 (39)	100 (41)
1	16,304 (32)	480 (33)	12,207 (32)	197 (32)	8195 (32)	69 (29)
2	7656 (15)	237 (16)	5799 (15)	94 (15)	3953 (15)	34 (14)
3	3960 (8)	106 (7)	3030 (8)	46 (8)	2123 (8)	22 (9)
4	3072 (6)	88 (6)	2313 (6)	38 (6)	1591 (6)	17 (7)
Age (years), mean (SD)	73.5 (8.9)	69.6 (9.9)	73.4 (9.0)	69.9 (10.3)	73.4 (9.0)	69.6 (10.6)
BMI (kg/m^2), mean (SD)	32.6 (6.5)	32.6 (6.1)	32.6 (6.5)	32.3 (6.0)	32.5 (6.4)	32.6 (6.0)
PROMs						
Preoperative OKS, mean (SD)	16.4 (7.6)	19.2 (7.8)	16.4 (7.6)	19.0 (7.9)	16.5 (7.7)	18.8 (8.2)
EQ-5D, mean (SD)	61.8 (20.5)	64.2 (20.4)	61.8 (20.5)	64.5 (20.0)	61.7 (20.5)	64.0 (19.9)
General health, *n* (%)						
0	36,509 (71)	909 (63)	27,727 (72)	367 (60)	19,045 (73)	142 (59)
1–3	8347 (16)	272 (19)	6087 (16)	133 (22)	3992 (15)	59 (24)
4–5	6262 (12)	268 (18)	4507 (12)	110 (18)	2907 (11)	41 (17)
Medical history, *n* (%)						
Gastrointestinal disease	14,360 (28)	360 (25)	10,786 (28)	159 (26)	7371 (28)	70 (29)
Osteoarthritis and other joint problems	13,378 (26)	269 (19)	10,042 (26)	112 (18)	6891 (27)	42 (17)
Mental health	6705 (13)	205 (14)	5151 (13)	99 (16)	3551 (14)	42 (17)
Respiratory diseases	13,379 (26)	383 (26)	10,016 (26)	146 (24)	6730 (26)	52 (21)
Cardiovascular diseases	41,694 (82)	1142 (79)	31,223 (81)	476 (78)	21,048 (81)	190 (79)
Thyroid problems	5597 (11)	140 (10)	4179 (11)	61 (10)	2853 (11)	19 (8)
Foot, hip or spinal pain	1944 (4)	46 (3)	1428 (4)	15 (2)	953 (4)	7 (3)
Coxarthrosis	2089 (4)	42 (3)	1545 (4)	20 (3)	1043 (4)	7 (3)
Neurological disorders	6629 (13)	219 (15)	5059 (13)	86 (14)	3500 (13)	40 (17)
Other arthrosis	4377 (9)	82 (6)	3271 (9)	37 (6)	2295 (9)	13 (5)
Polyarthrosis	3907 (8)	61 (4)	2965 (8)	23 (4)	2014 (8)	6 (2)
Spondylosis	2231 (4)	44 (3)	1706 (4)	22 (4)	1168 (5)	7 (3)

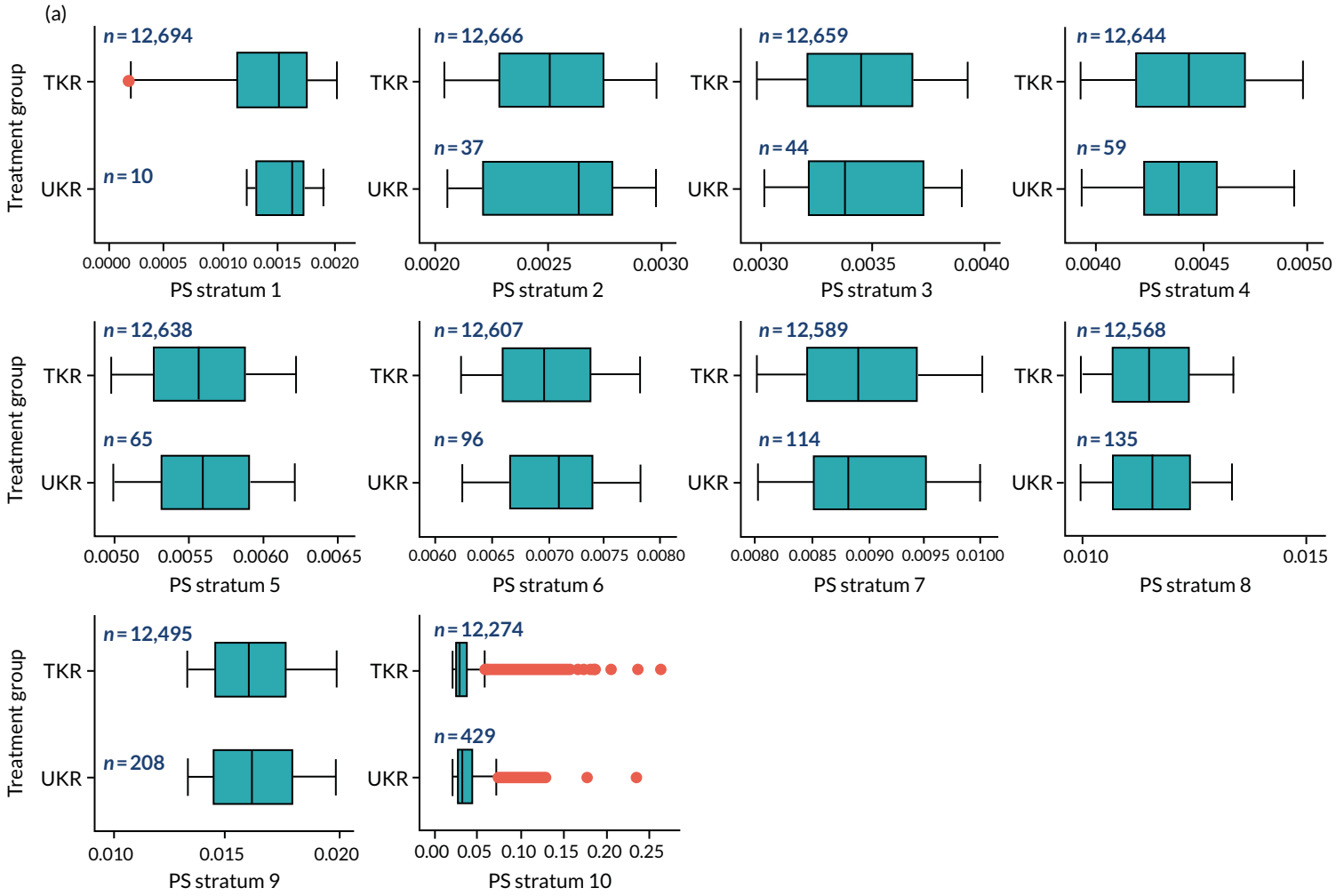


FIGURE 20 Box plot of the PS distribution for TKR and UKR in each stratum of the OKS cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method. (*continued*)

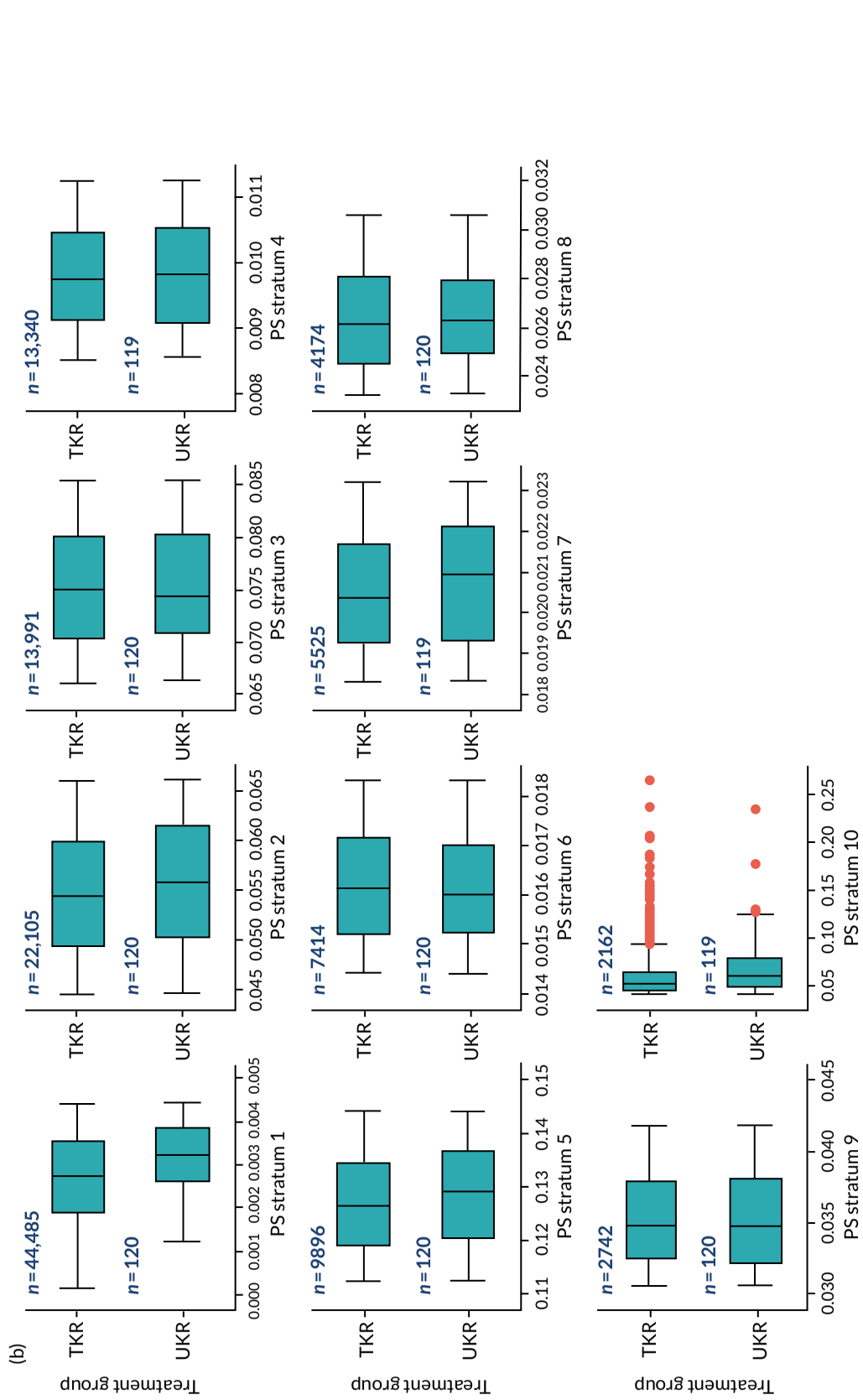


FIGURE 20 Box plot of the PS distribution for TKR and UKR in each stratum of the OKS cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method.

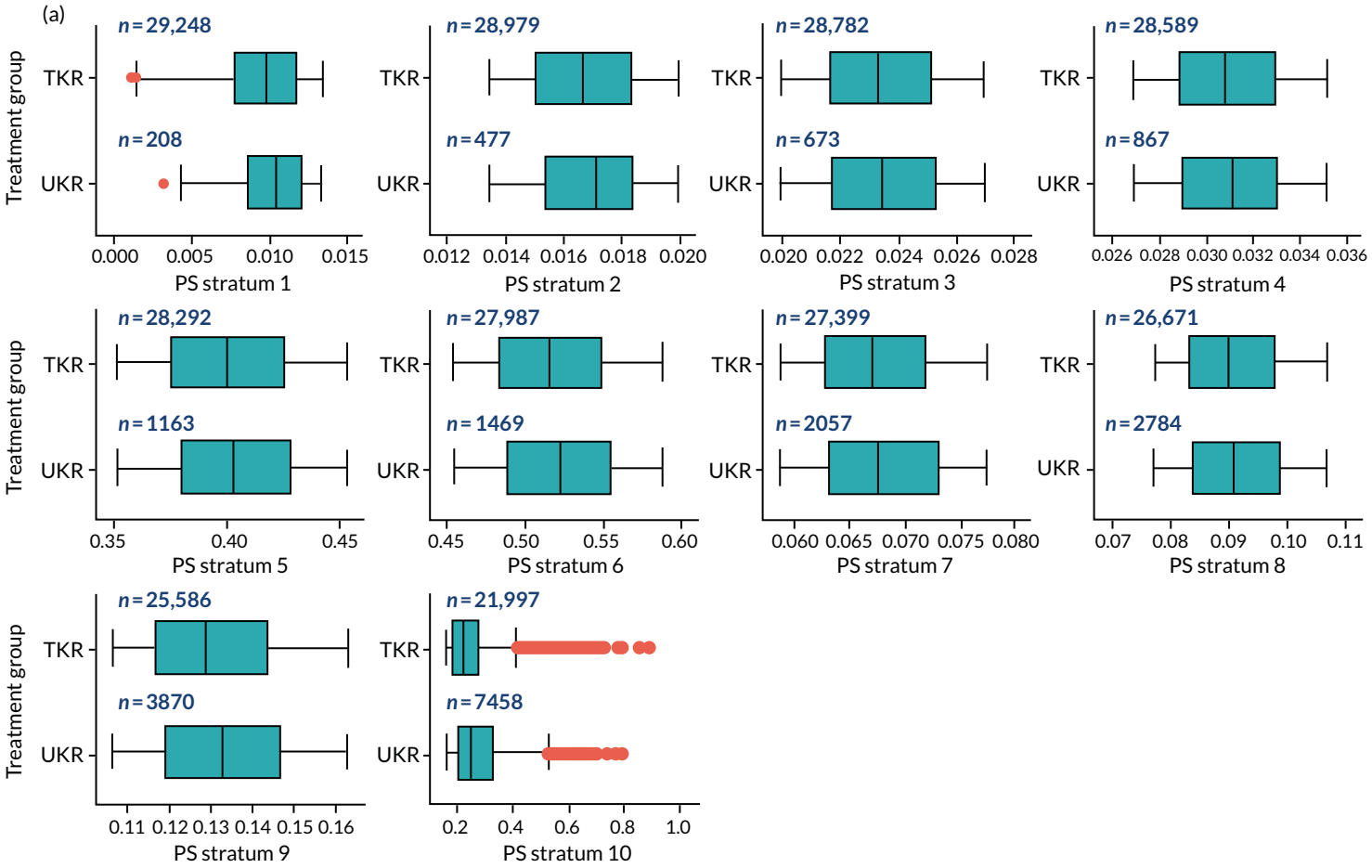


FIGURE 21 Box plot of the PS distribution for TKR and UKR in each stratum of the revision cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method. (*continued*)

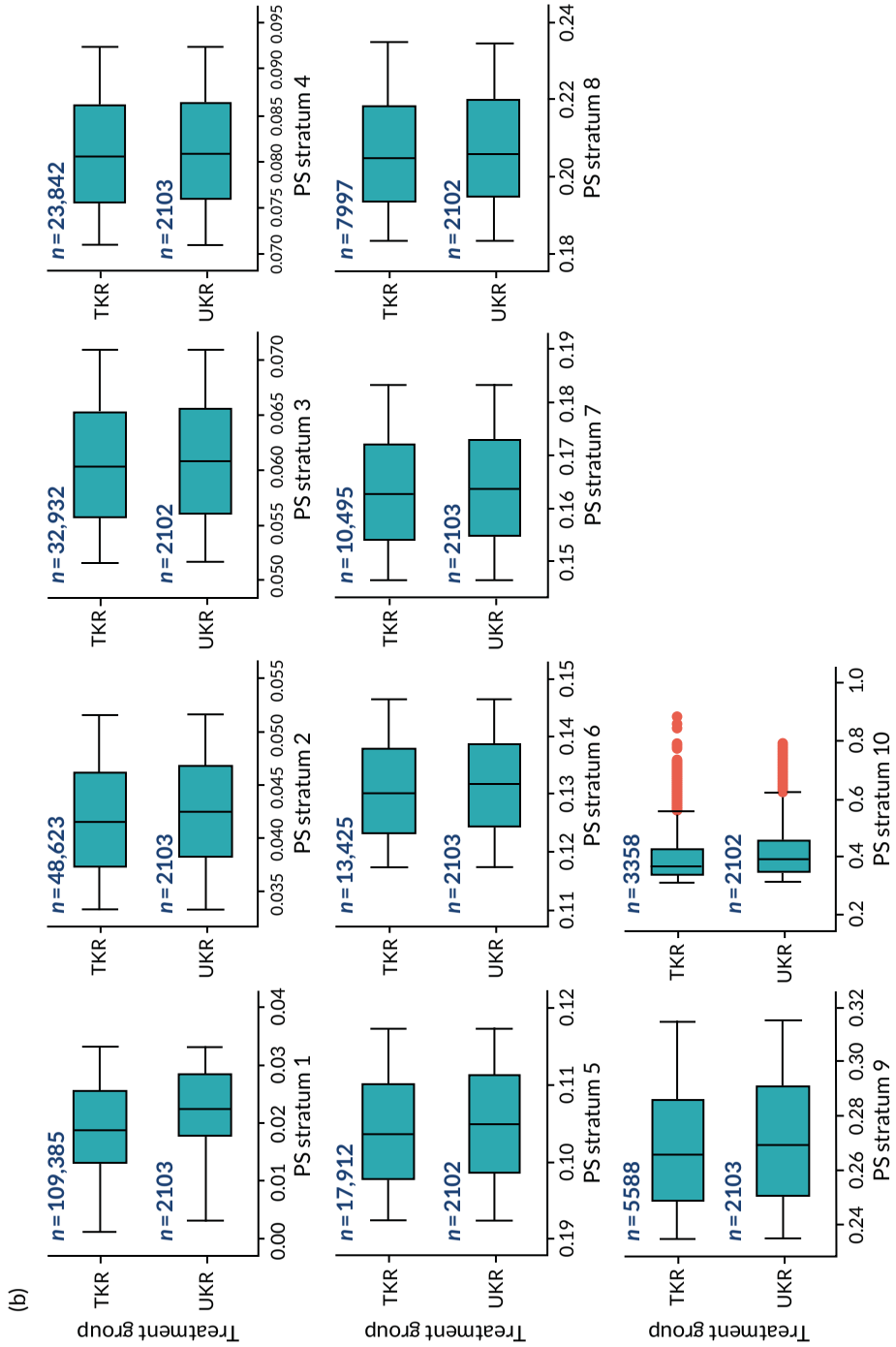


FIGURE 21 Box plot of the PS distribution for TKR and UKR in each stratum of the revision cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method.

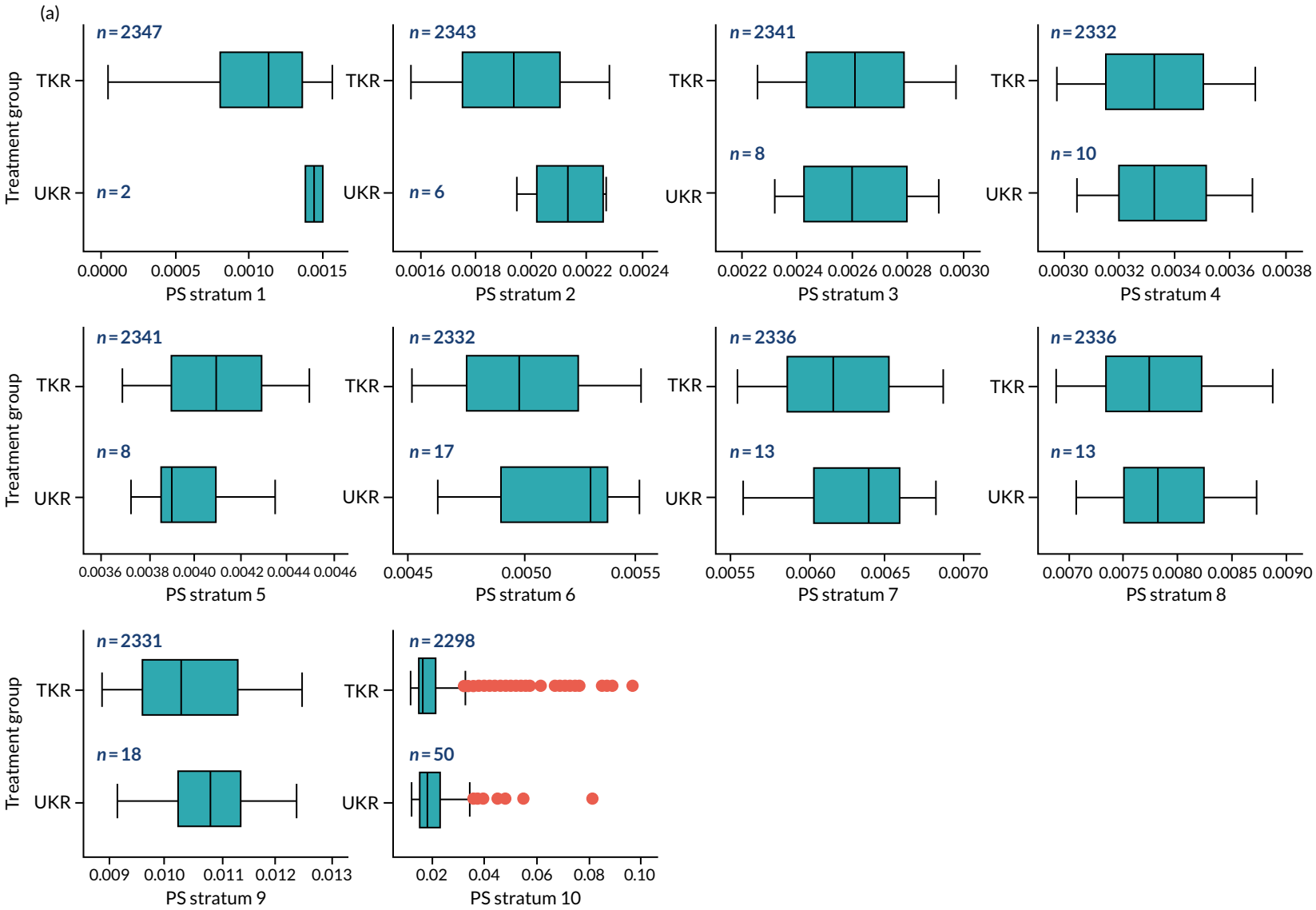


FIGURE 22 Box plot of the PS distribution for TKR and UKR in each stratum of the stage 2 OKS cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method. (*continued*)

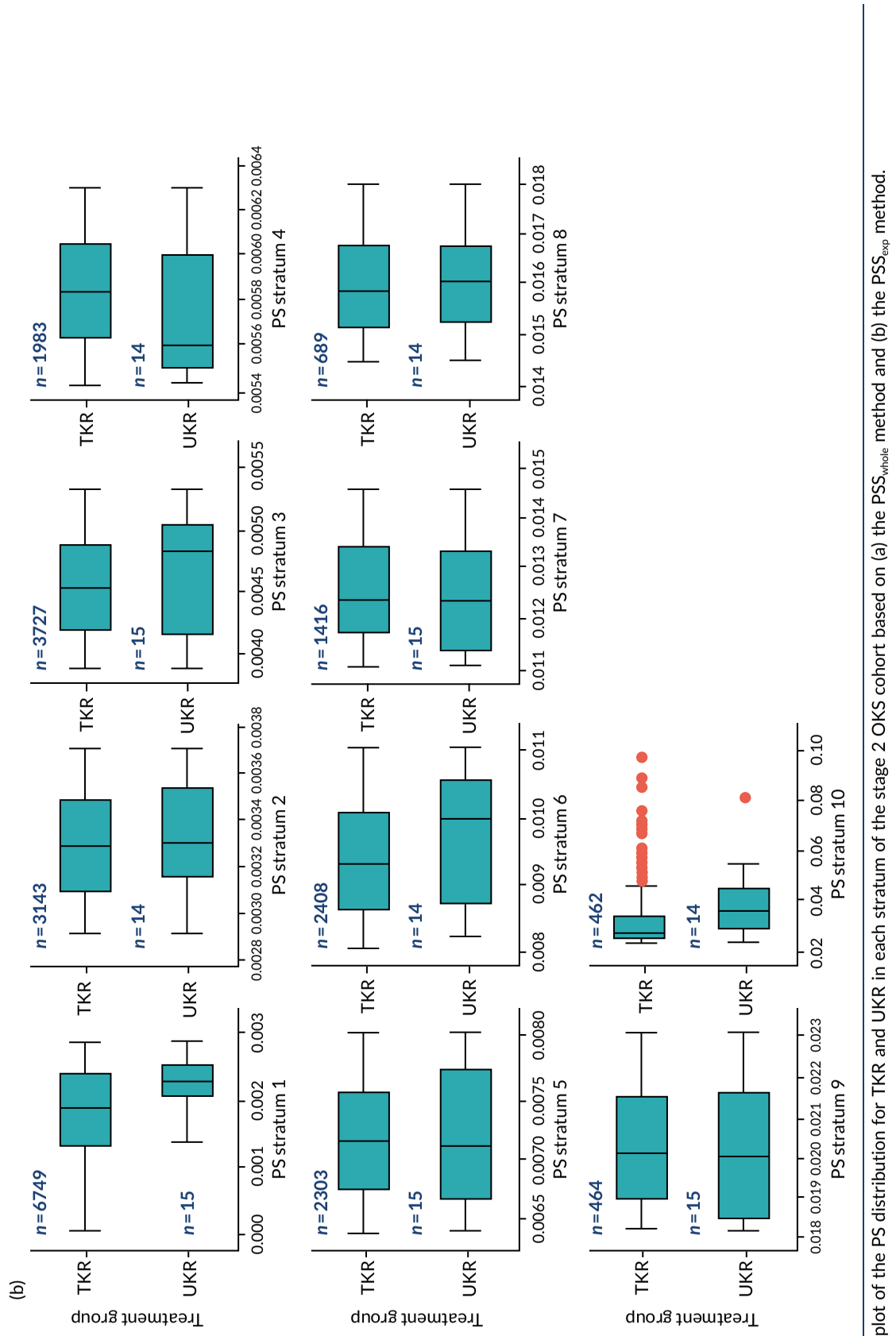


FIGURE 22 Box plot of the PS distribution for TKR and UKR in each stratum of the stage 2 OKS cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method.

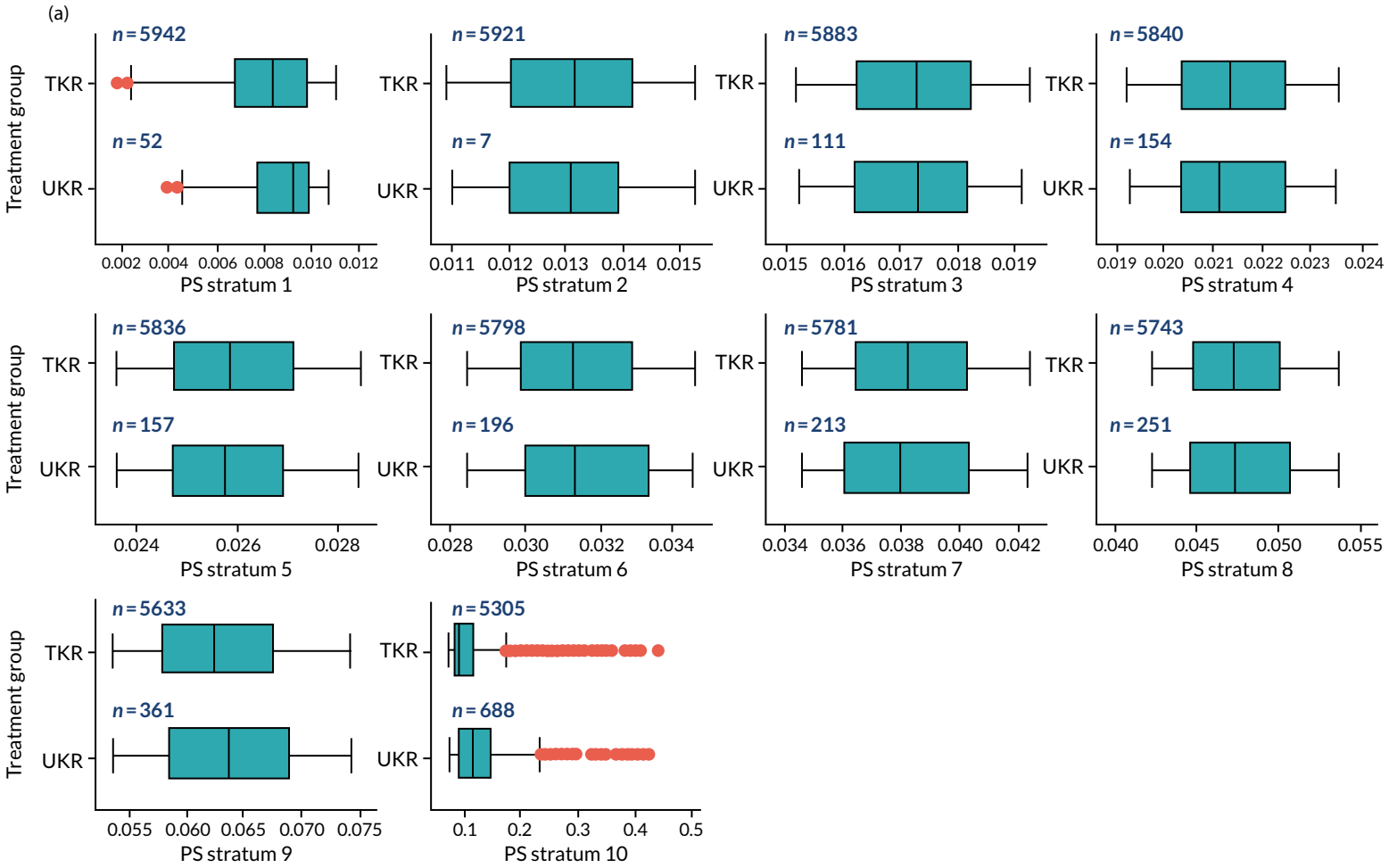


FIGURE 23 Box plot of the PS distribution for TKR and UKR in each stratum of the stage 2 safety cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method. (*continued*)

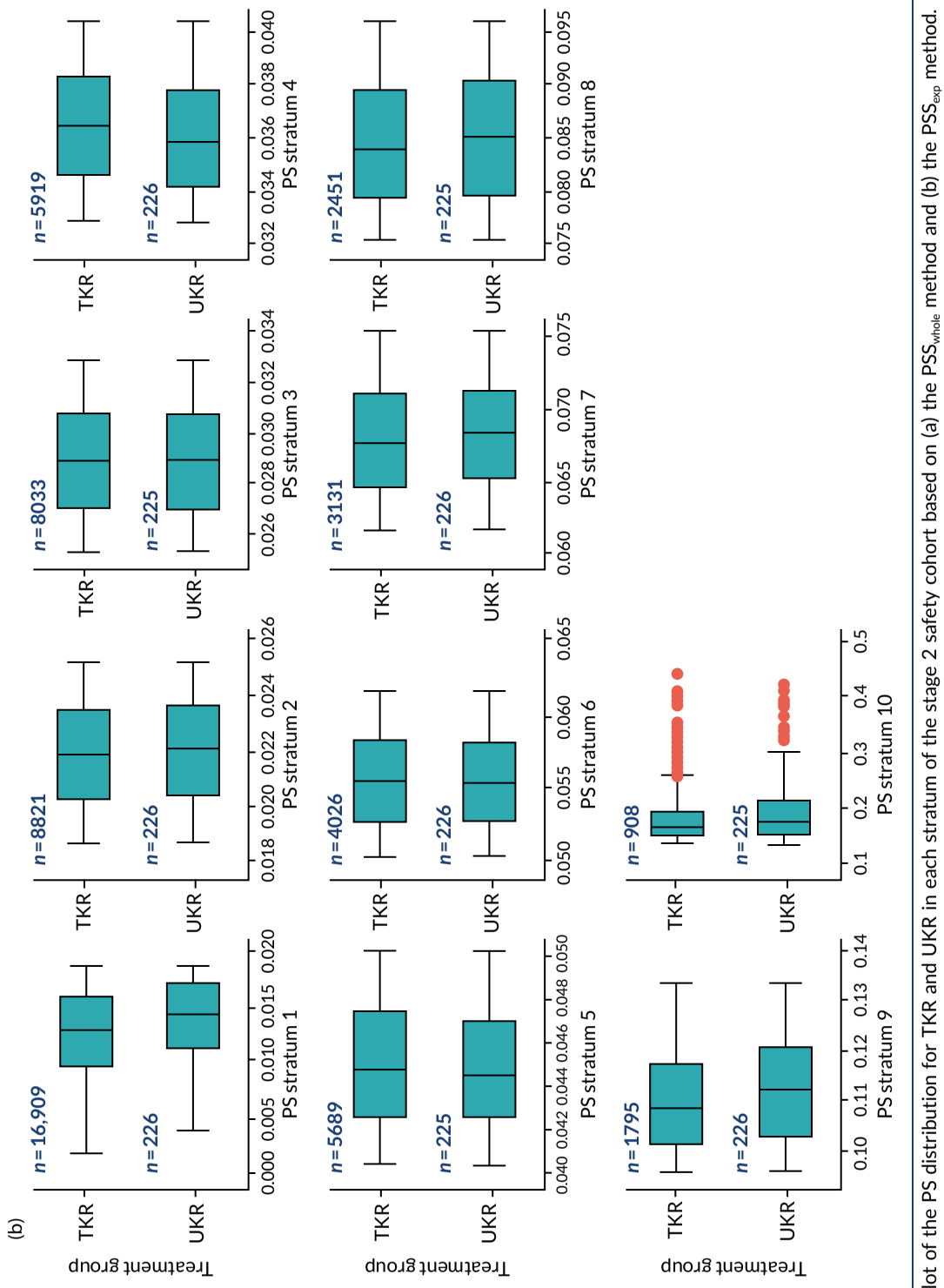


FIGURE 23 Box plot of the PS distribution for TKR and UKR in each stratum of the stage 2 safety cohort based on (a) the PSS_{whole} method and (b) the PSS_{exp} method.

Appendix 2 Code lists

Eligibility criteria code lists

TABLE 31 Cruciate ligament injury or knee injury ICD-10 codes

ICD-10 code	Description
S835	Sprain of cruciate ligament of knee
M232	Derangement of meniscus due to old tear or injury
M235	Chronic instability of knee
M236	Other spontaneous disruption of ligament(s) of knee
M238	Other internal derangements of knee
M239	Unspecified internal derangement of knee
S832	Tear of meniscus, current injury
S833	Tear of articular cartilage of knee, current
S837	Injury to multiple structures of knee
M22	Disorder of patella

TABLE 32 Rheumatoid arthritis or other inflammatory disorder ICD-10 codes

ICD-10 code	Description
M07	Psoriatic and enteropathic arthropathies
M08	Juvenile arthritis
M09	Juvenile arthritis in diseases classified elsewhere
M13	Other arthritis
M05	Seropositive rheumatoid arthritis
M06	Other rheumatoid arthritis
M02	Reactive arthropathies
M03	Post-infective and reactive arthropathies in diseases classified elsewhere

TABLE 33 Foot, hip and spinal pain ICD-10 codes

ICD-10 code	Description
M2555	Pain in hip
M2557	Pain in ankle and joints of foot
M5410	A diagnosis of radiculopathy, site unspecified
M5413	Radiculopathy, cervicothoracic region
M5414	Radiculopathy, thoracic region
M5415	Radiculopathy, thoracolumbar region
M5416	Radiculopathy, lumbar region
M5417	Radiculopathy, lumbosacral region
M5418	Radiculopathy, sacral and sacrococcygeal region
M5419	Radiculopathy, unspecified site
M543	Sciatica
M544	Lumbago with sciatica
M545	Lower back pain
M546	Pain in thoracic spine
M5480	Other dorsalgia, multiple sites in spine
M5483	Other dorsalgia, cervicothoracic region
M5484	Other dorsalgia, thoracic region
M5485	Other dorsalgia, thoracolumbar region
M5486	Other dorsalgia, lumbar region
M5487	Other dorsalgia, lumbosacral region
M5488	Other dorsalgia, sacral and sacrococcygeal region
M5489	Other dorsalgia, site unspecified
M5490	Other dorsalgia, multiple sites in spine
M5493	Unspecified dorsalgia, cervicothoracic region
M5494	Unspecified dorsalgia, thoracic region
M5495	Unspecified dorsalgia, thoracolumbar region
M5496	Unspecified dorsalgia, lumbar region
M5497	Unspecified dorsalgia, lumbosacral region
M5498	Unspecified dorsalgia, sacral and sacrococcygeal region
M5499	Unspecified dorsalgia, site unspecified
M0007	Staphylococcal arthritis, right ankle and foot
M0017	Pneumococcal arthritis, ankle and foot
M0027	Other streptococcal arthritis, ankle and foot
M0087	Arthritis due to other bacteria, ankle and foot
M0097	Pyogenic arthritis, unspecified, ankle and foot
M0107	Meningococcal arthritis, ankle and foot
M0117	Tuberculous arthritis, ankle and foot

TABLE 33 Foot, hip and spinal pain ICD-10 codes (*continued*)

ICD-10 code	Description
M0127	Arthritis in Lyme's disease, ankle and foot
M0137	Arthritis in other bacterial diseases classified elsewhere, ankle and foot
M0147	Rubella arthritis, ankle and foot
M0157	Arthritis in other viral diseases classified elsewhere, ankle and foot
M0167	Arthritis in mycoses, ankle and foot
M0187	Arthritis in other infectious and parasitic diseases classified elsewhere, ankle and foot
M7965	Limb pain, pelvic region and thigh
M7967	Limb pain, ankle and foot
M7905	Rheumatism, unspecified, pelvic region and thigh
M7907	Rheumatism, unspecified, ankle and foot
M7915	Myalgia, pelvic region and thigh
M7917	Myalgia, ankle and foot
M7925	Neuralgia and neuritis, unspecified, pelvic region and thigh
M7927	Neuralgia and neuritis, unspecified, ankle and foot
M7975	Fibromyalgia, pelvic region and thigh
M7977	Fibromyalgia, ankle and foot

TABLE 34 Foot, hip and spinal pain OPCS-4 codes

OPCS-4 code	Description
U503	Delivery of rehabilitation for joint replacement
W05	Prosthetic replacement of bone
W09	Extirpation of lesion of bone
W10	Open surgical fracture of bone
W11	Other surgical fracture of bone
W12	Angulation periarticular division of bone
W13	Other periarticular division of bone
W14	Diaphyseal division of bone
W16	Other division of bone
W17	Other reconstruction of bone
W18	Drainage of bone
W19	Primary open reduction of fracture of bone and intramedullary fixation
W20	Primary open reduction of fracture of bone and extramedullary fixation

continued

TABLE 34 Foot, hip and spinal pain OPCS-4 codes (*continued*)

OPCS-4 code	Description
W21	Primary open reduction of intra-articular fracture of bone
W22	Other primary open reduction of fracture of bone
W23	Secondary open reduction of fracture of bone
W24	Closed reduction of fracture of bone and internal fixation
W25	Closed reduction of fracture of bone and external fixation
W26	Other closed reduction of fracture of bone
W27	Fixation of epiphysis
W28	Other internal fixation of bone
W29	Skeletal traction of bone
W30	Other external fixation of bone
W31	Other autograft of bone
W32	Other graft of bone
W33	Other open operations on bone
W43	Total prosthetic replacement of other joint using cement
W44	Total prosthetic replacement of other joint not using cement
W45	Other total prosthetic replacement of other joint
W52	Prosthetic replacement of articulation of other bone using cement
W53	Prosthetic replacement of articulation of other bone not using cement
W54	Other prosthetic replacement of articulation of other bone
W55	Prosthetic interposition reconstruction of joint
W56	Other interposition reconstruction of joint
W57	Excision reconstruction of joint
W58	Other reconstruction of joint
W60	Fusion of other joint and extra-articular bone graft
W61	Fusion of other joint and other articular bone graft
W62	Other primary fusion of other joint
W63	Revisional fusion of other joint
W64	Conversion to fusion of other joint
W65	Primary open reduction of traumatic dislocation of joint
W66	Primary closed reduction of traumatic dislocation of joint
W67	Secondary reduction of traumatic dislocation of joint

TABLE 34 Foot, hip and spinal pain OPCS-4 codes (*continued*)

OPCS-4 code	Description
W86	Therapeutic endoscopic operations on cavity of other joint
W89	Other therapeutic endoscopic operations on other articular cartilage
W91	Other manipulation of joint
O09	Placement of bone prosthesis
O17	Secondary closed reduction of fracture of bone and internal fixation
O19	Other therapeutic endoscopic operations on other joint structure
O27	Other stabilising operations on joint
O29	Excision of bone
X05	Implantation of prosthesis for limb

TABLE 35 Knee surgery OPCS-4 codes

OPCS-4 code	Description
W69	Open operations on synovial membrane of joint
W71	Other open operations on intra-articular structure
W72	Prosthetic replacement of ligament
W73	Prosthetic reinforcement of ligament
W74	Other reconstruction of ligament
W75	Other open repair of ligament
W76	Other operations on ligament
W77	Stabilising operations on joint
W78	Release of contracture of joint
W80	Debridement of joint
W811	Excision of lesion of joint
W812	Removal of loose body from joint
W813	Drainage of joint
W814	Incision of joint
W816	Capsulorrhaphy of joint
W817	Insertion of therapeutic spacer into joint
W83	Endoscopic operations on articular cartilage
W84	Endoscopic operations on other joint structure
O18	Hybrid prosthetic replacement of knee joint using cement
O27	Other stabilising operations on joint
O29	Excision of bone

TABLE 36 Septic arthritis ICD-10 codes

ICD-10 code	Description
M0005	Staphylococcal arthritis and polyarthritis, pelvic region and thigh
M0006	Staphylococcal arthritis and polyarthritis, lower leg
M0015	Pneumococcal arthritis and polyarthritis, pelvic region and thigh
M0016	Pneumococcal arthritis and polyarthritis, lower leg
M0025	Other streptococcal arthritis and polyarthritis, pelvic region and thigh
M0026	Other streptococcal arthritis and polyarthritis, lower leg
M0085	Arthritis and polyarthritis due to other specified bacterial agents, pelvic region and thigh
M0086	Arthritis and polyarthritis due to other specified bacterial agents, lower leg
M0095	Pyogenic arthritis, unspecified, pelvic region and thigh
M0096	Pyogenic arthritis, unspecified, lower leg
M0105	Meningococcal arthritis, pelvic region and thigh
M0106	Meningococcal arthritis, lower leg
M0115	Tuberculous arthritis, pelvic region and thigh
M0116	Tuberculous arthritis, lower leg
M0125	Arthritis in Lyme's disease, pelvic region and thigh
M0126	Arthritis in Lyme's disease, lower leg
M0135	Arthritis in other bacterial diseases classified elsewhere, pelvic region and thigh
M0136	Arthritis in other bacterial diseases classified elsewhere, lower leg
M0145	Rubella arthritis, pelvic region and thigh
M0146	Rubella arthritis, lower leg
M0155	Arthritis in other viral diseases classified elsewhere, pelvic region and thigh
M0156	Arthritis in other viral diseases classified elsewhere, lower leg
M0165	Arthritis in mycoses, pelvic region and thigh
M0166	Arthritis in mycoses, lower leg
M0185	Arthritis in other infectious and parasitic diseases classified elsewhere, pelvic region and thigh
M0186	Arthritis in other infectious and parasitic diseases classified elsewhere, lower leg
M000	Staphylococcal arthritis and polyarthritis, multiple sites
M001	Staphylococcal arthritis and polyarthritis, shoulder region
M002	Staphylococcal arthritis and polyarthritis, upper arm
M008	Staphylococcal arthritis and polyarthritis, other site
M009	Staphylococcal arthritis and polyarthritis, unspecified site
M010	Pneumococcal arthritis and polyarthritis, multiple sites
M011	Pneumococcal arthritis and polyarthritis, shoulder region
M012	Pneumococcal arthritis and polyarthritis, upper arm
M013	Pneumococcal arthritis and polyarthritis, forearm
M014	Pneumococcal arthritis and polyarthritis, hand
M015	Pneumococcal arthritis and polyarthritis, pelvic region and thigh
M016	Pneumococcal arthritis and polyarthritis, lower leg
M018	Pneumococcal arthritis and polyarthritis, other site

TABLE 37 Patellofemoral damage or varus deformity ICD-10 codes

ICD-10 code	Description
M22	Disorder of patella
M2116	Varus deformity, not elsewhere classified, knee

Covariates included in the propensity score

TABLE 38 Charlson Comorbidity Index: AIDS ICD-10 codes

ICD-10 code	Description
B20	HIV disease
B21	HIV disease resulting in Kaposi's sarcoma
B22	HIV disease resulting in other specified diseases
B24	Unspecified HIV disease

HIV, human immunodeficiency virus.

TABLE 39 Charlson Comorbidity Index: metastatic ICD-10 codes

ICD-10 code	Description
C77	Secondary and unspecified malignant neoplasm of lymph nodes
C78	Secondary malignant neoplasm of respiratory and digestive organs
C79	Secondary malignant neoplasm of other and unspecified sites
C80	Malignant neoplasm without specification of site

TABLE 40 Charlson Comorbidity Index: moderate to severe liver diseases ICD-10 codes

ICD-10 code	Description
K704	Alcoholic hepatic failure
K711	Toxic liver disease with hepatic necrosis
K721	Chronic hepatic failure
K729	Hepatic failure, unspecified
K765	Hepatic veno-occlusive disease
K766	Portal hypertension
K767	Hepatorenal syndrome
I850	Oesophageal varices
I859	Oesophageal varices without bleeding
I864	Gastric varices
I982	Oesophageal varices with bleeding in diseases classified elsewhere

TABLE 41 Charlson Comorbidity Index: cancer ICD-10 codes

ICD-10 code	Description
C00	Malignant neoplasm of lip
C01	Malignant neoplasm of base of tongue
C02	Malignant neoplasm of other and unspecified parts of tongue
C03	Malignant neoplasm of gum
C04	Malignant neoplasm of floor of mouth
C05	Malignant neoplasm of palate
C06	Malignant neoplasm of other and unspecified parts of mouth
C07	Malignant neoplasm of parotid gland
C08	Malignant neoplasm of other and unspecified major salivary glands
C09	Malignant neoplasm of tonsil
C10	Malignant neoplasm of oropharynx
C11	Malignant neoplasm of nasopharynx
C12	Malignant neoplasm of pyriform sinus
C13	Malignant neoplasm of hypopharynx
C14	Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx
C15	Malignant neoplasm of oesophagus
C16	Malignant neoplasm of stomach
C17	Malignant neoplasm of small intestine
C18	Malignant neoplasm of colon
C19	Malignant neoplasm of rectosigmoid junction
C20	Malignant neoplasm of rectum
C21	Malignant neoplasm of anus and anal canal
C22	Malignant neoplasm of liver and intrahepatic bile ducts
C23	Malignant neoplasm of gallbladder
C24	Malignant neoplasm of other and unspecified parts of biliary tract
C25	Malignant neoplasm of pancreas
C26	Malignant neoplasm of other and ill-defined digestive organs
C30	Malignant neoplasm of nasal cavity and middle ear
C31	Malignant neoplasm of accessory sinuses
C32	Malignant neoplasm of larynx
C33	Malignant neoplasm of trachea
C34	Malignant neoplasm of bronchus and lung
C37	Malignant neoplasm of thymus
C38	Malignant neoplasm of heart, mediastinum and pleura
C39	Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathoracic organs
C40	Malignant neoplasm of bone and articular cartilage of limbs
C41	Malignant neoplasm of bone and articular cartilage of other and unspecified sites
C43	Malignant melanoma of skin

TABLE 41 Charlson Comorbidity Index: cancer ICD-10 codes (*continued*)

ICD-10 code	Description
C45	Mesothelioma
C46	Kaposi's sarcoma
C47	Malignant neoplasm of peripheral nerves and autonomic nervous system
C48	Malignant neoplasm of retroperitoneum and peritoneum
C49	Malignant neoplasm of other connective and soft tissue
C50	Malignant neoplasm of breast
C51	Malignant neoplasm of vulva
C52	Malignant neoplasm of vagina
C53	Malignant neoplasm of cervix uteri
C54	Malignant neoplasm of corpus uteri
C55	Malignant neoplasm of uterus, part unspecified
C56	Malignant neoplasm of ovary
C57	Malignant neoplasm of other and unspecified female genital organs
C58	Malignant neoplasm of placenta
C60	Malignant neoplasm of penis
C61	Malignant neoplasm of prostate
C62	Malignant neoplasm of testis
C63	Malignant neoplasm of other and unspecified male genital organs
C64	Malignant neoplasm of kidney, except renal pelvis
C65	Malignant neoplasm of renal pelvis
C66	Malignant neoplasm of ureter
C67	Malignant neoplasm of bladder
C68	Malignant neoplasm of other and unspecified urinary organs
C69	Malignant neoplasm of eye and adnexa
C70	Malignant neoplasm of meninges
C71	Malignant neoplasm of brain
C72	Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system
C73	Malignant neoplasm of thyroid gland
C74	Malignant neoplasm of adrenal gland
C75	Malignant neoplasm of other endocrine glands and related structures
C76	Malignant neoplasm of other and ill-defined sites
C81	Hodgkin's lymphoma
C82	Follicular lymphoma
C83	Non-follicular lymphoma
C84	Mature T/NK-cell lymphomas
C85	Other specified and unspecified types of non-Hodgkin's lymphoma
C88	Malignant immunoproliferative diseases and certain other B-cell lymphomas

continued

TABLE 41 Charlson Comorbidity Index: cancer ICD-10 codes (*continued*)

ICD-10 code	Description
C90	Multiple myeloma and malignant plasma cell neoplasms
C91	Lymphoid leukaemia
C92	Myeloid leukaemia
C93	Monocytic leukaemia
C94	Other leukaemia's of specified cell type
C95	Leukaemia of unspecified cell type
C96	Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue
C97	Malignant neoplasms of independent (primary) multiple sites

NK, natural killer.

TABLE 42 Charlson Comorbidity Index: renal diseases ICD-10 codes

ICD-10 code	Description
N18	Chronic kidney disease
N19	Unspecified kidney failure
N052	Unspecified nephritic syndrome with diffuse membranous glomerulonephritis
N053	Unspecified nephritic syndrome with diffuse mesangial proliferative glomerulonephritis
N054	Unspecified nephritic syndrome with diffuse endocapillary proliferative glomerulonephritis
N055	Unspecified nephritic syndrome with diffuse mesangiocapillary glomerulonephritis
N056	Unspecified nephritic syndrome with dense deposit disease
N057	Unspecified nephritic syndrome with diffuse crescentic glomerulonephritis
N250	Renal osteodystrophy
I120	Hypertensive chronic kidney disease with stage 5 chronic kidney disease or end-stage renal disease
I131	Hypertensive heart and chronic kidney disease without heart failure
N032	Chronic nephritic syndrome with diffuse membranous glomerulonephritis
N033	Chronic nephritic syndrome with diffuse mesangial proliferative glomerulonephritis
N034	Chronic nephritic syndrome with diffuse endocapillary proliferative glomerulonephritis
N035	Chronic nephritic syndrome with diffuse mesangiocapillary glomerulonephritis
N036	Chronic nephritic syndrome with dense deposit disease
N037	Chronic nephritic syndrome with diffuse crescentic glomerulonephritis
Z490	Preparatory care for renal dialysis
Z491	Extracorporeal dialysis
Z492	Other dialysis
Z940	Kidney transplant status
Z992	Dependence on renal dialysis

TABLE 43 Charlson Comorbidity Index: paraplegia ICD-10 codes

ICD-10 code	Description
G81	Hemiplegia and hemiparesis
G82	Paraplegia (paraparesis) and quadriplegia (quadriparesis)
G041	Tropical spastic paraplegia
G114	Hereditary spastic paraplegia
G801	Spastic diplegic cerebral palsy
G802	Spastic hemiplegic cerebral palsy
G830	Diplegia of upper limbs
G831	Monoplegia of lower limb
G832	Monoplegia of upper limb
G833	Monoplegia, unspecified
G834	Cauda equina syndrome
G839	Paralytic syndrome, unspecified

TABLE 44 Charlson Comorbidity Index: diabetes complications ICD-10 codes

ICD-10 code	Description
E102	Type 1 diabetes mellitus with kidney complications
E103	Type 1 diabetes mellitus with ophthalmic complications
E104	Type 1 diabetes mellitus with neurological complications
E105	Type 1 diabetes mellitus with circulatory complications
E107	Type 1 diabetes mellitus with multiple complications
E112	Type 2 diabetes mellitus with kidney complications
E113	Type 2 diabetes mellitus with ophthalmic complications
E114	Type 2 diabetes mellitus with neurological complications
E115	Type 2 diabetes mellitus with circulatory complications
E117	Malnutrition-related diabetes mellitus with multiple complications
E122	Malnutrition-related diabetes mellitus with kidney complications
E123	Malnutrition-related diabetes mellitus with ophthalmic complications
E124	Malnutrition-related diabetes mellitus with neurological complications
E125	Malnutrition-related diabetes mellitus with circulatory complications
E127	Malnutrition-related diabetes mellitus with multiple complications

continued

TABLE 44 Charlson Comorbidity Index: diabetes complications ICD-10 codes (*continued*)

ICD-10 code	Description
E132	Other specified diabetes mellitus with kidney complications
E133	Other specified diabetes mellitus with ophthalmic complications
E134	Other specified diabetes mellitus with neurological complications
E135	Diabetes mellitus with circulatory complications
E137	Other specified diabetes mellitus with multiple complications
E142	Unspecified diabetes mellitus with kidney complications
E143	Unspecified diabetes mellitus with ophthalmic complications
E144	Unspecified diabetes mellitus with neurological complications
E145	Unspecified diabetes mellitus with circulatory complications
E147	Unspecified diabetes mellitus with multiple complications

TABLE 45 Charlson Comorbidity Index: diabetes without complications ICD-10 codes

ICD-10 code	Description
E100	Type 1 diabetes mellitus with coma
E101	Type 1 diabetes mellitus with ketoacidosis
E106	Type 1 diabetes mellitus with other specified complications
E108	Type 1 diabetes mellitus with unspecified complications
E109	Type 1 diabetes mellitus without complications
E110	Type 2 diabetes mellitus with coma
E111	Type 2 diabetes mellitus with ketoacidosis
E116	Type 2 diabetes mellitus with other specified complications
E118	Type 2 diabetes mellitus with unspecified complications
E119	Type 2 diabetes mellitus without complications
E120	Malnutrition-related diabetes mellitus with coma
E121	Malnutrition-related diabetes mellitus with ketoacidosis
E126	Malnutrition-related diabetes mellitus with other specified complications
E128	Malnutrition-related diabetes mellitus with unspecified complications
E129	Malnutrition-related diabetes mellitus without complications
E130	Other specified diabetes mellitus with coma
E131	Other specified diabetes mellitus with ketoacidosis

TABLE 45 Charlson Comorbidity Index: diabetes without complications ICD-10 codes (*continued*)

ICD-10 code	Description
E136	Other specified diabetes mellitus with other specified complications
E138	Other specified diabetes mellitus with unspecified complications
E139	Other specified diabetes mellitus without complications
E140	Unspecified diabetes mellitus with coma
E141	Unspecified diabetes mellitus with ketoacidosis
E146	Unspecified diabetes mellitus with other specified complications
E148	Unspecified diabetes mellitus with unspecified complications
E149	Unspecified diabetes mellitus without complications

TABLE 46 Charlson Comorbidity Index: liver disease ICD-10 codes

ICD-10 code	Description
B18	Chronic viral hepatitis
K73	Chronic hepatitis, not elsewhere classified
K74	Fibrosis and cirrhosis of liver
K700	Alcoholic fatty liver
K701	Alcoholic hepatitis
K702	Alcoholic fibrosis and sclerosis of liver
K703	Alcoholic cirrhosis of liver
K709	Alcoholic liver disease, unspecified
K717	Toxic liver disease with fibrosis and cirrhosis of liver
K713	Toxic liver disease with chronic persistent hepatitis
K714	Toxic liver disease with chronic lobular hepatitis
K715	Toxic liver disease with chronic active hepatitis
K760	Fatty (change of) liver, not elsewhere classified
K762	Central haemorrhagic necrosis of liver
K763	Infarction of liver
K764	Peliosis hepatis
K768	Other specified diseases of liver
K769	Liver disease, unspecified
Z944	Liver transplant status

TABLE 47 Charlson Comorbidity Index: peptic ulcer ICD-10 codes

ICD-10 code	Description
K25	Gastric ulcer
K26	Duodenal ulcer
K27	Peptic ulcer, site unspecified
K28	Gastrojejunal ulcer

TABLE 48 Charlson Comorbidity Index: connective tissue disorder ICD-10 codes

ICD-10 code	Description
M05	Rheumatoid arthritis with rheumatoid factor
M32	Systemic lupus erythematosus
M33	Dermatopolymyositis
M34	Systemic sclerosis (scleroderma)
M06	Other rheumatoid arthritis
M315	Giant cell arteritis with polymyalgia rheumatica
M351	Other overlap syndromes
M353	Polymyalgia rheumatica
M360	Dermato(poly)myositis in neoplastic disease

TABLE 49 Charlson Comorbidity Index: pulmonary disease ICD-10 codes

ICD-10 code	Description
J40	Bronchitis, not specified as acute or chronic
J41	Simple and mucopurulent chronic bronchitis
J42	Unspecified chronic bronchitis
J43	Emphysema
J44	Other chronic obstructive pulmonary disease
J45	Asthma
J46	Status asthmaticus
J47	Bronchiectasis
J60	Coalworker's pneumoconiosis
J61	Pneumoconiosis due to asbestos and other mineral fibres
J62	Pneumoconiosis due to dust containing silica
J63	Pneumoconiosis due to other inorganic dusts
J64	Unspecified pneumoconiosis
J65	Pneumoconiosis associated with tuberculosis
J66	Airway disease due to specific organic dust
J67	Hypersensitivity pneumonitis due to organic dust
I278	Other specified pulmonary heart diseases
I279	Pulmonary heart disease, unspecified
J684	Chronic respiratory conditions due to chemicals, gases, fumes and vapours
J701	Chronic and other pulmonary manifestations due to radiation
J703	Chronic pulmonary disease

TABLE 50 Charlson Comorbidity Index: dementia ICD-10 codes

ICD-10 code	Description
F00	Dementia in Alzheimer's disease
F01	Vascular dementia
F02	Dementia in other diseases classified elsewhere
F03	Unspecified dementia
G30	Alzheimer's disease
F051	Delirium superimposed on dementia
G311	Senile degeneration of brain, not elsewhere classified

TABLE 51 Charlson Comorbidity Index: cerebrovascular disease ICD-10 codes

ICD-10 code	Description
G45	Transient cerebral ischaemic attacks and related syndromes
G46	Vascular syndromes of brain in cerebrovascular diseases
I60	Non-traumatic subarachnoid haemorrhage
I61	Non-traumatic intracerebral haemorrhage
I62	Other and unspecified non-traumatic intracranial haemorrhage
I63	Cerebral infarction
I64	Stroke, not specified as haemorrhage or infarction
I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
I66	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
I67	Other cerebrovascular diseases
I68	Cerebrovascular disorders in diseases classified elsewhere
I69	Sequelae of cerebrovascular disease
H340	Transient retinal artery occlusion

TABLE 52 Charlson Comorbidity Index: peripheral vascular disease ICD-10 codes

ICD-10 code	Description
I70	Atherosclerosis
I71	Aortic aneurysm and dissection
I731	Thromboangiitis obliterans
I738	Other specified peripheral vascular diseases
I739	Peripheral vascular disease, unspecified
I771	Stricture of artery
I790	Aneurysm of aorta in diseases classified elsewhere
I792	Peripheral angiopathy in diseases classified elsewhere
K551	Chronic vascular disorders of intestine
K558	Other vascular disorders of intestine
K559	Vascular disorder of intestine, unspecified
Z958	Presence of other cardiac and vascular implants and grafts
Z959	Presence of cardiac and vascular implant and graft, unspecified

TABLE 53 Charlson Comorbidity Index: congestive heart failure ICD-10 codes

ICD-10 code	Description
I43	Cardiomyopathy in diseases classified elsewhere
I50	Heart failure
I099	Rheumatic heart disease, unspecified
I110	Hypertensive heart disease with heart failure
I130	Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease
I132	Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease or end-stage renal disease
I255	Ischaemic cardiomyopathy
I420	Dilated cardiomyopathy
I425	Other restrictive cardiomyopathy
I426	Alcoholic cardiomyopathy
I427	Cardiomyopathy due to drug and external agent
I428	Other cardiomyopathies
I429	Cardiomyopathy, unspecified
P290	Neonatal cardiac failure

TABLE 54 Charlson Comorbidity Index: acute myocardial infarction ICD-10 codes

ICD-10 code	Description
I21	ST elevation myocardial infarction (STEMI) and non-ST elevation myocardial infarction (NSTEMI)
I22	Subsequent STEMI and NSTEMI
I252	Old myocardial infarction

TABLE 55 Osteoarthritis and other joint problems ICD-10 codes

ICD-10 code	Description
M17	Osteoarthritis of knee
M2580	Other specified joint disorders, unspecified joint
M2581	Other specified joint disorders, shoulder
M2582	Other specified joint disorders, elbow
M2583	Other specified joint disorders, wrist
M2584	Other specified joint disorders, hand
M2585	Other specified joint disorders, hip
M2587	Other specified joint disorders, ankle and foot
M2588	Other specified joint disorders, other site
M2589	Other specified joint disorder site NOS
M2590	Joint disorder NOS multiple sites
M2591	Joint disorder NOS shoulder region
M2592	Unspecified joint disorder, upper arm
M2593	Unspecified joint disorder, forearm
M2594	Unspecified joint disorder, hand
M2595	Unspecified joint disorder, pelvis and thigh
M2597	Unspecified joint disorder, lower leg
M2598	Joint disorder NOS ankle and foot
M2599	Unspecified joint disorder, other site

NOS, not otherwise specified.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library