Methods for Medical Device and Surgical Epidemiology: Applications in Knee Replacement and COVID-19 Related Tracheotomy

Albert Prats Uribe
St Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2022

Methods for Medical Device and Surgical Epidemiology: Applications in knee replacement and COVID-19 related tracheotomy

Candidate name: Albert Prats Uribe (St. Peter's College)

Thesis submitted for the degree of: Doctor of Philosophy

Although medical devices (MDs) and surgery have been part of medicine since its start, the methods for their study in epidemiology have not kept pace with the developments in pharmacoepidemiology. The recent increased attention from society to the potential harms of unsafe implants has led to improved legislation that call for a much closer evaluation of MDs. However, the evaluation of MDs and surgery presents unique caveats and challenges, such as the ascertainment of the indication of the surgery or how to include surgeon characteristics in the analysis, that can greatly influence outcomes.

To advance research on how one can overcome these challenges, I tested several methods. I used Propensity Score (PS) methods, namely PS matching, stratifications and inverse probability weighting (IPW); and Instrumental Variables (IVs), based on surgeon and hospital preference in the study of the effectiveness and safety of partial vs total knee replacement and evaluated them by comparing observational results to those from a randomised controlled trial. I applied the target trial framework to the study of the timing of tracheostomy in patients with COVID-19. I further studied the safety of knee replacements using the self-controlled case series method. I studied the potential heterogeneity by subgroups (on high-risk patients, by gender, by age, and by deprivation) on the safety and

effectiveness of knee replacements. I explored the effect of surgical volume on knee replacement outcomes. I finally used simulation studies to examine the flaws and challenges of preference-based IV.

I found that PS stratification and IPW may be able to minimise confounding and bias in the study of partial vs total knee replacement, provided that the study is carefully designed to consider both patient and surgeon characteristics. I also showed how preference-based IVs may not be fit for purpose for this use case.

Future research is needed to examine whether my findings are generalisable to other settings and whether other IVs can be used safely in MD and surgical research.

Declaration

The work presented in this thesis was undertaken by Albert Prats-Uribe and has not previously been submitted for the award of a degree by any university. All contributions by others have been acknowledged. This work was performed under the supervision of Prof Daniel Prieto-Alhambra, Dr Victoria Strauss and Prof Gary Collins, and completed within the Centre for Statistics in Medicine (CSM) in the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS) at the University of Oxford.

Acknowledgements

First, I want to express how grateful I am for the amazing supervision team that has supported me on the completion of this Thesis:

Daniel Prieto-Alhambra (University of Oxford), supervisor, friend, and mentor, has taught me how kindness and empathy united to curiosity and honesty are the most powerful tools for research;

Victoria Y Strauss (University of Oxford), beacon of methodological knowledge and always there to help, who has taught me to never stop pushing for excellence, without her support this Thesis would have been impossible;

and Gary S Collins (University of Oxford) for his encouragement and expert advice, and for always recommending the most useful literature.

Their combined efforts supervising different aspects of this DPhil have been invaluable to their completion. They have taught me the importance of rigorous, transparent, and relevant work, and have paved my path towards becoming an independent researcher.

In addition, I'm grateful to the UTMOST team, whose help and suggestions have strengthened this Thesis. Specially to Klara Berencsi, for the hard work of managing and cleaning the data, to Spyros Kolovos, for the methodological discussions in the safe room, to Victoria Strauss again, for teaching me all that I know about propensity scores, and to Danielle Robinson for introducing me to, and guiding me through the analysis of self-controlled case series.

To the COVID tracheostomy team, Francesc Xavier Avilés-Jurado, Isabel Vilaseca and all collaborators, for their hard work during very difficult times and for their positivity and encouragement.

To the COVID OHDSI Team, and our daily meetings during the pandemic. They gave me a sense of purpose during lockdown and taught me how important collaborative science is and how being part of a community can shine a light even in the darkest moments.

To my MPH thesis supervision team: Joan-Pau Millet, Àngels Orcau, and Joan A. Caylà who inspired my passion for research and public health. I am still applying all that I learnt from them.

To all the NDORMS staff and students who helped me, specially to Afsie Sabokbar and Sam Burnell for being the best support for graduate students.

To the whole Pharmaco- and Device Epidemiology research team, who have been colleagues, friends and teachers, for their brilliance and helpfulness. Thanks for always making me feel at home. Specially to Marta, Paloma, Spyros, Danielle, Ed, Jenny, Sam, and now Martí, for the coffees, pub afternoons and walks in the park.

To Lucy Wright for being my unofficial mentor, housemate and friend and our hours long breakfasts and talks. To Joan Albert, whose talks with inspired me and encouraged to keep pushing while being kind to myself. To Merce, Sandra, and Sara for always being there for me even if we only see each other few times a year. To Alexis Sentís, coR, friend and always tricking me into the most fun work, and to Alba Fernández-Sanlés, friend and an example to follow. To them and all my friends in Catalonia who had always made me felt loved and cared for.

To Carlos Martínez and Alberto Acedo, eternal flatmates, for continuing to endure my daily rants and taking part in the WhatsApp talk shows, even while being more than a thousand kilometres away. Without them, this thesis would have probably taken 3 years less, but what a boring experience it would have been.

To my parents Pilar and Jaume, and my brother, Arnau, for their everlasting love and support, and for always being there for me. To Gabi, girlfriend, lover, friend, companion, ukulele player, proof-reader, editor, and so many other things; for

being there in the good and not so good moments and for being a pillar of this thesis. To the rest of my family, specially to my late grandmas, who, even if they would have had no idea of what I work on, would have been so proud of me.

Finally, to all the patients who suffered from the diseases studied in this Thesis, their families, and caregivers. To the healthcare and other professionals involved in the care of these patients. Without them, and their data this thesis would have been impossible. It is my desire that this work can contribute back, to improve care for them and new patients.

Funding

The work on this dissertation was funded by a Fellowship from Fundación Alfonso Martín Escudero (FAME) and a grant from the Medical Research Council – Doctoral Training Programme (MRC-DTP) (MR/K501256/1, MR/N013468/1). I am grateful to FAME and the MRC for their support throughout the DPhil, and specially for their flexibility during the COVID-19 pandemic.

UTMOST project was funded by the NIHR HTA (project number 15/80/40) and was supported by the NIHR Biomedical Research Centre, Oxford. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Abbreviations

AD	Anno Domini
APACHE	Acute Physiology and Chronic Health Evaluation Score
ASA	American Society of Anesthesiology
ASMD	Absolute Standardised Mean Difference
ATE	Average Treatment Effects
ATT	Average Treatment Effects on The Treated
BC	Before Christ
BMI	Body Mass Index
BMJ	British Medical Journal
CE	European Conformity
CHARLSON	Charlson Comorbidity Index
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
COVID-19	Coronavirus Disease 2019
CRP	C-Reactive Protein
CT	Computed Tomography
DM	Diabetes Mellitus
EHDEN	European Health Data Evidence Network
EHR	Electronic Health Records
EMA	European Medicines Agency
EQ	EuroQol
ES	Effect Size

EU	European Union
EUDAMED	European Database for Medical Devices
FDA	Food And Drug Administration
FFDCA	Federal Food, Drug, And Cosmetic Act
HES	Hospital Episode Statistics
HR	Hazard Ratio
HTA	Health Technology Assessment
ICD	International Classification of Disease
ICIJ	International Consortium of Investigative Journalists
ICU	Intensive Care Unit
ID	Identifier
IDEAL	Idea, Development, Exploration, Assessment, Long-Term Follow-Up Collaboration
IMD	Index Of Multiple Deprivation
IMV	Invasive Mechanical Ventilation
INR	International Normalised Ratio
IPW	Inverse Probability Weighting
IQR	Interquartile Range
IRR	Incidence Rate Ratios
IV	Instrumental Variable
KR	Knee Replacement
LDH	Lactate Dehydrogenase
MD	Medical Device
MDD	Medical Device Directive

MDR	Medical Device Regulation
MERS	Middle East Respiratory Syndrome
MHRA	Medicines & Healthcare Products Regulatory Agency
MI	Myocardial Infarction
MICE	Multiple Imputation with Chained Equation
NHS	National Health Service
NIHR	National Institute for Health Research
NJR	National Joint Registry
OA	Osteoarthritis
OHDSI	Observational Health Data Sciences and Informatics
OKS	Oxford Knee Score
ONS	The Office for National Statistics
OPCS	Office Of Population Censuses & Surveys Codes
OR	Odds Ratio
OTI	Orotracheal Intubation
PAFI	Pao2/Fio2 Ratio
PCR	Polymerase Chain Reaction
PEEP	Positive End-Expiratory Pressure
PH	Proportional Hazards
PJI	Prosthetic Joint Infection
PKR	Partial Knee Replacement
PMA	Premarket Approval
PPE	Personal Protective Equipment
PROMS	Patient Reported Outcome Measures

PS	Propensity Score
PSA	Propensity Score Adjustment
PSM	Propensity Score Matching
PSS	Propensity Score Stratification
RCT	Randomised Controlled Trial
REF	Reference
RMSE	Root-Mean-Square Error As A Measure Of Accuracy
RR	Risk Ratio
SARS	Severe Acute Respiratory Syndrome
SCCS	Self-Controlled Case Series
SD	Standard Deviation
SE	Standard Error
SMD	Standardised Mean Difference
SOFA	Sequential Organ Failure Assessment Score
TKR	Total Knee Replacement
TOPKAT	Total Or Partial Knee Arthroplasty Trial
UDI	Unique Device Identification
UK	United Kingdom
US	United States of America
UTMOST	Risk-Benefit and Costs of Unicompartmental (Compared To Total) Knee Replacement For Patients With Multiple Co-Morbidities Project
VTE	Venous Thromboembolism

Table of Contents

Chapter 1

Introduction to Medical Device and Surgical Epidemiology and to the methods for their study in observational research

In this chapter I introduce the importance of Medical Device (MD) and surgical epidemiology. I describe how important safety concerns over the past decades have led to increased societal awareness of the issues concerning MD and how this has prompted changes in regulatory frameworks. I briefly explain the changes to regulatory frameworks in the EU, US, and UK and why a better understanding of methods is needed to address the new requirements arising from these regulations. After this, I focus on how MDs and surgical procedures are evaluated and the unique challenges that MDs and surgery pose to epidemiological research. I finally argue for the use of my main clinical case, the evaluation of Partial vs Total Knee Replacement, as a testing ground for methodology, and explain how I applied the lessons learned to an urgent COVID-19 question. Finally, I present the aims and structure of this dissertation.

1.1 Medical Devices and Surgery

MDs are widely used in the whole spectrum of medicine. Through history, humanity has used technological solutions to diagnose illnesses, improve disease symptoms or replace organ functions. There are records that date as early as the 5th millennia BC on the use of artificial dentures and flint drills to treat cavities.(1) Surgical instruments very similar to those used nowadays were found in the Roman buried city of Pompeii (79 AD).(2) This use of MD has continued to our days and boomed in the last century.

Definitions and classifications of MDs differ across regions and countries, as set by local and national laws and regulations. A common splitting point is whether the MD is meant to be implantable or not. Implantable devices are those placed inside the body during a procedure and meant to stay there for a long period. Examples of this kind of MDs include intraocular lenses, hip and knee implants, artery stents, and pacemakers. In 2020-2021 alone, NHS England has recorded over 700,000 inpatient procedures involving implants.(3) This number does not include ambulatory procedures such as contraceptive devices (contraceptive implant or intrauterine device), which are used by over 180,000 women in England.(4)

In this thesis, I will focus on implantable MDs, as they are widely used and pose the greatest medical risks and methodological challenges, but the methods and considerations that arise from these can be applied to all kinds of MDs and related procedures. Implantable MDs require procedures or surgeries in order to be inserted in and potentially replaced or retrieved from the body. Therefore MD closely intersects with surgical epidemiology.

Although MDs have been part of medicine since its start, the explosion of their use in the past century has led to an increasing need for safety and effectiveness evaluation. In the last decades, there have been worrying precedents of MDs causing harm: metal-on-metal hip replacements, (5) contraceptive implants, (6) or hernia meshes are some of the examples.(7) This has led to media interest in exposing the problem of unsafe implantable devices. Examples include a recent Netflix documentary, "The Bleeding Edge",(8) and the Implant Files from the International Consortium of Investigative Journalists (ICIJ),(9) the largest international healthcare investigation in history. These investigations revealed major flaws in how MD legislation worked in the European Union (EU) and the United States of America (US), leading to the approval of potentially dangerous MDs. This lax regulation and a rather small corpus of knowledge on MDs call for

quality research into and eventually guidance on the best methods to investigate the risk-benefit of MDs.

1.2 Regulatory framework and Device Classification

MDs are subject to more lenient regulation than drugs. Most MDs currently on the market, even implantable ones, have not undergone effectiveness or safety trials on the time of their approval, and too few have been assessed in observational post-marketing studies. (10)

The EU and US regulations set similar classifications based on risk for MDs, although with some differences for medium-risk MDs. The FDA classifies MDs in 3 classes: Class I, MDs that are low risk such as bandages, or stethoscopes; Class II, intermediate risk MDs, such as infusion pumps or CT scanners; and finally, Class III, high-risk MDs, such as implantable MDs like pacemakers or knee replacements. (11, 12) The EU classification is similar but separates class II MDs in two. The resulting categories are: i) Class I: low-risk MDs, which are further separated into sterile, measuring and other; ii) Class IIa: low to medium risk, used for less than 30 days, also classified in sterile and non-sterile; iii) Class IIb: medium to high-risk MDs; and iv) Class III: high risk MDs.(13)

Before the 1990s, it was up to each state in the EU to approve MDs to be used in their national space.(14) In June 1993, the Medical Device Directive (MDD) was passed to regulate MDs and allow them to be used across all member states after

earning a European Conformity (CE) mark in any member state.(15) MD approval was overseen by Competent Authorities, governmental agencies dedicated to regulating medical products. Low risk MDs could be just declared to the authorities, however, high risk MDs approval was outsourced to notified bodies, companies that specialise in evaluating products for CE, as that approval required more checks. Notified bodies have the responsibility of assessing if the MDs benefits outweigh the risks using clinical data, but the nature and quality of these studies is under discretion of the notified bodies, and there is no need to release that evidence to the public.(12) This situation contributed to safety issues and media scandals, such as an undercover investigation where reporters were able to get approval for a dangerous hip replacement, even after submitting as evidence for approval to the notified body a report that showed the prosthesis to release dangerous levels of toxic metals.(16)

In contrast, under the US regulation, manufacturers of high-risk MDs must demonstrate safety and effectiveness for new MDs to be approved and marketed. The US first regulated medical products in 1938 with the Federal Food, Drug, and Cosmetic Act (FFDCA). In 1976, the law was amended by Congress, under the *Medical Device Amendments to the FFDCA*, creating a specific regulation for MDs.

This established a risk-based classification system, new requirements, gave the FDA more powers, and created two pathways for MD regulation.(17) The first route for approval is used for new MDs and is called premarket approval (PMA). This route requires MD to prove efficacy and safety in clinical tests, but has been criticised for lacking exhaustivity and scrutiny(18-20). The second route for approval, 510(k), is designed to serve as a quick route for small modifications of previously approved MDs. This route was originally intended for MDs with fewer effectiveness and safety requirements, namely class I and II, such as surgical gloves and hearing aids. (21) However, as MDs quickly turned more complex, new implantable MDs appeared, and the 501(k) route was used for some high-risk MDs.(22) This was aggravated by the approval of the Medical Devices User Fee Act and its Least Burdensome Guidance in 2002, which mandated the FDA to use the most efficient route.(23) Despite this, most high-risk MDs approved in the US (about 80%) still go through PMA,(24) and the remaining are approved via 501(k). However, in a 2011 study looking at public FDA recalls of MDs due to potentially causing serious health problems or death, only 19% of the recalls were for MDs approved via PMA, with 71% having gone through the 501(k) route (and the rest being exempt).(21) In addition to this loophole, Hines J et al. found other 7 regulatory issues with the approval of MDs in the US. (24) These issues led the

Institute of Medicine to ask the FDA for a replacement to the 501(k) pathway.(25) The FDA responded with several reforms that were implanted in 2019.(26) The reformed pathway requires some pre-specified performance criteria, and call for the publication of the list of new devices approved based on approvals more than 10 years old, letting the "market" decide. but still fails to continue to monitor MD after approval and still allows for approvals via the 510(k) route. There is, therefore, a need to reinforce and expand post-marketing studies.(27)

In response to some of these problems and controversies, a new European regulation, of direct application without need for transposition in all member states, the Medical Device Regulation (MDR) EU 2017/745, was published in the Official Journal of the EU on 5 May 2017.(28) It was meant to become applicable on 26 May 2020, but, due to the COVID-19 pandemic, it was delayed and finally became applicable on 26 May 2021.(29) This regulation was not transposed to UK legislation due to the UK leaving the European Union. In February 2021, the UK Parliament passed the Medicines and Medical Devices Act 2021 that gives powers of registration of MD in the UK to the Medicines & Healthcare products Regulatory Agency (MHRA), but without adding additional requirements to those

in the Medical Devices Regulations 2002 that transposed the EU MDD from 1993. (30)

A crucial aspect of the new EU regulation is the creation of EUDAMED, an European database on MDs.(31) It registers actors, MDs, notified bodies, and evidence. Besides the creation of EUDAMED, this law imposes several requirements for high-risk MDs to all actors. Manufacturers must now report clinical evidence of both efficacy and safety in a summary of safety and clinical performance to be publicly available in EUDAMED. Health facilities must keep a record of all MDs implanted by their Unique Device Identification (UDI). The manufacturer must provide annual post-marketing safety surveillance reports and the EU has to establish systems to actively monitor data.(32) The European Medicines Agency (EMA) has received the mandate to support medical device expert panels, and to lead on crisis preparedness and management of MDs.(33)

In summary, several changes have occurred in the US and EU regulations. The FDA has "modernised" it's 501(k) pathway to require more documentation to approve a new device based on a previous one. It also plans to publish the list of devices approved based on approvals more than 10 years old, in a "market-based" approach. It also has launched the National Evaluation System for health

Technology (NEST), a centre to generate "Real World Evidence" throughout the medical device lifecycle using data from voluntary adhered medical providers. In contrast, the EU has increased the need for pre-approval clinical evaluations. This EU regulation has also made mandatory the creation of a UDI for all devices and the registering of those in a database, and now requires manufacturers to conduct post-market surveillance. These new requirements have increased the need for MD trials and observational research, requiring reliable and robust methodologies to study them.

1.3 Evaluation of Medical Devices

Policy changes in the past years have increased both regulators' and the public's awareness of the potential health risks, but also benefits, that MDs pose. This has resulted, as seen in the previous section, in an increased need for clinical evidence pre and post marketing. New regulation, and better data handling capabilities, e.g., linkage of registries to healthcare records, have helped increase the amount of data available or soon to be available on MDs. The need for evaluation of cost-effectiveness, safety and risk management of MDs is clearly there and a vast amount of data to answer these questions is now available.

However, knowledge of how to better address these questions is highly fragmented through disciplines. Historically, MDs and procedures have been studied as part of relatively isolated clinical areas. For example, knee and hip replacements would be studied in orthopaedic and rheumatology sciences and pacemakers and stents in cardiology sciences. But the increase in use and in complexity of MDs, combining not just inert materials but drugs or biological structures, and associated methodological challenges, highlight the need for an interdisciplinary academic discipline of MD science.(34)

IDEAL-D provides a first framework to think about integrated evaluation pathways across all development stages of MDs.(35) This framework, developed by expert consensus, describes a very useful flow structure in stages. These are 1) Idea, the first conception of a MD, 2a) Development, finding the optimal technique or design; 2b) Exploration, finding the outcomes of more widespread use and efficacy; 3) Assessment, comparison to the current standard of care and 4) Long-term studies, assessing the long-term effects and outcomes. They propose different studies for each stage, but fall short at the last stages, especially at stage 4, where they only propose to have a comprehensive disease-based registry or database. This is insufficient for the post-marketing requirements of MDD and for the more thorough study of safety, effectiveness, and costs as registries may only include information on some variables and may lack longitudinality. To resolve this, some of the IDEAL-D authors propose to link registries to routinely collected data.(36)

Methods to study MDs using registries and routinely collected data are, however, understudied and several challenges and pitfalls remain unsolved, as I will explain in **Section 1.7**. I focused this thesis on pinpointing the areas where this methodological vacuum exists. I selected areas that are of utmost importance for MD and surgical science and that can at least partially be answered with

observational data, namely stages 3 and 4 of the IDEAL-D framework. These areas correspond to three questions: whether the MD works, whether the MD is safe, and whether the MD has the same effect (benefit or harm) in all patients. In other terms: comparative effectiveness, safety, and heterogeneity of treatment effects. These objectives are closely interconnected, as they share methodologies and sometimes also outcomes.

1.4 Comparative effectiveness

Comparative effectiveness refers to the performance of MDs in clinical settings for the outcome they were intended to improve or prevent. Efficacy trials, also called explanatory trials, investigate the benefits of an intervention under controlled conditions. These trials are indispensable to first establish the utility of a MD. However, there is also a need to understand if MDs will also perform well in less than perfect conditions, as the conditions where MDs are deployed are usually different from the ones in efficacy trials, potentially leading to a reduction in external validity. Therefore, an efficacy trial should be followed by a large pragmatic trial, aimed at understanding how these MDs perform in clinical practice.(37) As previously discussed, many marketed MDs and procedures, both in the US and EU, have not undergone proper RCTs to assess effectiveness.

Although comparative trials are the optimal way to assess efficacy, constraints such as funding or difficulties of recruitment could prevent them from being performed. Also, investigating whether there are effect modifiers, such as structural, hospital or surgeon factors is usually out of the scope of trials. In this Thesis I analyse observational data on four use cases: i) the timing of elective tracheotomy for COVID-19 in **section 2.2**, where the emergency situation called for

a quick evaluation of the available data; ii) the effectiveness of knee replacements on high surgical risk patients, a population usually excluded from trials, as seen in **section 4.1;** iii) the impact of health inequities and structural disparities on knee replacement effectiveness as seen in **section 4.2**; and iv) the modifying effects of surgeon experience and volume on knee replacement outcomes, as seen in **section 4.3**. Cases I and II would probably be well-suited for a trial, although structural reasons, such as limited resources, prevented them from occurring. Case III and IV are examples of health services epidemiology, where although a trial could help answer the question, are probably better suited to an observational study, as our interest lie not so much on the intervention but on the processes that modulate the intervention effectiveness. In cases such as these, there is a need for rigorous approaches to the analysis of comparative effectiveness using observational data.

1.5 Safety

The study of safety is probably one of the most important undertakings of observational epidemiology in MD. The increase in post-marketing safety surveillance requirements that will be in place with new regulations on MDs calls for robust methods. There is an increasing amount of good quality data on MDs from registries and other routinely collected data sources.(38) These data, together with the availability of long-term follow-up via electronic health records, has offered the opportunity to study safety issues that would have been missed if we were to rely only on clinical trials.(39) Safety is a long studied subject in pharmacoepidemiology and there are well-established standards on how to perform studies on drug safety. (40) However, there is little information on how to perform observational safety studies of MD and the problems they present.

In this Thesis I applied standard pharmacoepidemiologic methods to both short-term and long-term safety of knee replacements including cohort (**section 3.1**) as well as and self-controlled designs (**section 3.2)**.

1.6 Treatment Heterogeneity

Effect modification and interaction are two very important phenomena in pharmacoepidemiology. Effect modification is an effect that changes depending on another variable. If two variables have a causal effect on the outcome, we have an interaction.**(41)** This difference in effects is important in MD, as factors like surgeon and hospital variables, and patient features like deprivation or gender can have a great impact.

In **Chapter 4**, I explore in-depth different effect modifiers of knee replacements. First, I apply the methods that replicated trial results on **Chapter 2** to those patients who would have been excluded from the trial. This helps assess if this MD is equally effective and safe in these patients, as shown in **section 4.1**. Furthermore, I explore if there are differences in effectiveness and safety of partial vs total knee replacements according to gender, age and deprivation in **section 4.2.** Finally, I explore one of the most striking cases of effect modification, the impact of surgeon volume on outcomes, in **section 4.3**.

1.7 Medical device and surgical research caveats

MDs, in particular implantable ones, present some particularities that make their study challenging. These challenges must be carefully considered while designing and performing MD studies.

The first challenges relate to the characteristics of the MDs themselves. First, some MDs are used in very specialised environments, like the Extracorporeal Membrane Oxygenation in the ICU, and that makes it very difficult to properly power analyses for some outcomes without tapping into national or international registries or EHR. This is also a problem for other more widely used MDs, but that undergo several modifications in short periods of time. This could mean that the same MD could be substantially different as time passes, with only a few people receiving each iteration. If detailed information about the MD itself is not available, as is often the case when using health records, effectiveness estimates may be an average of all MDs grouped and safety issues could be overlooked. Having access to detailed registries that incorporate brand, model, and version, or to unique MD identifiers could help mitigate these potential downfalls.

Another relevant issue with MDs is finding a suitable comparator for a MD. This is the case when, for example, the alternative to the MD is a less invasive procedure

or pharmacological treatment. Other issues happen when the chosen exposure and comparator have slightly different indications de facto, e.g., when a MD is only implanted if the patient is considered fit enough to undergo the surgery. This calls for careful thinking and design of the comparability of the chosen exposures and acknowledgement of all possible biases that arise from it.

Another key problem when studying medical devices is how to disentangle the effects of the devices from the effect of the procedures. This is the case for the previous caveat, if a comparator device has a less invasive procedure or for differences between surgeons and procedures explained in the next paragraphs. It is also important to note that some devices can be implanted using different procedures, making their study even more difficult.

Ascertaining when a patient is exposed could also be challenging in MD epidemiology. The same MD could result in different definitions of exposure. For example, we could consider a patient exposed after they had had a defibrillator implanted, but also after a defibrillator cardioversion shock. This also could happen when considering drug-releasing implants. Another issue with exposure definition, very common when using routinely collected data, is to pinpoint the exact site where the MD is implanted. Laterality of MDs, such as the side where a

knee replacement or intraocular lenses have been implanted, or detailed localisation, such as which coronary artery has a stent been applied to, are usually not well recorded. Another type of usually missing information is the type of device or surgery, such as in total and partial knee replacement, where more detailed information on the device is only available in registries. This may lead to biases or hinder a whole study, so the definition of exposure ought to be contemplated in the design phase.

Another caveat related to the exposure is that for some MDs there is a large lag between the decision of implanting them and the actual operation. This could be the case with high-risk patients who might die or have complications that prevent them from having the MD implanted in the preoperative stage, or if the allocated MD is deemed impossible to implant after surgical inspection. This generates a problem when identifying observations based on the treatment, named immortal time bias.(42) If one only selects patients who have undergone a procedure (rather than all patients that were eligible), there is a period of time when some outcomes of interest could not have occured. The presence of this bias is very frequent in MD epidemiology, and should be considered and corrected, when possible, in the design stage.

A final consideration regarding exposure is the moment when the exposure ceases. Although not always possible, it is important to have information on when and whether an implantable MD had to be removed, as it could mean it led to worse outcomes for the patient than for those whose MD is still implanted. This should be considered in MD studies, carefully planning how the end of exposure will be analysed.

There are several biases relating to patient characteristics. The classic confounding by indication (40) is also present in MD epidemiology, being of special importance disease severity. Some MDs are implanted only when pharmacotherapy has failed, or different MDs may be used for different disease stages. This can be a problem with routinely collected data, as measures of severity may not be recorded. However, when recorded, we have design and analytical tools to minimise confounding. Another potential bias related to the patient is the "healthy user bias", where patients who receive the implant are generally healthier than those who do not. This could be the case for very invasive operations, where the surgical risk does not outbalance the long-term beneficial effects of the MD. A similar effect occurs in RCTs where excluding the sicker patients can harm generalisability, in

contrast to observational studies, where all patients are usually included potentially leading to confounding by indication.

A key consideration, which has much less effect in the study of drugs, is by whom and where the procedure is performed. Characteristics of the hospital or clinic where the procedure is performed such as size, clinical level, teaching status, local deprivation, or distance to the site of residency of the patient can affect the outcomes and should be taken into consideration.

The neighbourhood where a patient lives, the access to social support and care, and other potential sources of inequity such as gender, age, ethnicity, and social class can be major determinants of outcomes. MDs are much more sensitive than drugs to factors such as access to post-operatory care, right to sick leave after a surgery, or capacity to perform long-term rehabilitation or self-care. These factors must always considered and studied in order to improve care and wider policies related to MDs.

Surgeon characteristics have, sometimes, a much higher impact than prescriber characteristics in the study of drugs. Volume, the number of operations a centre or a physician performs per year, can have a great impact on some outcomes through different mechanisms such as technical specialisation, dedicated processes, or

better patient selection. This is highly related with learning curves for some procedures, where outcomes improve after an operator or a team has done a certain number of operations. This highlights the importance of recording good information both on the operator and the wider provider, which is sometimes missing in health records.

The final consideration relates to the outcomes we study. The study of long-term outcomes, which are of importance in implants especially for safety, requires good follow-up data, through providers and levels of care. This is not usually recorded on registries, so the use of linked healthcare data is necessary. Shorter term outcomes such as those related to post-operative care require less time of follow-up but more detail on dates, such as dates of outcome diagnosis or dates of discharge of the hospital, in addition to the procedure date.

Target Trial framework

To minimise the risk of confounding and some biases described in this section, it is helpful to have a pre-specified protocol and have thought about the effect that intervention allocation and timings could impact results. A good tool to do this is the "target trial" framework.(43) This consists in trying to mimic as much as

possible the timings and conditions of a randomised experiment. To do so, one must think about what the perfect trial to answer the research question at hand and mimic it with observational data. This makes biases apparent and makes us articulate the trade-offs one accepts when designing the study. Although not all the studies on this thesis have not been designed following explicitly this framework, I found it a good tool to uncover potential biases, so I have added a discussion based on the framework.

1.8 Clinical applications: Knee Replacement and Tracheotomy

This Thesis focuses on some of the aforementioned problems and particularities, using as an example the knee replacement (KR). KR is one of the most used implants in the world. In the US over 1,000,000 KR are performed yearly, and more than 650,000 are performed in the EU.(44, 45) A knee replacement is usually performed when the knee is damaged and substantial pain and reduction of mobility exist. This damage is usually a consequence of different diseases and conditions, the most common being knee osteoarthritis. Other diseases that can lead to a KR include rheumatoid arthritis, haemophilia, gout, or traumatic injuries. (46) Although there are other surgeries, the main types of knee replacement are Total knee replacement (TKR) and Partial Knee replacement (PKR). (47-49) TKR replaces the whole knee, and all the knee compartments involved. PKR only replaces the affected compartment of the knee, but it is much less common. (50) This difference also makes the replacement procedure for PKR less invasive, making difficult to separate the effect of the device from the effect of the procedures.

There has been extensive previous research both in TKR and PKR. From 2000, more than 26,000 articles have been published on knee replacement.(51) Recently,

results from a large RCT have shown that their efficacy in reducing pain and improving mobility is similar. I tried to mimic the settings of the RCT in **Chapter 2** to study different methods to reduce confounding. But there are still important clinical questions to resolve, such as comparative safety, the risk of complications, or the effect of surgeon volume of outcomes, that I aimed to answer in **Chapters 3 and 4**. The high use, the availability of well-recorded registry data, and the readiness of solid research for comparison makes knee replacements a perfect test ground for MD observational methods.

The COVID-19 pandemic gave me the opportunity to apply the expertise gained in the study of knee replacements to important pressing clinical questions. The need arose to perform an observational study about tracheotomy, the results of which are shown in **Chapter 2.1**. A tracheotomy is a surgical procedure that implants a tracheotomy tube in the trachea to bypass upper respiratory system and allow direct access to the breathing tube.(52) In COVID-19 treatment, a tracheotomy is sometimes needed for the patient management in the ICU, when the patient needs invasive mechanical ventilation for a prolonged period. (53, 54) I aimed to assess the best timing for this procedure, as this was disputed. (55) This study was a good

opportunity to apply the target trial framework (43) to MDs/procedures and explore different methods to analyse time to event data.

1.9 Aims

Given the lack of evidence on which are the best methods to address observational research of MDs and surgical procedures, I focused to test some of them on the study of knee replacement and tracheotomy. The overall aim of this thesis was to evaluate the effects both in terms of effectiveness and safety of these interventions from observational data. To do that I use a target trial framework and a range of methods to reduce confounding when needed: propensity score matching, stratification, inverse probability weighting and instrumental variable estimation and the self-controlled case series design.

The second aim was to find in simulation studies why preference-based Instrumental Variables failed in the case studies, and whether that failure was related to caveats and problems common in Medical Devices and Surgery such as multilevel structures and surgeon level effects.

The work presented in this thesis arises from a careful evaluation of methods applied to a rich dataset of knee replacement patients and a registry of tracheotomies, and their evaluation in simulation studies.

1.10 Structure of thesis

The objectives for this thesis are distributed in six chapters as follows:

1 – Introduction

2 – Comparative Effectiveness

In this chapter, I focus on effectiveness of Partial Knee Replacement on the UTMOST study, and on the effectiveness of early tracheotomy to:

a) Evaluate the suitability of observational methods, Instrumental Variables (IV) (preference, volume, area and time based) and Propensity Scores (PS) (adjustment, stratification, matching and inverse probability weighting), to assess comparative effectiveness of Partial Knee Replacement (PKR) vs Total Knee Replacement (TKR).

b) Find out which methods out of IV estimation and PS matching, stratification, and IPW better replicate in an observational setting the results of a pragmatic trial that answered the same question.

c) Use the state-of-the-art methods to quickly answer pressing questions such as the optimal timing of tracheotomy surgery in COVID-19 patients requiring invasive ventilation.

3 – Comparative Safety

In this chapter, I focus on safety of knee replacements to:

a) Evaluate PKR vs TKR safety outcomes using PS matching, stratification, IPW and IVs.

b) Evaluate the suitability of the Self-Controlled Case Series (SCCS) design for analysing the risk of complications after PKR or TKR.

4 – Treatment Heterogeneity

In this chapter, I focus on conducting subgroup analysis of comparative effectiveness and safety of knee replacements to:

a) Assess the effectiveness and safety of PKR compared to TKR on patients with high surgical risk, using the methods validated in Chapter 2.

b) Elucidate if there are differences in the effectiveness and safety of PKR between gender, age, and socioeconomic factors.

c) Quantify the modifying effect of surgeon volume on the effectiveness and safety of PKR and TKR.

5 – Methods Research

In this chapter, I use simulation to test the performance of different methods:

a) Test and evaluate the performance of preference based IV estimation in the setting from Chapter 2, in presence of non-normal outcomes and confounding.

b) Test and evaluate the performance of preference based IV estimation in a surgical multilevel setting.

c) Test and evaluate different multilevel PS strategies to minimise confounding.

6 – Conclusion

In this chapter, I summarise the overarching results of the previous chapters and discuss the potential implications and limitations of the findings for MD and surgical epidemiology.

Chapter 2

Effectiveness - Assessing the best methods in observational data replicating a trial of PKR vs TKR and the trial emulation framework for the study of COVID-19 tracheotomy timing

Comparative effectiveness, as pointed in **section 1.4**, can refer to all kinds of benefits, or benefit-cost ratios between different interventions or MDs. In this Chapter, I will focus on the benefits, how to evaluate which of several interventions produces the desired effect using routinely collected data.

Randomised Controlled Trials are the best method to assess effectiveness of implantable MDs, but these are still uncommon due to different factors.(56) Some examples of these are costs, time needed until study completion, ethical considerations, and other feasibility issues.(57, 58)

The use of observational routinely collected data, while not being the gold standard, constitutes an efficient and fast way of evaluating effectiveness. The main predicament of observational research is always the possibility of confounding by indication. Confounding could introduce bias or produce spurious results. Thus, we need methods to minimise confounding.

In this chapter I will aim to accomplish two objectives. First, to try to understand how the available methods to deal with confounding, which we borrow from drug effectiveness research, (40) perform in a MD setting. For this, I will compare the results of instrumental variables and propensity score methods to the results of a clinical trial studying total vs partial knee replacement. Second, I will demonstrate the usefulness of the target trial framework (59) for the evaluation of surgical interventions. To do so, I will apply the framework to the study of the optimal timing of tracheotomy in COVID-19 patients that need mechanical ventilation.

2.1 Assessing the best methods to replicate a trial of partial vs total knee replacement

UTMOST was a project designed to evaluate effectiveness, safety, and treatment heterogeneity for partial vs total knee replacement, and some of its results and methods will be spread across the relevant chapters of this dissertation.(60) But, as this is the first time it appears, I present a full introduction of the study and of all its objectives next:

Introduction

This study focuses on one of the MDs that is used most frequently and for the longest period: Knee Replacement. This MD is used for the treatment of osteoarthritis (OA) of the knee. OA is one of the most common diseases of the knee and produces pain, swelling and stiffness.(61, 62) Patients can be managed conservatively using drugs to relieve pain and discomfort, but when these symptoms have an impact on their quality of life and cannot be controlled pharmacologically, knee replacement surgery is indicated.(63)

Knee replacement is the most common implantable MD in the UK, with 303,960 procedures between 2015 and 2017. (50) Total knee replacement (TKR), the most common procedure, greatly improves function and quality of life.(47, 48) Another

strategy, when possible, is to use a partial knee replacement (PKR), which only replaces the affected compartment of the knee.(49) Fewer than 9% of knee replacements in the UK are partial, (50) but it is estimated that up to 47% of patients would be suitable for a PKR.(64) There is controversy on which treatment is the most indicated for these cases. TKR is a simpler operation, thought to be less prone to early problems and failure, and it is believed that PKR would eventually fail and require revision surgery, which involves a TKR procedure. (65) In contrast, PKR is cheaper and results in faster recovery, (66) fewer complications, (67) and superior function. (68) However, there is little evidence to inform this choice. To resolve this matter, a recent RCT has been conducted: TOPKAT (NIHR HTA – 08/14/08: Total or Partial Knee Arthroplasty Trial).(69)

This well-designed multi-centre RCT has successfully recruited, randomized, and followed participants up for a total of five years. The results from TOPKAT have been reported in the form of a scientific manuscript in the Lancet, (70) and in an NIHR HTA report.(71) TOPKAT demonstrates a small (<2 points in Oxford Knee Score, OKS) short-term (1-year) benefit in patient reported outcomes associated with the use of PKR compared to TKR, but no difference in the longer term (five years).

The quality and internal validity of MD RCTs such as TOPKAT is unquestionable, but two key issues limit their usefulness for determining the comparative risk-benefit in actual practice conditions:

- Limited external validity: only patients with ASA (American Society of Anesthesiologists) grades 1-2 were eligible for enrolment in TOPKAT. This excluded patients with multiple comorbidities. National Joint Registry (NJR) reports suggest that about one in six candidates (16.7%) for Knee Replacement (KR) surgery are ASA grade 3 or worse, (72) not eligible for inclusion in TOPKAT.
- Length of follow-up and statistical power: due to the cost and difficulty of primary data collection, surgical RCTs are sometimes underpowered to detect rare events. This limits the availability of data on key (usually rare and long-term) safety outcomes, such as complications (revision, systemic infection, wound infection, cardiovascular disease, and venous thromboembolism).

There is therefore a need to complement the results from TOPKAT with good quality data on the performance of such different surgical approaches for the

increasing group of multi-morbid patients requiring knee surgery, which TOPKAT cannot provide.

Observational data from the NJR can provide insights into the impact of these different types of knee replacement for all patients because detailed surgery information is mandatory in the UK. In a recent Lancet paper, the authors tried to answer some of these questions, using propensity score matching to minimise bias.(73, 74) In this manuscript, the authors acknowledged that unmeasured confounders (such as unrecorded conditions, disease severity, or drug use) could at least partially explain the study findings since propensity score matching could only account for measured confounders. Such unresolved bias can however be minimised with other more novel and robust pharmaco-epidemiological analytical methods, such as instrumental variables,(75) inverse probability of treatment weights, or propensity score stratification.(76-78)

These methods have recently been applied in observational drug/medicines comparative safety and/or effectiveness research. Indeed, the US FDA and colleagues from a number of academic institutions are working on the replication of previous drug RCTs using observational methods in an attempt to demonstrate their usefulness for drug and vaccine safety and comparative effectiveness

research.(79-81) These methods have been used to compare the performance of different surgical procedures or MDs, specially PS matching, but rarely have been tested in a trial replication study.(82) We need a better understanding of the performance of these different methods for comparative effectiveness/safety studies for the evaluation of surgical and MD alternatives using (observational) routinely collected data.

Furthermore, this need becomes more pressing with the introduction of new regulations for MDs, covered in **section 1.2**, which require a more comprehensive evaluation of implantable devices including orthopaedic prostheses. (83) Methodological research and guidelines on the use of real-world data is needed to inform the post-marketing use and safety of MDs.

Objective

To study the validity of different epidemiology analytical methods – traditionally used in drug and vaccine studies to minimise confounding – for the assessment of alternative surgical procedures. I will use knee replacement (PKR compared to TKR) amongst ASA grade 1-2 (eligible for TOPKAT) as a use case for demonstration, where the previously mentioned TOPKAT RCT will be used as a "gold standard" for comparison.

2.1.1 Methods

Study Design, data sources and study population

Retrospective cohort study. All patients who had undergone a first primary total (TKR) or partial (PKR) knee replacement in the National Joint Registry (NJR) from 2009 until December 2016 were included. The NJR cohort was linked to the Hospital Episode Statistics (HES). Finally, this resulting dataset was also linked to the Patient Reported Outcome Measures (PROMs) database. A description of these databases and their linkage follows. ***Figure 2.1*** shows data sources usage in the study.

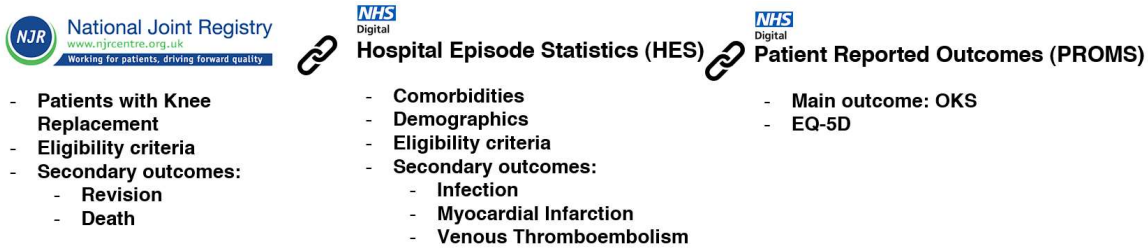


Figure 2.1 - Data sources used for the UTMOST study.

NJR

The National Joint Registry (NJR) for England, Wales, Northern Ireland and the Isle of Man is a registry that collects information on joint replacement surgery of hip, knee, ankle, elbow and shoulder and monitors in real time the outcomes achieved by these MDs. It includes both primary and revision surgeries. The NJR

started collecting data for knee and hip replacements in April 2003. This collection has become mandatory since April 2011. At first, compliance was low, e.g., 43% in 2004, but rolls to 95% in 2015. Recent studies show that actual data is complete and accurate. (84, 85) As for knee replacements, NJR contains more than 1 million verifiable primary knee replacements up to 31 December 2017 with a maximum follow-up of 14 years.(86) Osteoarthritis is the most common reason for these surgeries, being the cause of 85% of all the knee replacements and the most common procedure type. Revision surgeries are linked to the primary operation using patient unique identifiers. There are 33,292 first revisions linked to primary knee replacements.(86) NJR also has mortality data, obtained from external linkage to the Office of National. A data request on NJR knee replacement taking place until the end of 2016 was granted in October 2017 (NJR Application Reference RSC2016/13).

HES

Hospital Episode Statistics (HES) is a database covering NHS hospitals and independent sectors that provide NHS services in England. HES includes admitted patient care data from 1997, outpatient data from 2003, accident and emergency data from 2007 and diagnostic imaging data from 2012. HES data are submitted by

hospitals for reimbursement purposes. Although HES is an administrative dataset, its inpatient records have been used extensively for research purposes in recent years. (87) These records have high accuracy and completeness for musculoskeletal procedures and outcomes.(88)

HES impatient records provides information on hospital diagnoses and procedures, administrative details (date of admission and discharge), and basic sociodemographic data (such as region, age and gender) for each hospital admission recorded. Diagnoses are coded using the International Classification of Disease Version 10 (ICD-10), and procedures are coded using the Office of Population Censuses & Surveys Version 4 (OPCS-4).

PROMs Database

Patient Reported Outcome Measures database (PROMs) records routinely collected quality of life questionnaires for some elective surgeries performed in England before and after the surgeries: knee and hip replacement, varicose vein and groin hernia surgeries. It was introduced in 2009 by the National Health Service (NHS) England.(89, 90) The PROMS Database contains patients' self-completed questionnaires at the time before surgery and at 6 months after surgery. Pre-operative questionnaires are administered at the hospital and post-operative

questionnaires are sent out to patients six months post operation and collected by post.

For knee replacement, the focus of the study, PROMs records Oxford Knee Score (OKS). I used this measure as the main outcome, as it is the same measure TOPKAT uses. This score is used to measure self-reported knee pain and function. It consists of 12 questions and five possible responses, making up to a score ranging from 0 to 48.(91, 92)

Another well-known and widely used PROM recorded in PROMS database is EuroQoL (EQ-5D-3L). EQ-5D-3L is a quality-of-life measure with two parts: the EQ-5D index and EQ-5D general health scale scale. EQ5D-index contains five questions and three possible responses for each of five sub-scales: mobility, self-care, daily activities, pain or discomfort, and depression or anxiety.(93) The raw score obtained has been weighted in the database, according to UK preferences, to represent the whole UK society, resulting into a score (EQ-5D utility index) ranging from -0.59 (worst state) to 1.00 (best state). The EQ-5D general health scale scale is a patient self-assessment on their health in general with a score of 0 (the worst imaginable) to 100 (best imaginable).(89)

Data linkage

NJR patients were matched to HES in a deterministic fashion, requiring the same information on patient identifiable fields. HES was matched to PROMS in a probabilistic manner. This data linkage was approved and conducted by NHS Digital (DARS-NIC-172121-G0Z1H-v0.11).

Target Trial and sample

The target population were the patients in NJR who had a record of a primary TKR or PKR from 2009 to 2016. For the replication of results to work, the most important factor is to get as close as possible to what would have been the trial population. I tried to apply the eligibility criteria for this as closely to the TOPKAT criteria I could with the available data. Operationalization for each of the inclusion and exclusion criteria is shown in ***Table 2.1***. Clinical code lists for this exclusion criteria can be found in the ***Appendix 2.1*****.**

In brief, subjects were excluded if they had previous inflammatory arthritis, foot, hip or spinal pathology, or previous knee surgery. They were also excluded if they had an operative ASA of 3 or more.

TOPKAT criteria	UTMOST Criteria
TOPKAT surgery type	
TKR or PKR and considered eligible for both	Only TKR/PKR recorded in the surgery type were included.
Trial participants	
Patients participated 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/PKR was included if there were multiple knee replacement surgeries in the NJR.
-	Patients received surgeries before 31 Dec 2016 to allow post-operative 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 post-operative OKS collected were excluded in the OKS cohort.
Inclusion criteria	
Medial compartment osteoarthritis with exposed bone on both femur and tibia	Data unavailable- clinical assessment was not recorded in the NJR.
Functionally intact anterior cruciate ligament	Patients with a record of previous cruciate ligament injury in HES were excluded.
Full thickness and good-quality lateral cartilage present	Data unavailable- clinical assessment was not recorded in the NJR.
Correctable intra-articular varus deformity	Data unavailable- clinical assessment was not recorded in the NJR.
Medically fit showing an ASA of 1 or 2	Patients with ASA of 1 or 2 in NJR were included.
Clinical exclusion criteria	
E.1. Require revision knee replacement surgery	Not applicable as only primary procedures of TKR/PKR were included.
E.2. Have rheumatoid arthritis or other inflammatory disorders	Patients with a record of rheumatoid arthritis or other inflammatory disorders were excluded.
E.3. Are unlikely to be able to perform required clinical assessment tasks	Clinical assessment was not recorded in NJR
E.4. Have symptomatic foot, hip or spinal pathology	Patients with a record of foot, hip, or spinal pain in 1 year prior were excluded.
E.5. Previous knee surgery other than diagnostic arthroscopy and medial menisectomy	Patients with a record of prior knee surgery were excluded.
E.6. Previously had septic arthritis	Patients with a record of septic arthritis were excluded.
E.7. Have significant damage to the patella-femoral joint especially on the lateral facet.	Patients with a record of patellofemoral damage were excluded.

Table 2.1 Inclusion and exclusion criteria and its correspondence with TOPKAT (RCT)

The operation date recorded in 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. In the context of timing of the target trial timings, I was not able to have information at time of randomisation, the time when the decision of the surgery was made, and if there was differential attrition from that point to the surgery.

In addition to the TOPKAT exclusion/inclusion criteria, I also excluded patients with bilateral / lateral / patellofemoral knee replacement according to NJR, as they were not eligible for PKR.

Patients without index of multiple deprivation (IMD) data were excluded due to the implausibility to impute IMD with the available variables.

Finally, patients who opted-out to participate from research were excluded, with an updated list of patients obtained from NHS Digital in October 2019.

For this chapter, I also restricted the cohort to those patients who had been part of the PROMS database and had post-operative OKS recorded.

Outcome: Oxford Knee Score

The primary outcome of the trial was post-operative Oxford Knee Score (OKS). The OKS is a PROM that measures patient perceived knee pain and function using 12 questions and five possible responses, summing up to a score ranging from 0 to 48, with 48 being the best possible outcome. This questionnaire is specific for assessing function and pain after TKR surgery and it has been validated and shown to be an effective measure. (94, 95) For this project, I used as outcome OKS as collected at 6-12 months after surgery. This is similar timing to the target trial, 1-year post randomisation OKS in TOPKAT

Methods to minimise confounding

In the last decades, several methods have been proposed to address the issue of confounding in epidemiology. Propensity Scores (PS) were proposed by Rosenbaum and Rubin in 1983 to minimise confounding, and since then have arisen as one of the most used methods to do so, also in surgical literature.(82) These have evolved, and several new techniques have improved their performance, and are widely accepted as one of the best ways for minimising confounding.(40) However, their main limitation is the inability to address unmeasured confounding. Another method, called Instrumental Variable (IV)

analysis can theoretically address this, assuming several very stringent conditions are met.(77) I aimed to test several of these methods against the results from the RCT.

In **section 2.1.4**, I explain and test a range of PS methods, and produce their respective treatment effect estimates. The tested methods are Propensity Score matching, PS stratification, and Inverse Probability weighting. I also tested in this section conditioning on all selected confounders and non-adjusted regression. IVs used and methods for them are explained and presented in a different **subsection, 2.1.3**. In this section I present several potential instrumental variables based on surgeon preference, hospital preference, geographical location and on calendar time.

Sensitivity Analyses

The trial, TOPKAT, included only experienced surgeons who had done more than 10 PKR surgeries. To replicate this in a sensitivity analysis, I restricted the cohort to those patients whose surgeries were performed by surgeons who had done more than 10 surgeries of the same type, that is PKR or TKR, in the previous year.

Comparing to TOPKAT

The choice of agreement criterion to assess whether a method is deemed valid is not straightforward. For these studies I used 5 agreement criteria to assess whether a method could be seen as capable of replicating the results from the TOPKAT RCT. These criteria were: 1) coverage, 2) whether the treatment effect estimate is inside the 95% CI of the TOPKAT estimate, 3) the chi-square for heterogeneity test with a p-val>=0.05; 4) an I2 for heterogeneity below 40%; having a small between method variance (Tau2); 5) and statistical significance agreement, defining as equivalent results those with the same direction and significance as TOPKAT. Agreement criteria is further discussed in **section 2.1.5**.

2.1.2 Descriptive Results

Source cohort

Figures 2.2 and 2.3 show the patient flow chart for this study. There were 878,105 records of TKR or PKR in NJR database. Of these, 9,320 were duplicates for all variables and were discarded. 868,785 unique observations remained. From those, a total of 553,567 records had been linked to HES. This large drop in numbers is inherent to HES, which only covers England data, as opposed to NJR, covering the whole UK. Patients on HES linked to multiple NJR IDs, NJR patients linked to multiple HES IDs, inconsistent dates matching and opt out patients were removed. Up to this point, the data management was done by someone else, and I was presented with a unique merged database, as stated in the **disclosure/footnote.** At this point there were 514,435 patients who undertook TKR and 39,132 patients who had an PKR surgery in the linked database according to NJR records.

I removed duplicate surgeries, those done in the same patient, same knee and same day, totalling 2,099. This study focused on the first incident knee replacement. This led to exclusion of a further 5,141 patients who had a bilateral knee replacement surgery, and, for 74,601 patients who had surgery in both knees on different dates, I eliminated the second surgery. Patients whose surgeries were carried out after

2016, and therefore were not present on the PROMS database were also excluded (n=3,696). Patients that were recorded as dead before the operation date were also removed (n=432). In addition, a total of 5,402 patients were excluded due to patellofemoral or lateral knee replacements. I excluded 4,567 patients without IMD data, due to the impossibility to impute this variable. Finally, I excluded 52 people who had inconsistent age, defined as an age difference greater than 3 years between HES and NJR.

After the application of these exclusion criteria, a total of 425,284 and 32,293 patients were eligible for further analyses. This cohort will be the *source cohort* for further analyses in **Chapters 4 and 5**.

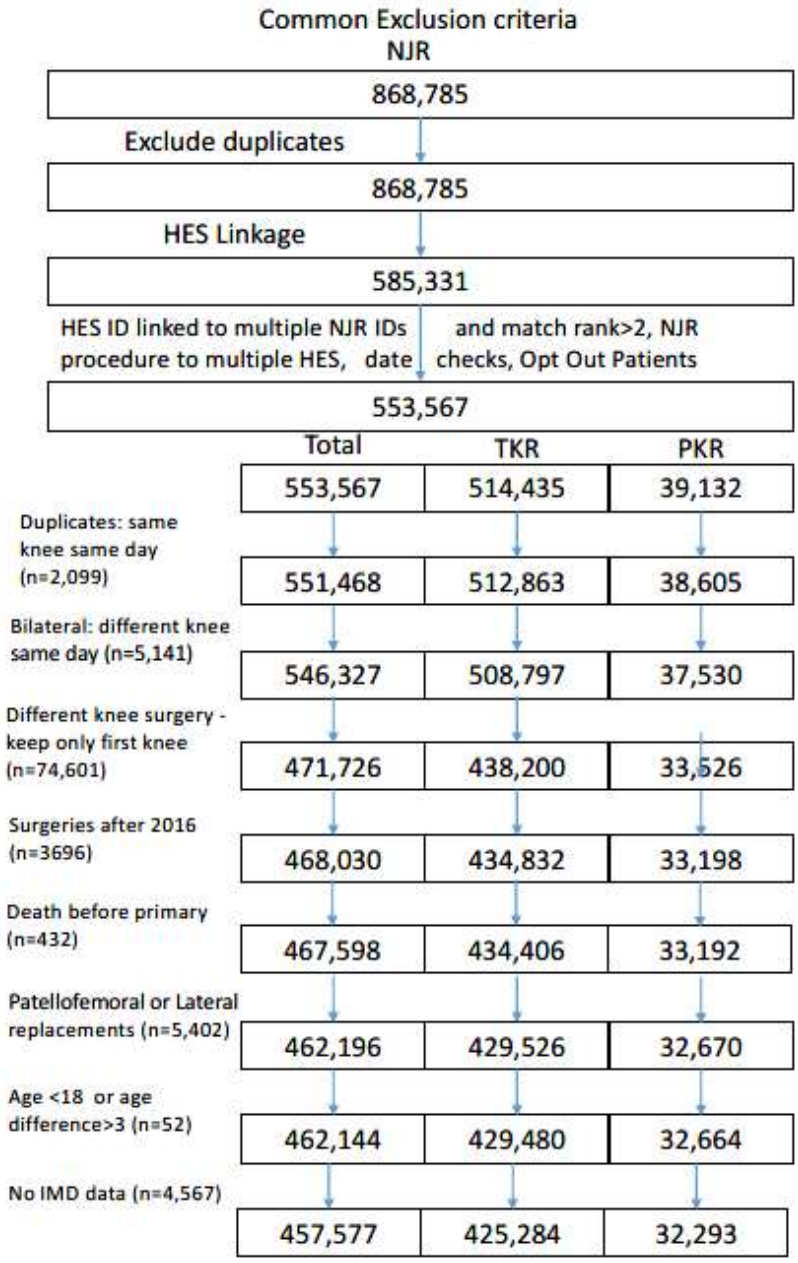


Figure 2.2. Main Patient Flowchart. Data cleaning and inclusion criteria for the whole project.

OKS cohort

In **this chapter and chapter 4**, I applied additional eligibility criteria to try to mimic the TOPKAT RCT. These resulted in the following exclusions: 77,074 patients (72,183 TKR and 2,891 PKR) who had a pre-operative ASA score greater than 2. These patients will constitute the population for the analyses in **Section 5.1.3 .** 79,571 TKR and 8,376 PKR patients, were excluded due to the TOPKAT clinical eligibility criteria transported to this study in ***Table 2.1***. The remaining patients, a total of 273,530 TKR and 21,026 PKR recipients were included for the safety analyses shown in **Chapter 4** (*whole cohort*). Of these, 1,197 PKR and 125,834 TKR patients had a recorded post-operative OKS in PROMs database and could be included for analysis in this chapter (*OKS cohort*). This is shown in ***Figure 2.3.***

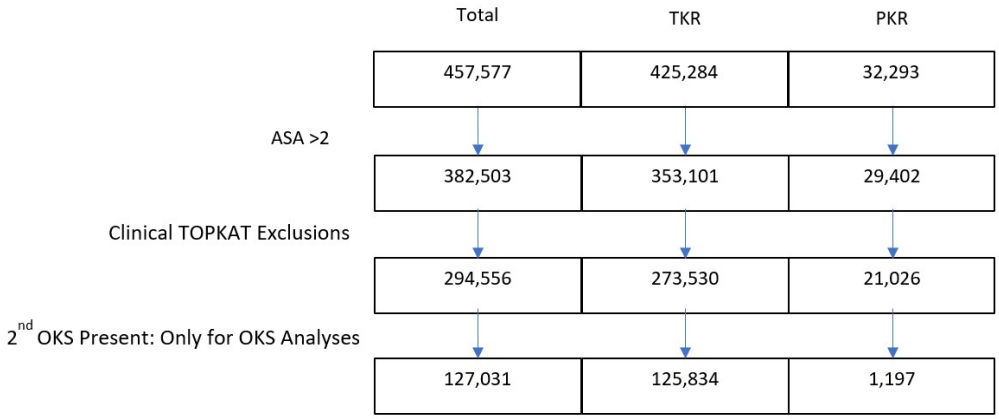


Figure 2.3. Patients eligible for stage 1. Application of ASA 2 and TOPKAT exclusion criteria.

Patient characteristics

Table 2.2 shows the patient-level characteristics in the OKS cohort. As shown in ***Appendix Table 2.2***, patient characteristics in the OKS cohort and in the whole cohort were generally comparable. However, PKR patients in the OKS cohort had less comorbidity than those in the revision cohort.

Several differences between patients receiving TKR and those patients receiving PKR are worth mentioning. Patients receiving TKR were mostly females (56%), while patients receiving PKR were predominantly males (52%). Those undergoing PKR tended to be younger than those having TKR (mean (SD): 64.9(9.4) vs 70.4(8.6) years old) and had a lower preoperative ASA score (20% of PKR were P1 – Fit and Healthy vs 11% of TKR). They also had a lower Charlson index score (69% had a score of 0 vs 76% in TKR). The comorbidities in which they differed the most were other joint problems (12% in PKR vs 19% in TKR), cardiovascular disease (43% in PKR vs 58% in TKR), and gastrointestinal disease history (15% in PKR vs 20% in TKR). Finally, there was a clear difference between the pre-operative OKS, where it was almost 2 points lower for TKR (OKS mean (SD): 19.7(7.6) in TKR vs 21.9(7.5) in PKR).

N(%) or mean (SD)	OKS cohort			
	TKR (N=125,834)	%	PKR (n=1,197)	%
Gender				
F	70671	56	576	48
M	55163	44	621	52
Rural Index				
Urban	92052	73	844	71
Town and fringe	15730	13	164	14
Village	12637	10	138	12
Isolated	5415	4	51	4
IMD				
Least deprived 10%	14168	11	149	12
Less deprived 10-20%	15194	12	137	11
Less deprived 20-30%	15435	12	142	12
Less deprived 30-40%	15405	12	138	12
Less deprived 40-50%	14611	12	164	14
More deprived 10-20%	8628	7	102	9
More deprived 20-30%	10110	8	84	7
More deprived 30-40%	11621	9	123	10
More deprived 40-50%	13557	11	106	9
Most deprived 10%	7105	6	52	4
ASA				
P1 - Fit and healthy	13849	11	242	20
P2 - Mild disease not incapacitating	111985	89	955	80
Charlson Comorbidity				
0	86474	69	915	76
1	26733	21	224	19
2	8357	7	41	3
3+	4270	5	17	1
Age	70.4	8.6	64.9	9.4
BMI	30.4	5.0	29.6	4.7
PROMS pre-operative OKS	19.7*	7.6*	21.9*	7.5*
EQ-5D general health scale	70.0*	19.2*	71.1*	19.0*
EQ-5D Index				
Excellent	88778	71	604	50

N(%) or mean (SD)	OKS cohort			
	TKR (N=125,834)	%	PKR (n=1,197)	%
1	1433	1	33	3
2	10398	8	181	15
3	17504	14	271	23
4	6886	5	94	8
Poor	835	1	14	1
Gastrointestinal Disease	25142	20	174	15
Other Joint Problems	23578	19	149	12
Mental Health	11421	9	101	8
Respiratory Diseases	17078	14	147	12
Cardiovascular Diseases	73382	58	515	43
Thyroid Problems	9742	8	80	7
Foot, hip, spinal pain	1519	1	15	1
Coxarthrosis	4395	3	25	2
Neurological Disorders	7491	6	67	6
Other Arthrosis	5930	5	41	3
Polyarthrosis	7520	6	29	2
Spondylosis	3501	3	17	1

Table 2.2: Baseline patient-level characteristics for patients who received TKR and PKR surgeries in the OKS cohort.

Surgeon characteristics

Half of the patients receiving TKR were operated by surgeons who had done 50 or more TKR surgeries in the previous year, and less than 10% of the surgeries were performed by surgeons that had done less than 10 TKR in the previous year. Conversely, for patients receiving PKR the median number of surgeries of the same type done by the same surgeon in the previous year was 9. Only 48.9% (582/1197) PKR patients had received a surgery performed by a high-volume surgeon. I will explore this further in **Section 5.1.2**

Discussion

After comparing patients receiving TKR and PKR, there were important differences between the two treatment groups. Patients receiving PKR were on average 5.5 years younger, and consistently healthier, with lower ASA score, Charlson index and prevalence of comorbidities. This is in line with the belief of some surgeons that this surgery is more suitable to younger and healthier patients. (96) However, PKR patients in the OKS cohort had less comorbidity than those in the revision cohort (***Appendix Table 2.2*** shows both). This could have been produced by a healthy volunteer effect, where healthier patients would be more likely to engage more with the PROMS programme. (97) These results suggest indication bias, pointing to the need for methods to control confounding. The methods I used for this are explained in the following sections. The differences on surgical volume could also impact the results of the analyses, and, in addition to the sensitivity analyses I have explored this in **section 2.1.4** and **section 5.1.2.**

2.1.3 Instrumental Variables

Introduction

Contrary to PS based methods, which are prone to residual confounding as they can account only for measured confounders, (98) IV estimation, under certain assumptions, can account for both observed and unobserved variables.(99) To perform these methods we rely on the existence of an 'instrument', a variable that: a) we have measured or observed that predicts the exposure; b) is only related to the outcome through the exposure; and c) is independent of all confounders, known and unknown. *Figure 2.4* shows a directed acyclic graph exemplifying it. This approach mimics randomised clinical trials more closely. One could think of random treatment allocation as an instrument: it is a strong predictor of the exposure or treatment, it does not have an effect on the outcome other than through treatment (if the RCT is double-blinded), and it is independent of confounders (in this case through being randomly assigned). If we are interested in Average Treatment Effects (ATE), it should also fulfil a fourth condition: effect homogeneity, where the effect of the treatment is constant among individuals, or where there are not unmeasured effect modifiers. An alternative of this condition is monotonicity, where the number of defiers (individuals who would receive the

opposite treatment to the one they have been assigned independently of the assigned treatment) is negligible.(77) In this section I will be applying Instrumental Variables analyses to UTMOST effectiveness outcomes.

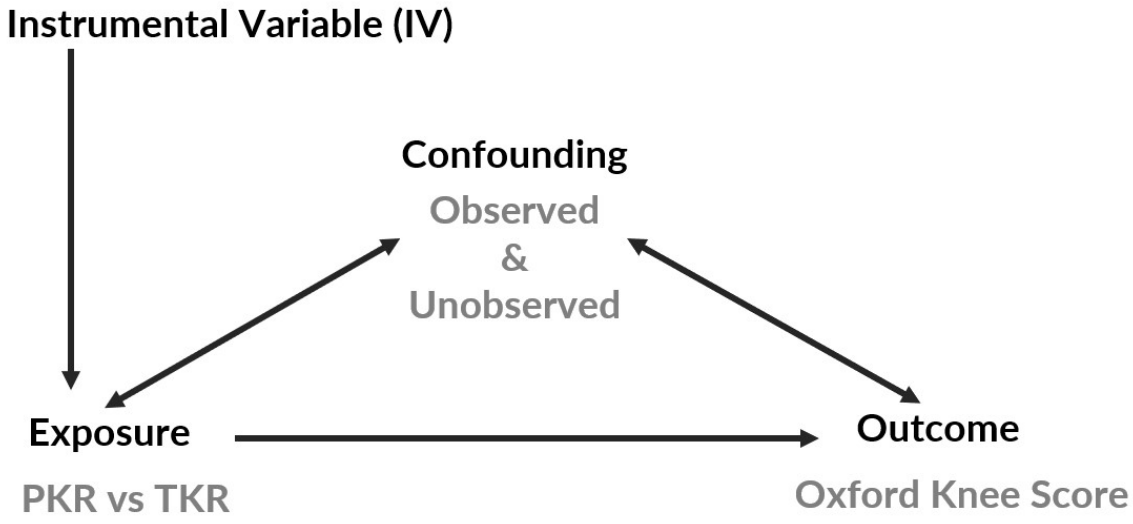


Figure 2.4. Directed Acyclic Graph of Instrumental Variables

Methods

For this study, a range of instruments were proposed, trying to mimic the ones usually proposed and used for pharmacoepidemiologic studies. These were constructed and then tested against the underlying IV assumptions:

- Preference-based instruments:

    - Surgeon preference for PKR

- surgical PKR volume

- and hospital PKR volume

- Geographical location

- Calendar time, as in date of surgery

IV construction

Lead surgeon preference for PKR was calculated as follows: data of eligible surgeries was sorted in an increasing order of date of operation, and surgeon preference was calculated based on the last 20/30/50 consecutive procedures as n of PKR / total procedures. That is, for each patient, I observed the surgeons' previous (20 or 30 or 50) knee replacement surgeries and calculated their preference as the proportion of those that were PKR. This proportion of PKR was then used as an instrumental variable applied at the patient level to account for changes in preference over time. I used the same approach for **consultant surgeon** and **surgical unit** preferences.

Surgeon experience and **hospital volume** were estimated based on the number of knee replacement procedures undertaken by/in each of the surgeons/centres identified in the NJR in the previous year and in the whole study period.

For **geographical location**, two instrumental variables were constructed. Patient region of residence and region of treatment prevalence of PKR were used to construct the instrument. Regional instruments have been previously used for the evaluation of surgical techniques using observational data elsewhere.(100)

Finally, **calendar time** was constructed based on the recorded date of surgery. I determined secular trends in PKR surgery in the NJR data and evaluated if there was an inflexion point when PKR uptake took off, to use it as an instrument: 0 for surgeries before that moment, 1 for surgeries afterwards. This method has been used in pharmaco-epidemiology in situations where uptake of a medication changes after launch or marketing/production discontinuation of a given drug or drug class.(101)

IV assumptions and diagnostics

IV relies on three strong assumptions (102, 103), as shown in ***Figure 2.4*** 1) there is a strong association between IV and the exposures; 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. For this study I will not evaluate the extra assumption of effect homogeneity, as the proportion of defiers (patients who would get PKR if assigned to TKR and TKR if assigned to PKR) is supposed to be

low, fulfilling monotonicity. The first assumption can be tested by the F-statistic value from the first-stage linear regression. A rule of thumb to hold the assumption is when the F-statistic value is greater than 10 or an odds ratio greater than 2 (104, 105). The two latter assumptions are unverifiable or not directly testable as they involve unobservable variables(103), but circumstantial evidence suggests that these assumptions are met.(106) Particularly with regards to assumption 3, i.e. IV is independent of confounders, region of residence, date of surgery, surgeon, surgeon preference, and hospital allocation are often used as they are distributed randomly and thus independent of confounders.(106) Some of these may not hold to careful examination of the assumptions, as region, date of surgery, and hospital allocation depend on other potential confounders such as deprivation or surgical volume. A formal falsification test based on the standardized difference was used to test this assumption, using a cut-off point of 10% for the standardized difference in means or proportions of confounders between IV groups (107). Any instruments that violate this assumption would be considered not valid and therefore excluded from the IV analyses.

IV outcome model

I used a linear two-stage least squares model, with the instrumental variable binarised at median. (108) In this method, two regressions are performed. In a first stage, a model estimates the effect of the IV of interest on the treatment (PKR or TKR). On the second stage, the predicted treatment from model one is used to compare the outcome (OKS) between predicted PKR and predicted TKR.

Software and packages

Al analyses were performed using STATA 15 (109). For the calculation of preference IVs I used the tssmooth command. Balance was calculated using pbalchk. Instruental variable regression was performed with ivregress with a 2sls estimator and xtivreg with a re estimator.

Results

Eligible Patients

The population for primary analyses consisted of 127,031 patients (125,834 TKR and 1,197 PKR recipients) as previously seen in ***Figure 2.2***. A total of 294,556 patients (273,530 TKR and 21,026 PKR recipients) were used to construct the IV in this section and safety outcomes for **Chapter 3**. Baseline characteristics for these cohorts are presented in **section 2.1.2 and 3.1.2**.

In the process of creating IVs, additional patients had to be excluded. For instance, surgeon-based preference for PKR was estimated based on previous ten surgeries, and therefore naturally excluded the initial ten patients in the dataset ***(Table 2.3)*.**

Surgeon ID	Patient ID	Date of surgery	Treatment	Preference for PKR	Binary IV
12345	1	Jan 2010	TKR	n/a	n/a
12345	2	Jan 2010	PKR	n/a	n/a
12345	3	Feb 2010	TKR	n/a	n/a
12345	4	Feb 2010	TKR	n/a	n/a
12345	…				
12345	20	Mar 2010	TKR	n/a	n/a
12345	21	Mar 2010	TKR	0.20	High
12345	22	Apr 2010	TKR	0.20	High
12345	23	Apr 2010	TKR	0.10	Low
12345	…				
12345	56	Nov 2010	PKR	0.11	Low
12345	57	Nov 2010	TKR	0.14	High
12345	58	Jan 2011	TKR	0.15	High

Table 2.3. Example on the construction of preference-based instrumental variables: preference for PKR in the previous eligible 5 surgeries. All data in the table is fake and not

true patient data. A median preference of 0.12 in the dataset has been assumed for binarization of the instrument in this example.

This process resulted in the exclusion of 20, 30, and 50 patients per surgeon respectively from the construction of three surgeon preference instruments. The higher the number of surgeries required for the estimation, the higher the number of excluded participants. For the example, estimating lead surgeon-based preferences based on 20, 25 and 50 surgeries led to exclusion of 17,857 (161 of which PKR), 25,141 (232 PKR), and 39,243 (383 PKR) patients, respectively. As expected, over 99% of these participants were receiving TKR. Specific numbers of exclusions across IVs are presented later in this **Chapter 3** and in C**hapter 4**.

IV construction

The following IVs were developed and tested against the IV assumptions: Surgeon preference for PKR; hospital and region-based preference; volume, area, and calendar time-based instruments.

Surgeon preference for PKR

The importance of surgeon preference for PKR is clear and aligns with the growing use of physician prescription preference as an IV in drug safety research. The NJR presented different categories of surgeons: Lead surgeon, Consultant Surgeon, and

Surgical Unit; and I calculated preference IVs for the three categories. The following steps were followed:

1. Sorting of patients by surgeon (Lead, Consultant or Surgical Unit) pseudonymised identifiers as provided by NJR and NHS digital
2. Sorting of patients by date of operation within each surgeon id/cluster
3. Exclusion of the X (20, 30 or 50) first surgeries performed by each surgeon, following the numbers described above
4. Computing of patient-level preference for PKR, based on proportion of patients within the previous X (20, 30, or 50) surgeries who had received a PKR
5. Estimation of surgeon preference, then binarised as high vs low preference (for PKR) using the instrument-specific median preference as a cut-off

This method accommodates time-varying preference, as the preference is estimated on a set number of previous patients operated by the same surgeon, instead of all past data available. ***Table 2.4*** shows descriptive values for surgeon preference. The mean preference for PKR seems to be around 6% for the whole cohort. For PKR patients, however, the preference of their surgeons seems much higher, with a mean of around 20%.

Total								
Instrument		**mean**	**sd**	**p50**	**p10**	**p25**	**p75**	**p90**
Lead surgeon	**Last 20**	5.9%	11.4%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	6.1%	11.0%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	6.2%	10.8%	0.0%	0.0%	0.0%	8.0%	22.0%
Consultant surgeon	**Last 20**	5.8%	11.1%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	5.9%	10.7%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	6.0%	10.4%	0.0%	0.0%	0.0%	8.0%	20.0%
Surgical Unit	**Last 20**	6.5%	8.8%	5.0%	0.0%	0.0%	10.0%	15.0%
	Last 30	6.6%	8.2%	3.3%	0.0%	0.0%	10.0%	16.7%
	Last 50	6.5%	7.7%	4.0%	0.0%	2.0%	10.0%	16.0%
TKR								
Instrument		**mean**	**sd**	**p50**	**p10**	**p25**	**p75**	**p90**
Lead surgeon	**Last 20**	5.8%	11.2%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	5.9%	10.8%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	6.1%	10.6%	0.0%	0.0%	0.0%	8.0%	20.0%
Consultant surgeon	**Last 20**	5.7%	10.8%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	5.8%	10.5%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	5.9%	10.3%	0.0%	0.0%	0.0%	8.0%	20.0%
Surgical Unit	**Last 20**	6.5%	8.7%	5.0%	0.0%	0.0%	10.0%	15.0%
	Last 30	6.5%	8.1%	3.3%	0.0%	0.0%	10.0%	16.7%
	Last 50	6.5%	7.6%	4.0%	0.0%	2.0%	10.0%	16.0%
PKR								
Instrument		**mean**	**sd**	**p50**	**p10**	**p25**	**p75**	**p90**
Lead surgeon	**Last 20**	21.6%	18.7%	15.0%	0.0%	5.0%	30.0%	50.0%
	Last 30	20.9%	16.5%	16.7%	3.3%	6.7%	30.0%	46.7%
	Last 50	20.7%	15.6%	18.0%	2.0%	8.0%	30.0%	42.0%
Consultant surgeon	**Last 20**	21.3%	19.9%	15.0%	0.0%	5.0%	30.0%	45.0%
	Last 30	21.0%	18.4%	16.7%	3.3%	6.7%	30.0%	46.7%
	Last 50	19.5%	16.3%	16.0%	2.0%	8.0%	28.0%	42.0%
Surgical Unit	**Last 20**	12.4%	12.7%	10.0%	0.0%	5.0%	20.0%	30.0%
	Last 30	12.2%	11.8%	10.0%	0.0%	3.3%	16.7%	26.7%
	Last 50	12.2%	11.4%	8.0%	2.0%	4.0%	16.0%	26.0%

Table 2.4. Mean, SD, median, and percentile 10, 25, 75 and 90 of preference-based instruments (proportion of previous N surgeries performed by the same surgeon that were PKR).

Other preference-based Instrumental Variables

Following a method analogous to the one used to estimate surgeon preference, hospital and region preference for PKR were also built and tested for use. These values were computed based on the first 20/30/50 surgeries performed in each hospital and given region, respectively.

Volume-based Instrumental Variables

Volume-based variables were estimated in two different ways: (1) Total number of surgeries, both PKR and TKR, done by the same surgeon (Lead Surgeon, Consultant Surgeon or Surgical Unit) during the whole cohort study period, and (2) Total number of surgeries done by the same surgeon in the previous year. These variables were then binarised into high vs. low volume surgeons, based on a median split.

Area-based Instrumental Variables

Both area of treatment and area of residence were computed and tested for use. High versus low intake of PKR was calculated based on the median prevalence of use of PKR across regions, based on full cohort data. For both treatment and residence, the government office region was used, as listed in the HES dataset.

Calendar Time

I verified whether there were meaningful changes over time in the use/uptake of PKR in our analytical dataset, which would justify using calendar time as an IV. There was an increase in PKR use from 2011 but it was minimal, from an average of 6.64% before 2011 to 7.33% afterwards, as displayed in ***Figure 2.5***. Therefore, even though calendar time has been useful in a number of drug safety studies (e.g., covering a period where conditions of use change substantially) the lack of secular trends in our database precluded any further efforts to build or test a calendar time-based IV.

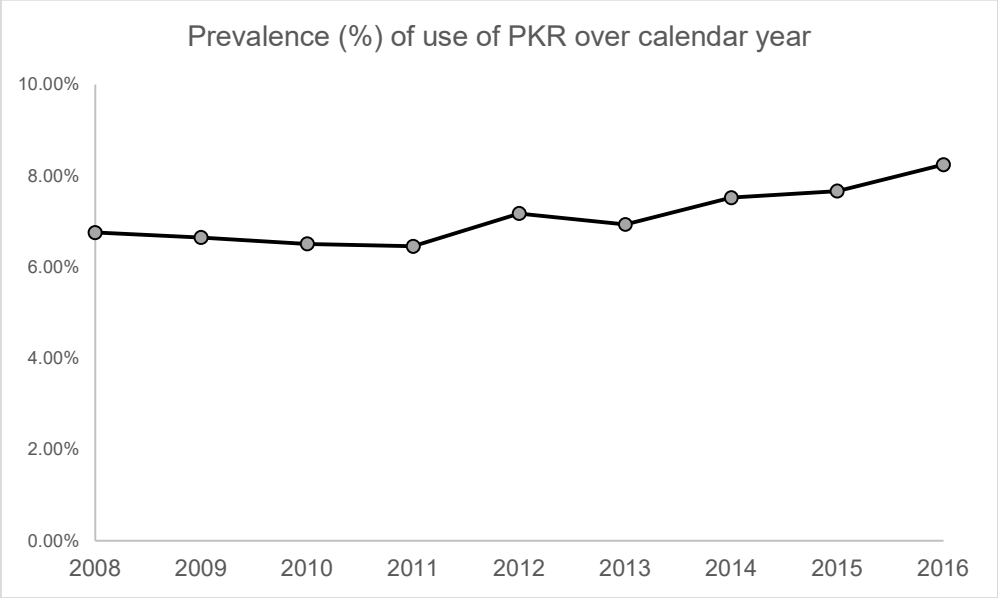


Figure 2.5 – Secular trends in the use of PKR vs TKR in the analytical dataset

Instrumental Variable selection

After creating the proposed IVs, I created an IV "shortlist" for the final analyses, based on instrument diagnostics. These diagnostics were in concert with the first and third IV assumptions outlined in **Section 2.1.3**:

1st: The IV must be strongly associated with the exposure of interest

3rd: The IV must be independent of confounders.

Of the many approaches previously used to characterise the strength of the association between an IV and an exposure of interest, perhaps the most intuitive is the use of the F-statistics and the Odds Ratio (OR) obtained from a logistic regression model (with instrument as an independent variable and outcome as dependent variable). Following the evidence from previous simulation studies (110), I used OR of 2.0 as the cut-off to decide whether to shortlist each of the estimated instruments and made sure F-statistic was greater than 10.

Although the third assumption is mostly untestable, as it involves unobserved confounders, one can test it for known/recorded confounders. If there is a violation in known confounders this would already dismiss the use of that potential IV. As a falsification test for this assumption, Absolute Standardised Mean Difference (ASMD) in the known confounders between high and low preference groups were

estimated. A threshold of ≤0.10 ASMD was used to select instruments that result in balanced groups — that is, independent of confounders, following Ali MS et al.(111) The presence of imbalance greater than 0.10 ASMD for any of the known confounders meant that the IVs were not shortlisted for the final analysis.

Results from these falsification tests for each of these instruments, including summary diagnostic estimates and the decision whether to shortlist for analyses, are reported in ***Table 2.5*****.**

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

Table 2.5 Shortlisting of Instrumental Variables, % of excluded patients, Odds Ratio of the association with PKR, F-statistic, Maximum ASMD in the falsification tests and shortlisting status

Preference-based Instrumental Variables

Surgeon-based preference

As previously discussed, surgeon-level preferences were estimated based on the 20, 30, and 50 previous surgeries. This was estimated for the three surgeon categories: Lead, Consultant and Surgical Unit.

All the estimated surgeon-based preference instruments were strongly associated with the exposure as shown in ***Table 2.5***. ORs ranged from 3.82 [95% CI 3.71-3.93] – surgical unit preference based on 20 previous surgeries, to 29.54 [2.50-31.80] – lead surgeon preference based on the previous 30.

With regards to independence of confounders, unacceptable imbalance (SMD>0.1) for at least one known confounder was identified in most of the surgeon-based preference instruments. Socio-economic deprivation was the most commonly imbalanced confounder. Balance on baseline characteristics for each of the pre-specified confounders stratified by binary instrument status are reported in detail in the following tables **(*Tables 2.6 to 2.8*).**

Covariate	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.033	0.027	0.026
Age at primary surgery	0.037	0.041	0.042
Body mass index	0.012	0.019	0.017
IMD socio-economic status	0.097	0.089	0.083
Pre-operative 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/spine 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 injection/s	0.015	0.002	0.001

IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference

Table 2.6. Covariate balance for a selected list of confounders, stratified by binary lead surgeon preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Covariate	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.031	0.029	0.026
Age at primary surgery	0.012	0.019	0.021
Body mass index	0.012	0.018	0.010
IMD socio-economic status	**0.108**	0.098	0.091
Pre-operative 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/spine 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 injection/s	0.015	0.006	0.016

IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference

Table 2.7. Covariate balance for a selected list of confounders, stratified by binary consultant surgeon preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Covariate	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.016	0.030	0.032
Age at primary surgery	0.038	0.038	0.048
Body mass index	0.000	0.011	0.010
IMD socio-economic status	**0.136**	**0.114**	**0.126**
Pre-operative 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
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/spine 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 injection/s	0.033	0.031	0.025
IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference			

Table 2.8. Covariate balance for a selected list of confounders, stratified by binary surgical unit preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Other preference-based Instrumental Variables

Hospital and region preference for PKR were not shortlisted as they were not strong enough and had imbalance for socio-economic status.

Volume-based Instrumental Variables

Similar to preference, surgeon volume was estimated for lead, consultant surgeon and surgical unit. None of these instruments proved strong enough for further analyses as based on the pre-specified threshold of OR>2.0. Indeed, ORs for these instruments ranged from 0.73 [0.65-0.82] for surgical unit based on the previous year, to OR 1.20 [1.07-1.35] for lead surgeon overall experience.

In terms of confounding, all of these instruments resulted in acceptable imbalances, with all SMDs well below the pre-specified threshold. The highest SMDs seen ranged from 0.059 for lead surgeon experience to 0.092 for surgical unit total experience.

Given the violation of the first assumption, none of the volume-based instruments were taken forward for further analyses of the effect of PKR on OKS.

Area-based Instrumental Variables

Two area-based instruments were estimated, one based on the patient's area of residence, and another one based on the hospital/treatment centre where the knee replacement operation took place. Both did not reach enough instrument strength, with OR 1.37 [1.22-1.53] for the former and 1.67 [1.48-1.88] for the latter. In addition, neither of both area-based instruments reduced confounding for known variables, with maximum imbalances of SMD 0.16 for pre-operative OKS and 0.14 for IMD (socio-economic status).

Therefore, none of the area-based IVs were selected for further analyses.

Calendar Time

As explained above, no calendar time could be identified for use as instrumental variable, as no strong changes in secular trends of PKR uptake were identified in the study period.

Shortlisted IVs

After applying the pre-specified criteria, a total of five instrumental variables were taken forward for analysis:

- Lead surgeon preference for PKR:

- Based on the previous 20 surgeries

- Based on the previous 30 surgeries

- Based on the previous 50 surgeries

- Consultant surgeon preference for PKR:

- Based on the previous 30 surgeries

- Based on the previous 50 surgeries

The outcome results obtained for each of these five instruments are detailed in the following section.

IV results

Two-stage regression results for the association between PKR and post-operative OKS (primary outcome) for the five selected instruments and compared to TOPKAT are shown in ***Figure 2.6***.

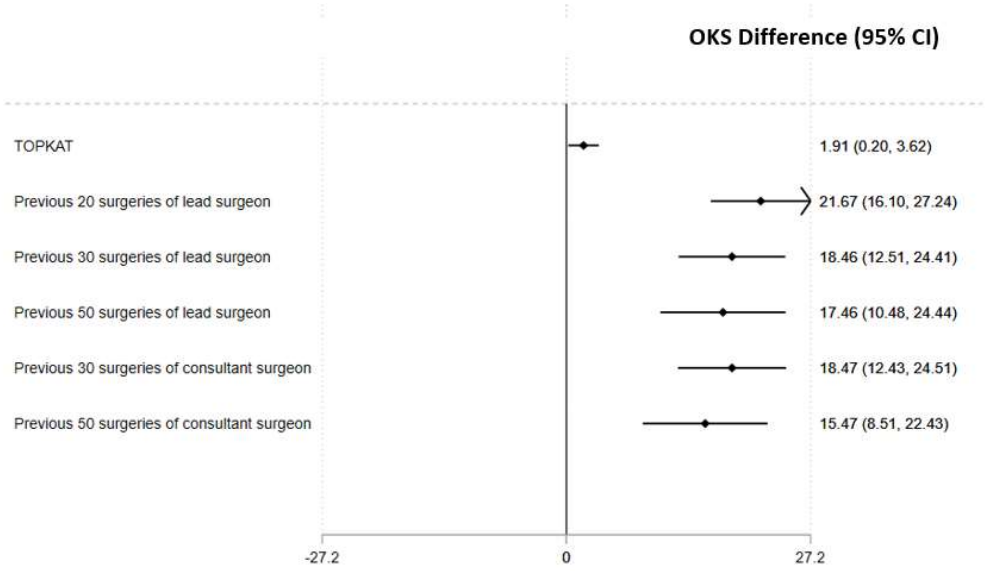


Figure 2.6 – Association between PKR (compared to TKR) and post-operative OKS in the TOPKAT trial (top) and in the 5 shortlisted instrumental variable analyses

Discussion

Overall, this method had several shortcomings that need to be considered carefully. Five out of a total of 17 potential instruments tested passed the diagnostic tests and were eligible for two-stage regression analyses. Of the other

12, four failed due to residual confounding (one or more variables with an imbalance, indicated by ASMD >0.1), 6 failed due to instrument weakness (OR <2 or F-stat <10 in the association between instrument and PKR exposure), and the two area-based instruments failed both diagnostics.

Surgeon preference-based instruments that failed due to unresolved confounding did so because of an imbalance in socio-economic status. This suggests that surgeon preference for PKR is associated with socio-economic status, maybe through geography. Previous studies have reported on the heterogeneity in use of PKR nationally, and on the determinants of surgeon (49) and patient choice. (112) There is evidence that inequality exists in the access (provision vs need) to knee replacement generally. (113) However, no data is available on the potential heterogeneity in access to PKR nationally and/or globally, and on its impact on patient outcomes. A recent study has reported geographical variation in outcomes of primary hip replacement, with surgical volume (both by surgeon and by hospital) associated with better patient outcomes and PKR (vs TKR) associated with a reduced risk of complications. (114)

All the volume-based as well as the regional/area-based instruments tested failed to reach enough strength in their association with the use of PKR. Area-based

instruments also resulted in residual imbalances in terms of pre-operative OKS and socio-economic status. I was not able to use calendar time as an IV since there was no sudden temporal change in PKR uptake.

None of the tested instrumental variables yield results comparable to those obtained from the TOPKAT trial. The reasons underlying this are unclear but could include the violation of untested assumptions (e.g. a direct association between surgeon preference for PKR and the postoperative OKS), the presence of residual confounding for unobserved variables, or the non-normal distribution of the primary outcome of interest (OKS). Methodological research is needed to try to understand the performance and the possible sources of bias in the use of IVs, which can lead to these implausible results. I explore these further in a simulation study presented in **section 2.2**.

2.1.4 Propensity Score methods

Introduction

Propensity Score methods are used to summarise covariate information and minimise differences between treatment groups of patients (36). We can think of PS as being the probability of getting a specific treatment of interest by each patient. It is usually calculated using a logistic regression with a set of available potential confounders. In this section I will be applying Propensity Score analyses to UTMOST effectiveness outcomes.

Methods

Variables included in the propensity score model

In this case, variables that can act as confounders were selected based on clinical expertise and previous literature. The 18 variables used are presented in ***Table 2.9.*** I extracted the variables sex and socio-economic status (Index of Multiple Deprivation [IMD] and rural index) from HES. Charlson Comorbidity Index, and the rest of the co-morbidities were retrieved from HES data with a 3-year lookback period. Data on pre-operative PROMS and EQ5D were extracted from the PROMs database.

Covariate	Data source	Description
Sociodemographic and clinical factors		
Age	NJR	Age at operation
Gender	NJR	Gender
Rural Urban	HES	The official statistic of the classification of rural and urban area: urban; town and fringe; village; isolated
IMD	HES	Index of multiple deprivation. Patients' deprivation status in percentile.
BMI	NJR	Calculated from height and weight
PROMS pre-operative OKS	PROMS	Self-reported pre-operative OKS score, ranging from 0 to 44.
PROMS EQ-5D	PROMS	Self-reported pre-operative EQ-5D VAS, ranging from 0 to 100.
PROMS General health	PROMS	Self-reported pre-operative general health, ranging from 0 (excellent) to 5 (poor).
Charlson comorbidity	HES	The Charlson comorbidity number recorded in HES (the code list is shown in Appendix Table 2.3) 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 three years before the operation.
Osteoarthritis and other joint problems	HES	An ICD-10 code for other joint problems in HES in the three years before the operation (code list is shown in Appendix Table 2.3).
Mental health	HES	An ICD-10 code starting with "H" (mental health) in HES in the three 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 three years before the operation.
Cardiovascular disease	HES	An ICD-10 code starting with "I" (cardiovascular disease) in HES in the three years before the operation.
Thyroid problems	HES	An ICD-10 code starting with "E0" (thyroid problems) in HES in the three years before the operation.
Foot, hip, spinal pain	HES	An ICD-10 code for foot, hip, or spinal pain problems in HES in the three years before the operation (code list shown in Appendix Table 2.3).
Coxarthrosis	HES	An ICD-10 code starting with "M16" (hip osteoarthritis) in HES in the three 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 three years before the operation.
Other arthrosis	HES	An ICD-10 code starting with "M19" (other arthrosis) in HES in the three years before the operation.
Polyarthrosis	HES	An ICD-10 code starting with "M15" (polyarthrosis) in HES in the three years before the operation.
Spondylosis	HES	An ICD-10 code starting with "M47" (spondylosis) in HES in the three years before the operation.

Table 2.9. A description of patient-level covariates included in the PS model

Missing data

I imputed missing data on BMI, EQ5D general health scale and preoperative PROMS using multiple imputation with chained equation (MICE) and 10 imputations. For BMI and EQ5D general health scale the model used for imputation was linear, for PROMS, the model was Poisson distributed. I used the rest of the variables used in the PS model and the outcome for imputation. I evaluated the results of the imputation process to make sure that extreme or non-plausible values were generated.

Propensity Score Adjustment

PS adjustment is one of the most common ways of using propensity scores for observational research in surgical literature.(82) This method uses the predicted values as a covariate in the outcome model, in addition to the treatment. (82, 115) It constitutes the most straightforward approach, but it assumes that the association between PS and the outcome is linear and that no interaction exists between PS, exposure, and outcome. To account for some of this potential of misspecification of the true relation between the PS and the outcome I explored both linear as well as non-linear adjustment using fractional polynomial regression.(116) This technique assesses the best fitting polynomial form of the propensity score using likelihood

ratio tests between the model with and without the interaction terms. (117) Although this method can estimate the average treatment effect on the treated (ATT) and the average treatment effect on the exposed (ATE), increases the difficulty of interpretation, as a non-linear relationship between PS and the outcome does not have a straightforward interpretation.(118) No covariate imbalance check was required in this method.

Propensity Score Matching

In surgical literature, the most common way to use PS is matching. (82) This consists of using the PS generated to match patients who received PKR to those who received TKR with the same PS, that is, similar chance of getting an PKR. In my case I used a maximum calliper width of 0.02 standard deviations (SD). (119) By matching on similar PS, PKR and TKR groups should be on average comparable regarding the measured confounders, and therefore exchangeable. This concept tries to mimic an RCT where, through randomisation, we achieve similar characteristics between treatment and control arms. Therefore, any differences in outcome between PS exposed (PKR) and unexposed (TKR) matched patients is equivalent to the treatment effect.

This method has been proven to be effective at minimising confounding by indication in pharmaco-epidemiological (drug safety and comparative effectiveness) studies (120). However, it does not estimate ATE. As PS matching searches for patients in PKR and TKR based on their probability of PKR, those with extremely high of PKR or TKR may not be matched to a patient with similar probabilities in the other treatment group. Data from these patients are not used for the analyses. This usually drops patients who would never have received an PKR, and lets us estimate ATT. (121) For this method I assessed covariate balance, a diagnostic method explained in following sections.

Propensity Score Stratification

The third approach I used was PS stratification. This method uses all participants in the dataset, and ranks them based on their estimated PS. Then, they are separated into equal groups (e.g., tenths, fifths, etc…), or strata. Within each stratum, both exposed and unexposed patients have similar PS, such that a similar distribution of confounders between treatment groups. This allows for calculating treatment effects separately for each stratum, on patients with similar confounder distribution. These effects, as many as strata created, are then pooled by averaging with a weight of the strata proportions. In this study, I averaged treatment with a

weight of strata proportions and estimated standard errors using a Jacknife approach.(122, 123) PS stratification done this way estimates the effect of treatment in the target population, ATE, as stratum-specific treatment effects are weighted by the proportion of total subject within that stratum.(124) This PS approach has some advantages compared to the previous presented methods. First, PS stratification overcomes the pitfall of PS adjustment of being sensitive to potential non-linear associations between PS and the outcome, which could render the whole analysis invalid. Secondly, unlike PS matching, stratification does not need to exclude any patients, resulting in not reducing sample size, which leads to gains in precision.

Simulation studies have shown that stratifying into a higher number of strata results in less chance of bias. But, in some cases where one of the studied treatments is rare, having a large number of strata means that the ones on the tails will be dominated by a given treatment.(125) A solution to this conundrum is to stratify based only on the infrequent exposure population. This leads to better bias reduction in simulations,(126) and by weighting the estimates by the stratum size it still produces ATE. For this study, I tested separately PS stratification in 10 strata based on the distribution of PS in the whole dataset (PS tenths) and based on the

distribution of PS amongst PKR recipients. For this method, I assessed PS distribution in each stratum and covariate valance, as further explained in the diagnostics section.

Propensity Score Weighting

Lastly, I used Inverse Probability Weighting (IPW). For this, a weight is calculated for each patient: the inverse of the probability of the treatment they had received. In this case, for patients receiving PKR their weight would be 1/PS, and for those receiving TKR 1/(1-PS). (121, 125, 127) This gives more importance to those patients who received a treatment that they had low chances to be prescribed. This generates a pseudo-population where the distribution of confounders is similar between treated and untreated by upweighting those with rare traits in their treatment group.

This method is very similar to what is usually done in surveys to achieve representative samples of a given population, by giving weights to certain groups, and so survey sampling weights methods are used in the estimation of treatment effects. IPW usually estimates ATE as the previously discussed PS stratification and PS adjustment.

This method is not without limitations, one of the most relevant in this study being the presence of rare exposures. The low frequency of an exposure or patients with very low probability of being treated can lead to enormous weights than can have an exaggerated effect on the treatment estimates. To resolve this caveat, I used weight stabilisation, which use the marginal probability of the treatment of interest instead of 1 in the weight numerator. (125, 127) This also scales the sample size back to the original, where the classic weights double the sample size. As diagnostics, I assessed the distribution of weights and the covariate imbalance.

Unadjusted and fully adjusted regression

To compare PS matching, stratification and IPW with other commonly used strategies of analysis I also included two non-PS strategies. The first one, the outcome model without any covariate adjustment, looking at the "crude" OKS difference between TKR and PKR. The second strategy consists in a fully adjusted model, where I used all the variables selected to create the propensity as regressors in the outcome model.

Propensity score diagnostics

In terms of diagnostics, I assessed the PS distributions visually to be sure that treated and untreated PS distributions overlap. For propensity score stratification, I also compared the distribution of PS between TKR and PKR per strata using box plots to check whether they were similar, an important assumption of this method, and the absolute numbers and prevalence of PKR in each stratum. For IPW, I looked at the range of stabilised weights.

For matching, stratification and IPW I assessed covariate imbalance, that is, if the method was able to balance the prevalence or mean value of the identified confounders between PKR and TKR. This is done using a cut-off of 0.1 in Absolute Standardised Mean Differences (ASMD). When ASMD was higher than the cut-off, for a given confounder, that variable was included as a covariate and adjusted for in the outcome model.(119) I checked balance in each imputed dataset for PS matching and IPW and in each imputed dataset and stratum for PS stratification methods.

Outcome model

All models used linear regression for the estimation of differences in post-operative OKS including a lead surgeon cluster level in the regressions to account for non-independence of patients of a same surgeon. This is equivalent to the

analytical strategy used in the TOPKAT RCT. To account for the existence of 10 imputations, I ran the analyses separately for each imputed dataset and pooled the estimators and their errors using Rubin Rules. (128)

Software and packages

I used STATA 15 (109) for all analyses and Microsoft Excel for pooling imputations results. In addition to the base packages, I used psmatch2, propwt, and pbalchk. Matching and stabilised IPW were computed with psmatch2 and propwt, and added to the mixed effects regression (meglm) as sampling weights (pweights). The PS stratification was coded by me and applied using survey weights with the svy command. I explored non-linear terms for the continuous variables on the PS and on the PS adjusted results using Fractional polynomial regression (fp command).

Results

Imputation results

BMI had 33,062 missing values, pre-operative OKS 1,224 and EQ5D general health scale 12,785. BMI had values ranging 10 to 60, pre-operative OKS ranged 0 to 48 and EQ5D general health scale 0 to 100. Some implausible values (7 out of a million

values) were generated for BMI, and there were truncated at the previous max and minimums. After imputing missing data using MICE, BMI, pre-operative OKS and EQ5D general health scale had a similar mean and standard deviation than the original dataset *Table 2.10*. There were no differences in medians and IQR.

	Pre-imputation		Post-imputation	
	Mean	**SD**	**Mean**	**SD**
BMI	30.40	5.04	30.42	5.05
Preoperative OKS	19.70	7.58	19.70	7.56
EQ5D general health scale	70.05	19.22	69.97	19.23

Table 2.10. Pre and post imputation mean and SD values of imputed variables

Propensity Score calculation

For the PS calculation, the logistic regression predicting treatment yield the coefficients shown in ***Appendix Table 2.4*** for the first imputed dataset with very similar estimates for the rest. The strongest predictors of treatment were basal general health, and foot hip, or spinal pain. There were differences between PKR and TKR patients in terms of predicted PKR PS, especially in small PS values. ***Figure 2.7*** shows these PS distributions. There was enough overlap to perform PS techniques.

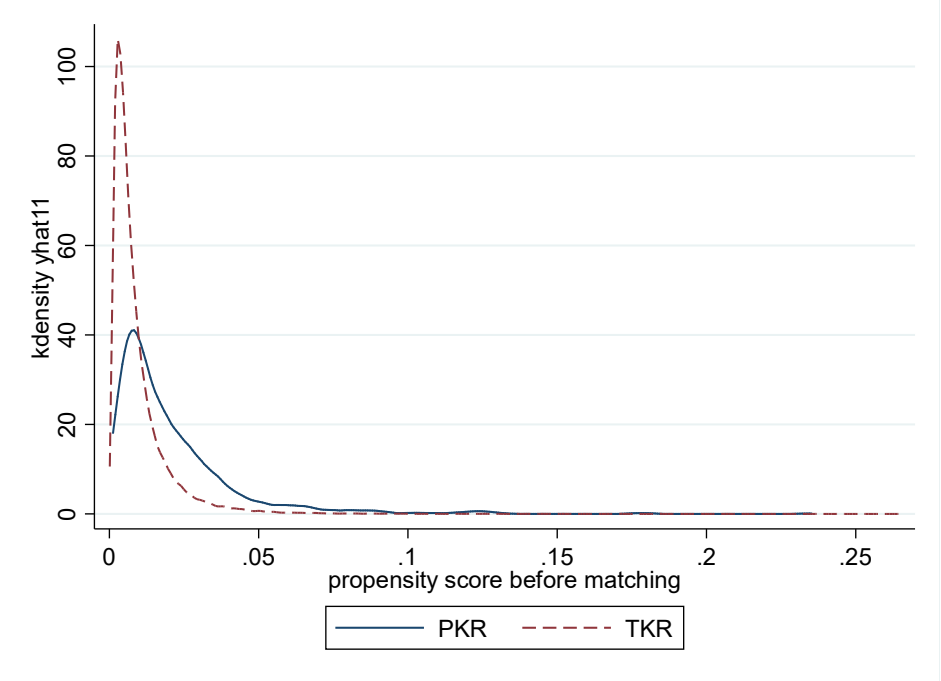


Figure 2.7 – Propensity score distribution in the OKS cohort. 1st Imputation

Propensity Score Diagnostics

In the following section I discuss covariate balance assessments for propensity score matching, propensity score stratification, and inverse probability weighting methods and other specific metrics particular to each one.

Matching

I matched 1,197 PKR patients to 5,652 TKR patients based in PS for the effectiveness analyses. This analysis excluded 120,182 TKR patients, which were not matched to a PKR patient. After matching, differences in PS distributions disappeared ***Figure 2.8***.

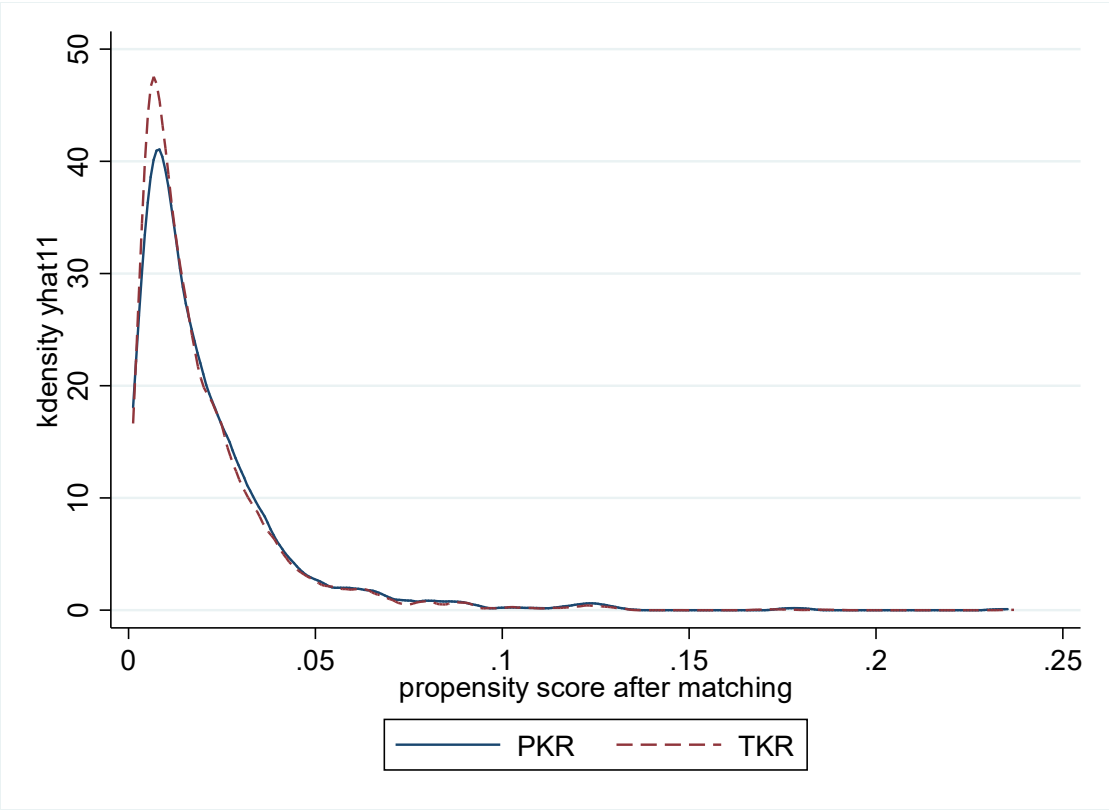
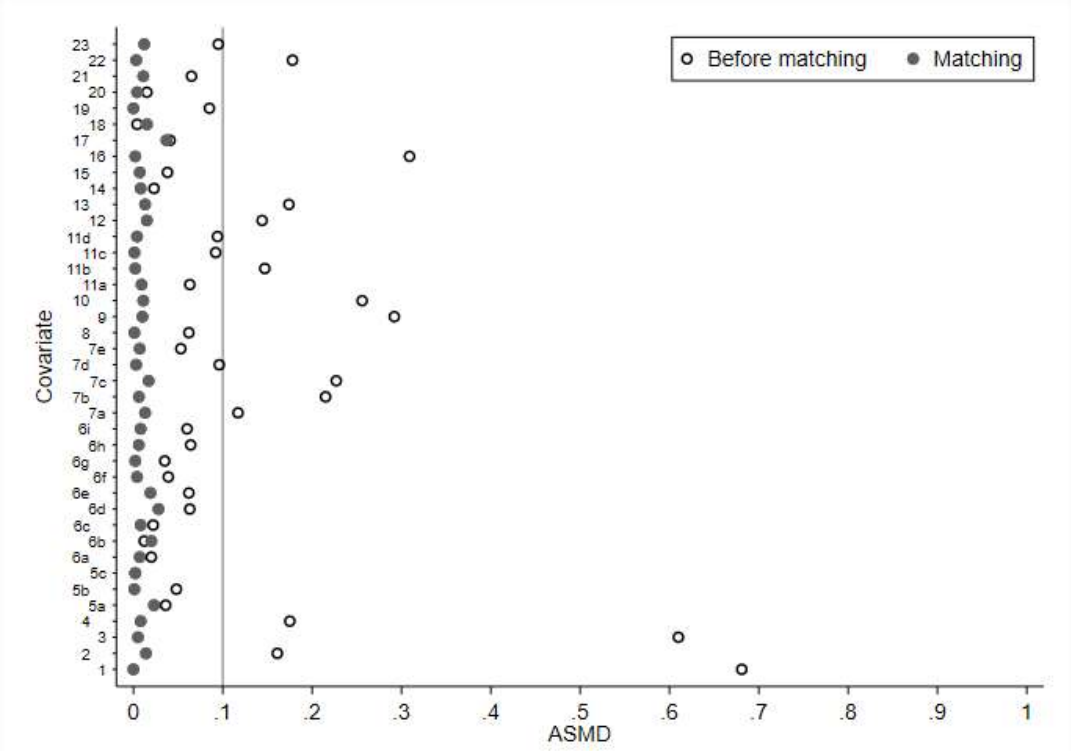


Figure 2.8 – Propensity score distribution in the OKS cohort after matching. 1st Imputation.

There were no significant differences in any of the studied baseline characteristics between PKR and TKR after matching, as ASMD was less than 0.1 for all those variables, as shown in ***Figure 2.9***.



Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 2.9 – Absolute standardised mean difference (ASMD) of PS variables in the OKS cohort after matching. 1st Imputation.

Baseline characteristics for the whole OKS cohort (before matching) and the matched cohorts (after PS matching) are detailed in ***Appendix Table 2.5*** for comparison. As matching restricts the sample to patients who could be matched,

there were changes in the characteristics of the included TKR patients, becoming more "comparable" to PKR recipients: healthier, younger and with a higher proportion of men.

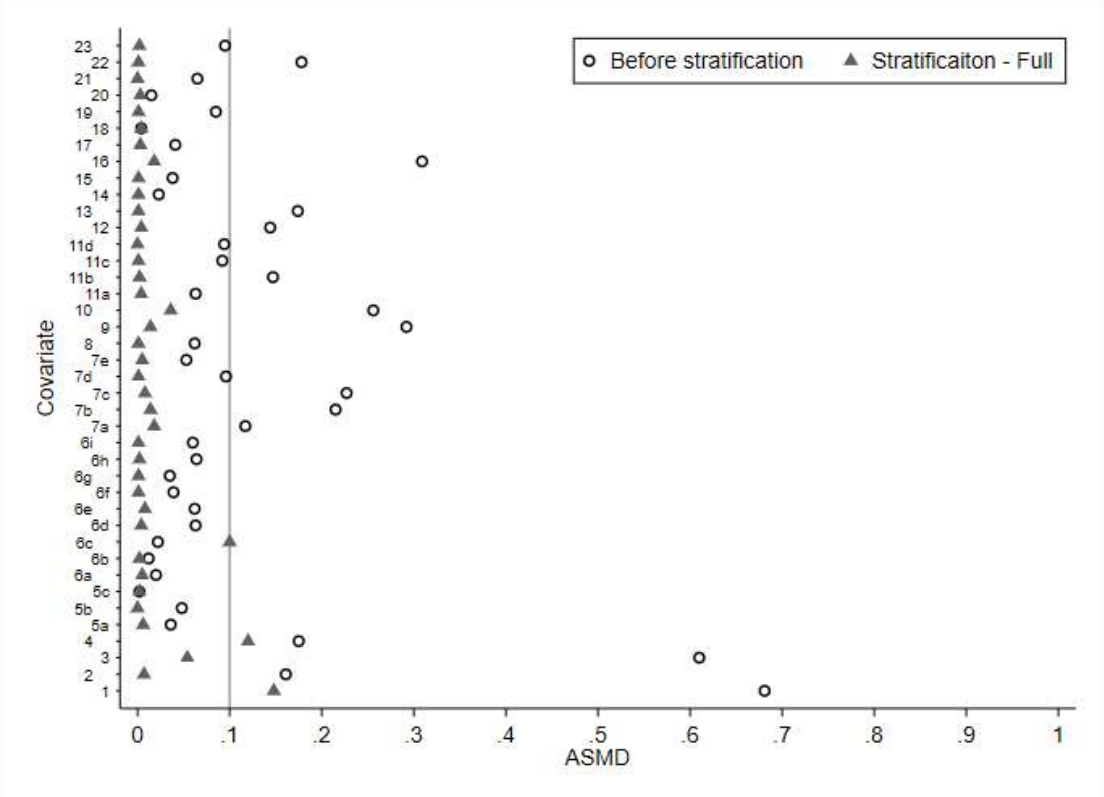
Stratification

Ten strata were created in two ways for the OKS cohort: based on deciles of the distribution of the estimated PS in the whole cohort (abbreviated as PSSwhole) and based on the distribution of the PS in the PKR cohort (abbreviated as PSSexp). These methods did not exclude any patients.

Stratifying by the whole population, PSSwhole, resulted in covariate imbalances in six of the ten strata: stratum 1-6 were dominated by TKR patients and had <1% PKR patients. In contrast, PSSexp ensured equal proportions of PKR and TKR recipients between and within strata and resulted in similar PS distributions between PKR and TKR, ***Supplement Figure 2.1*****),** suggesting a good overall balance.

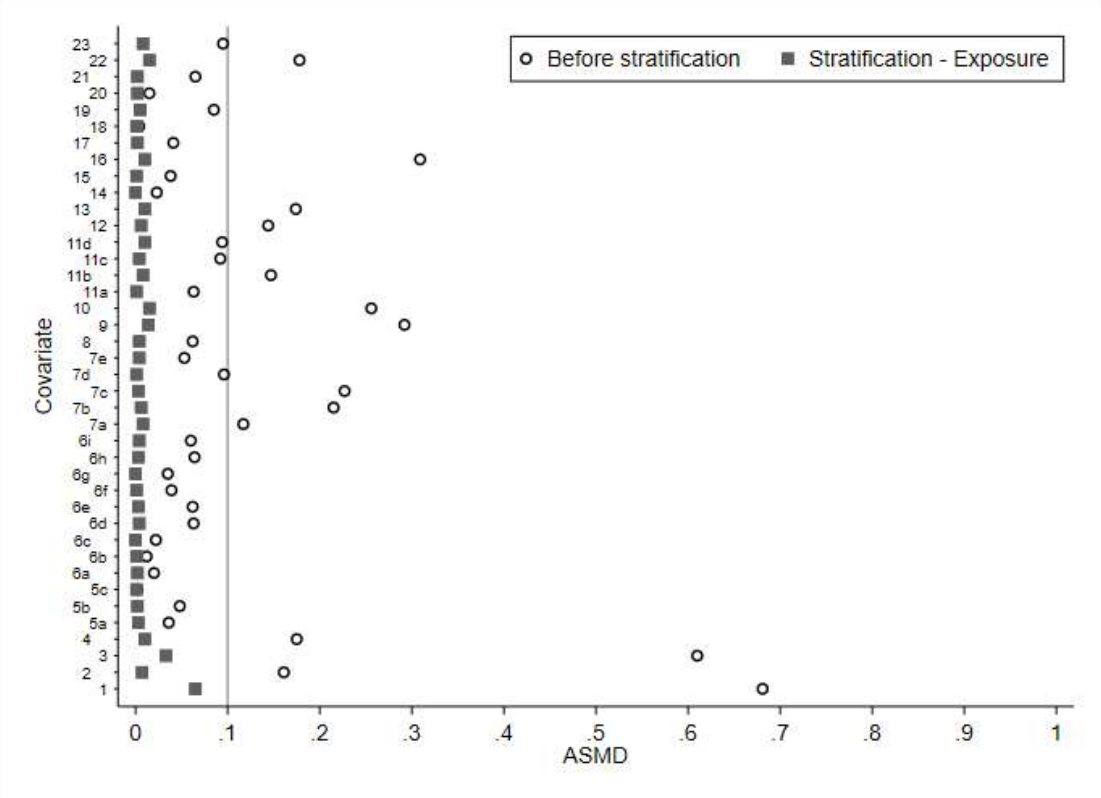
Figure 2.10 shows the mean ASMD for each of the confounders across strata in the OKS cohort using PSSwhole as a measure of covariate balance. There were some imbalances even after stratification: overall PS and BMI remained imbalanced between TKR and PKR patients. Within strata, covariate balance was not always achieved, especially in stratum 1-6, which, as explained, had few PKR patients.

PSSexp stratification provided balance for all covariates with an average ASMD ≤0.1 across strata (***Figure 2.11***). This method also had a good covariate balance in most strata.



Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 2.10 – Propensity score distribution in the OKS cohort before and after stratification by the full cohort. 1st Imputation.



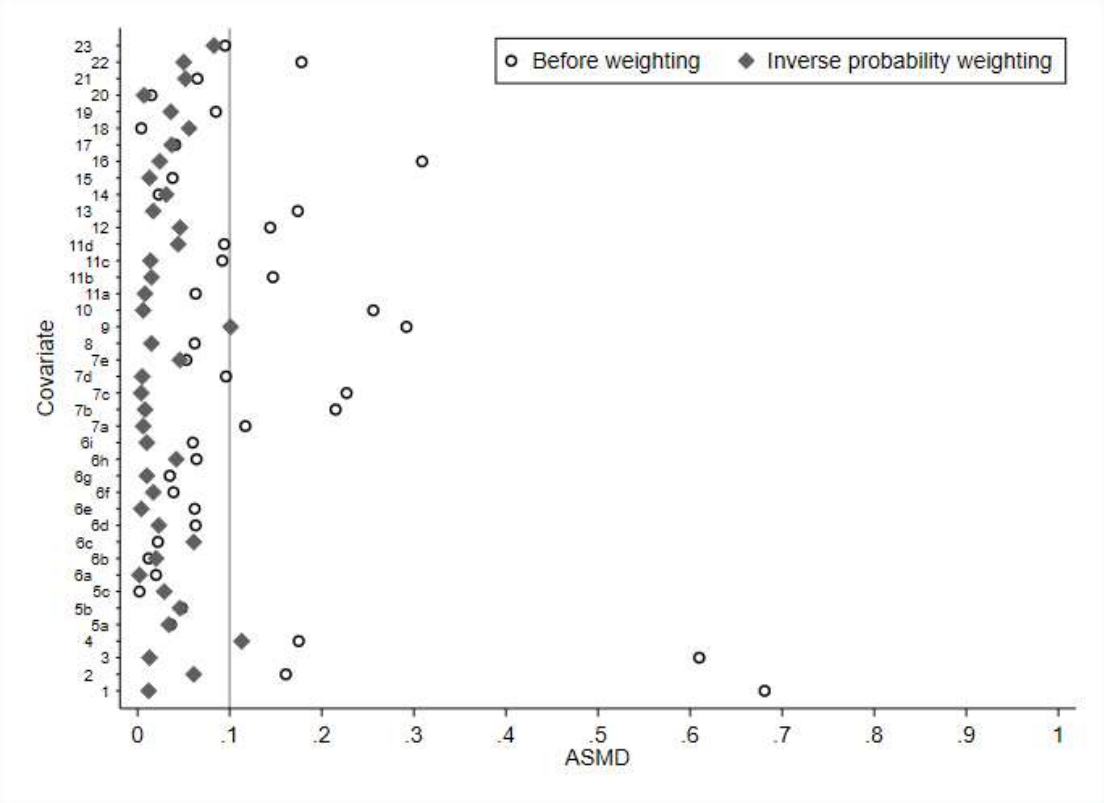
Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 2. 11 – Propensity score distribution in the OKS cohort after stratification by the exposure. 1st Imputation.

IP Weighting

As explained, I created stabilised weights for each patient, which results in a new pseudo-population. For the PKR patients, stabilised weights ranged from 0.04 to 7.90 (IQR: 0.37, 1.30) with a mean of 1, while TKR patients were given stabilised weights ranging from 0.99 to 1.35 (IQR: 0.99, 1.00) with a mean of 1.

As for covariate balance, PKR patients had a similar distribution in all the covariates, except for BMI, with an ASMD just above 0.1 (***Figure 2.12***). This imbalance is of limited clinical relevance: mean BMI for PKR were 29.87 kg/m2 and 30.43 kg/m2 for TKR subjects. We can consider that IPW worked to minimise confounding to an acceptable degree based on the pre-specified threshold (ASMD ≤0.1) with the only exception of BMI, which was adjusted for in the final analyses.



Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 2. 12 – Propensity score distribution in the OKS cohort after IPW. 1st Imputation.

Main Results

First, I calculated pre- and post-operative OKS, crude, and adjusted using the following methods: PSM, IPW, PSSwhole, PSSexp, PSA (both linear and non-linear adjustment) and adjusted by all covariates. For comparison purposes, ***Table 2.11*** presents these results and the same values for the TOPKAT trial. At baseline, the mean OKS for patients receiving TKR and PKR were similar to those reported in TOPKAT in most of the analyses. Crude/unadjusted means of pre-operative OKS for patients receiving TKR and PKR differed by two points at baseline (19.68 [SD: 7.56] for TKR and 21.88 [SD: 7.52] for PKR). TKR participants in this study were similar to those in the TOPKAT trial, where pre-operative mean OKS was 19.0 [SD: 7.2]. However, my PKR participants differed by about 3 points compared to those in TOPKAT (mean pre-operative OKS 18.8 [SD: 7.0]). PSSwhole and PSSexp (and covariate and PS adjustment, both linear and non-linear), resulted in the same baseline OKS as in the crude/unadjusted analysis.

As for the PS matching strategy, mean pre-operative OKS was similar between the TKR and PKR recipients: 21.96 [SD: 7.76] and 21.88 [SD: 7.52] respectively. However, both PKR and TKR groups differed further from participants in the trial, with pre-operative OKS >2 points higher in the PS-matched cohort vs TOPKAT.

The pseudo-population created in IPW also had similar baseline pre-operative OKS for TKR (average 19.70 [SD: 7.20]) compared to PKR (average 20.41 [SD: 7.42]), and closer to the one seen in the TOPKAT trial.

Post-operatively, around 6-12 months after the surgery, there was a large improvement in OKS compared to baseline OKS. All the applied methods found a favourable treatment effect for PKR compared to TKR, although for most the 95% Confidence Interval for the estimate included the null effect (0) in the cases of PSM, IPW, PSSwhole, PSA (linear and nonlinear) and fully conditioning on PS variables (abbreviated as fully adjusted). Only PSSexp found a difference (in terms of statistical significance) between PKR and TKR, with a point estimate (95%CI) of 0.76 (0.15 to 1.36). The TOPKAT estimate, used as gold standard, was 1.91 (0.20 to 3.62).

		TKR mean (SD)	PKR mean (SD)	Mean difference / Effect size (95% CI)
TOPKAT	Pre-op.	19.0 (7.2)	18.80 (7.0)	
	Post-op.	35.1 (10.3)	36.9 (9.9)	1.91 (0.20, 3.62)
Crude	Pre-op.	19.68 (7.56)	21.88(7.52)	-
	Post-op.	35.80 (9.35)	36.74 (9.77)	0.76 (0.22, 1.29)
PSM	Pre-op	21.96 (7.76)	21.88 (7.52)	-
	Post-op	36.71 (9.14)	36.74 (9.77)	0.27 (-0.38, 0.92)
IPW	Pre-op	19.70 (7.57)	20.41 (7.42)	-
	Post-op	35.80 (9.35)	36.64 (9.50)	0.58 (-0.19, 1.35)
PSS_{whole}	Pre-op	19.68 (11.64)	21.88 (7.94)	-
	Post-op	35.80 (11.35)	36.74 (10.13)	0.56 (-0.03, 1.16)
PSS_{exp}	Pre-op	19.68 (13.30)	21.88 (7.77)	
	Post-op	35.80 (12.31)	36.74 (9.87)	**0.76 (0.15, 1.36)**
PS_{Alin}	Pre-op.	19.68 (7.56)	21.88(7.52)	-
	Post-op.	35.80 (9.35)	36.74 (9.77)	0.14 (-0.39, 0.68)
$PS_{Anonlin}$	Pre-op.	19.68 (7.56)	21.88(7.52)	-
	Post-op.	35.80 (9.35)	36.74 (9.77)	0.10 (-0.44, 0.63)
Fully Adjusted	Pre-op.	19.68 (7.56)	21.88(7.52)	-
	Post-op.	35.80 (9.35)	36.74 (9.77)	0.06 (-0.43, 0.55)

IPW: inverse probability weighting; PS: propensity score; PSM: propensity score matching; PSS_{whole}: PS stratification based on the whole cohort, PSS_{exp}: propensity score stratification based on the exposure cohort; PS_{Alin}: Propensity score linear adjustment; $PS_{Anonlin}$: Propensity score non-linear adjustment; SD: standard deviation; TKR: total knee replacement; PKR: partial knee replacement.

Table 2.11. Pre- and post-operative Oxford Knee Score (OKS) in TOPKAT, UTMOST and for each PS method

Sensitivity Analyses

As sensitivity analyses, and to better mimic the conditions in the TOPKAT trial, I restricted the cohort to those patients whose surgeon had done at least 10 surgeries of the same type in the previous year. From the 127,031 patients, this sensitivity analysis included 602 out of 1,197 (50.2%) PKR and 114,871 out of 125,834 (91.3%) TKR (***Table 2.12***). At the surgeon level, this resulted in the inclusion of 2,625 out of 3,895 (67.4%) surgeons that had performed a TKR but only 164 out of 452 (36.3%) lead surgeons that had performed a PKR. This shows how most PKR are performed by low volume surgeons (as defined by TOPKAT inclusion criteria for surgeons). Whether this has an effect on outcomes is further explored in ***Section 5.1.2***. Baseline patient characteristics for participants operated by such high-volume surgeons are reported in ***Appendix Table 2.6.***

	OKS cohort			
	Patients		Surgeons	
	TKR	PKR	TKR	PKR
All	125,834	1,197	3,895	452
≥10 surgeries previous year	114,871 *91.3%*	602 *50.3%*	2,625 *67.4%*	164 *36.3%*

Table 2.12. Number of participants and surgeons according to surgeon volume.

PS matching resulted in a sample of 602 PKR matched to 2,934 TKR and an imbalance of Urban-Rural index of Town and Fringe (0.106). IPW resulted in imbalanced variables: sex (0.102), BMI (0.141), Urban-Rural index of Town and Fringe (0.125), IMD decile n3 (0.102), polyarthritis (0.155) and spondylosis (0.105). Full cohort stratification resulted in imbalance of PS (0.136), and exposed stratification resulted in optimal balance for all variables. As sample size for these analyses was very small, I did not try to condition on unbalanced variables. Non-linear PS analysis did not converge, and therefore it is not shown.

Alter excluding surgeries performed by low volume surgeons, there was an even larger improvement in OKS compared to baseline OKS. The favourable treatment effect for PKR compared to TKR was even higher than in the main analyses and the 95% Confidence Interval for the estimate did not include the null effect (0). The point estimates are shown in ***Table 2.13***.

		TKR mean (SD)	PKR mean (SD)	Mean difference / Effect size (95% CI)
TOPKAT	Pre-op.	19.0 (7.2)	18.80 (7.0)	-
	Post-op.	35.1 (10.3)	36.9 (9.9)	1.91 (0.20, 3.62)
Crude	Pre-op.	19.7 (7.6)	22.1 (7.6)	-
	Post-op.	35.8 (9.3)	37.6 (9.0)	1.36 (0.61, 2.12)
IPW	Pre-op	19.7 (7.6)	20.5 (7.6)	-
	Post-op	35.8 (9.3)	37.5 (8.7)	1.32 (0.32, 2.33)
PSS_{whole}	Pre-op	19.7 (7.6)	22.1 (7.6)	-
	Post-op	35.8 (9.3)	37.6 (9.0)	1.37 (0.54, 2.20)
PSS_{exp}	Pre-op	19.7 (7.6)	22.1 (7.6)	-
	Post-op	35.8 (9.3)	37.6 (9.0)	1.37 (0.54, 2.20)

IPW: inverse probability weighting; PS: propensity score; PSM: propensity score matching; PSS_{whole}: PS stratification based on the whole cohort, PSS_{exp}: propensity score stratification based on the exposure cohort; PS_{Alin}: Propensity score linear adjustment; $PS_{Anonlin}$: Propensity score non-linear adjustment; SD: standard deviation; TKR: total knee replacement; PKR: partial knee replacement.

Table 2.13. Pre- and post-operative Oxford Knee Score (OKS) in TOPKAT, UTMOST and for each PS method for the sensitivity cohort.

Discussion

Propensity Score methods seem able to produce reasonable estimates, closer to those of the trial. These methods yielded a small 1.3-point additional increase in post-operative OKS for PKR compared to TKR. This is consistent with some previous studies including TOPKAT,(70, 129) showing similar effectiveness between PKR and TKR, with an average OKS difference of <3, below the known threshold for clinical significance.(92, 95) I will formally compare them to the trial in the next section. As for covariate balance, matching was able to produce good balance, but reduces sample size and makes the sample closer to the PKR patients, estimating ATT. PS Stratification seems like a promising method for estimating ATE, and gets better balance when stratifying by the exposure when having a treatment much less frequent than the other. IPW also produces effective covariate balance. In sensitivity analyses, restricting sample to higher volume surgeons, there was higher imbalance in all methods, except for PS stratification based on the exposure.

In the diagnostics of Propensity Score methods, I used ASMD to check balance and graphical comparisons of the distribution fo the PS and the IP weights. These diagnostics were prespecified in a protocol. (130) I

didn't examine the balance in higher order moments and interactions or compared variances and distributions between PS groups. This could have led to unresolved confounding, especially if some of the confounders had a quadratic term or an interaction in the true propensity score.(127, 131)

2.1.5 Comparing to TOPKAT

In this section I will be comparing the different Instrumental Variables analyses and Propensity Score UTMOST effectiveness outcome results from the previous sections to those from the TOPKAT trial.

Agreement criteria

There are several ways of assessing if a method or study was able to replicate a trial.(132) For this study, I selected 5 agreement criteria to assess the validity of each method to replicate the results from the TOPKAT RCT, as described below.

Coverage: deemed as "passed" if the treatment effect estimate of the studied methods was inside the 95% CI of the TOPKAT estimate.

Statistical Significance Agreement: to pass this test, the treatment effect estimate must be significant, and that effect must be in the same direction as TOPKAT.

The remaining 3 agreement criteria are based on heterogeneity measures of a random effects metanalysis of the TOPKAT estimates with each method's results:

Chi-square for heterogeneity test: I considered the RCT was replicated if the test yielded a p-value >0.05, that is, that we cannot find heterogeneity

I2 for heterogeneity: I considered the results I^2 below 40% as equivalent to agreement with the trial results.(133)

Tau2 (τ^2): this represents variance between two estimates. There is no clear cut-off variance. I considered a relatively small between method variance (Tau2) as having successfully replicated the trial. (134)

Instrumental Variable Results

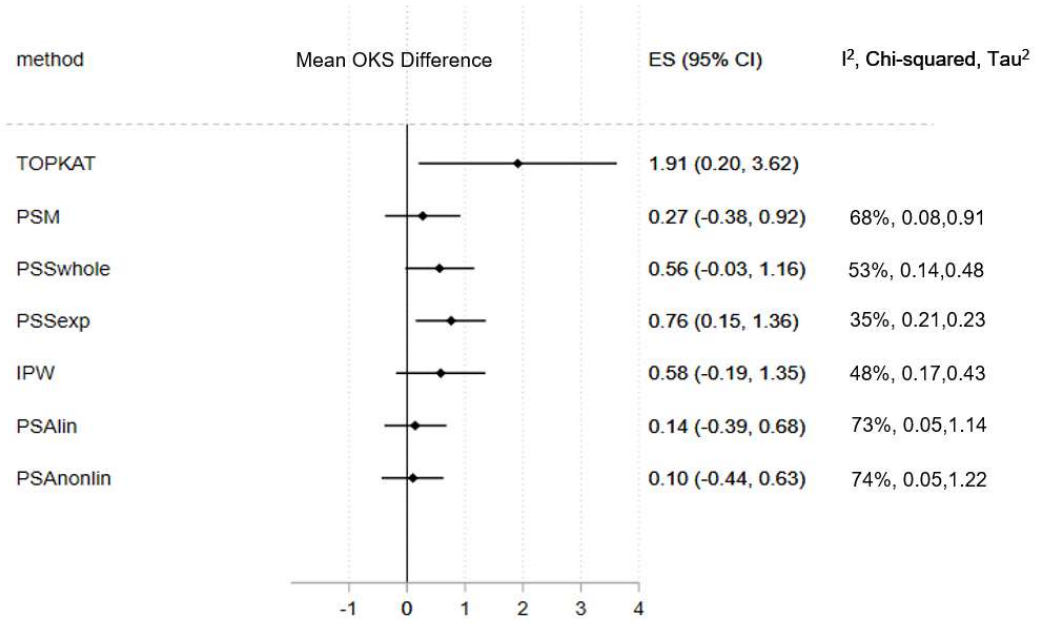
As illustrated in ***Table 2.14***, all the proposed instruments' estimates departed greatly from the results obtained in TOPKAT. None of the estimates, or their confidence intervals, fulfil the first criteria, that is overlap with the main estimate (or with the upper or lower limit of the confidence interval) obtained from TOPKAT. The second criteria were fulfilled for all estimates, as PKR had a positive effect size, and they did not include 0 in their confidence intervals. Results from the random effects metanalysis show how these results are very different 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 20 surgeries); I^2 ranging from 92.7% to 97.7% for the same instruments respectively; and all Q^2 <0.001.

Preference Instrument		OKS Difference (95% CI)	T^2	I^2	Chi^2 test p-value
	TOPKAT	1.91 (0.20, 3.62)			
Lead surgeon	**last 20 surgeries**	21.67 (16.10, 27.24)	190.88	97.70%	<0.001
Lead surgeon	**last 30 surgeries**	18.46 (12.51, 24.41)	132.02	96.40%	<0.001
Lead surgeon	**last 50 surgeries**	17.46 (10.48, 24.44)	114.17	94.40%	<0.001
Consultant surgeon	**last 30 surgeries**	18.47 (12.43, 24.51)	131.94	96.30%	<0.001
Consultant surgeon	**last 50 surgeries**	15.47 (8.51, 22.43)	85.27	92.70%	<0.001

Table 2.14. Consistency of results obtained from instrumental variable analyses compared with TOPKAT findings

Propensity scores

Propensity score methods performed much better than Instrumental Variables in this setting. Although all methods produced point treatment effects estimates slightly lower than TOPKAT, all of them, except PSA, were included within the 95% CI of the TOPKAT trial. As shown in ***Figure 2.13,*** all methods except for PSA adjustment had a chi-square p>0.05. The closest estimates to TOPKAT were achieved by PSSwhole and PSSexp, with the smallest τ^2, 0.23 and 0.48 respectively. PSSexp was the only methods with small heterogeneity, I2 <40%, followed by IPW, which showed moderate heterogeneity I2=48% and $\tau^2 = 0.43$. PSM estimate, 0.27, was very close to the lower CI of TOPKAT estimate of 0.20. It also had high heterogeneity (I2=68% and a $\tau^2 = 0.91$). PSA had even wider heterogeneity with TOPKAT results, suggesting that neither PSM nor PSA methods were capable of replicating the TOPKAT trial findings.



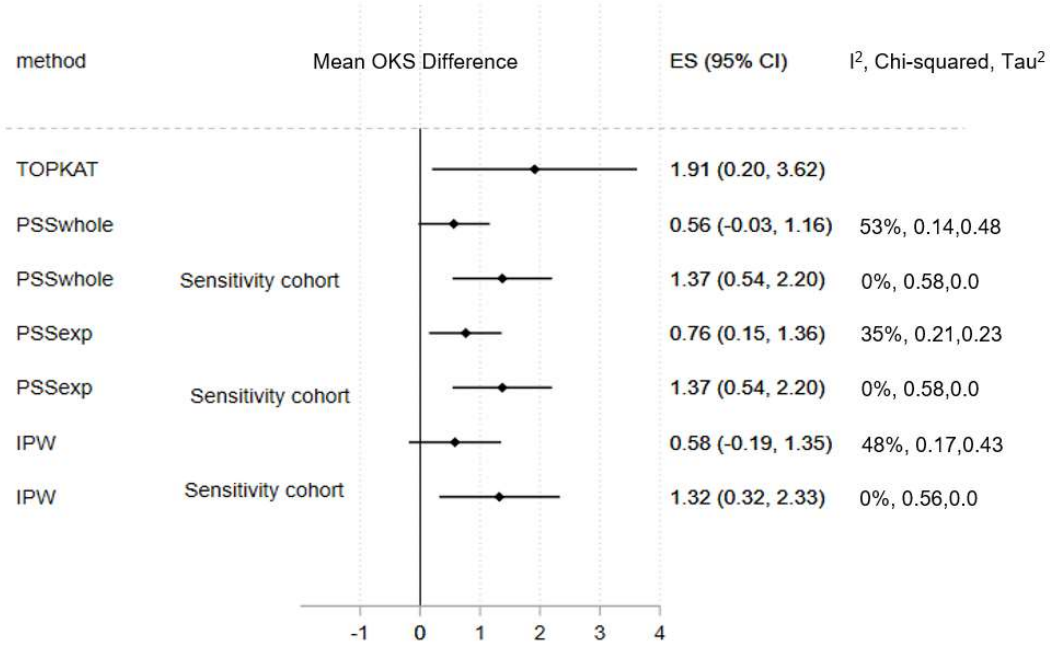
Note: PSM: propensity score matching; IPW: inverse probability weighting; PSSwhole: propensity score stratification based on the whole cohort; PSSexp: propensity score stratification based on the exposure cohort; PSAlin: propensity score linear adjustment; PSAnonlin: propensity score nonlinear adjustment with PS0 and ln(PS)0.

Figure 2.13 – Forest plot of the post-operative Oxford Knee Score (OKS) effect size (ES) for TOPKAT and each of the tested propensity score methods, with heterogeneity measures (I2, Chi-squared, and Tau2).

Sensitivity PS analyses

I used the previously 'validated' methods (IPW, PSSwhole and PSSexp) in the sensitivity cohort, that is patients operated by "high volume surgeons". This cohort was closer to the patients in the TOPKAT trial. The results from this sensitivity analysis compared to those from the main analysis (in the whole OKS cohorts) show how restricting the analysis to surgeons eligible for the trial resulted in

estimates closer to those seen in TOPKAT (***Figure 2.14***). The obtained treatment effect estimates, and their 95%Cis, are fully covered within the 95% CI of the TOPKAT estimate, and heterogeneity drops dramatically, with all three I2 now dropping to 0% and Tau2 of 0.



Note: PSM: propensity score matching; IPW: inverse probability weighting; PSSwhole: propensity score stratification based on the whole cohort; PSSexp: propensity score stratification based on the exposure cohort; PSAlin: propensity score linear adjustment; PSAnonlin: propensity score nonlinear adjustment with PS0 and ln(PS)0.

Figure 2. 14 - Forest plot of the post-operative Oxford Knee Score (OKS) effect size for TOPKAT and each of the validated methods in the whole 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 (I2, Chi-squared, tau2).

Discussion

In this chapter I demonstrated how some of the 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. (40) Unsurprisingly, methods that estimated Average Treatment Effects were the closest to TOPKAT results. (70) PS stratification based on the distribution of the PS in the PKR (exposed) cohort passed all the proposed criteria with PS stratification based on the whole cohort and IPW obtaining borderline results. PS adjustment was an exception to this: not able to validate, probably as it has strong assumptions on the relationship between PS and the outcome.(118)

However, sensitivity analysis restricted to patients operated by surgeons eligible for TOPKAT (based on their previous volume of operations of the same type as the index surgery performed) resulted in findings much closer to those seen in the RCT. In these analyses, all three methods (PS stratification based on the whole cohort, PS stratification based on the PKR cohort, and IPW) were deemed valid. This points to the need to think and apply strict inclusion criteria for surgeons when trying to emulate a trial with routinely collected data. In addition, this demonstrates previously described issues with the 'transportability' of effects seen

in surgical trials to routine practice settings, where less surgeons might obtain different findings to those participating in the aforementioned RCTs. The results seen in the sensitivity show how surgeon volume contributes to the differences in treatment effect observed in the main UTMOST cohort analysis, as well as in TOPKAT. This is fundamental to understand some of the reasons underlying any observed differences between "real world" analyses and surgical RCTs. These results point the need for more research on how to address surgeon/operator confounders on procedure epidemiology, especially when exclusion of low volume operators is not feasible. Following this question, I explore this further later in **Chapter 5** by evaluating different strategies to generate Propensity Scores, including multilevel modelling and incorporating surgeon variables in the PS. This is a fairly new subject and calls for further methodological research into the replicability of trials in the fields of surgery, procedures and MD epidemiology.

2.2 Timing of elective tracheotomy and duration of mechanical ventilation amongst patients admitted to intensive care with severe COVID-19

2.2.1 Introduction

With the irruption of the COVID-19 pandemic, there was a pressing need to rigorously evaluate the use of technologies and procedures, being used on a clinical setting without or with little evidence. This was of absolute certainty for pharmacotherapy, where political influence and bad science drove the use of medicines in the first months of the pandemic. But also, there was a need for procedures for which we had good evidence of efficacy but not much evidence on how timing or other modifications could affect this efficacy. And all this in a context of scarcity of resources and extraordinary pressure on health systems.

In this section I aim to study one of the questions that were crucial in this context regarding tracheotomy. Tracheotomy is the most used procedure for patients in the intensive care unit (ICU). Around 10% to 24% of patients with invasive mechanical ventilation (IMV) require this procedure if they need prolonged respiratory support.(54) The best moment to perform this procedure is still disputed, and there is substantial variation in the type and timing between

practices.(55) Two systematic reviews found that performing an early tracheotomy may reduce the lengths of IMV and ICU care required, but with little certainty for specific subgroups. (135, 136)

Tracheotomy has been crucial for the management of COVID-19, as around 3% of hospitalised COVID-19 patients get into respiratory failure and need IMV. (53)

Tracheotomy is mainly indicated when long-term intubation is required, when there is a need for better management of secretions, when sedation needs to be reduced, in progression to weaning, and to prevent laryngeal oedema. It is thought that tracheotomy reduces length of IMV, and ICU stay on these patients. This lower ICU stay is crucial both for the patient and for the healthcare system, particularly during the COVID-19 pandemic.

At the pandemic start, concerns were raised about the infection risk of the surgical team. The lack of enough personal protective equipment (PPE) or the diversion of surgeons to emergency rooms made some hospitals unable to perform these surgeries early (137). The difficulties and risks of moving patients to the operation room were additional drawbacks. All these factors led scientific societies to issue recommendations regarding how to perform tracheotomy safely,(138) although

they mostly drew on SARS-CoV-1 and MERS experience and on expert opinion and were not consistent on several points (137, 139-142).

Since then, an observational study seemed to point that a tracheotomy is a safe procedure even at the patient's bedside, following appropriate PPE recommendations. (143) In addition, previous studies suggested that early tracheotomy might reduce time to weaning and ICU length of stay. (136, 143)

Objectives

As length of stay has a crucial role both in the patient's recovery and in the proper functioning of a hospital during a pandemic situation, I evaluated the effect of tracheotomy timing (early vs late) in the weaning and mortality rates of COVID-19 ICU patients who require a tracheotomy during IMV via a target trial framework.

2.2.2 Methods

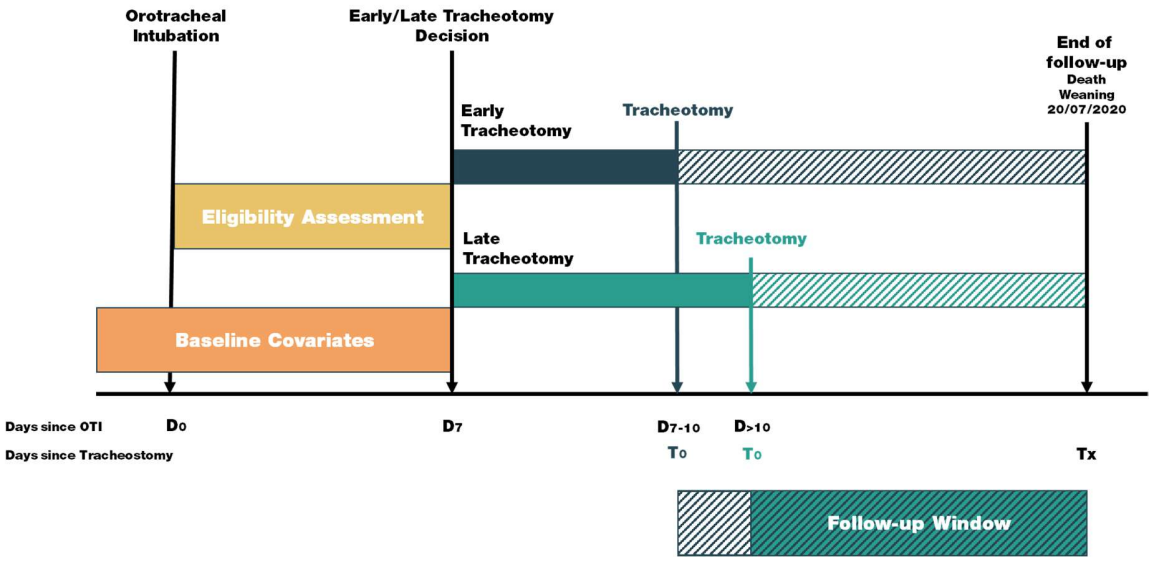
Study design, data sources and population

Prospective cohort. I followed all patients receiving a tracheotomy between 11 March 2020 and 20 July 2020 in 36 hospitals in Spain. The local ethics committee approved the study protocol and waved informed consent given the observational nature of the study. A collection form was sent to each centre at the beginning of the study. Researchers from each hospital collected the data from ICU admission until the date of weaning or death or end of study (July 20th). I included patients requiring IMV and subsequent tracheotomy, suffering respiratory failure caused by a PCR-confirmed SARS CoV-2 infection**.** Patients with missing tracheotomy date, orotracheal intubation date, or outcome date were excluded. I also excluded patients with missing data on age or sex or with a tracheotomy performed in the first 7 days after orotracheal intubation were also excluded. This last exclusion criterion was added to account for setting day 7 as the target randomisation time.

Target Trial framework

To minimise confounding and bias, I used the trial emulation framework. (43) Unfortunately I wasn't able to influence data collection and I could only apply it at the analysis design stage. This consists in trying to mimic as much as possible the

timings and conditions of a randomised experiment. The exposure, or treatment strategy, was late or early tracheotomy. The "randomisation" time, (D7) was 7 days after the initiation of IMV (D0), as this is the date when a decision is made on whether a patient will be having a tracheotomy. As it would be done in a cohort or a trial, I only considered the baseline characteristics (at D7) before or on this date. One of the pitfalls of this this study is that I only have information on participants that actually received a tracheotomy, and not for everyone eligible at D7. This made it impossible to perform an intention-to-treat analysis, so I only performed a per-protocol analysis. Participants were followed-up from the day of the tracheotomy until death, weaning or 20 of July of 2020, whichever was sooner. I censored patients who had not died or weaned by the end of the follow-up. These timings are described in ***Figure 2.15***.



OTI: orotracheal intubation

Figure 2.15 – Target trial description

Outcomes

As the main outcome I used time to weaning, defined as days from tracheotomy to weaning from IMV and weaning in 14 days. Secondary outcomes included death, defined as days from tracheotomy to death and mortality in 14 days, and rates of intraoperative bleeding (excessive bleeding that difficult standard tracheotomy or requiring additional haemostatic measures), postoperative bleeding (bleeding that required revision of stoma) and ventilatory complications (air leak).

Exposure and covariates

My main exposure variable was timing of tracheotomy: early versus late. I defined 'Early' as tracheotomy on day 7 to 10 after orotracheal intubation, and 'late' as on day 11 or later. This definition was agreed with an expert group of clinicians (the TraqueoCOVID Group from the Spanish Society of Otorhinolaryngology) and it is consistent with literature.(135) Sex and year of birth were acquired at hospital admission. I included comorbidities thought to increase risk of severe COVID-19: hypertensive disease, immunosuppression, heart failure, autoimmune disease, chronic obstructive pulmonary disease (COPD), pregnancy, diabetes mellitus, neuromuscular disease, and ischaemic heart disease. Pronation cycles were registered as start and end of pronation. Measures of PaO2/FiO2 ratio (PAFI) and positive end-expiratory pressure (PEEP) at intubation (D0), 7 days after intubation (decision date, D7), and at tracheotomy (T0) were obtained. Clinical scores were only possible to obtain at ICU admission: APACHE II and SOFA scores. Surgery related variables recorded included international normalised ratio (INR), use of anticoagulants, use of vasoactive drugs, presence of secretion problems, and indication at surgery. Blood analytical parameters recorded included total lymphocyte and leukocyte count, INR, D-dimer, ferritin, lactate dehydrogenase,

and C-reactive protein at admission. These variables at baseline are obtained from the electronic medical records and recorded by the researchers.

Statistical analyses

I described the proportion or mean and standard deviation of each variable by exposure, included/excluded status and for the whole sample. I calculated weekly incidence rates (events per 100 person-day) and cumulative incidence on the whole period of weaning and death. These incidences were computed overall and stratified by early versus late tracheotomy. I plotted the incidence and survival functions and explored visually the assumption of proportional hazards.

I used multivariable Cox model to estimate cause-specific hazard ratios (csHRs) of weaning and death for early versus late tracheotomy, by running a separate model for each outcome and censoring if the competing outcome occurred.(144) I also used a mixed effects Weibull regression using hospital of treatment as clusters to account for the differences in practice in each hospital. To take into account the competing risk of death on the risk of weaning, I used Fine-Gray analysis. (145) To analyse the relative risk of intraoperative and postoperative bleeding and ventilatory complications I used Poisson models. A second model was done by using a mixed effects Poisson model with hospital as a random intercept. All

models were repeat-adjusted for age and sex. All models were further adjusted for age, sex, APACHE, SOFA, PAFI, PEEP, and pronation days. These variables were selected in consultation with clinicians to assess what could constitute a confounder.

I imputed missing baseline variables using multiple imputation with chained equations. I used predictive mean matching with 5 k nearest neighbours (146) for continuous variables, and logistic models for binary variables, generating 100 imputed datasets.(147)

As a sensitivity analysis, I tested for interactions between tracheotomy timing and clinically important variables. The tested variables included median age, sex, APACHE II, SOFA, PEEP, PAFI, and days of pronation.

2.2.3 Results

As shown in ***Figure 2.16***, 794 participants were recorded, 98 were excluded. Forty were excluded due to general exclusions, with 27 missing or inconsistent dates of exposure or outcome and 13 missing age. A further 58 patients were excluded due to having a tracheotomy before my set index date, at least 7 days after OTI as these patients are more likely to have different indications for tracheotomy. ***Appendix Table 2.7*** shows the baseline characteristics of these groups of patients. There are little differences in the general exclusions, but the very early tracheotomy group is quite different: Fewer days of pronation (a mean of 1.8 in this group compared to 5.8 in the included patients), more chance of the tracheotomy indication was secretions management (17% vs 11%), a higher proportion of ischaemic cardiopathy (26% vs 11%), and a much higher SOFA score (8.4 vs 6.3). This confirms that these patients potentially have different indications and are in a much more critical state.

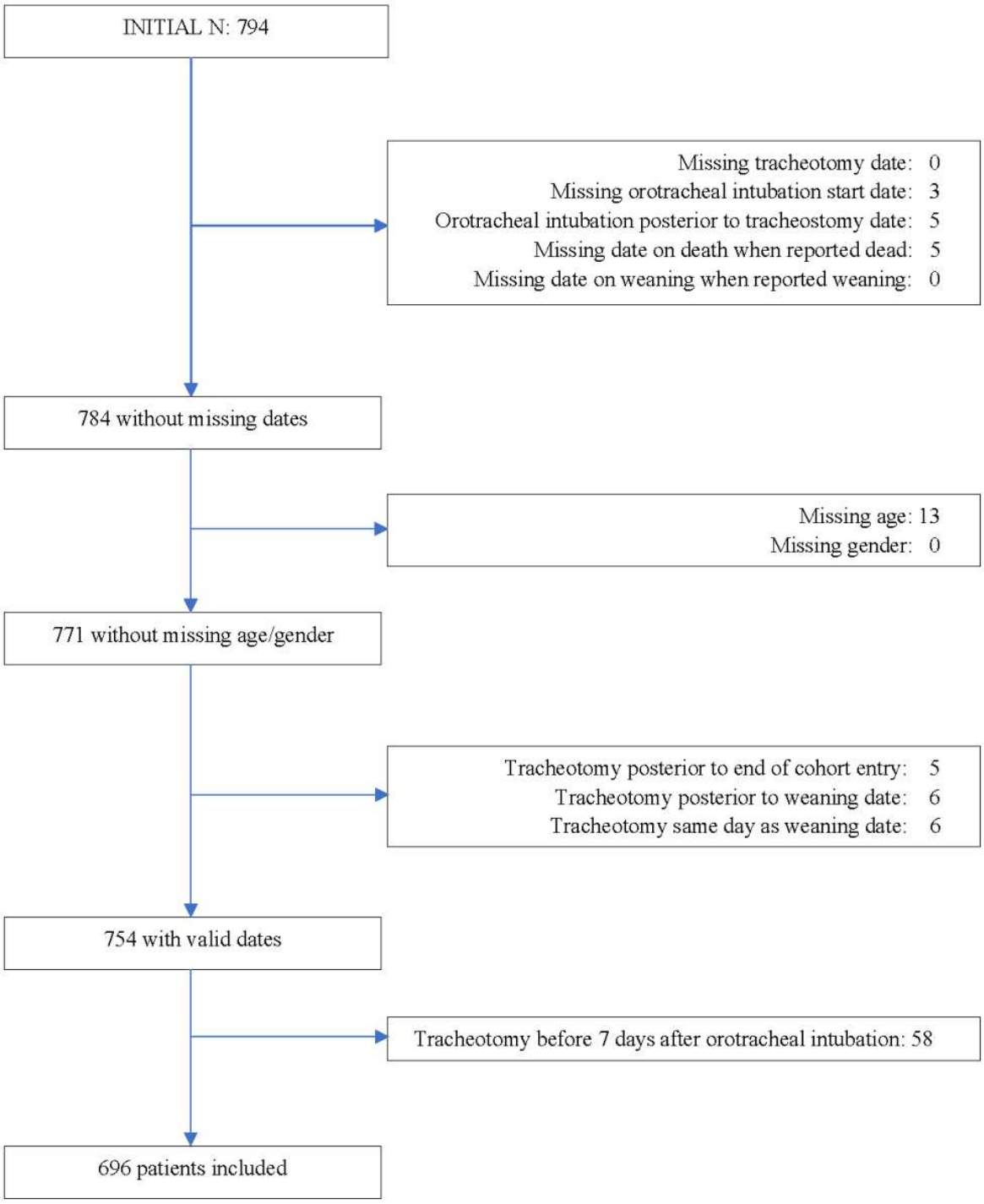


Figure 2.16 – Inclusion and exclusion of study participants.

Imputation results

PAFI, PEEP, APACHE II, anticoagulant use, and comorbidity data (High Blood Pressure, Immunodepression, Cardiac insufficiency, Autoimmune disease, COPD, Diabetes Mellitus, Neuromuscular disease, Ischemic cardiopathy) had missing data as shown in ***Table 2.16***. Pre-imputation and post-imputation values are similar for all variables as shown in ***Table 2.15***.

	Pre-imputation		Post-imputation	
Continuous variables	Mean	SD	Mean	SD
PEEP	9.67	3.00	9.66	3.00
PAFI	140.92	69.36	139.82	69.08
APACHE II	15.05	6.55	15.05	6.54

Binary Variables	Pre-imputation	Post-imputation
High Blood Pressure	46.6%	46.6%
Immunodepression	7.1%	7.1%
Cardiac insufficiency	3.5%	3.5%
Autoimmune disease	5.8%	5.8%
COPD	7.2%	7.2%
Diabetes Mellitus	21.6%	21.6%
Neuromuscular disease	1.4%	1.5%
Ischemic cardiopathy	9.4%	9.4%
Anticoagulant drug	54.5%	55.4%

PAFI: PaO_2/FiO_2 ratio; PEEP: positive end expiratory pressure; COPD: Chronic obstructive pulmonary disease

Table 2.15 – Pre and post imputation values (Mean and standard deviation for continuous and prevalence for binary) of imputed variables

Descriptive results

Baseline characteristics of the study participants can be found in ***Table 2.16***, overall and by tracheotomy timing. I included 696 patients, 142 (20.4%) of whom received an early tracheotomy. They were mostly men (69.1%) and had a mean age of 63 years old. There was little difference between early and late tracheotomy recipients, namely PAFI at ICU admission (139.2 for late tracheotomised vs 153.8 for early) although it turned out to be quite similar at day 7 (182.6 for late tracheotomised vs 183.9 for early) and use of anticoagulant drugs.

	Total patients	Late (>10d after OTI)	Early (<=10d after OTI)
	N=696	N=554	N=142
Sex, female	30.9%	29.4%	36.6%
Age (years)	63.0 (10.2)	63.0 (10.4)	63.2 (9.2)
Tobacco consumption			
Never	74.6%	74.2%	76.1%
Smoker	16.1%	15.3%	19.0%
Missing	9.3%	10.5%	4.9%
Smoking Index (pack/year)	3.4 (12.6)	3.3 (12.6)	3.4 (12.6)
Missing	16.8%	17.1%	15.5%
Weight (Kg)	83.1 (15.5)	83.0 (15.1)	83.2 (17.3)
Missing	19.7%	19.7%	19.7%
Height	168.6 (9.1)	168.8 (9.0)	167.9 (9.3)
Missing	21.8%	22.0%	21.1%
BMI	29.3 (5.4)	29.2 (5.2)	29.8 (6.3)
Missing	23.7%	23.5%	24.6%
Comorbidities			
High Blood Pressure	46.6%	44.6%	54.2%
Immunosuppression	7.0%	7.6%	4.9%
Heart failure	3.4%	3.4%	3.5%
Autoimmune disease	5.7%	6.1%	4.2%
COPD	7.2%	7.0%	7.7%
Pregnancy	0.4%	0.5%	0.0%
DM	21.6%	20.8%	24.6%
Neuromuscular disease	1.4%	1.4%	1.4%
Ischemic cardiopathy	9.3%	8.8%	11.3%
APACHE II	15.1 (6.6)	15.3 (6.7)	11.2 (6.1)
Missing	18.2%	16.8%	23.9%
SOFA	6.1 (3.6)	6.0 (3.4)	6.7 (4.4)
Missing	21.8%	22.6%	19.0%
INR at tracheotomy	1.6 (2.1)	1.5 (1.9)	1.8 (2.7)
Missing	17.8%	18.6%	14.8%
PAFI at intubation	142.1 (70.0)	139.2 (69.2)	153.8 (72.1)
Missing	0.0%	0.0%	0.0%
PAFI at day 7	182.9 (73.6)	182.6 (74.7)	183.9 (69.8)
Missing	13.2%	14.6%	7.7%
PAFI at tracheotomy	192.7 (69.3)	195.0 (69.7)	184.1 (67.4)
Missing	7.6%	8.7%	3.5%

	Total patients	Late (>10d after OTI)	Early (<=10d after OTI)
PEEP at intubation	12.6 (5.1)	12.6 (5.6)	12.5 (3.2)
Missing	10.8%	12.3%	4.9%
PEEP at day 7	11.1 (7.8)	11.3 (8.6)	10.6 (3.3)
Missing	13.9%	15.2%	9.2%
PEEP at tracheotomy	9.7 (3.0)	9.5 (2.9)	10.6 (3.4)
Missing	0.0%	0.0%	0.0%
Complications: Ventilator problems	13.9%	14.4%	12.0%
Missing	1.1%	1.4%	0.0%
Anticoagulant treatment	56.2%	60.1%	40.8%
Vasoactive drugs at tracheotomy	40.4%	40.8%	38.7%
Missing	12.8%	13.9%	8.5%
Vasoactive drugs at OTI	52.3%	52.7%	50.7%
Missing	4.9%	4.9%	4.9%
Secretion problems			
No	73.7%	72.0%	80.3%
Increase pressure	12.6%	13.2%	10.6%
Obstruction	3.9%	3.4%	5.6%
Missing	9.8%	11.4%	3.5%
Indication tracheotomy			
Prolonged mechanical ventilation	81.6%	83.2%	75.4%
Secretions management	10.2%	9.6%	12.7%
Other	8.0%	7.0%	12.0%
Missing	0.1%	0.2%	0.0%
Lymphocyte count	5304.9 (26886.6)	6167.0 (29741.4)	1961.1 (9140.1)
Missing	1.9%	2.0%	1.4%
INR	1.6 (2.3)	1.5 (2.0)	2.1 (3.0)
Missing	8.5%	8.8%	7.0%
D-Dimer	1515.0 (1734.6)	1511.5 (1733.5)	1528.3 (1746.4)
Missing	22.1%	22.6%	20.4%
Ferritin	1367.7 (1312.6)	1385.0 (1298.7)	1300.2 (1369.6)
Missing	24.9%	24.9%	24.6%
LDH	599.0 (705.3)	571.5 (564.8)	70.4 (1081.7)
Missing	13.8%	14.3%	12.0%
Leukocyte count	4538.0 (9721.8)	4497.1 (10380.7)	4697.2 (6581.9)
Missing	1.6%	1.6%	1.4%

	Total patients	Late (>10d after OTI)	Early (<=10d after OTI)
Lymphocytes	62.0 (180.1)	64.5 (186.3)	52.6 (154.2)
Missing	1.4%	1.8%	0.0%
CRP	20.9 (22.7)	21.2 (22.5)	20.0 (23.7)
Missing	44.1%	47.3%	31.7%

COPD: Chronic obstructive pulmonary disease, DM: diabetes mellitus, INR: International normalised ratio, PAFI (PaO2/FiO2), PEEP: positive end-expiratory pressure, OTI: orotracheal intubation, LDH: lactate dehydrogenase, CRP C-reactive protein.

Table 2.16 – Baseline characteristics in tracheotomised patients stratified by tracheotomy timing after imputation. Percentage for categorical variables or mean (SD) for continuous.

One of the variables of interest is pronation. ***Appendix Table 2.8*** shows the frequency of pronation, days of pronation, and whether pronation finished after or before tracheotomy in the two groups. At baseline, the proportion of patients pronated and how many days these patients were pronated were similar between the late and early tracheotomy groups. Participants with late tracheotomies stayed more days pronated (9.5 days in late vs 6.8 in early). 50% of those with a late tracheotomy finished their last pronation cycle before tracheotomy, in contrast with 33% of those with an early tracheotomy. They also had more days of pronation before tracheotomy, 9.0 days in late tracheotomy versus 5.3 days in patients with early tracheotomies.

In ***Table 2.17*** I present the time to weaning or death or censoring, the time to weaning and time to death for each treatment group. The median follow-up time

from decision to weaning or death was 27 days for late tracheotomy and 13 days for early tracheotomy. When looking at time from tracheostomy to weaning or death the median time is similar, 12 days for patients with early tracheotomy compared to 13 days for late tracheotomy. Among those who were successfully weaned, participants weaned in 11 days since tracheotomy and 19 days since orotracheal intubation. Patients who received a late tracheotomy group weaned in 12 and 29 days, respectively.

	Early Tracheotomy			Late Tracheotomy		
Days to weaning or death or censoring	**Median**	**p25**	**p75**	**Median**	**p25**	**p75**
Since Tracheotomy (T0)	12	6	22	13	7	26
Since Decision (D7)	13	8	24	23	16	37
Since Orotracheal Intubation (D0)	20	15	31	30	23	44
Days to weaning (those who wean)						
Since Tracheotomy (T0)	11	6	17	12	7	21
Since Decision (D7)	12	8	20	22	16	33
Since Orotracheal Intubation (D0)	19	15	27	29	23	40
Days to death (those who die)						
Since Tracheotomy (T0)	13	7	23	14	7	25
Since Decision (D7)	16	9	26	24	17	35
Since Orotracheal Intubation (D0)	23	16	33	31	24	42
p25: lower quartile, p75: upper quartile						

Table 2.17 – Days to weaning by tracheotomy timing, since orotracheal intubation and since tracheotomy.

Rates of weaning before the end of the follow-up differed between early and late tracheotomy. 360 of the 554 participants in the late tracheotomy group had a successful weaning before the end of follow-up (3.1 patients weaning per 100 patient-days [2.8 to 3.4]). In the early tracheotomy group, 102 out of 142 had a successful weaning (3.9 patients weaning per 100 patient-days [3.1 to 4.7]). Mortality rates were almost equal between groups. In the late tracheotomy group, 170 patients died out of 504, with a mortality rate of 0.32 deaths per 100 patient-days 95%CI (0.23 to 0.44). In the early tracheotomy group, 38 patients died out of 142, with a rate of 0.32 deaths per 100 patient-days 95%CI (0.23 to 0.44). ***Figure 2.17*** shows these cumulative rates.

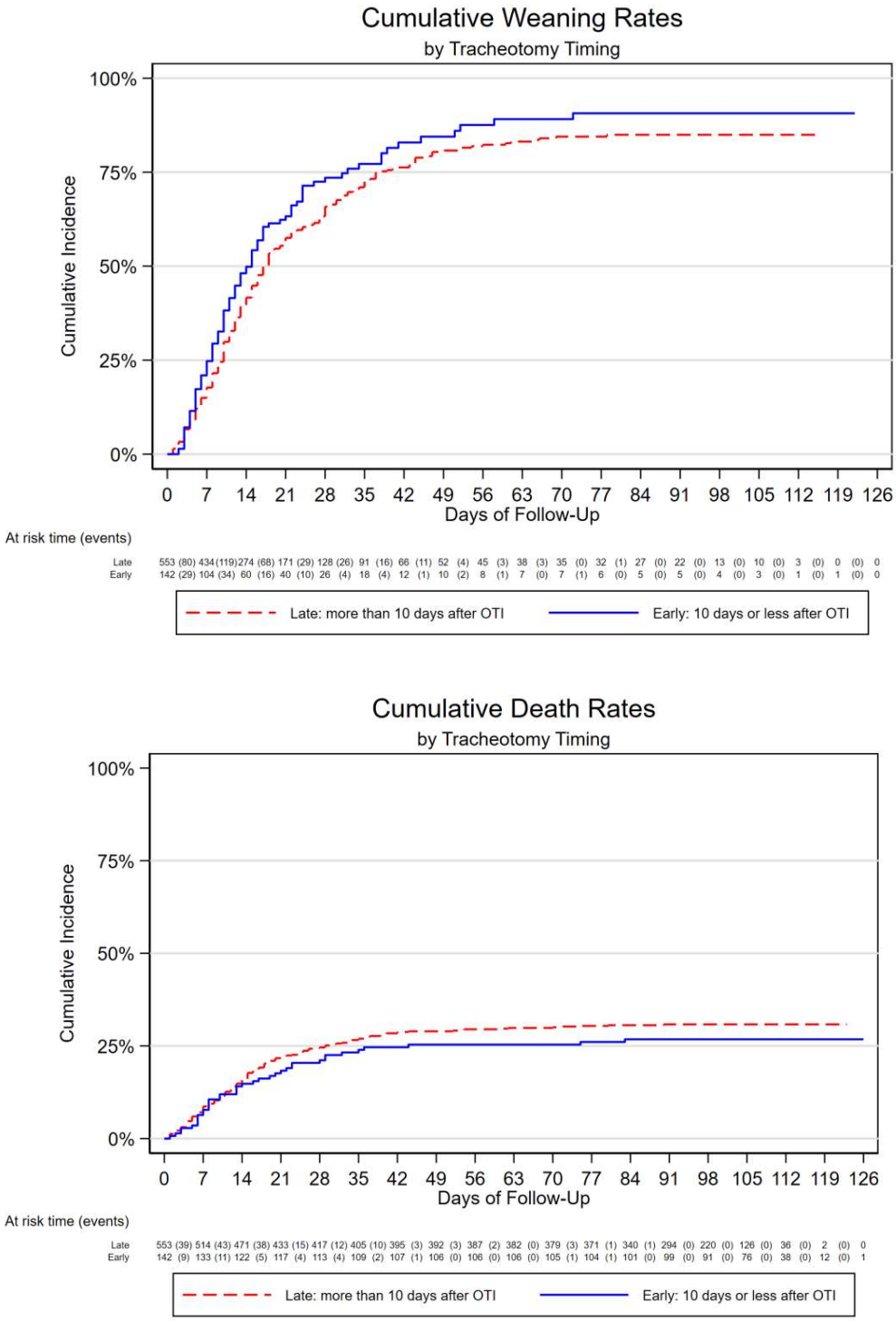


Figure 2.17 – Cumulative incidence of weaning and death by tracheotomy timing.

Regression results

As shown in *Table 2.18*, in the Cox regression the early tracheotomy group had a higher probability of weaning after tracheotomy, csHR 1.25 95%CI (1.00 to 1.56). The estimate remained the same after adjusting for age and sex, csHR of 1.25 (1.00 to 1.56), and for the clinically relevant variables, csHR of 1.31 (1.02 to 1.81). After considering the competing risk of death, the estimates of weaning remain almost unchanged. But, when taking into account the possible clustering by hospitals using the hospital as a random intercept, the weaning estimates are greatly attenuated, with an unadjusted csHR of 1.07 (0.84 to 1.36), age-sex adjusted of 1.10 (0.86 to 1.40), and fully adjusted of 1.02 (0.96 to 1.33). The 14-day weaning relative risk estimate was 1.21 (0.92 to 1.60) with the Poisson regression and 1.15 (0-86-1.55) in the mixed effects Poisson.

The unadjusted estimate for mortality in the Cox regression was of 0.85 (95% CI: 0.60 to 1.21), also not attenuated after adjusting for age and sex, 0.87 (0.61 to 1.25), or for the selected clinical variables: 0.91 (0.56 to 1.47). It remained changed after accounting for hospital clustering.

	Weaning		Death	
Early vs Late Tracheotomy	**csHR**	**95% CI**	**csHR**	**95% CI**
Cox	1.25	(1.00 , 1.56)	0.85	(0.60 , 1.21)
Age and Gender Adjusted	1.25	(1.00 , 1.56)	0.87	(0.61 , 1.25)
Fully Adjusted	1.28	(0.95 , 1.71)	0.77	(0.48 , 1.25)
	csHR	**95% CI**	**csHR**	**95% CI**
Mixed effects Weibull	1.07	(0.84 , 1.36)	0.82	(0.57 , 1.19)
Age and Gender Adjusted	1.10	(0.86 , 1.40)	0.83	(0.57 , 1.21)
Fully Adjusted	1.02	(0.96 , 1.33)	0.70	(0.42 , 1.18)
	sdHR	**95% CI**		
Fine and Gray Crude	1.22	(0.98 , 1.52)		
Age and Gender Adjusted	1.21	(0.97 , 1.50)		
Fully Adjusted	1.31	(0.98 , 1.77)		
	RR		***RR***	**95% CI**
14-day Poisson	1.21	(0.92 , 1.60)	0.93	(0.58 , 1.50)
14-day Mixed effects Poisson	1.15	(0.86 , 1.55)	0.99	(0.60 , 1.62)

csHR: cause-specific hazard ratio, CI: confidence interval. The fully adjusted model included PAFI (PaO2/FiO2), PEEP (positive end-expiratory pressure), SOFA (Sequential Organ Failure Assessment Score) APACHE II (Acute Physiology And Chronic Health Evaluation II Score), and pronation days as covariates. F-G: Fine and Gray, sdHR: subdistribution hazard ratio RR: relative risk, CI95: Confidence interval 95%

Table 2.18 Associations of tracheotomy timing with time to weaning and time to death. Cox, Mixed effects Weibull, Fine and Gray and 14d Poisson.

As for complications, my study was underpowered to detect any differences between early and late tracheotomy. Early tracheotomy had a risk ratio for intraoperative complications of 0.57 (95% CI: 0,24 to 1,34) compared to late tracheotomy and 1.17 (0.83 to 1.66) for post-operative complications. The results did not change considerably after adjusting for age and sex, fully adjusting or after considering the multilevel structure. These results are presented in ***Table 2.19***.

Early vs Late Tracheotomy	Early	Late	Crude		Age-Sex Adjusted		Fully Adjusted	
			RR	95% CI	RR	95% CI	RR	95% CI
Intraoperative	6	41	0.57	(0.24 , 1.34)	0.58	(0.25 , 1.37)	0.57	(0.24 , 1.34)
Bleeding	4	20	0.78	(0.27 , 2.28)	0.81	(0.28 , 2.36)	0.78	(0.27 , 2.28)
Ventilatory problems	4	22	0.71	(0.24 , 2.05)	0.72	(0.25 , 2.08)	0.71	(0.24 , 2.05)
Postoperative	41	136	1.17	(0.83 , 1.66)	1.21	(0.86 , 1.72)	1.17	(0.83 , 1.66)
Bleeding	33	102	1.26	(0.85 , 1.87)	1.30	(0.87 , 1.92)	1.26	(0.85 , 1.87)
Ventilatory problems	8	34	0.92	(0.42 , 1.98)	0.97	(0.45 , 2.10)	0.92	(0.42 , 1.98)

Mixed Effects	Crude		Age-Sex Adjusted		Fully Adjusted	
Early vs Late Tracheotomy	**RR**	**95% CI**	**RR**	**95% CI**	**RR**	**95% CI**
Intraoperative	0.70	(0.28 , 1.73)	0.72	(0.29 , 1.79)	0.22	(0.27 , 1.73)
Bleeding	0.86	(0.28 , 2.68)	0.89	(0.28 , 2.84)	0.42	(0.48 , 3.69)
Ventilatory problems	0.93	(0.30 , 2.89)	0.94	(0.30 , 2.92)	0.44	(0.05 , 3.67)
Postoperative	1.18	(0.81 , 1.72)	1.22	(0.83 , 1.78)	1.49	(0.97 , 2.31)
Bleeding	1.27	(0.84 , 1.94)	1.31	(0.85 , 1.99)	1.82	(1.12 , 2.97)
Ventilatory problems	0.89	(0.39 , 2.06)	0.93	(0.40 , 2.15)	0.88	(0.36 , 2.17)

RR: relative risk, CI: confidence interval. The fully adjusted model included PAFI (PaO2/FiO2), PEEP (positive end-expiratory pressure), SOFA (Sequential Organ Failure Assessment Score), APACHE II (Acute Physiology And Chronic Health Evaluation II Score), and pronation days as covariates.

Table 2.19 Associations of tracheotomy timing with incidence of intraoperative and postoperative complications.

Sensitivity Analyses

I also explored possible interactions between tracheotomy timing and clinically important variables. Although PAFI and pronation days had a very significant interaction with tracheotomy timing in the Cox regression, this interaction completely disappeared when accounting for clustering in a mixed effects regression. There were no different interactions in the mortality outcomes.

2.2.3 Discussion

Introduction

In this emergency study, I tried to evaluate the optimal timing of tracheotomy for ICU-admitted patients with COVID-19 who required IMV. The results show little difference between early tracheotomy and late tracheotomy in terms of weaning. It does not seem to have any differential effects on mortality or complications. Considering these results, an early tracheotomy might be indicated when clinically feasible, as at equal weaning days after a tracheotomy, advancing it a couple of days could release ICU space.

Clinical Implications

The main aim of this study was to assess whether the early tracheotomy was as good as a late tracheotomy in days of weaning in COVID-19 patients. This timing can be affected by several factors in a clinical setting. Some of those are cause of IMV, severity of disease, neurological status, complications, and the patient's chances of recovery. In this study, as we restricted the sample to COVID-19 patients, the cause of IMV is severe respiratory failure. These patients may have high pronation requirements, to what I paid special attention.

The optimal timing for patients receiving IMV in the ICU is yet a matter of discussion in non-COVID scenarios. Two large randomised controlled trials did not found differences between late and early tracheotomy in complications, pneumonia, or mortality. (148, 149) But, neither of these trials had days of mechanical ventilation as an outcome. In terms of discharge from ICU, in a Cochrane review and a more recent metanalysis, early tracheotomy had a higher probability of discharge at day 28 compared to late tracheotomy (>10d). (135, 148)

Another important factor deciding to perform an early tracheotomy is pronation. This position is widely used in COVID-19 critical patients to improve the ventilation/perfusion quotient. But, if the patients have a tracheotomy, it makes pronation more difficult to perform, and increases the risk of displacing the cannula. Some guidelines advise for delaying or nut doing altogether the tracheotomy if prone manoeuvres are required, but it is not a contraindication(137, 140). In this study, I also found that late tracheotomy group had more pronation days than early tracheotomised overall. But there was little difference in patients who were pronated at the time of the decision, where 61.3% of the early tracheotomy group were proned vs 65% of the late group. On the early group, 25% continued pronation after tracheotomy without major incidences and similar

weaning times. In this case, it does not seem to be a definitive factor for the decision or to affect outcomes.

Early tracheotomy has also been thought to increase the risk of complications in critically ill patients. At the start of the pandemic, there was a high mortality rate for COVID-19 patients, and that made doctors less prone to perform invasive procedures such as tracheotomy. These patients had higher rates of pronation and anticoagulant use, that could further increase complication risk. In non-COVID-19 patients, rates of complications range between 6 and 27% (150, 151). But, in COVID patients seems to be much higher, up to 55.3% had postoperative complications within the first 30 days. (152) These complications included infection (36%), haemorrhage (19%), and subcutaneous emphysema (9%). In this study, I found a much lower complication rate of 4% for early tracheotomy and 11% for late tracheotomy, signalling that it is a much safer procedure than it was previously thought.

Best practices for tracheotomy in critically ill COVID-19 patients were suggested during the early stages of the pandemic, but they were mostly focused on preventing surgeon infection. They put the focus on delayed tracheotomy, (140) type of tracheotomy, (152) protective equipment (139, 141), or where best to

perform a tracheotomy. (153) As the focus was on the procedure, patient outcomes as weaning outcomes, total days of IMV, or mortality were not taken into account.

Tracheotomy protocols in different hospitals also varied around the world during the COVID-19 pandemic, with little evidence for timing. (154) But earlier tracheotomy might facilitate the weaning process and reduce the length of mechanical ventilation required, subsequently leaving more ICU beds available for other patients. (143) I found that early tracheotomy does not seem to implicate longer IMV times, making the total length of stay in the ICU shorter, without increasing complications or mortality rates.

Although not exempt from methodological pitfalls, which I discuss in the next section, this study suggests that quicker weaning and ICU discharge for COVID-19 patients may be possible with early tracheotomy without an increased risk of complications or mortality. This quicker weaning can be a useful tool for keeping free beds on the ICU and achieving better planning.

Methodological implications

For this study I tried to apply the knowledge gained of MD/surgical epidemiology during the past years such as the use of clear timelines on a target trial framework,

how to address confounding and how to take into account the differences in practice between centres.

One of the most common problems in surgical observational studies is the lack of a clear timeline of events and the lag between a decision to perform certain procedure is made and the procedure itself. This creates a problem, especially in settings like this one, when the timing of procedures is different, and there is a high mortality. To overcome it, I proposed to think about timings in a trial emulation framework (59) randomising at day 7, which is approximately when the necessity of doing a tracheotomy arises. I only considered clinical variables as possible confounders if they were measured on day 7 or before. This helped to avoid common errors as adding possible mediators to the model, such as ventilatory or biochemical criteria on the day of the tracheotomy.

The nature of the data collection and initial design, already done when I decided to analyse the data, prevents from interpreting as causal the effects of tracheotomy timing on total days of IMV. These is due to the lack of information on patients who died between day 7 of IMV and the time the tracheotomy was supposed to happen, creating a high risk of immortal time bias.(42) Although the observed differences in total IMV duration post-tracheotomy described are due to the

tracheotomy timing, it is quite possible that the differences in the overall IMV duration are artificially inflated by immortal time bias.(42) The best way to address this research question is an RCT, with an intention-to-treat analysis. But, during a pandemic this would have been quite challenging to perform at the necessary scale.

As for confounding, I described several clinical and ventilatory criteria that, apart from institutional factors, are thought to be the main factors that inform the decision about when to perform a tracheotomy. I predefined, with the help of clinicians the variables that would most likely influence both the decision and the outcomes: age, sex, APACHE, SOFA, PAFI, PEEP, and pronation days. However, these variables were very similar between timing groups at the index date, day 7 of IMV and on the measurements before day 7. I decided to take the approach of getting unadjusted estimates, conditioned for age and sex, and further adjusting for the proposed variables. The estimates were not attenuated after the adjustments. As an academic exercise I calculated a propensity score (PS), and the overlap between groups was almost total as shown in ***Appendix Figure 2.2***. I further used this PS to perform matching and inverse probability weighting, which,

although improved balance, rendered estimates and confidence intervals almost identical to the crude regression.

This study had data of multiple centres which differed on their protocols, so substantial variations in the use of early vs late tracheotomy, on procedures used, on weaning criteria and possibly on outcomes are expected. This improves the external validity of the study, as it spans through a wide variety of settings and practices. But it could introduce problems related to the clustering of exposures and outcomes. The preference for late or early tracheotomy is very dependent on the hospital, and it is not unreasonable to think that weaning practices also are. In this sample also the number of patient that each hospital contributes is not equal or balanced. This is why after correcting for this clustering the estimates get attenuated, probably signalling a strong hospital effect on outcomes.

In this study, I used three of the most used methods to analyse time to event data and to take into account competing risks.(155) In this setting, competing risks are of great importance, as almost 30% of the patients included died before weaning. One of the most used method for resolving this issue is the analyses of subdistribution hazards as proposed by Fine and Gray. (145) This method considers competing risks, such as death in this case, by not excluding the patients

from the risk pool. This is useful when estimating risks for the patient, but it could appear that probability of weaning is decreased if the risk of dying is increased. This is the reason why one can use cause-specific Hazards, like the ones produced by the Cox and the Weibull regressions here, that condition on the competing event. But hazards are conditional on survival to the time of study, which can be affected by treatment, making generally impossible to interpret hazard ratios as causal effects.(155) Because of this, I decided to include the relative ratio (RR) of weaning and death in 14 days, with the timeline based in clinical suggestion. This is probably the easiest interpretable estimate, considering the ratio of risks without eliminating the competing event. In the study, there seems to be no differential effect of treatment or other variables on death, making the three estimates very similar.

Chapter 3

Safety - Comparative safety of TKR and PKR and Post operative risks of KR using a self-controlled case series

Comparative safety constitutes one of the most important undertakings of MD or surgical epidemiology. This consists of comparing possible harmful effects of an intervention. Safety should be considered when assessing the suitability of an intervention and weighted against the benefits.

Although large pragmatic Randomised Controlled Trials can detect some safety signals, they are usually underpowered to detect differences in the occurrence of even common (1 in 10 to 1 in a 100) events. The presence of safety issues could totally render a procedure unsuitable for medical practice if the risks of it exceed the benefits. Therefore, safety research constitutes the epitome of why observational research is useful and needed in surgical and medical procedure sciences. However, confounding by indication remains an essential caveat in comparative safety studies.

In this chapter I continue to explore the methods I benchmarked in **Chapter 2** to study effectiveness outcomes and apply what I have learnt there (in **Section 3.1**). I also introduce a novel study design particularly well suited for exploring safety of an intervention: the self-controlled case series, in **Section 3.2**.

3.1 Comparative rates of revision, complications, and mortality in the study of partial vs total knee replacement

This chapter follows the analyses of **section 2.1.** Here I apply the validated methods to the *safety cohort.* This allows me to assess outcomes that the TOPKAT RCT was not powered to look at. The difference in objectives led me to present these results in a different section. Here, I do not formally compare to the results of the trial, and patients included are different: the cohort in **Chapter 2** only includes patients who had participated in PROMS. Also, although revision is arguably an effectiveness outcome, I considered it better suited for this section as it shares population.

Introduction

TOPKAT has provided gold standard evidence on the effectiveness of PKR vs TKR for medial compartmental knee osteoarthritis, but it did not show differences in safety events, as they were uncommon. This creates a need to characterise the safety profiles and understand the risks of complications, such as revision, systemic and wound infection, myocardial infarction (MI), venous thromboembolism (VTE), and mortality.

In observational studies, PKR has a lower risk of post-operative complications such as MI, VTE and mortality. (129) But, these studies have also found an increased risk of revision, between 2 and 3 times, consistently using NJR data, (73) in a meta-analysis, (129) and a multinational analysis led by OHDSI and EHDEN collaborators.(156) In TOPKAT, PKR was not associated with an excess risk of revision, contradicting all this previous observational research.(70) This disagreement could well be explained by residual confounding in the observational studies. It could also be due to differences between TOPKAT and the observational studies in the selected population of patients or surgeons (or both) included in the RCT vs population-based cohort analyses. This calls for more evidence on these differences. To shed light on the subject, I studied it in the context of a trial "replication". The population selected tries to mimic the one on the trial, and the methods to minimise confounding were able to replicate the effectiveness estimators from the trial, as seen in **Chapter 2**.

Objectives

To study the comparative safety of PKR vs TKR in terms of short-term complications (90-day post-operative venous thromboembolism, myocardial

infarction, and prosthetic joint infection) and long-term (5-year) revision and death, among those eligible for TOPKAT and using the methods that were able to replicate RCT results validated in **Chapter 2**.

3.1.1 Methods

Study design, data sources and population

Cohort study, including all patients who had undergone a first total or partial knee replacement between 2009 and 2016 in the UK. Further particulars can be found in **section 2.1.1.**

Sample and target population

I used data from the same target population described in **section 2.1.1** with the only difference of having no need to exclude patients without post-operative OKS information or that did not take part of the PROMs database in this chapter.

Outcomes

In this chapter I looked at revision, post-operative complications, and mortality. Patients were followed up from the date they had surgery (index date) until the earliest of: end of enrolment in the database (31/Dec/2016), death, or five years after index date. I censored patients if a surgery was performed in the contralateral knee, as it was not possible to adjudicate complications to one of the two surgeries.

Complications

I looked at three post-operative complications: myocardial infarction (MI), venous thromboembolism (VTE), and prosthetic joint infection (PJI). I identified these complications within 90-days after primary surgery using ICD-10 codes recorded as primary diagnosis in the HES dataset. These codes are shown in ***Appendix table 3.1***. They were pre-specified based on previous research and reviewed by clinicians.(157)

Revision

Revision occurring within 5-years from primary joint replacement surgery was extracted from the NJR dataset, where the collection of this information is mandatory. I studied both 5-year risk and time to revision within 5 years.

Mortality

Mortality records were obtained through linkage of the NJR dataset to the Office for National Statistics (ONS) data. Death analyses were censored at revision as death was not registered in NJR data if a revision occurred first. I studied both 5-year mortality and time to death within 5 years.

Methods to minimise confounding

In the previous chapter I studied two groups of methods to minimise confounding in observational studies: I) Propensity Scores (PS), widely used and accepted, but incapable of dealing with unmeasured confounding,(82) and II) Instrumental Variable (IV) analysis, which theoretically addresses this under very strong assumptions.(77)

I had previously found IVs led to major bias and implausible estimators in the effectiveness cohort with a continuous outcome. In this chapter I applied this method to a larger cohort, to a binary outcome, and with a potentially different confounding structure. The effect of IMD or of pre-operative knee score may not have the same effect on short term complications, mortality or revision as on pain and functionality. These differences made me think it could be worth testing and presenting IV results, although they did not replicate the trial.

As for Propensity Scores, I continued to use the PS matching, PS stratification, and Inverse Probability weighting (IPW). I also used conditioning on all selected confounders and non-adjusted regression to see how much they divert from the latter.

Sensitivity analyses

As in Chapter 3, I included as sensitivity analyses only those experienced surgeons who had done more than 10 PKR surgeries. As the cohort size here allows, and as volume/experience seems to be a critical factor for revision risk, I did further analyses looking at 30, 50 surgeries. These analyses are discussed in **Chapter 4**.

Comparison to TOPKAT

TOPKAT was underpowered to detect these outcomes. The only estimate that could be produced would be relative risk of 5-year revision, but with high uncertainty, and very wide confidence intervals. Because of this, I did not apply any formal test to compare to TOPKAT but discussed the qualitative differences in point estimates.

3.1.2 Descriptive results

Revision/Safety cohort

For the revision and safety analyses I used the cohort of patients with the TOPKAT eligibility criteria, without excluding those without post-operative OKS, as shown in *Figure 2.3* resulting in a total of 273,530 TKR and 21,026 PKR patients.

Patient characteristics

Patients in the cohort used for this chapter seem slightly less healthy than the subset that had a post-operative OKS. This cohort, as the effectiveness cohort, also had differences between patients who had undergone TKR vs PKR. Patients receiving PKR were mostly males (52%) compared to TKR where they were mostly females (57%) and they were younger (mean 64.3 years old in PKR vs 70.2 for TKR). Patients with PKR had a lower ASA score (11% vs 21% rated as fit and healthy), less comorbidities (69% vs 73% with a 0 Charlson Comorbidity Index. PKR patients were less likely to have other joint problems (13% vs 18% in TKR) and cardiovascular disease (46% vs 58%). More patient characteristics are shown in *Table 3.1.*

Stage 1	Safety cohort			
N(%) or mean (SD)	TKR (n=273,530)	%	PKR (n=21,026)	%
Gender				
F	155267	57	10016	48
M	118263	43	11010	52
Rural Index				
Urban	203938	74	14607	70
Town and fringe	32573	12	2698	13
Village	26012	10	2596	12
Isolated	11007	4	1125	5
IMD				
Least deprived 10%	29339	11	2917	14
Less deprived 10-20%	31518	12	2871	14
Less deprived 20-30%	31946	12	2669	13
Less deprived 30-40%	32593	12	2480	12
Less deprived 40-50%	31209	11	2456	12
More deprived 10-20%	20502	7	1224	6
More deprived 20-30%	23357	9	1415	7
More deprived 30-40%	26174	10	1917	9
More deprived 40-50%	29479	11	2156	10
Most deprived 10%	17413	6	921	4
ASA				
P1 - Fit and healthy	30224	11	4394	21
P2 - Mild disease not incapacitating	243306	89	16632	79
Charlson Comorbidity				
0	187509	69	15408	73
1	58781	21	4134	20
2	17834	7	996	5
3	6172	3	308	1
4	3234	1	180	1
Age	70.2*	8.9**	64.3	9.5
BMI	30.5*	5.1**	30.0	4.9
PROMS pre-operative OKS	19.3*	6.8*	21.3*	6.2*

Stage 1	Safety cohort			
N(%) or mean (SD)	TKR (n=273,530)	%	PKR (n=21,026)	%
EQ5D health scale	69.2*	19.4*	69.7*	19.2*
EQ5D index				
Excellent	161904	59	6546	31
1	43913	16	6643	32
2	30058	11	4400	21
3	26008	9	2217	10
4	10024	4	834	4
Poor	1623	1	386	2
Gastrointestinal Disease	52029	19	3621	17
Other Joint Problems	49941	18	2696	13
Mental Health	25823	9	2380	11
Respiratory Diseases	37754	14	2827	13
Cardiovascular Diseases	157504	58	9592	46
Thyroid Problems	20724	8	1249	6
Foot, hip, spinal pain	3096	1	205	1
Coxarthrosis	8966	3	381	2
Neurological Disorders	16435	6	1208	6
Other Arthrosis	12818	5	708	3
Polyarthrosis	15935	6	675	3
Spondylosis	7378	3	349	2

Table 3.1: Baseline patient-level characteristics for patients who received TKR and PKR surgeries in the safety cohort.

Surgeon characteristics

Surgeon volume for TKR was similar to the effectiveness cohort, with 50% of the TKR patients being operated by surgeons with a volume higher than 50 surgeries

in the past year, but is slightly different for PKR with a median of 15 surgeries on the previous year for the PKR surgeons (9 in the effectiveness cohort).

Cumulative Incidence of Outcomes

Overall, 852 patients who received a PKR (4.1%) and 4,090 patients who received a TKR (1.5%) had to undergo a revision surgery within 5 years of the primary surgery. Mortality was greater in the TKR group, where 14,260 patients (5.21%) died in the 5-year follow-up compared to the PKR group, where 517 patients died in 5 years (2.46%). As for complications, in the 3 months following the primary operation, 62 patients who had an PKR (0.29%) had a VTE event vs 1,750 (0.64%) on those who had a TKR. 26 patients (0.12%) in the PKR group had an MI vs 572 in the TKR group (0.21%). 17 PKR patients (0.08%) had a PJI vs 338 (0.12%) TKR patients. ***Table 3.2*** shows revision and death rates at different time points.

	N	5 year Incidence (95% CI)	N	1 year Incidence (95% CI)	N	90 day Incidence (95% CI)
Revision surgery						
TKR	4,090	1.50% (1.45% , 1.54%)	1,051	0.38% (0.36% , 0.41%)	210	0.08% (0.07% , 0.09%)
PKR	852	4.05% (3.79% , 4.33%)	183	0.87% (0.75% ,1.00%)	26	0.12% (0.08% , 0.18%)
All-cause mortality						
TKR	14,260	5.21% (5.13% , 5.29%)	1,948	0.73% (0.69% , 0.76%)	487	0.18% (0.16% , 0.19%)
PKR	517	2.46% (2.25% , 2.67%)	65	0.31% (0.24% , 0.39%)	11	0.05% (0.03% , 0.09%)

Table 3.2: Outcome rates of Revision and death for patients who received TKR and PKR surgeries.

Discussion

As discussed in the previous chapter, patients in this cohort seem to have slightly more comorbidities than patients with recorded post-operative OKS. Patients undergoing PKR seem to be fitter and younger than TKR patients. Also, patients receiving PKR in this cohort seem to have been operated by higher volume surgeons, a difference that may be explained by geographical differences both on the use of PKR and the uptake of PROMS.(97, 114) I explore this further in **Chapter 4**.

3.1.3 Instrumental Variables

Introduction

As explained previously, in **Chapter 2**, instrumental variable methods, after fulfilling certain very astringent assumptions about the instrument, can produce accurate estimates and deal with measured and unmeasured confounding.(99) After the pitfalls and problems that arose in **Chapter 2** with the use of this method, I still present the analyses, as the method may have performed better with binary/time to event and with safety rather than effectiveness outcomes. In this cohort there is also more patients, and IV analysis may be better powered.

Methods

I used the same instruments proposed in **Chapter 2.** These instruments had been constructed in the whole cohort, but only evaluated in the subset of patients who had OKS reported. Therefore, I show the evaluation for the full cohort on this section Instrumental variable regression was performed with ivpoisson and ivprobit.

Results

Eligible patients

A total of 294,556 patients (273,530 TKR and 21,026 PKR recipients) were included to construct the IVs and to analyse the outcomes. Baseline characteristics are presented in **section 3.1.2**.

The exclusion of patients in the process of creating surgeon/hospital/region preference IVs, explained in **section 3.1.3,** led to the exclusion of 56,099 (19%), of which 3,167 PKR (15%), 78,831 (27%), of which 4,416 PKR (21%), and 111,484 (38%) patients, of which 6,674 PKR (32%), for 20, 30 and 50 surgeries performed by the same lead surgeon. Detailed proportions of excluded patients per each IV are shown in ***Table 3.4.***

Instrumental Variable construction

The constructed IVs are the same as for **Chapter 2** analyses. Preference based: surgeon preference for PKR, hospital preference for PKR, region Preference for PKR; volume based, which was binarised: Total number of surgeries done by the same surgeon/surgical unit in the previous year or in the whole study period; and area of treatment and area of residence intake of PKR. Calendar time was already

discarded because of being nonviable (as there has not a time with a steep change in the uptake of PKR) in **Chapter 2**. Further detail is shown in **section 2.1.3.**

For this cohort, surgeon preference is different than for the **Chapter 2** cohort. ***Table 3.3*** shows descriptive values for surgeon preference without restricting for patients with Oxford Knee Score. The total preference for PKR is around 7-8% compared to 6% on the *effectiveness cohort*. The preference of surgeons who performed PKR is 26-27% compared to 20-21% in the previous cohort.

Total								
Instrument		mean	sd	p50	p10	p25	p75	p90
Lead surgeon	Last 20	7.6%	13.5%	0.0%	0.0%	0.0%	10.0%	25.0%
	Last 30	7.6%	13.1%	0.0%	0.0%	0.0%	10.0%	26.7%
	Last 50	7.8%	12.7%	2.0%	0.0%	0.0%	10.0%	26.0%
Consultant surgeon	Last 20	7.3%	13.1%	0.0%	0.0%	0.0%	10.0%	25.0%
	Last 30	7.4%	12.7%	0.0%	0.0%	0.0%	10.0%	23.3%
	Last 50	7.5%	12.3%	2.0%	0.0%	0.0%	10.0%	24.0%
Surgical Unit	Last 20	7.1%	9.6%	5.0%	0.0%	0.0%	10.0%	20.0%
	Last 30	7.1%	9.0%	3.3%	0.0%	0.0%	10.0%	16.7%
	Last 50	7.1%	8.6%	4.0%	0.0%	2.0%	10.0%	18.0%
TKR								
Instrument		mean	sd	p50	p10	p25	p75	p90
Lead surgeon	Last 20	5.9%	11.3%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	6.0%	11.0%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	6.2%	10.7%	0.0%	0.0%	0.0%	8.0%	20.0%
Consultant surgeon	Last 20	5.8%	11.0%	0.0%	0.0%	0.0%	5.0%	20.0%
	Last 30	5.9%	10.6%	0.0%	0.0%	0.0%	6.7%	20.0%
	Last 50	6.0%	10.3%	0.0%	0.0%	0.0%	8.0%	20.0%
Surgical Unit	Last 20	6.5%	8.7%	5.0%	0.0%	0.0%	10.0%	15.0%
	Last 30	6.5%	8.1%	3.3%	0.0%	0.0%	10.0%	16.7%
	Last 50	6.5%	7.6%	4.0%	0.0%	2.0%	10.0%	16.0%
PKR								
Instrument		mean	sd	p50	p10	p25	p75	p90
Lead surgeon	Last 20	27.7%	20.1%	25.0%	5.0%	10.0%	40.0%	55.0%
	Last 30	27.4%	19.0%	23.3%	3.3%	13.3%	40.0%	53.3%
	Last 50	26.7%	17.8%	24.0%	6.0%	12.0%	38.0%	50.0%
Consultant surgeon	Last 20	26.8%	20.1%	25.0%	5.0%	10.0%	40.0%	55.0%
	Last 30	26.6%	19.2%	23.3%	3.3%	10.0%	40.0%	53.3%
	Last 50	26.1%	18.3%	24.0%	4.0%	12.0%	38.0%	52.0%
Surgical Unit	Last 20	15.8%	15.2%	10.0%	0.0%	5.0%	20.0%	35.0%
	Last 30	15.7%	14.5%	13.3%	3.3%	6.7%	20.0%	36.7%
	Last 50	15.6%	13.9%	12.0%	2.0%	6.0%	20.0%	34.0%

Table 3.3. Mean, SD, median, and percentile 10, 25, 75 and 90 of preference-based instruments.

Instrumental Variable selection

Same IV selection criteria described in **Section 2.1.3** were applied here and presented in ***Table 3.4***.

Unit	IV	% additionally excluded	Odds ratio (95% CI)	F-statistic	Max ASMD	Short-listed
Lead surgeon	Last 20 preference	19.4%	21.84 (20.64 , 23.11)	24,292	0.096	X
	Last 30 preference	26.8%	29.00 (26.94 , 31.19)	19,714	0.089	X
	Last 50 preference	37.8%	28.91 (26.86 , 31.13)	19,375	0.109	
Consultant surgeon	Last 20 preference	13.1%	18.93 (17.95 , 19.95)	23,831	0.108	
	Last 30 preference	18.6%	25.43 (23.76 , 27.21)	19,933	0.100	X
	Last 50 preference	28.4%	24.60 (23.03 , 26.28)	20,2945	0.111	
Surgical unit	Last 20 preference	2.5%	3.81 (3.70 , 3.93)	9,054	0.137	
	Last 30 preference	3.8%	4.17 (4.03 , 4.31)	8,314	0.115	
	Last 50 preference	6.2%	4.61 (4.45 , 4.78)	8,896	0.125	
Lead surgeon	Total experience	0%	1.29 (1.26 , 1.33)	318	0.074	
	Yearly experience	0%	1.26 (1.22 , 1.30)	259	0.075	
Consultant surgeon	Total experience	0%	1.25 (1.22 , 1.29)	243	0.064	
	Yearly experience	0%	1.22 (1.19 , 1.26)	197	0.058	
Surgical unit	Total experience	0%	1.24 (1.20 , 1.27)	224	0.091	
	Yearly experience	0%	1.17 (1.14 , 1.21)	128	0.078	
Area residence		0%	1.91 (1.85 - 1.96)	2,058	0.200	
Area treatment		0%	1.96 (1.90 - 2.02)	2,109	0.144	
CI: confidence interval; ASMD: absolute standardised mean difference						

Table 3.4 Shortlisting of Instrumental Variables, % of excluded patients, Odds Ratio of the association with PKR, F-statistic, Maximum ASMD in the falsification tests and shortlisting status

Preference based Instrumental Variables

The estimated surgeon-based preference instruments based on the previous 20, 30, and 50 surgeries were strong. The association with the exposure ranged from an OR of 3.81 [95%CI 3.70 to 3.93] in surgical unit preference based on the 20 previous surgeries, to an OR of 29 [26.9 to 31-2] in lead surgeon preference based on the previous 30 surgeries.

Most of the instruments led to residual imbalance (ASMD>0.1) for at least one of the known (recorded) confounders in the dataset. Index of Multiple Deprivation was the most imbalanced variable on most of the instruments, especially when the instrument represented bigger populations or areas. For example, for surgical unit the ASMD of IMD was 0.137, 0.115, and 0.125 for instruments based on the previous 20, 30 and 50 surgeries respectively. ***Tables 3.5 to 3.7*** show ASMD for each of the pre-specified confounders.

Covariate	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.034	0.027	0.032
Age at primary surgery	0.035	0.041	0.037
Body mass index	0.011	0.018	0.013
IMD socioeconomic status	0.096	0.089	**0.109**
Preoperative OKS	0.021	0.019	0.016
Myocardial infarction	0.001	0.008	0.001
Heart failure	0.007	0.008	0.004
Peripheral artery disease	0.007	0.007	0.006
Cerebrovascular disease	0.005	0.008	0.011
Dementia	0.006	0.012	0.009
Respiratory/pulmonary disease	0	0	0.005
Peptic ulcer	0.001	0	0.009
Mild liver disease	0.005	0.001	0.002
Severe liver disease	0.025	0.022	0.03
Diabetes	0.016	0.012	0.014
Diabetes with complications	0.011	0.014	0.004
Hemi/paraplegia	0.003	0.004	0.004
Chronic kidney disease	0.002	0.002	0.001
Solid tumours/malignancies	0.009	0.011	0.013
Metastatic cancer	0.006	0.007	0.008
Foot/hip/spine pain	0.021	0.034	0.035
Previous arthroscopy	0.011	0.014	0.011
Hip osteoarthritis	0.02	0.014	0.016
Previous knee washout	0.014	0.016	0.016
Hip replacement	0.014	0.002	0.016
Previous knee injection/s	0.034	0.027	0.032
IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference			

Table 3.5. Covariate balance for a selected list of confounders, stratified by binary lead surgeon preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Confounder	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.029	0.029	0.037
Age at primary surgery	0.012	0.018	0.014
Body mass index	0.013	0.018	0.011
IMD socio-economic status	**0.108**	**0.1**	**0.111**
Pre-operative OKS	0.017	0.018	0.014
Myocardial infarction	0.01	0.011	0.005
Heart failure	0.006	0.012	0.01
Peripheral artery disease	0.001	0.004	0.006
Cerebrovascular disease	0.011	0.009	0.011
Dementia	0.005	0.01	0.006
Respiratory/pulmonary disease	0.006	0.009	0.007
Peptic ulcer	0.007	0.003	0.011
Mild liver disease	0.012	0.012	0.002
Severe liver disease	0.025	0.022	0.028
Diabetes	0.016	0.012	0.018
Diabetes with complications	0.009	0.012	0.006
Hemi/paraplegia	0	0.002	0.003
Chronic kidney disease	0.002	0.002	0.009
Solid tumours/malignancies	0.009	0.011	0.009
Metastatic cancer	0.005	0.006	0.006
Foot/hip/spine pain	0.015	0.025	0.022
Previous arthroscopy	0.008	0.013	0.012
Hip osteoarthritis	0.03	0.025	0.031
Previous knee washout	0.012	0.013	0.011
Hip replacement	0.015	0.006	0.01
Previous knee injection/s	0.029	0.029	0.037
IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference			

Table 3.6. Covariate balance for a selected list of confounders, stratified by binary consultant surgeon preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Confounder	ASMD based on 20 previous surgeries	ASMD based on 30 previous surgeries	ASMD based on 50 previous surgeries
Sex	0.014	0.03	0.032
Age at primary surgery	0.034	0.037	0.048
Body mass index	0.001	0.009	0.008
IMD socio-economic status	**0.137**	**0.115**	**0.125**
Pre-operative OKS	0.006	0.009	0.007
Myocardial infarction	0.009	0	0.006
Heart failure	0.014	0.005	0.009
Peripheral artery disease	0.002	0.01	0.004
Cerebrovascular disease	0.002	0.006	0.008
Dementia	0.009	0.001	0.001
Respiratory/pulmonary disease	0.003	0.005	0.009
Peptic ulcer	0.013	0.01	0.016
Mild liver disease	0.018	0.013	0.016
Severe liver disease	0.018	0.022	0.024
Diabetes	0.001	0.006	0.004
Diabetes with complications	0.009	0.006	0.012
Hemi/paraplegia	0.004	0.004	0.006
Chronic kidney disease	0.021	0.02	0.021
Solid tumours/malignancies	0.006	0.005	0.003
Metastatic cancer	0.001	0.004	0.004
Foot/hip/spine pain	0.025	0.026	0.025
Previous arthroscopy	0.011	0	0.002
Hip osteoarthritis	0.041	0.038	0.042
Previous knee washout	0.011	0.009	0.007
Hip replacement	0.031	0.03	0.023
Previous knee injection/s	0.014	0.03	0.032
IMD: Index of multiple deprivation; OKS: Oxford Knee Score; ASMD: absolute standardised mean difference			

Table 3.7. Covariate balance for a selected list of confounders, stratified by binary surgical unit preference for partial knee replacement estimated based on the previous 20, 30, and 50 surgeries

Volume-based Instrumental Variables

Surgeon volume was also estimated for the lead, consultant, and surgical unit level. Volume was calculated as number of surgeries performed in the previous year and total volume of PKR interventions any time before surgery. None of these were strongly associated to the type of surgery received considering the threshold of OR>2.0. Odds Ratio were weak, ranging from 1.17 [1.14 to 1.21] to 1.29 [1.26 to 1.33]. Hence, none of these variables were taken forward as instrumental variables.

Area-based Instrumental Variables

None of the two area-based instrumental variables tested reached enough instrument strength to be considered as useful instrumental variables. The one based in the area of residence of the patient had an OR of 1.91 [1.85 to 1.96] and the one based on the area where the surgery was performed had an OR of 1.96 [1.90 to 2.02]. Neither variable achieve balance with ASMD higher to 0.1, 0.2 for area of residence and 0.14 for area of treatment.

Calendar time Instrumental Variable

As stated in **Chapter 2**, no calendar time was identified as an instrumental variable as there were no extreme changes in trends of PKR uptake across the study period.

Shortlisted Instrumental Variables

After applying the strength and imbalance pre-specified criteria, 3 instrumental variables were short-listed:

- Lead surgeon preference for PKR
    - Based on the previous 20 surgeries
    - Based on the previous 30 surgeries
- Consultant surgeon preference for PKR
    - Based on the previous 30 surgeries

These were taken forward as potential instrumental variables for outcome analysis in the following section.

Outcome model results

As shown in ***Table 3.8***, all three shortlisted instrumental variables had consistent estimates. Risk of revision of the prosthesis in the following 5 years was twice as high for PKR patients compared to TKR patients. Risk ratio was 2.20 95%CI(1.45 to 3.33) for the IV previous 20 surgeries from the same lead surgeon, 2.22 (1.39 to 3.54) for the IV previous 30 surgeries from the same lead surgeon, and 1.88 (1.19 to 2.99) for the IV previous 30 surgeries from the same consultant surgeon.

Mortality rate in 5 years seemed to be lower for PKR. Risk ratio was 0.46 95%CI(0.36 to 0.59) for the IV previous 20 surgeries from the same lead surgeon,

0.49 (0.38 to 0.65) for the IV previous 30 surgeries from the same lead surgeon, and 0.54 (0.41 to 0.70) for the IV previous 30 surgeries from the same consultant surgeon.

As for 3-month complications, PKR had a lower risk of VTE with RR of 0.43 (0.22 to 0.85), 0.44 (0.21 to 0.93), and 0.33 (0.16 to 0.70) respectively for the three IVs: preference in the 20 previous surgeries of the same lead surgeon, preference in the 30 previous surgeries of the same lead surgeon and preference in the 30 previous surgeries of the same consultant surgeon. For prosthetic joint infection the IV results were 0.38 (0.08 to 1.73), 0.67 (0.13 to 3.43), and 0.61 (0.12 to 3.02), and for myocardial infarction RR was 0.96 (0.31 to 2.95), 1.02 (0.30 to 3.53), and 0.84 (0.24 to 2.90).

	last20 lead		last30 lead		last 30 consultant	
Outcome	**RR**	**95%CI**	**RR**	**95%CI**	**RR**	**95%CI**
5-year Revision	2.20	(1.45 , 3.33)	2.22	(1.39 , 3.54)	1.88	(1.19 , 2.99)
5-year Mortality	0.46	(0.36 , 0.59)	0.49	(0.38 , 0.65)	0.54	(0.41 , 0.70)
3-month VTE	0.43	(0.22 , 0.85)	0.44	(0.21 , 0.93)	0.33	(0.16 , 0.70)
3-month PJI	0.38	(0.08 , 1.73)	0.67	(0.13 , 3.43)	0.61	(0.12 , 3.02)
3-month MI	0.96	(0.31 , 2.95)	1.02	(0.30 , 3.53)	0.84	(0.24 , 2.90)

Table 3.8 – Association between PKR (compared to TKR) and 5-year revision, 5-year mortality, 3-month venous thromboembolic events, 3-month prosthetic joint infection, and 3-month myocardial infarction in the 5 shortlisted instrumental variable analyses

Discussion

Instrumental variables seem to work better in this population, or, at least, they don't yield implausible results. The results in the selection phase are similar to the *effectiveness* cohort. A lower number of proposed instruments passed the proposed diagnostics, 3 out of 17 tested. Of the other 14, 8 of the instruments failed due to observed residual confounding: they were preference-based and area of residence. This imbalance happens mostly in socioeconomic variables and in pre-operative OKS, showing a potential unequal geographical and socioeconomic distribution of the preference for PKR or TKR and their outcomes as discussed in **Chapter 2**. The 6 instrumental variables based in volume did not pass the strength requisite (OR>2 in the relationship between instrument and PKR exposure). Area-based instruments also failed this diagnostic test.

The outcome results from the three selected instrumental variables analyses showed how PKR patients have more than double risk of needing a revision of the prosthesis in 5 years, half the risk of dying in the following 5 years and a 60-70% reduction of 90-day VTE. They show a very high uncertainty on the results for PJI

and MI, probably due to lack of power to detect differences, as they are much less common outcomes.

These results are consistent with those that resulted from applying the validated PS methods (IPW and PS stratification) to this cohort and their clinical implications are discussed in **section 3.1.4**. What these results also show, is how, in this cohort, instrumental variables seem to produce accurate results, similar to those of validated methods. This difference with the effectiveness cohort in **section 2.1.2**, where the results were extreme and implausible, could have several explanations.

In this chapter, we have a larger cohort that comprises all patients who underwent knee replacements and would have been eligible for the trial. The cohort in the previous chapter is much smaller, as it is restricted to patients who were offered and completed PROMs questionnaires. This sample could have been different, or there might have been an unmeasured confounder that is related to the IV and to the OKS, but not to revision, mortality, and complications. The outcome distribution and regression form here, being a binary outcome modelled with a Poisson regression may be less sensitive to bias than a non-normal outcome (OKS) modelled with a linear regression. This is discussed and tested in two simulation studies shown in **Chapter 5**.

3.1.4 Propensity scores

Introduction

In this chapter, I used the same methods as in **Chapter 2**, with a focus on those that replicated similar effect size as TOPKAT main outcome. This section only describes the differences in the study settings between this chapter and Chapter 2.

Methods

PS methods, missing data, and diagnostics

Patient variables included in PS model are the same as described in ***Table 2.9***. Similarly to **Chapter 2**, I imputed missing data using MICE (10 imputations) with all covariates and outcomes. I used our previously "validated" PS methods, stratification and weighting, and the non-validated adjustment, matching, and fully adjusted and unadjusted regression as explained in **section 2.1.4**. As in **Chapter 2,** I used ASMD with a cut-off of 0.1 as a diagnostic, comparing patient level characteristics of TKR vs PKR receivers after matching, stratification and IPW. Where a variable remained imbalanced (ASMD>0.1) I accounted for it by including the non-balanced covariate in the subsequent outcome analyses.

Outcome model

I used two different strategies to model outcomes. The first was a Poisson model, for 5-year revision, 5-year mortality and 90-day venous thromboembolism (VTE), prosthetic joint infection (PJI) and myocardial infarction (MI). I also used the lead surgeon as a cluster in a random effects model. This is the standard causal inference approach. Although its estimates cannot be interpreted causally, I decided to add an analysis using cause-specific survival for revision and mortality as time to event is also a relevant outcome for most of these events. To account for the uncertainty in the imputation, I analysed each of the 10 imputed data sets separately and the estimates for each imputed dataset were combined using Rubin rules (128).

Results

Imputation Results

Among the 294,556 patients, BMI had 85,940 missing values. Pre-operative OKS 144,404, EQ5D health scale 158,956. BMI had values ranging 10 to 60, pre-operative OKS ranged 0 to 48 and EQ5D general health scale 0 to 100. BMI imputation with linear regression generated some implausible values (18 out of 3 million values), so I truncated them at the previous max and minimums. After imputing missing data

using MICE, BMI, pre-operative OKS and EQ5D general health scale had a similar mean and standard deviation than the original dataset as shown in ***Table 3.9***. There were no differences in medians and IQR.

	Pre-imputation		Post-imputation	
	Mean	**SD**	**Mean**	**SD**
BMI	30.43	5.10	30.44	5.09
Preoperative OKS	19.26	7.68	19.44	6.81
EQ5D general health scale	69.15	19.69	69.28	19.47

Table 3.9. Pre and post imputation mean and SD values of imputed variables

Propensity Score calculation

Logistic regression predicting treatment coefficients for the PS calculation are shown in ***Sup Table 3.2*** for the first imputed dataset. The rest of the imputed datasets had very similar estimates. Strongest predictors of treatment included basal general health, and deprivation. There were differences between PKR and TKR patients in terms of predicted PKR PS, especially in small PS values. ***Figure 3.1*** shows the PS distributions. There was enough overlap to perform PS techniques.

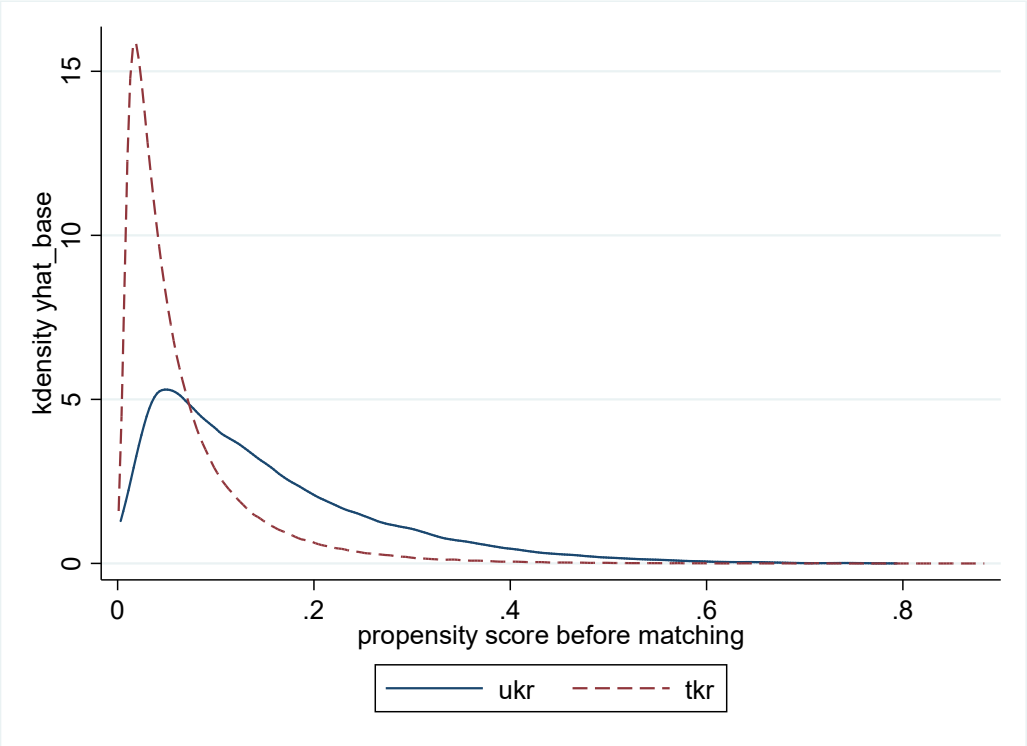


Figure 3.1 – Propensity score distribution in the Safety cohort. 1st Imputation.

Propensity Score Diagnostics

In this section I discuss covariate balance, and specific diagnostic metrics for each method, PS matching, stratification and inverse probability weighting.

Matching

I matched 21,026 PKR patients to 92,071 TKR patients, based on the patient-variable safety cohort propensity score. After matching, 181,459 TKR patients were excluded, as they could not be matched to any PKR patient. The propensity score differences seen in ***Figure 3.1*** disappeared after matching, as shown in ***Figure 3.2***. These patterns were similar in all imputed datasets.

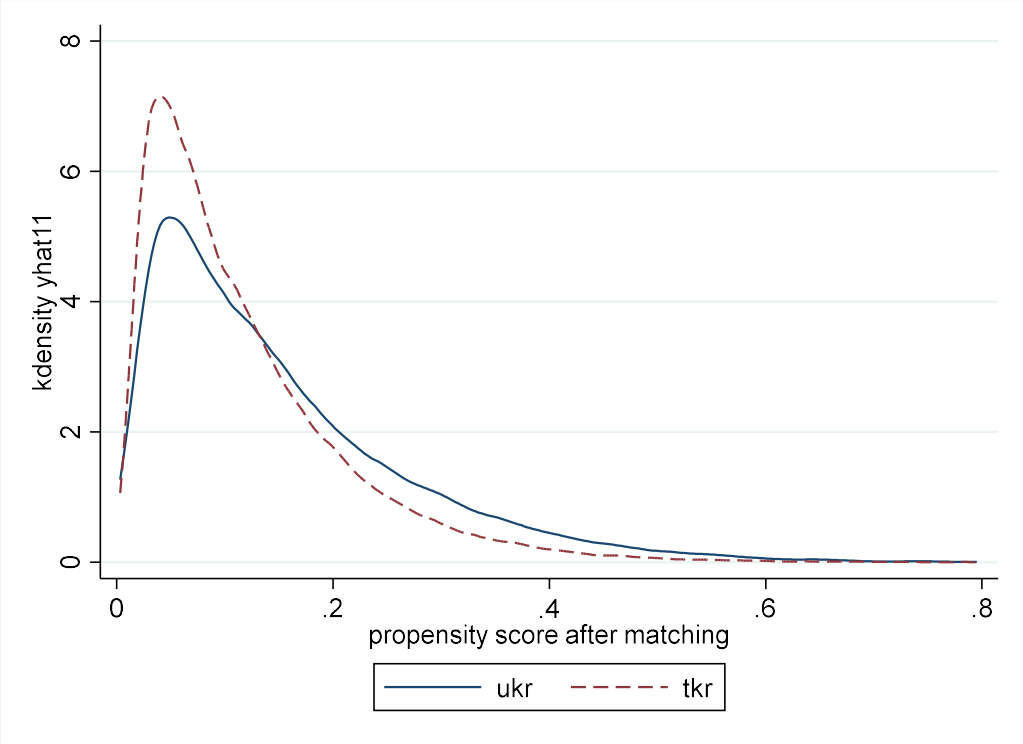
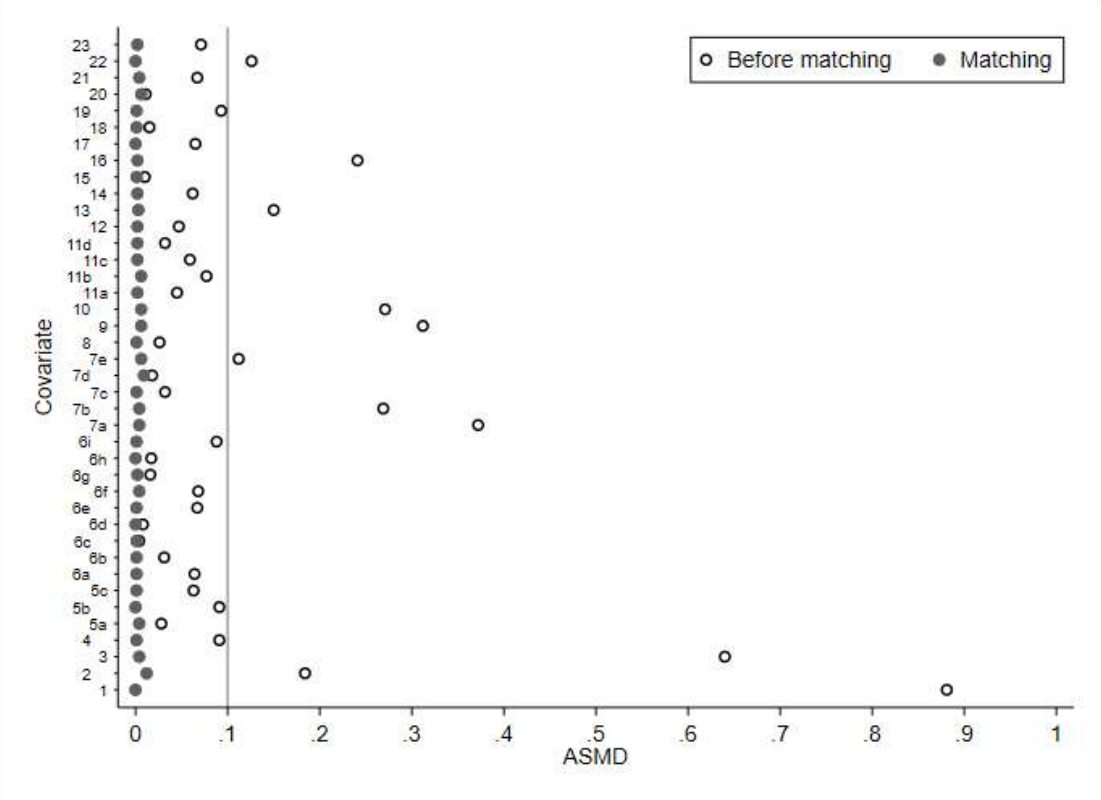


Figure 3.2 – Propensity score distribution after matching in the Safety cohort. 1st Imputation.

Baseline characteristics between PKR patients and TKR patients in the matched cohort were fairly similar. Patient-level covariate balance was excellent after matching, having an ASMDs ≤ 0.1. This can be visually assessed in ***Figure 3.3***. The matched cohort and the full cohort characteristics are shown in ***Sup Table 3.3***, where one can see differences in the TKR patients between the unmatched and matched cohorts, as matching selects TKR patients similar to those who received an PKR.



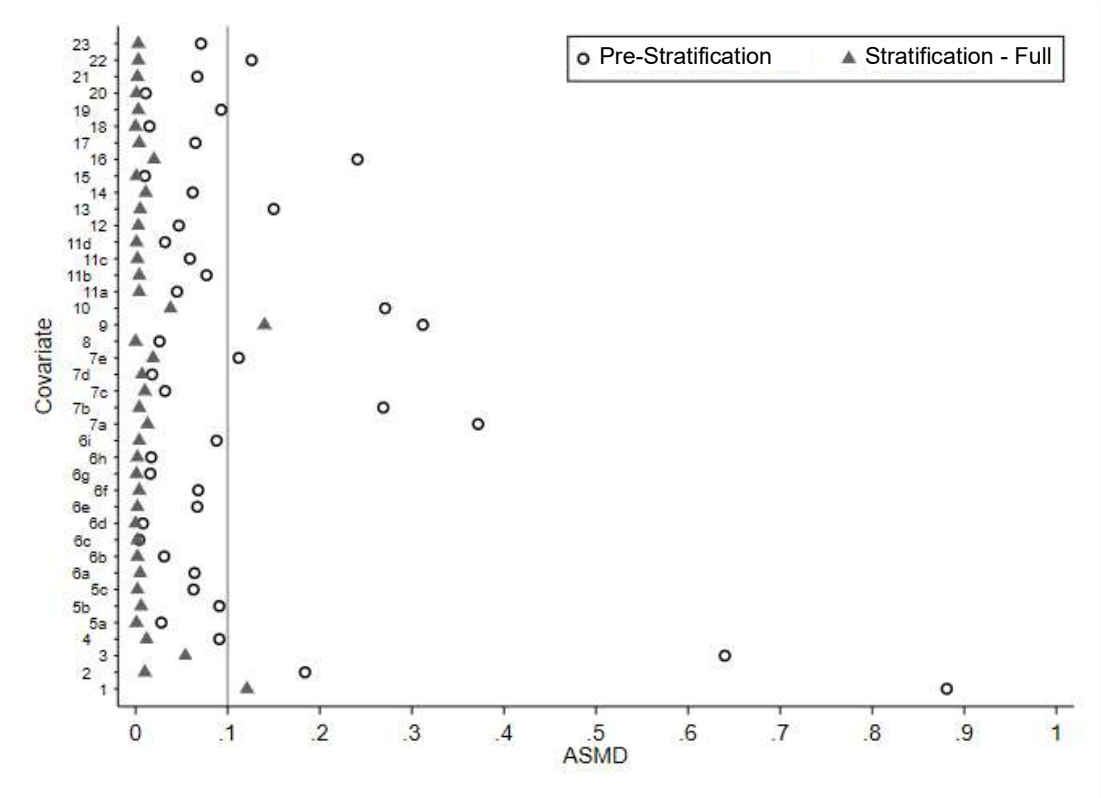
Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 3.3 – Absolute standardised mean difference (ASMD) of PS variables in the OKS cohort after matching. 1st Imputation.

Stratification

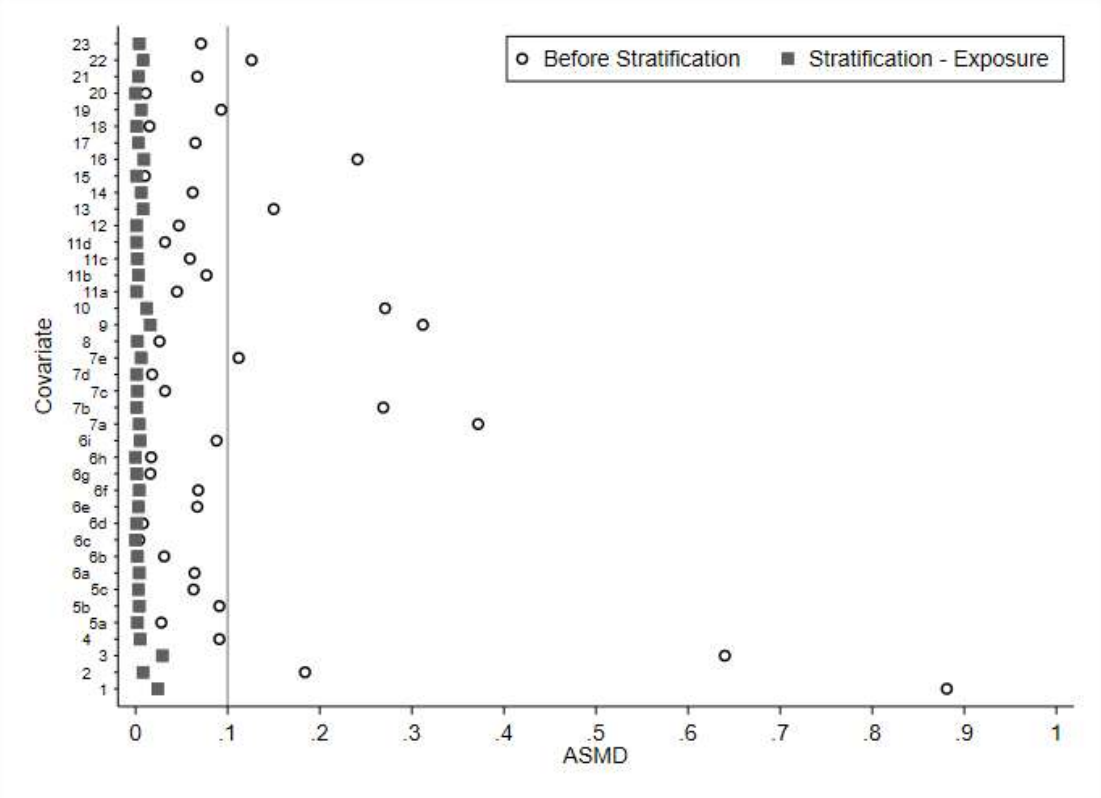
Using the PS based on the patient-level variables I used two different ways to partition the full cohort into 10 strata: based on the deciles of the whole cohort (PSSwhole); and based on the deciles of the PS for the PKR patients (PSSexp). Unlike the previous method, I did not exclude any patient.

Both methods showed good PS distributions as shown in **Sup Figure 3.1**. Here the PSSwhole method performed better than for the effectiveness cohort of **Chapter 2**. There was evidence of imbalance in some strata of gender, age, BMI, IMD, pre-operative general health, EQ-5D index, pre-operative OKS, Charlson Index, and some comorbidities. But, taking into account all strata, overall, the only imbalanced covariate was pre-operative OKS, with an ASMD of 0.14 as seen in **Figure 3.4.** This variable was used as a regressor in the effect estimation for this method. PSSexp performed much better, having no imbalance across any of the strata overall as shown in **Figure 3.5**. Minor imbalances were seen in some strata for a few covariates.



Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 3.4 – Absolute standardised mean difference (ASMD) of PS variables in the OKS cohort after stratification by the whole cohort. 1st Imputation.

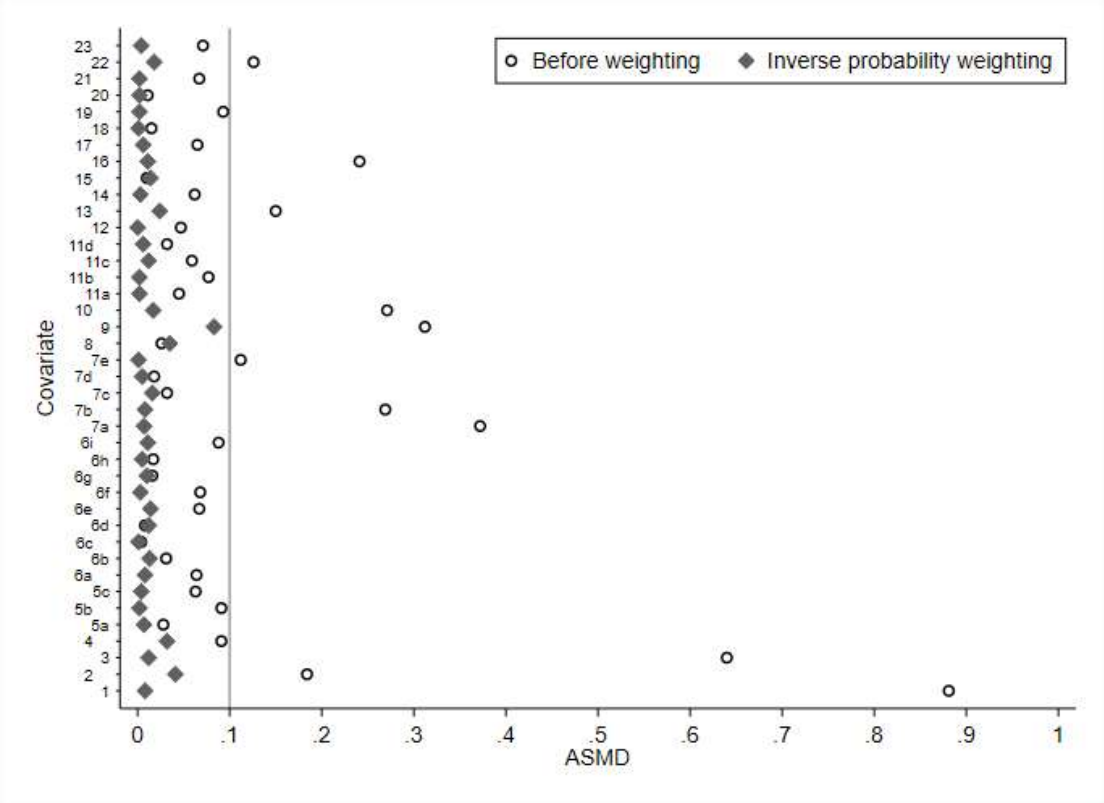


Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 3.5 – Absolute standardised mean difference (ASMD) of PS variables in the OKS cohort after stratification by the exposed cohort. 1st Imputation.

IP Weighting

The resulting pseudo-population after applying the stabilised weights had weights ranging from 0.09 to 23.19 in PKR patients and from 0.93 to 7.93 in TKR patients with a mean weight of 1. IPW managed to balance all observed confounders in this cohort as seen in **Figure 3.6**. When adding lead surgeon as a random effect variable in the PS model, stabilised weights ranged from 0.92 to 57.89 in TKR and from 0.12 to 118.65 in PKR. This approach did not achieve balance in age (ASMD=0.20), or in cardiovascular disease (ASMD=0.12). Neither did it achieve balance for propensity score (ASMD= 0.24). This is shown in **Figure 3.6.**



Note: 1 Overall propensity score; 2 Males; 3a Rural index urban (≥10,000); 3 Age; 4 Body mass index; 5a Town and fringe; 5b Village; 5c Isolated; 6a Less deprived 10-20%; 6b Less deprived 21-30%; 6c Less deprived 31-40%; 6d Less deprived 41%-50%; 6e More deprived 10-20%; 6f More deprived 21-30%; 6g More deprived 31-40%; 6h More deprived 41-50%; 6i Most deprived; 7a General health=1; 7b General health=2; 7c General health=3; 7d General health=4; 7e General health=5; 8 Pre-operative quality of life measure (EQ-5D); 9 Pre-operative OKS; 10 ASA=2, mild diseases; 11a Charlson index=1; 11b Charlson index=2; 11c Charlson index=3; 11d Charlson index=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, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 3.6 – Absolute standardised mean difference (ASMD) of PS variables in the OKS cohort after IPW. 1st Imputation.

Main Results

Revision

As shown earlier, 852 (4%) PKR Patients and 4,090 (1%) TKR patients underwent revision surgery within 5 years of the index procedure. Of these, 852 and 1,383 were included in the PSM matched PKR and TKR cohorts respectively.

As seen in **Table 3.10**, all the methods showed an increase in risk, more than double, of 5-year revision for PKR when compared to TKR patients. The most extreme results were achieved with PSSwhole (adjusted by pre-operative OKS) and PSSexp, showing an approximate 3-fold increase, with RR 3.07 95%CI (2.80 to3.36) and 2.88 95%CI (2.63 to 3.15) respectively. Unadjusted regression achieved the same point estimate as Stratification by the exposed, 2.88 95%CI (2.66 to 3.13). IPweigthing and PS matching had lower estimates of 2.17 (1.93 to 2.45) and 2.09 (1.87 to 2.34) respectively. Similarly, PS adjusted, and fully adjusted regression had results of 2.18 (1.99 to 2.37) and 2.08 (1.91 to 2.27).

	Poisson , RR				Weibull PH , csHR			
	5y Revision		5y Death		5y Revision		5y Death	
	Est	95%CI	Est	95%CI	Est	95%CI	Est	95%CI
IPW	2.17	(1.93, 2.45)	0.62	(0.55, 0.70)	2.20	(1.95, 2.48)	0.63	(0.56, 0.72)
PS match	2.09	(1.87, 2.34)	0.64	(0.57, 0.71)	2.20	(1.97, 2.46)	0.69	(0.62, 0.76)
PSS whole	3.07	(2.80, 3.36)	0.48	(0.44, 0.53)	2.90	(2.65, 3.17)	0.47	(0.43, 0.51)
PSS exp	2.88	(2.63, 3.15)	0.46	(0.42, 0.51)	2.90	(2.65, 3.17)	0.47	(0.43, 0.51)
Adj	2.18	(1.99, 2.37)	0.64	(0.58, 0.70)	2.28	(2.09, 2.48)	0.67	(0.61, 0.73)
NonAdj	2.88	(2.66, 3.13)	0.46	(0.42, 0.51)	2.90	(2.67, 3.14)	0.47	(0.43, 0.51)
Covariate Adjusted	2.08	(1.91, 2.27)	0.64	(0.58, 0.70)	2.22	(2.04, 2.42)	0.70	(0.64, 0.77)

	Poisson , RR					
	MI		VTE		PJI	
	Est	95%CI	Est	95%CI	Est	95%CI
IPW	0.71	(0.41, 1.23)	0.53	(0.37, 0.77)	0.72	(0.34, 1.50)
PS match	0.87	(0.56, 1.36)	0.52	(0.39, 0.68)	0.55	(0.32, 0.94)
PSS whole	0.59	(0.40, 0.88)	0.49	(0.38, 0.63)	0.69	(0.43, 1.12)
PSS exp	0.59	(0.40, 0.88)	0.48	(0.37, 0.61)	0.65	(0.41, 1.06)
Adj	0.78	(0.52, 1.17)	0.53	(0.41, 0.69)	0.59	(0.36, 0.98)
NonAdj	0.59	(0.40, 0.87)	0.48	(0.37, 0.61)	0.65	(0.40, 1.07)
Covariate Adjusted	0.80	(0.53, 1.20)	0.52	(0.40, 0.68)	0.65	(0.40, 1.08)

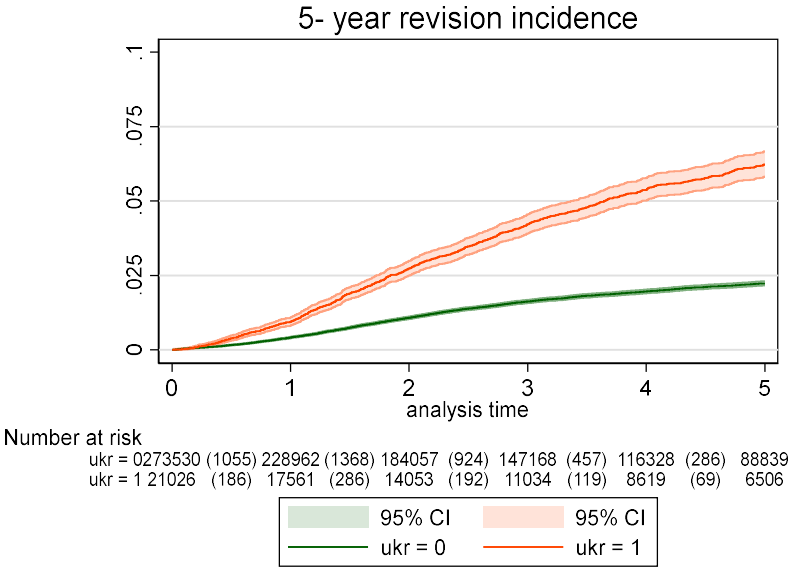
Note: RR: Risk Ratio; PH: Proportional Hazards; IPW: inverse probability weighting; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposure cohort; PS match: propensity score matched; Adj: propensity score adjusted; NonAdj: Crude estimate; Covariate Adjusted: Adjusted by all covariates.

Table 3.10 – Association between PKR (compared to TKR) and 5-year revision, 5-year mortality, 3-month venous thromboembolic events, 3-month prosthetic joint infection, and 3-month myocardial infarction in the 5 shortlisted PS analyses. Mixed-effects Poisson and Weibull proportional hazards regressions with cause specific hazard ratios.

Figure 3.7 shows the unadjusted cumulative incidence curves for revision over 5 years of follow up, where PKR has a clear higher rate of revision. Analysing survival with a Cox proportional hazards regression, estimates were almost identical to the Poisson regression.

Death

Within the same time, 517 (2.5%) and 14,260 (5.2%) PKR and TKR participants died. Even after stratification and weighting, the estimates show a reduction to half the mortality in PKR than TKR. This is probably not causal, and further explained in the discussion. IPW yielded an estimate of 0.64 (0.57 - 0.73), similar to matching, with 0.65 (0.59-0.72). Stratification estimates were lower, with stratification by the full cohort showing a RR of 0.50 (0.45 - 0.54), and by the exposure of 0.48 (0.43 - 0.52). Not adjusting yielded an estimate similar to stratification, 0.46 (0.42 - 0.51). PS adjustment and covariate adjustment had similar estimates to matching, 0.65 (0.59 - 0.71) and 0.65 (0.59 - 0.71) respectively. **Figure 3.7** shows the unadjusted cumulative incidence curves for death over 5 years of follow up, where PKR had a much lower rate of death. Analysing survival with a Cox proportional hazards' regression, estimates were almost identical to the Poisson regression.



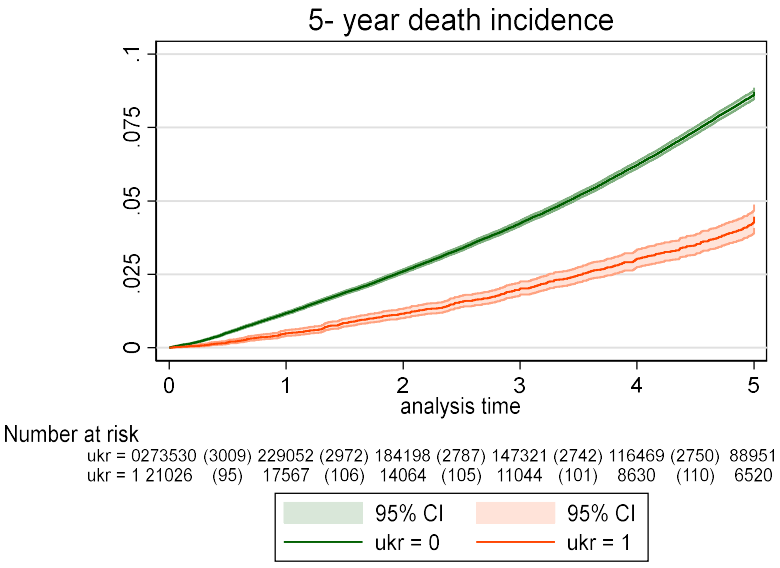


Figure 3.7 - Cumulative incidence of revision and death by type of knee replacement.

Complications

I observed a decrease rate of complications at 90 days. Patients who received PKR had a 30-20% decrease in point rates of MI, but with wide uncertainty. Stratification had the most optimist estimators, 0.59 95%CI(0.40 to 0.88) stratifying for the full cohort, and 0.59 (0.40 to 0.88) stratifying by the exposure, as well as the crude estimate 0.59 (0.40 to 0.87). The rest of estimates had confidence intervals too wide to conclude anything: IPW 0.71 (0.41 to 1.23), matching 0.87 (0.56 to 1.36), Adjusted by PS 0.78 (0.52 to 1.17), and covariate adjusted 0.80 (0.53 to 1.20).

PKR seems to decrease to almost half the incidence of VTE. This is very consistent among methods, with estimates of 0.53 (0.37 to 0.77) for IPW, 0.52 (0.39 to 0.68) for PS matching , 0.49 (0.38 to 0.63) stratifying by the full cohort, 0.48 (0.37 to 0.61) stratifying by PKR, 0.53 (0.41 to 0.69) adjusting by PS, 0.48 (0.37 to 0.61) non-adjusted, and 0.52 (0.40 to 0.68) conditioning on all variables.

For PJI there is a decrease in PKR receivers, again with high uncertainty in some methods. It ranges from 0.55 (0.32 to 0.94) in PS Matching, to 0.72 (0.34 to 1.50) in IP weighting. Stratification by the full cohort showed a RR of 0.69 (0.43 to 1.12), similar to stratifying by the exposure 0.65 (0.41 to 1.06), and to non-adjusting 0.65

(0.40 to 1.07) and covariate adjusting 0.65 (0.40 to 1.08). PS adjustment yielded a RR of 0.59 (0.36 to 0.98).

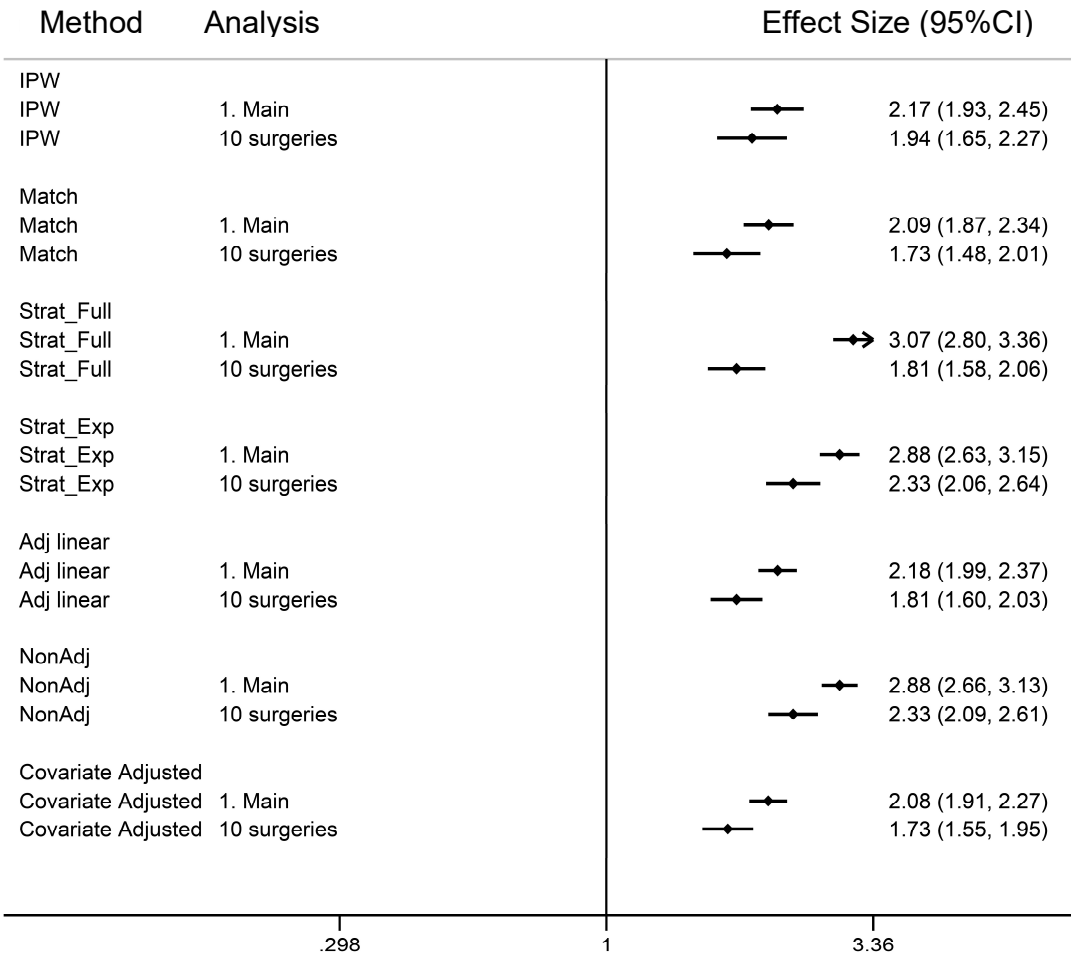
Sensitivity Analyses

I examined the association between PKR (vs TKR) and 5-year revision and death for the cohort of people whose surgery was performed by high volume surgeons, similar to those who were included in the trial. This restricted the analysis to 248,785/273,530 (91.0%) TKR patients, and restricted much more in PKR, to 13,334/21,026 (63.4%) patients. High volume surgeons included 3,001/4,597 (65.3%) surgeons who had done 10+ TKR in the previous year, and only 474/1,462 (32.4%) who had performed 10+ PKRs. The details on this cohort are further explored In **Section 4.1**.

The proportion of patients undergoing revision decreased in patients who received an PKR by a high volume surgeon: 4.1% PKR participants underwent revision surgery in the 5-year window following their index procedure, compared to 3.3% of those operated by surgeons with 10+ PKR in the previous year. Mortality results did not change substantially between the full cohort and the sensitivity one.

All methods achieved excellent balance and diagnostics, except PS stratification by the whole cohort, which imbalanced the propensity score, subsequently adjusted

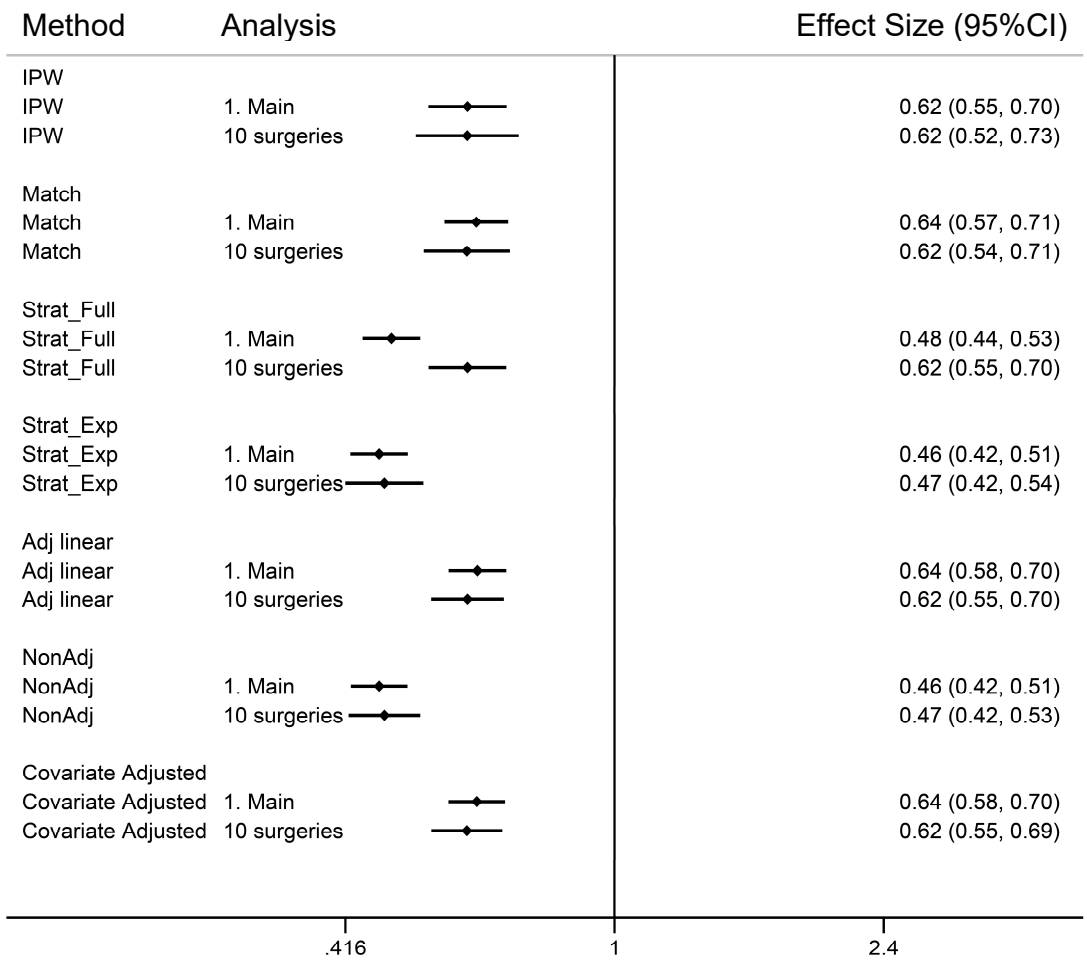
by it in the analyses. Revision in the following 5 years relative risk estimates for PKR went down noticeable when looking at this higher volume cohort, with a RR of 1.94 (1.65 – 2.27) for IP weighting, 1.73 (1.48 to 2.01) for matching, 1.79 (1.58 to 2.06) for stratification by the full cohort and 2.33 (2.06 to 2.64) for stratification by the exposed. Similar results were achieved by non-adjusted analyses with a RR of 1.79 (1.60 to 2.03), covariate adjusting with a RR of 2.33 (2.09 to 2.61) and PS adjustment, that yielded a RR of 1.72 (1.55 to 1.95). This is shown in **Figure 3.8.**



Note: IPW: inverse probability weighting; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposure cohort; PS match: propensity score matching; Adj: propensity score adjusting; NonAdj: Crude estimate; Covariate Adjusted: Adjusted by all covariates.

Figure 3.8 - Forest plot of 5 year revision effect size for PKR vs TKR and each of the validated methods in the whole 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.

Figure 3.9 shows the same analyses for mortality. Restriction to experienced surgeons' cohort did not have a substantial effect on 5-year mortality following surgery for PKR vs TKR. There is still a ~50% reduction in mortality in PKR than in TKR. IPW 0.59 (0.52 - 0.73) , matching 0.60 (0.54 - 0.71), Stratification by the full cohort 0.61 (0.55 - 0.70), by the exposure 0.47 (0.42 - 0.54), not adjusting 0.61 (0.55 - 0.70) PS adjustment 0.47 (0.42 - 0.53) and covariate adjustment 0.60 (0.55 - 0.69).



Note: IPW: inverse probability weighting; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposure cohort; PS match: propensity score matching; Adj: propensity score adjusting; NonAdj: Crude estimate; Covariate Adjusted: Adjusted by all covariates.

Figure 3.9 - Forest plot of 5 year effect size for death in PKR vs TKR and each of the validated methods in the whole 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.

As for complications, the power reduces considerably to be able to interpret these estimates, but results seem similar. The VTE point estimate seems slightly lower, with even more reduction in VTE rates in this cohort of patients, but such differences could be due to uncertainty. The results for these analyses are reported in **Table 3.11**.

	MI		VTE		PJI	
	Est	**95%CI**	**Est**	**95%CI**	**Est**	**95%CI**
IPW	1.00	(0.55 , 1.83)	0.44	(0.27 , 0.71)	0.71	(0.33 , 1.50)
PS match	0.97	(0.58 , 1.61)	0.43	(0.30 , 0.60)	0.78	(0.42 , 1.43)
PSS whole	0.92	(0.58 , 1.46)	0.44	(0.32 , 0.62)	0.79	(0.45 , 1.39)
PSS exp	0.72	(0.46 , 1.13)	0.40	(0.29 , 0.56)	0.83	(0.48 , 1.43)
Adj	0.92	(0.58 , 1.46)	0.44	(0.31 , 0.63)	0.79	(0.45 , 1.41)
NonAdj	0.72	(0.46 , 1.13)	0.40	(0.28 , 0.57)	0.83	(0.47 , 1.44)
Covariate adjusted	0.94	(0.59 , 1.48)	0.44	(0.31 , 0.63)	0.86	(0.48 , 1.51)

Note: IPW: inverse probability weighting; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposure cohort; PS match: propensity score matched; Adj: propensity score adjusted; NonAdj: Crude estimate; Covariate Adjusted: Adjusted by all covariates.

Table 3.11 – Association between PKR (compared to TKR) and 3-month venous thromboembolic events, 3-month prosthetic joint infection, and 3-month myocardial infarction for patients operated by surgeons with +10 surgeries of the same type in the previous year. Mixed-effects Poisson regression.

Discussion

After using propensity score methods, I obtained results similar to those of instrumental variables. PKR seems to have higher revision rates, up to 2-3 times the rates of those receiving a TKR. Results for death rates were consistently lower, with a reduction of about 30 - 50% in mortality among those who received a PKR compared to TKR. Complication rates were also lower, with a half reduction of the incidence of VTE in the 90 days following surgery, and a 30-55% reduction of infection. There is also an uncertain 10-30% reduction on MI, but the study is probably underpowered to make any assertions about it.

These results add to the discrepancies between randomized controlled trials, such as TOPKAT,(71) where little difference is found in terms of revision need for PKR vs TKR, and observational analyses where results are quite similar to the ones seen in the present analyses.(129, 158) This could be explained due to the low power of TOPKAT for this outcome, as the estimates of this Chapter lie inside the trial confidence intervals. The differences between controlled settings and academic centres (the RCT efficacy estimate) and routine practice (effectiveness) can also be playing a part. These differences could be in both individual and environmental

factors, one of them could be the surgeon volume. (159) The sensitivity analyses reinforce this explanation.

Another option could be that I was unable to effectively minimise all confounding. Although I have used state of the art methods, validated for effectiveness outcomes against the results of the trial, the confounding structure may be different for this outcome. This is reinforced by the results of lower mortality in PKR, which raise doubts regarding whether I was able to control for health status in this cohort, and whether mortality acts as a competitive risk.

As for short-term (90-day) complications, PKR seems safer than TKR in terms of having an occurrence of VTE. MI and PJI rates were also lower, but with wide confidence intervals. These findings are supported by results in observational studies (158) and strengthen the case of PKR being first choice for some osteoarthritis patients more prone to suffer these complications. (160)

In the sensitivity analyses, I looked only at the cohort of patients operated by surgeons that did more than 10 surgeries of the same type over the previous year. The patients who received an PKR from high volume surgeons have a slight reduction of revision rates, with estimates of RR of 1.73 to 1.94. The rest of the outcomes show similar rates. This is further discussed in **Chapter 5**, where I

explore the differences in patient characteristics and outcomes according to surgeon volume. This raises methodological issues on how to get to a real estimate of outcomes when surgeon characteristics have a deep impact on them, potentially confounding or interacting with the treatment effects. These issues are explored for Propensity Score methods in **Chapter 5.**

3.2 UTMOST short term complications - a SCCS

Introduction

As shown in the previous section Partial Knee Replacement (PKR) seems to have a decreased rate of complications at 90 days compared to Total Knee Replacement (TKR) such as MI and VTE. Other observational studies have found similar results. Burn et al analysed over 32,000 PKR and more than 250,000 TKR participants after propensity score matching and found a 50% reduction in the risk of 90-day post-operative venous thromboembolism, but no significant reduction in the risk of infection.(156) These results are also consistent with a meta-analysis published in the BMJ in 2019, which found that PKR was associated with a 60% reduction in the risk of post-operative venous thromboembolism compared with TKR. (129) Although these studies use the best methods to control confounding and carefully plan the methodology, there is still potential for bias.

There is also the issue of risk communication. In the previous section I have shown incidence rates of complications and information to calculate risk differences. However, this information is not enough to present information to the patient on the risks of having a knee replacement as the populations that undergo TKR and PKR are very different. To add better context, it is important to ascertain the risk

had participant not undergone an intervention. This could be achieved by running a big placebo-controlled trial. However, that would require an unfeasible number of participants. Another option would be to compare the risk with previous underlying incidence of complications. This approach would be very sensitive to indication bias.

To solve this problem, I propose to use a self-controlled case series (SCCS) approach. This method has been extensively used and was developed to assess the risk of rare complications associated with vaccines.(161, 162) Recently, the method has been extended to other several other uses, where the aim was to quantify the increase of risk of having a certain intervention compared to not having it.(163) I propose to apply this methodology to answer what is the relative increase of risk of myocardial infarction (MI) and venous thromboembolism (VTE) after having a knee replacement (either PKR or TKR).

Methods

Study Design, data sources and study population

I used a self-controlled case series study (SCCS) methodology, a within person comparison.(161, 164) Participants were all adults with complications who

underwent a first primary TKR or PKR in the National Joint Registry (NJR) from 2009 until December 2016. As in the previous chapters, the NJR cohort was linked to Hospital Episode Statistics (HES) in England, which contains information on hospital diagnoses and procedures and sociodemographic data.

Self-controlled case series

The SCCS is a method used to examine the occurrence of acute outcomes and their relationship with a short-term exposure, or a short-term increase of risk after an exposure. The most representative example of this is the investigation of vaccine safety. SCCS compares the chance of a given outcome on the period before the exposure, used as control, and the period after the exposure. (164, 165) This method has several advantages: it is only based on cases of the outcome, providing good estimates of relative incidence and inherently controlling for all confounders as each patient is compared with oneself. Time varying confounders, such as age, can be incorporated into the analyses. However, this approach would only be used if the risk is relatively small and to produce estimates of absolute incidence.

The most important assumption when applying this method is that the outcome does not increase or decrease the likelihood of the exposure or future outcome occurrences. (164-167) This assumption is considered in this study by adding a

wash-out period before the exposure, where an occurrence of the outcomes could have delayed the knee replacement. This assumption could also be broken if the event increases mortality, as is the case with MI. I performed a sensitivity analysis restricting the analysis to those alive the whole study period, to study if it would considerably change the estimates.

Another assumption is that events must be independently recurrent or rare. As events are not rare, I also tested in sensitivity analyses by checking whether considering only the first outcome.

Inclusion and exclusion criteria

I selected patients present in the NJR using the inclusion/exclusion criteria shown in ***Section 2.1.2*** and kept both TKR and PKR patients. I selected patients with an MI or VTE present in HES.

Outcomes and Exposures

The outcomes of interest were myocardial infarction (MI) and venous thromboembolism (VTE). I identified these complications in the year before and in the year after primary surgery. I searched the primary code of the HES dataset for the same ICD-10 specified earlier in this chapter in ***Appendix table 3.1***. I did not

perform this method for prosthetic joint infection, as the method requires that the events can happen before the exposure. The exposure was defined as a first primary knee replacement as identified in the NJR.

Statistical analyses

I described, using mean and SD or percentage, sociodemographic and clinical variables for patients with each complication. I used a SCCS to determine the association of a knee replacement and MI or VTE. **Figure 3.10** shows primary and secondary timings for analyses. Primary analysis compares the risk between month 0 to 3 post surgery vs both month 6 to 12 pre-surgery and 3 to 12 post surgery. I left 6 months wash out before surgery to account for the time where having one of these events could have prevented the patient from getting the knee replacement. On a secondary analysis a second period of interest, 3 to 6 months post-surgery was added to assess if risk of these complications decreases after 3 months.

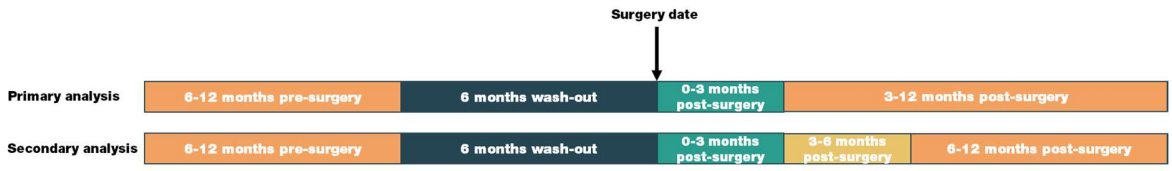


Figure 3.10 – Timings of the primary and secondary analyses of the SCCS.

As for statistical analysis I performed a conditional fixed-effects Poisson regression (xtpoisson with the option fe in STATA) to control for the correlation between periods of the same individual, offset by the natural logarithm of risk periods, producing relative risk (RR) for each post-operatory period. I included age in 5-year groups as a time-varying covariate. To check for possible differential effects between subgroups I performed interactions with gender, ASA, and type of knee replacement (PKR vs TKR). Those with a p-value for the interaction <0.1 were further analysed stratified. As sensitivity analyses, I repeated the analyses taking into account only the first event of a patient for each complication, to check that the assumption of independence of recurrent events was not broken. Similarly, I repeated the analyses excluding those patients who died during the risk window, to check that the event does not significantly increase the probability of death. Patients with more than a knee surgery were kept only for the first surgery and followed up until the earliest of date of subsequent surgery or end of follow up. Patients with bilateral surgeries were not included.

Results

As shown in Figure 2.2, from 868,785 NJR entries, 457,577 were included, as patients with a first knee replacement surgery. Of these, 6,947 patients had an MI 6-12 months pre-surgery or 0 to 12 months post-surgery and 4.227 had a VTE on the same period.

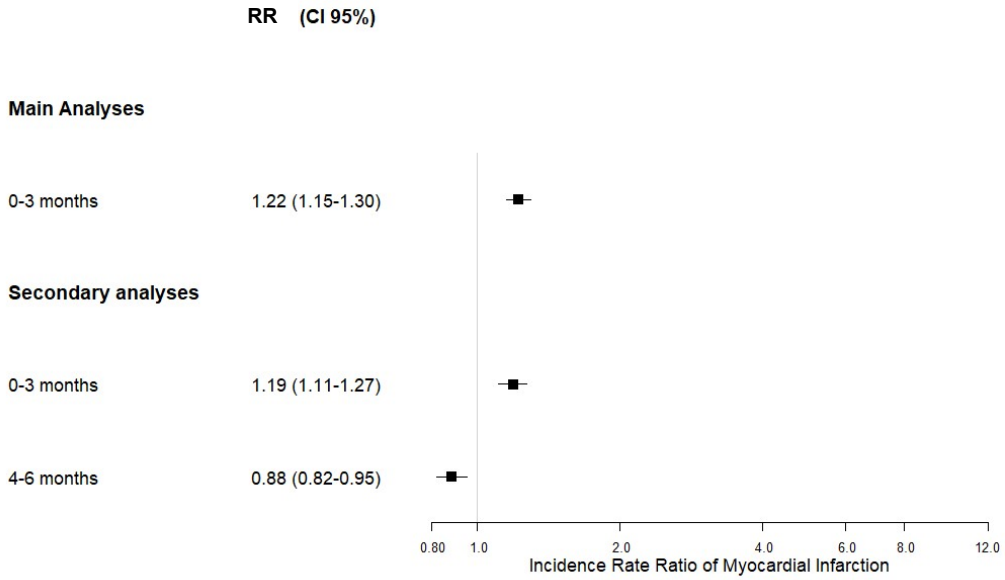
Table 3.12 shows the characteristics of those who had an MI or a VTE before or after surgery and are included in the analysis vs the whole all knee replacements. In general patients who had an MI or a VTE were more likely to have a TKR (94.7% for MI and 96.8% for VTE vs 92.9% in the whole cohort). Although patients with a knee replacement or with a VTE were mostly women (57%), patients with an MI were mostly men (61%). Both MI and VTE patients were most likely to live in more deprived areas. Patients with an MI had higher ASA scores and Charlson Index. Patients with MI or VTE were slightly older, with a mean of 71.3 years old for VTE and 72.4 years old for MI compared to 69.8 years old for all knee replacements. They both had a one point lower pre-operation Oxford Knee Score (OKS) scale, although not clinically significant and generally lower EQ5D measures. VTE and MI patients had generally more comorbidities, especially cardiovascular disease (88% for MI, 71.5% for VTE vs 61% for all knee replacements).

Stage 1	All Knee Replacements		had an MI		had a VTE	
N(%) or mean (SD)	(n=457,577)	%*	(n=6,947)	%	(n=4,227)	%
Knee Replacement						
TKR	425,284	93	6,577	95	4,090	97
PKR	32,293	7	370	5	137	3.2
Gender						
F	258,904	57	2,737	40	2,410	57
M	198,673	43	4,210	60	1,817	43
Rural Index						
Urban	342,734	75	5,305	76	3,230	76
Town and fringe	54,020	12	801	12	464	11
Village	42,924	9	606	9	386	9
Isolated	17,899	4	235	3	147	4
IMD						
Least deprived 10%	47,670	10	690	10	382	9
Less deprived 10-20%	51,892	11	782	11	450	11
Less deprived 20-30%	52,522	11	764	11	419	10
Less deprived 30-40%	53,237	12	788	11	462	11
Less deprived 40-50%	51,744	11	777	11	468	11
More deprived 10-20%	35,832	8	578	8	391	9
More deprived 20-30%	39,674	9	614	9	387	9
More deprived 30-40%	44,266	10	715	10	431	10
More deprived 40-50%	49,216	11	708	10	484	12
Most deprived 10%	31,524	7	531	8	353	8
ASA						
P1 - Fit and healthy	45,196	10	294	4	275	67
P2 - Mild disease not incapacitating	337,307	74	4,184	60	3,042	72
P3 - Incapacitating systemic disease	73,680	16	2,389	34	893	21
P4 - Life threatening disease	1,394	<1	80	1	17	<1
Charlson Comorbidity						
0	280,386	61	2,690	39	2,334	55
1	111,792	24	2,106	30	1,122	27
2	39,072	9	1,061	15	409	10
3	15,754	3	562	8	198	5
4	10,573	2	528	8	164	4
Age	69.8	9	72.4	9	71.3	9
BMI	30.8	5	30.5	5	31.8	6
PROMS pre-operative OKS	18.5	8	17.6	8	17.1	8

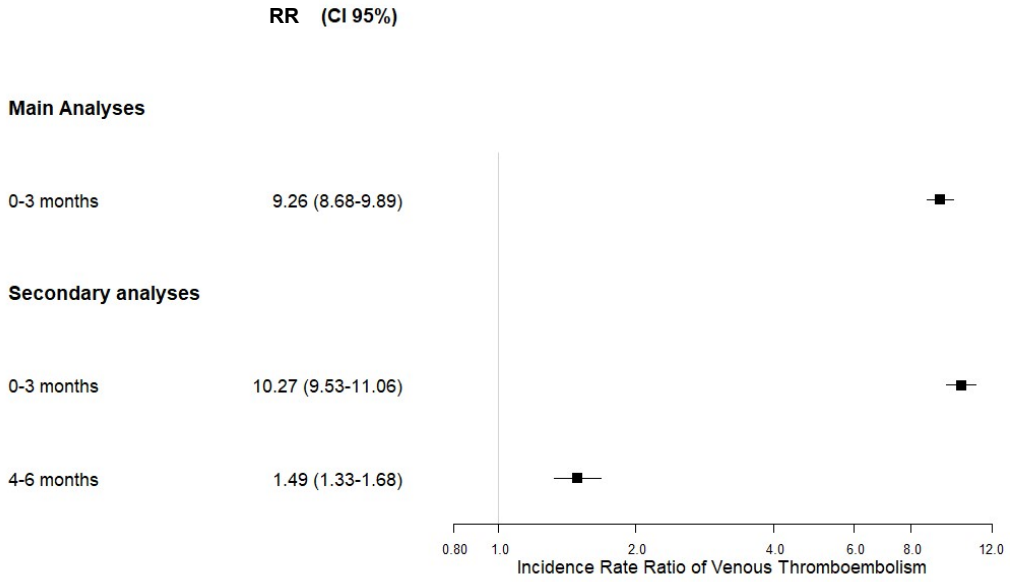
Stage 1	All Knee Replacements		had an MI		had a VTE	
N(%) or mean (SD)	(n=457,577)	%*	(n=6,947)	%	(n=4,227)	%
EQ5D general health scale	39	32	35.8	32	34.7	33
EQ5D index						
Excellent	171,264	72	2502	68	1493	69
1	2,180	1	18	1	20	1
2	15,993	7	197	5	159	7
3	29,533	12	482	13	266	12
4	15,124	6	391	11	193	9
Poor	2,646	1	95	3	31	1
Gastrointestinal Disease	102,039	22	2,134	31	1,165	28
Other Joint Problems	99,904	22	2,020	30	1,098	26
Mental Health	53,387	12	960	14	488	12
Respiratory Diseases	77,172	17	1,652	24	988	23
Cardiovascular Diseases	279,241	61	6,098	88	3,024	72
Thyroid Problems	38,648	9	644	9	425	10
Foot, hip, spinal pain	31,905	7	528	8	332	8
Coxarthrosis	15,667	3	259	8	166	4
Neurological Disorders	36,289	8	785	11	423	10
Other Arthrosis	28,593	6	790	11	369	9
Polyarthrosis	29,083	6	471	7	314	7
Spondylosis	15,603	3	312	5	180	4

Table 3.12 – Baseline patient-level characteristics for patients who received a knee replacement surgery, and who had a VTE or an MI.

Results of the main and secondary analyses are shown in **Figure 3.11**. Knee replacement was associated with having an MI in the 90 days after the surgery, with a 1.22 95%CI (1.15 to 1.30) increase in the incidence rate when compared the incidence rate prior to the surgery. The incidence rate of VTE was higher post knee replacement when compared to that prior surgery with an RR of 9.26 95%CI (8.68 to 9.89). In secondary analyses, the increase in the 90 days after surgery for both events remained high and fell on 90 to 180 days in both cases: 0.88, 95%CI(0.82 to0.95) for MI and 1.49, 95%CI(1.33 to 1.68).



Myocardial infarction



Venous Thromboembolism

Figure 3.11 – Forest Plot of primary and secondary analyses for the risk of MI and VTE 90 days (primary analysis) or between 90 and 180 days (secondary analysis) after surgery.

Table 3.13 shows the results of the sensitivity analyses. Patients with more than 1 MI represent 40% (2,669 patients), and with more than 1 VTE 30% (1,254 patients). Patients who died in the risk window were 191 (3%) for MI and 105 (2%) for VTE. There are minimal differences both in the primary and secondary analyses in terms of the estimates after repeating the analyses only considering the first event for each patient and after excluding the patients who died in the risk window. This suggests that assumptions of the SCCS method hold.

		Primary Analyses	*Secondary Analyses*	
Outcome	***Sensitivity analyses***	***0 – 3 months***	***0 – 3 months***	***3 – 6 months***
MI	***Original***	*1.22 (1.15, 2.30)*	*1.19 (1.11, 1.27)*	*0.88 (0.82, 0.95)*
	First event	*1.24 (1.16, 1.32)*	*1.21 (1.13, 1.29)*	*0.89 (0.82, 0.96)*
	No death	*1.33 (1.16, 1.32)*	*1.31 (1.24, 1.37)*	*0.93 (0.88, 0.99)*
VTE	***Original***	*9.26 (8.68, 9.89)*	*10.27 (9.53, 11.06)*	*1.49 (1.33, 1.68)*
	First event	*9.50 (8.89, 10.15)*	*10.63 (9.85, 11.45)*	*1.53 (1.36, 1.73)*
	No death	*8.88 (8.40, 9.38)*	*10.10 (9.48, 10.75)*	*1.62 (1.47, 1.79)*

Figure 3.13 – Results of the primary and secondary analyses for the risk of MI and VTE 90 days (primary analysis) or between 90 and 180 days (secondary analysis) after surgery for each selected stratum.

There were no strong interactions between the risk window and ASA or gender, with p-values 0.2 and 0.6 respectively in the primary analyses for MI. The interaction with type of surgery had a p-value < 0.1. As for VTE, gender, ASA and PKR all had p-values lower than 0.1 for the interaction.

Table 3.14 shows the results of the stratified analyses for these outcome-variable pairs. The rate of MI was increased in the 3 months after the surgery in the TKR group, with an RR of 1.25 (1.17-1.33). The rate of VTE was also greatly increased in the TKR group, RR of 9.43 (8.82-10-08), and for patients receiving PKR, with an RR of 5.51 (3.87 – 7.84). ASA patient groups also had a differential increase of VTE incidence after a surgery, with an RR of 9.72 (9.03, 10.48) for ASA 1-2 and 7.80 (6.79, 8.96) for ASA 3-4. There was little difference in gender, RR of 9.72 (8.91-10.61) in women vs 8.69 (7.87-9.60) in men.

		Primary Analyses	*Secondary Analyses*	
Outcome	***Strata***	***0 – 3 months***	***0 – 3 months***	***3 – 6 months***
MI	***Surgery Type***			
	PKR	*0.73 (0.53, 1.02)*	*0.71 (0.51, 1.00)*	*0.86 (0.63, 1.17)*
	TKR	*1.25 (1.17, 1.33)*	*1.22 (1.14, 1.30)*	*0.88 (0.82, 0.95)*
VTE	***Gender***			
	Women	*9.72 (8.91, 10.61)*	*10.59 (9.60, 11.69)*	*1.41 (1.20, 1.65)*
	Men	*8.69 (7.87, 9.60)*	*9.83 (8.79, 11.01)*	*1.60 (1.34, 1.90)*
	ASA			
	1-2	*9.72 (9.03, 10.48)*	*10.95 (10.06, 11.92)*	*1.57 (1.38, 1.80)*
	3-4	*7.80 (6.79, 8.96)*	*8.22 (7.05, 9.58)*	*1.24 (0.97, 1.60)*
	Surgery Type			
	PKR	*5.51 (3.87, 7.84)*	*6.13 (4.15, 9.08)*	*1.52 (0.86, 2.68)*
	TKR	*9.43 (8.82, 10.08)*	*10.45 (9.69, 11.27)*	*1.49 (1.32, 1.68)*

Figure 3.14 – Results of the sensitivity analyses for the risk of MI and VTE after surgery for each selected stratum.

Discussion

I showed that knee replacement is associated with a 22% increase of the rate of having an MI in the 90 days after the surgery. Nine times higher incidence in VTE with 90 days after the operation was observed when compared to pre-operation. This increase in rate is lower for patients undergoing PKR, with a 5.5x increase in VTE, and falls quickly after the 90 days for all knee replacements. There seems to be no increase of MI in patients with a PKR.

This study shows an increased rate of MI in the first 3 months after surgery, especially after a TKR. Other research has also found an increased rate of MI after a TKR, with rates of MI between 0.1% to 0.8%, consistent with the estimates on the previous section. (168-170) Regarding timing, a previous study found a 31-fold MI risk increase during the first two weeks after surgery that remained elevated for the following 6 weeks.(171) However, this study matched TKR patients to patients without TKR, an approach prone to huge biases. I overcome the limitations of finding a suitable index date and adjusting confounding by doing a SCCS thus getting much more credible estimates of risk.

I also showed a large increase in the risk of VTE. It is widely known that lower extremity surgery is a risk factor for VTE: more than 2% of patients have a VTE

after a knee replacement. (172) Having a VTE is associated with a higher in-hospital mortality and prolonged hospitalisation.(173) To prevent this, antithrombotic prophylaxis it is routinely prescribed to patients who undergo these procedures.(172) My study results reinforce the needs of communicating the risk to patients and to take every possible measure to prevent it.

The increase in risk seems to be higher in women and ASA 1-2 patients, probably due to lower risk of pre-operative VTE. A striking result is the different increase between TKR and PKR in risk of VTE. Although not directly comparable, the increase in risk of VTE in PKR was almost half of the observed for TKR. This could be due to the healthier and younger profile of PKR patients in this study. But it is in line with previous research and the results of the previous section, which also quantify a ~50% reduction of VTE in PKR compared to TKR.(174)

This study has several strengths. Notably the very large sample size, difficult to achieve in any other type of study. This study involved almost 7,000 patients who had an MI and a knee replacement and more than 4,000 with a VTE and a knee replacement. This has allowed me to also check for interactions and calculate risks stratified by type of surgery, gender, and ASA. This study nicely complements the previous section on comparative safety; there I produced absolute risks, where

here I produce unconfounded relative risks, giving a global view on the risk of complications after a knee replacement. Another strength lies in the fact that this SCCS analysis is a within person method, which greatly reduces the possibility of confounding. This differs from the PS analyses, where residual confounding could remain.

However, this study is not without limitations. The analyses could be affected by a differential capture of MI and VTE before and after surgery, as I only recorded HES data. This is probably more worrying on VTE, where the occurrence before surgery could be underestimated if it was treated ambulatorily. For MI the risk of bias is negligible, as most infarctions would be treated in hospital. There are also two classic limitations of the method. The SCCS assumes independence of events, and of risk of death. Although this is likely not to be completely true, the effect seems negligible: in the sensitivity analyses looking at only the first occurrence of each event, and in the analyses on those who survived the analysis window, the results were very similar to the main ones. The second limitation comes from the assumption that the event does not prevent exposure. To avert this, I have used a 6-month wash-out period, a time where we believe the occurrence of an MI or a

VTE could have prevented the surgery. This period was discussed with clinical experts and tried out with other wash-out periods, being this the optimal.

Chapter 4

Treatment Heterogeneity - Effectiveness and safety of PKR vs TKR in patients with high surgical risk, according to demographics, and to surgeon volume

Treatment heterogeneity or heterogeneity of treatment effect can refer to any difference in the direction or magnitude of the treatment effect that is not random but associated with patient factors, surgeon and hospital characteristics and pre and post operative care.(175) For example, some drugs are thought to work better in men than in women, or some genetic factors can directly make some treatments, such as codeine, inefficacious.

Although clinical trials report effects stratified by some factors, as gender, most of the current evidence is based on an average of the effect for all trial participants. But if the treatment has differential effects, and this average effect does not apply to some groups, we may end up overtreating, undertreating or harming some of the patients.

Clinical trials have limited use to assess treatment heterogeneity as strict eligibility criteria used in the trial would mean study population between groups are homogeneous and therefore less useful to evaluate treatment heterogeneity.

Observational research or pragmatic trials, on the other hand, can provide valuable insight due to potentially larger size and more representative inclusion of participants. In observational research, challenges remain related to confounding by indication and other known sources of bias during the evaluation of treatment effects in observational data.

In this chapter I will perform subgroup analyses to assess heterogeneity of treatment effects on the use of partial knee replacement (PKR). First, I will apply the methods validated in **Chapter 2** to the cohort of patients who would have been excluded from the TOPKAT trial due to severe systemic disease. This will help evaluate if PKR is also useful and harmless as a Total Knee replacement in these patients. Second, I will explore if there are differential effects of PKR according to gender or age. And lastly, I will explore one of the most striking observations from the previous **Chapters:** the observed differences in treatment effect associated with PKR according to the number of surgeries of the same type performed by the treating surgeon.

4.1 UTMOST Stage 2: Applying validated methods to RCT excluded

Introduction

As shown in the TOPKAT trial (71) and in **Chapters 2**, partial knee replacement (PKR) is as effective as Total Knee Replacement (TKR) for treating osteoarthritis. In **Chapter 3** I showed how, although PKR may be associated with an increased risk of revision, the rate of post-operative complications is much lower. This speaks in favour of using this less invasive and potentially safer surgery on patients with severe systemic disease. These patients, who I define as having an ASA<3, represent 1 in every 6 patients undergoing a knee replacement in the UK.(72) This proportion and the absolute number of patients is very likely to increase due to the aging of the population.(176) Furthermore, there will be a growing backlog of these surgeries on the coming years due to the disruption of regular care caused by the COVID/19 pandemic. The benefits for this group of patients can be particularly high, considering that, additional to the decreased rate of complications, PKR also has faster rehabilitation times.(177)

Patients with severe systemic disease are not, however, often included in clinical trials, and they were excluded from TOPKAT.(178) In this section I used routinely collected data as an alternative strategy to compare the outcomes of PKR and TKR

in patients with severe systemic disease. To minimise confounding, I applied the methods that were able to replicate the results of the controlled trial in **Chapter 2**.

Objectives

The aim of this chapter is to evaluate the risks and benefits of PKR compared with TKR for patients with severe systemic disease. To do this, I applied the validated methods in **Chapter 2** to this population.

Methods

Study design, data sources and population

Cohort study. I included all patients who had undergone a first total (TKR) or partial knee replacement (PKR) between 2009 and 2016 in the UK, present in the NJR linked to HES and PROMs. This is further explained in **section 2.1.1.** *Figure 2.2* and **Section 2.1.2** show how I arrived to the main cohort of patients with TKR or PKR. For this section I further restricted the analysis to those with an ASA grade 3 or 4, who would have been excluded from the TOPKAT RCT and have multiple pre-existing comorbidities. As I was not trying to replicate the trial, inclusion criteria were different to those described in *Table 2.1*. I excluded patients who were not eligible for PKR such as those with previous cruciate ligament injury, or

inflammatory arthritis. In terms of target trial, I tried to keep the same timings and considerations explained in **Chapter 2**.

Outcomes

In this section I will be looking at both effectiveness and safety outcomes. As in TOPKAT and in **Chapter 2** I also used the post-operative OKS. OKS is a Patient Reported Outcome Measure (PROM) that measures patient perceived knee pain and function using 12 questions and five possible responses, summing up to a score ranging from 0 to 48, with 48 being the best possible outcome. (92)

For the revision/safety analyses, I looked at the same outcomes explained in **Chapter 3**: 5-year risk of revision, and 5-year mortality. I also looked at the following 90-day post-operative complications: Myocardial Infraction (MI), Venous Thromboembolism (VTE) and Prosthetic Joint Infection (PJI). Code lists for all these outcomes were generated based on previous experience and reviewed by two clinical epidemiologists including myself, and are shown in ***Table 3.1***. Participants were followed up from their index surgery date until the earliest of: end of 2016, revision surgery (for complications analyses only), the date of a surgery for the other knee, death, or end of five years after primary surgery.

Methods to minimise confounding

PS methods, missing data, and diagnostics

Here I used the PS methods that were able to effectively minimise confounding in **Chapter 2:** Propensity Score Stratification, both by the full cohort (PSSwhole) and by the exposure (PSSexp), and Inverse Probability Weighting (IPW). Missing data was imputed using MICE creating 10 datasets with the same technique as in **Chapter 2 and 3**.

I used the same diagnostics as in the previous chapters. I compared patient level characteristics of TKR vs PKR receivers after applying the methods using ASMD with a cut-off of 0.1. Remaining imbalances (ASMD>0.1) in patient-level characteristics were accounted for by adjusting for the non-balanced covariate in the subsequent outcome regression analyses.

Outcome model

For the analysis of difference in post-operative OKS between PKR and TKR I used a multi-level linear regression clustered by lead surgeon as the outcome model. For the analysis of 5-year revision and 5-year death, I used multilevel cause-specific

survival models clustered on lead surgeon. Patients were censored either when they had revision or at death (a competing event).

For 90-day post-operative complications, I compared the cumulative incidence of complications between PKR and TKR. I also used a Poisson model clustered on lead surgeon to calculate Relative Risks (RR). I did not considered mortality as a competing risk in these analyses as mortality rates are very low during these first 90 days.

All outcome analyses were conducted in each of the imputed datasets and their estimators and standard errors pooled using Rubin Rules. (128)

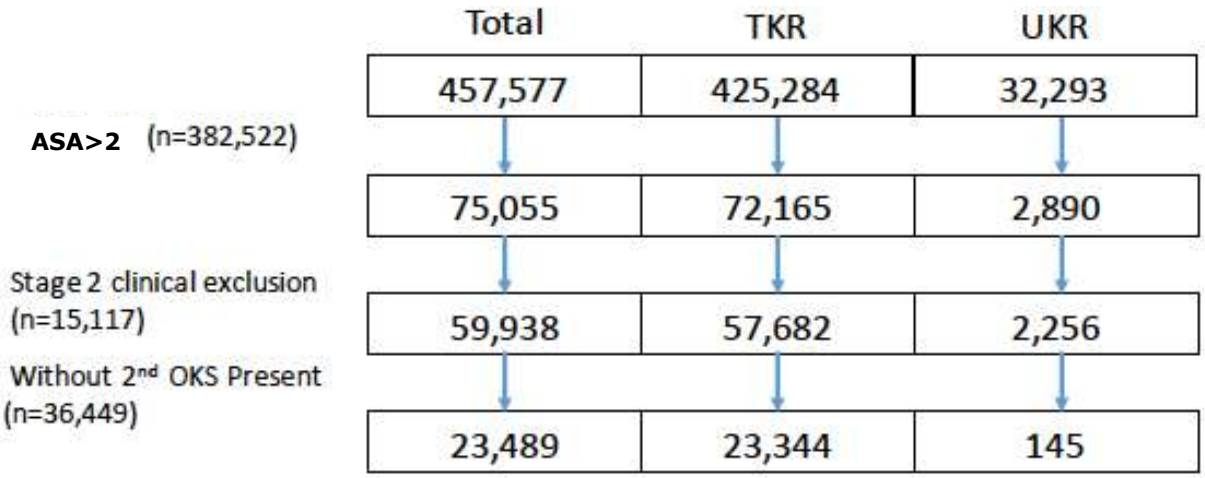
Sensitivity analyses

Sensitivity analyses were conducted for 5-year revision and death by restricting analyses to surgeries performed by lead surgeons with at least 10 surgeries of the same types in the previous year. The low number of patients with registered OKS prevented me from conducting the sensitivity analyses in this cohort. Furthermore, the low number of people with 90-days complications prevented me from doing the sensitivity analyses for these outcomes.

Descriptive results

Cohorts

Following from the source cohort presented in section 2.1.2 of 457,577 patients (32,293 PKR and 425,284 TKR), 383,522 patients (353,119 TKR and 29,403 PKR) had an ASA grade 1-2 and were excluded for this chapter (***Figure 4.1***). I excluded a further 15,117 as they did not fulfil the inclusion criteria. A total of 57,682 TKR and 2,256 PKR recipients were included for revision, death, and complication analyses. Of these, only 145 PKR and 23,344 TKR patients had recorded PROMs and hence were used for the post-operative OKS analyses.



Note: ASA: American Society of Anaesthesiologists physical status classification system; OKS: Oxford Knee Score; TKR: total knee replacement; PKR: partial knee replacement.

Figure 4.1: ASA 3-4 specific eligibility criteria and resulting patient flow.

Patient characteristics

Baseline characteristics for safety and OKS cohorts are shown in ***Table 4.1***. Patients who received a TKR and had OKS data did not differ substantially from the patients in the safety cohort. There were differences between the safety and the OKS cohort in terms of baseline characteristics for PKR patients. PKR patients in the *OKS cohort* had a lower Charlson Comorbidity Index (34% vs 38% have a CI of 0) and they were more likely to live in the countryside (22% vs 16% with rural index of 3 or 4) than PKR patients in the *safety cohort*.

Regarding differences between PKR and TKR, in the safety cohort, patients who received PKR were younger (mean age [SD]: 69[10] vs 73.5[8.9]) and more likely to be men (57% vs 44%). Geographically the patients who received a PKR were more likely to live in the countryside (16% vs 11% with a rural index of 3/4) or in a least deprived area (26% vs 18%). PKR patients were less likely to have a history of other joint problems (19% vs 26%).

The same differences in sex, rurality and deprivation seemed to be true for the OKS cohort. In this cohort, PKR patients had lower Charlson comorbidity scores than those who received TKR (34% vs 39% with Charlson 0). The proportion of

PKR and TKR patients who had a post-operative OKS recorded also differed greatly: 145/2,256 (6.4%) for PKR vs 23,344/57,682 (40.5%) for TKR.

Table 4.1: Baseline patient-level characteristics for patients who received TKR and PKR surgeries with ASA 2 or 3.

Stage 1	Safety cohort				Effectiveness cohort			
N (%) or mean (SD)	TKR (n=57,682)		PKR (n=2,256)		TKR (N=23,334)		PKR (n=145)	
Sex								
Females	32086	56	978	43	12683	54	68	47
Males	25596	44	1278	57	10661	46	77	53
Rural Index								
1	44296	77	1629	72	17626	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								
Least deprived 10%	4784	8	309	14	2026	9	16	11
Less deprived 10-20%	5756	10	274	12	2464	11	20	14
Less deprived 20-30%	6281	11	246	11	2634	11	10	7
Less deprived 30-40%	6298	11	230	10	2683	11	20	14
Less deprived 40-50%	6391	11	268	12	2617	11	18	12
More deprived 10-20%	5400	9	163	7	2011	9	11	8
More deprived 20-30%	5570	10	166	7	2143	9	8	6
More deprived 30-40%	5857	10	231	10	2307	10	18	12
More deprived 40-50%	6205	11	230	10	2616	11	11	8
Most deprived 10%	5140	9	139	6	1843	8	13	9
ASA								
P3 - Incapacitating systemic disease	56625	98	2232	99	22973	98	142	98
P4 - Life threatening disease	1057	2	24	1	371	2	3	2
Charlson Comorbidity								
0	22672	39	863	38	9162	39	50	34
1	18369	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	73.5*	8.9*	69.0*	10.0*	73.5*	8.6*	69.8*	10.2*
BMI	32.6*	6.4*	32.6*	6.1*	32.6*	6.3*	32.6*	6.1*
PROMS pre-operative OKS	16.4*	7.6*	19.2*	8.0*	17.0*	7.6*	19.4*	8.6*
PROMS EQ5D Health Scale	61.8*	20.5*	63.7*	20.5*	62.7*	20.1*	63.7*	22.2*
PROMS EQ5D Index	0.3*	0.3*	0.4*	0.3*	0.3*	0.3	0.4*	0.3*
PROMS General Health								

Stage 1	Safety cohort				Effectiveness cohort			
N (%) or mean (SD)	TKR (n=57,682)		PKR (n=2,256)		TKR (N=23,334)		PKR (n=145)	
0	40968	71	1399	62	16522	71	78	54
1	9563	17	430	19	4052	17	35	24
2 +	7151	12	427	19	2770	12	32	22
Gastrointestinal Disease	16270	28	584	26	6741	29	36	25
Other Joint Problems	15064	26	420	19	6196	27	35	24
Mental Health	7503	13	326	14	2819	12	14	10
Respiratory Diseases	15186	26	622	28	6024	26	37	26
Cardiovascular Diseases	47105	82	1745	77	19269	83	110	76
Thyroid Problems	6354	11	204	9	2630	11	8	6
Foot, hip, 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

Note: SD: standard deviation; ASA: American Society of Anaesthesiologists physical status classification system; BMI: body mass index; IMD: index of multiple deprivation; OKS: Oxford Knee Score; PROM: patient-reported outcome measure.
* mean (SD) is presented.

Surgeon characteristics

In this high risk cohort, surgeon volume followed a similar distribution as in **Chapter 2 and 3**. Volume of surgeons performing TKR, both in the effectiveness and the safety cohort was similar: a median of 47 TKR surgeries in the previous year for safety cohort, and 43 for the effectiveness cohort. As for the PKR, 50% of the PKR patients were operated by surgeons with a volume higher than 15 surgeries in the past year in the safety cohort and 16 in the effectiveness cohort.

Outcome Rates

In the effectiveness subcohort, patients who received an PKR had a slightly higher postoperative OKS than those who had a TKR, mean (SD) 33.1(10.5) vs 31.7(10.5). In the safety cohort, 90 patients who received a PKR (4.0%) and 847 patients who received a TKR (1.5%) had a revision surgery within 5 years after the primary surgery. Mortality was slightly higher in the TKR group, where 6,401 patients (11.1%) died in the 5-year follow-up after surgery, compared to 164 patients (7.3%) following a PKR. As for complications, in the 3 months following the primary operation, 6 patients who had an PKR (0.27%) had a VTE event vs 460 (0.80%) among those who had a TKR. 8 patients (0.35%) in the PKR group had an MI vs

282 in the TKR group (0.49%). 4 PKR patients (0.18%) had a PJI vs 120 (0.21%) TKR patients.

Discussion

There was little evidence of selection bias in our analyses of PROMs, with baseline characteristics overall similar when comparing patients with PROMs available vs the full cohort: in TKR they are very similar, and there are minor differences in the PKR cohort, the most likely to introduce bias, the difference in rurality. This could point to a slightly different geographical uptake of PROMS, but also could be due to the low numbers of the effectiveness cohort.(179) As for the differences between TKR and PKR patients, my findings were consistent with what I have shown in the ASA<3 cohorts in previous chapters: PKR surgeries were used mostly in fitter and younger patients living in a least deprived area.

Surgeon volume does not seem to differ between the effectiveness and safety cohorts, but as seen in the **previous Chapters**, the surgeons performing PKR had much less volume of PKR than those performing TKR.

As for outcomes, post-operative OKS seem to be higher in PKR, similar to what we've seen in low surgical risk patients. Revision rates for PKR and TKR are almost identical to those seen in **Chapter 3**. Mortality rates are much higher, double for TKR and 3 times the PKR of **Chapter 3** patients. This is expected, as these patients have a much higher basal risk of death. Also, the differences in

mortality narrow between PKR and TKR but they are still significantly lower for PKR. The risk of VTE remains similar in these PKR patients compared to less critical patients, but the risk of VTE in TKR increases. Risk of PJI and MI are higher than in **Chapter 3**, as expected but they do not seem to differ much between TKR and PKR.

Propensity score Analyses

Results

Imputation Results

Among the 59,938 patients in the *safety cohort*, BMI had 16,954 (28.3%) missing values. Pre-operative OKS had 30,531 (50.9%) missing values, and EQ5D general health scale had 33,634 (56.1%) missing values. Among the 23,489 patients in the *effectiveness cohort*, BMI had 6,111 (26.0%) missing values. Pre-operative OKS had 269 (1.1%) missing values, and EQ5D health scale had 2,601 (11.1%) missing values. For both cohorts, BMI had values ranging 10 to 60, pre-operative OKS ranged 0 to 46 and EQ5D general health scale 0 to 100.

There were no implausible values generated. After imputing missing data using MICE, BMI, pre-operative OKS and EQ5D general health scale had a similar mean and standard deviation than the original cohorts as shown in ***Table 4.2***. There were no differences in medians and IQR.

Table 4.2: Pre and post imputation mean and SD values of imputed variables.

	Pre-imputation		Post-imputation	
Effectiveness cohort	**Mean**	**SD**	**Mean**	**SD**
BMI	32.6	6.3	32.6	6.3
Preop OKS	17.0	7.6	17.0	7.6
EQ5D general health	62.8	20.1	62.7	20.1
Safety Cohort				
BMI	32.6	6.4	32.6	6.4
Preop OKS	16.4	7.6	16.5	7.7
EQ5D general health	61.8	20.5	61.9	20.5

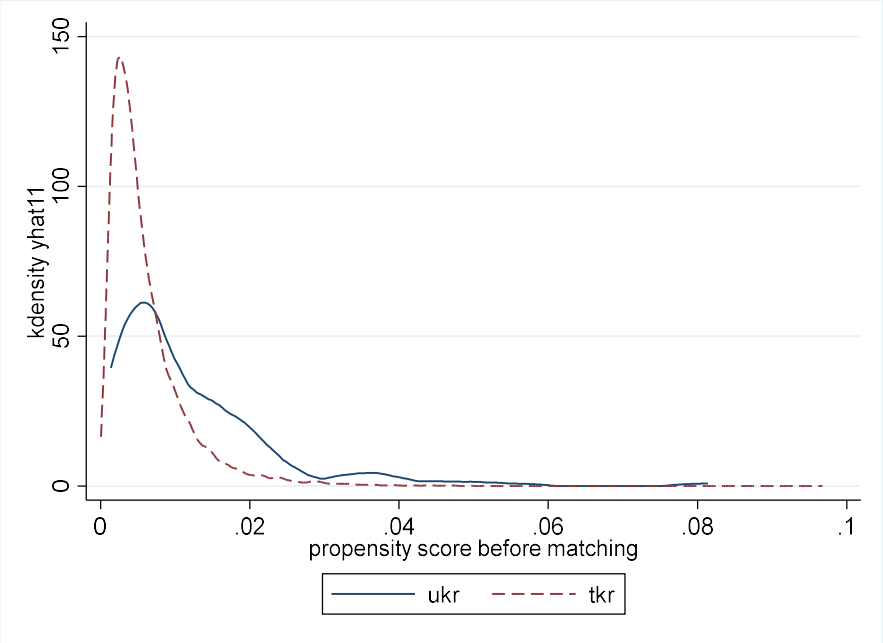
Propensity Score calculation

Annex Table 4.1 shows logistic regression coefficients for the PS in the first imputed dataset. These regressions have been done separately for the effectiveness and safety cohort. The rest of the imputed datasets had very similar estimates.

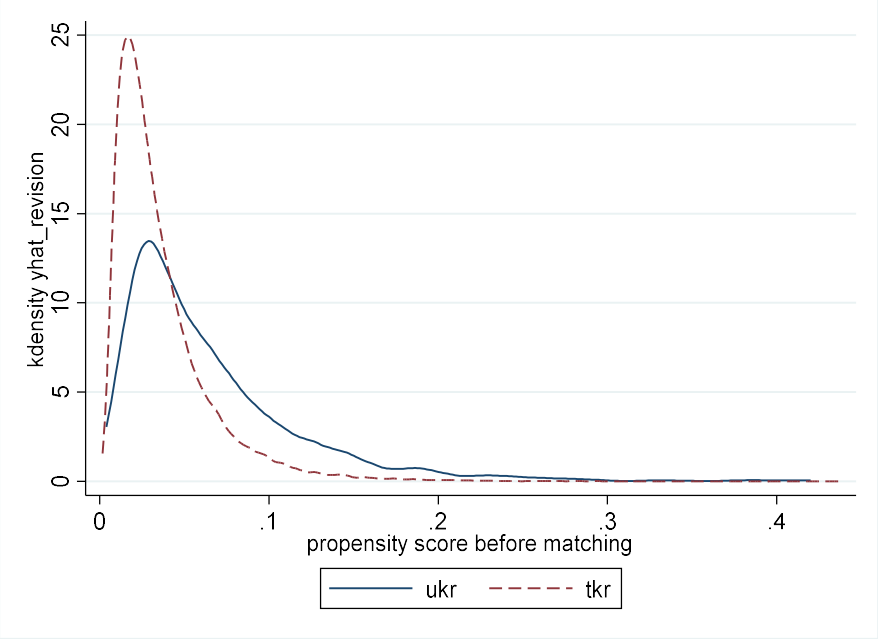
As the numbers of patients in some categories were very low for the effectiveness analyses, I grouped some variables. I recoded Rural-urban into 2 categories: Urban vs Rural. I recoded IMD into groups of similar size: 2 less deprived (Less deprived 30% and less deprived 30-50%), and 2 more deprived (more deprived 30-50%, Most deprived 30%). Charlson Comorbidity Index was recoded to 0, 1, 2, and 3 or more.

In the safety cohort, the strongest predictors of PKR included gender, age, basal general health, comorbidities, and deprivation. In the effectiveness cohort, age, general health, and comorbidities were strong predictors.

PS distributions for PKR and TKR were similar between cohorts and to those of stage 1. There was less overlap between PKR and TKR patients in terms of predicted PKR PS in low PS values. ***Figure 4.2*** shows the PS distributions.



Effectiveness cohort



Safety cohort

Figure 4.2 – Propensity score distribution. 1st Imputation.

Propensity Score Diagnostics

Stratification based on the distribution of PS in the whole cohort

In the effectiveness cohort, PSSwhole resulted in variable imbalance in some strata. This could be due to the differences in PS distribution present in some strata, as shown in ***Appendix Figure 4.1***, probably driven by the small number of patients in some strata. Within each strata covariate balance was not always achieved, for some of the imputed datasets. All covariates in all datasets had an ASMD >0.1 in stratum 1, containing the lowest tenth of propensity score values. This could be to the low number of PKR patients in this stratum, defined by its low probability of having an PKR: there were only 2 (0.09%). Strata 2-5 had between 6 to 17 PKR patients per stratum, with covariate imbalance in some datasets. Covariate balance was much better in strata 6-10, which had a higher probability of PKR and therefore more PKR patients included. But, looking at balance across strata, average ASMD was ≤0.1 for each variable, as shown in ***Figure 4.3***, indicating that good balance was achieved for all individual covariates across strata.

The safety cohort had much better covariate balance within-strata, probably as the number of patients in each strata was greater. The PS distribution for PKR patients in each stratum was similar to that for TKR patients in the same stratum (***Appendix***

Figure 4.2). As in the effectiveness cohort there were less patients in the first strata (lower PS fifths). Overall, good balance was achieved for all covariates across strata, as seen in *Figure 4.4*.

Stratification based on the distribution of PS in the exposed (PKR)

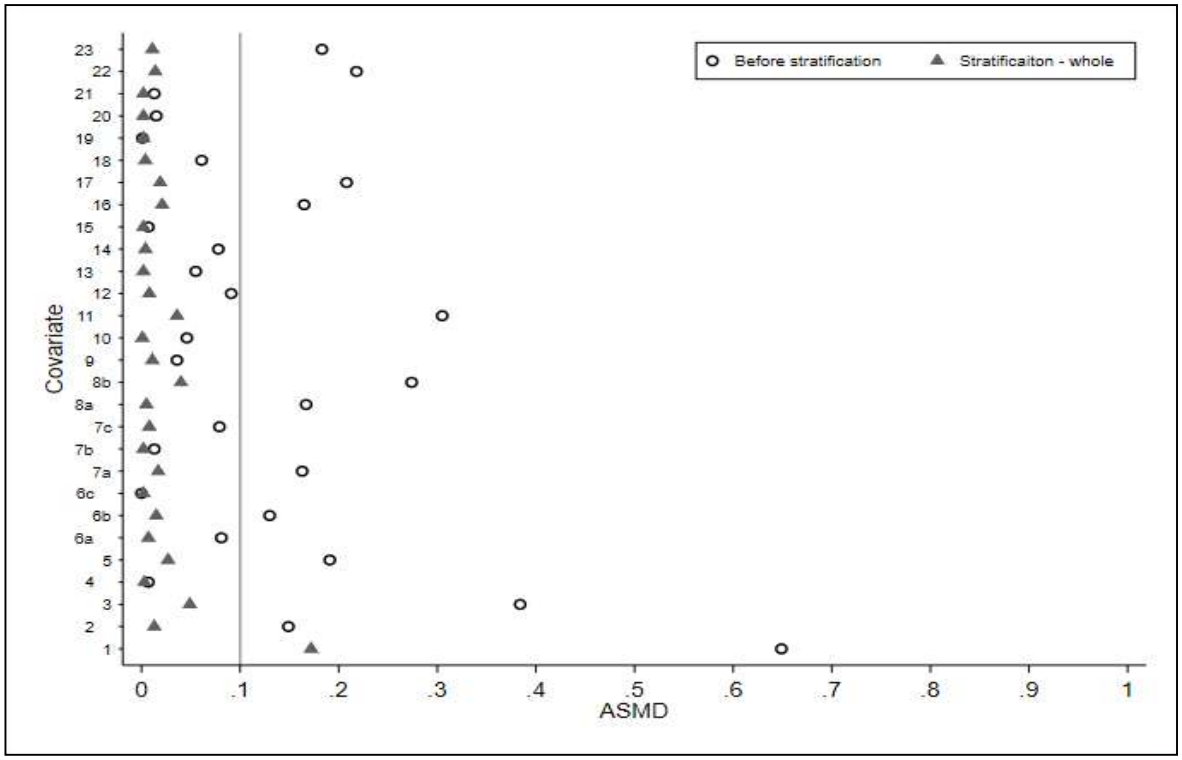
PSSexp bases the stratification only on the distribution of PS in the PKR group, therefore guaranteeing a similar number of PKR patient in each strata. Probably because of this, it obtained better balance in the PS distribution between PKR and TKR when compared to PSSwhole **(See *Appendix Figure 4.1*)**. Using this method, I was able to achieve better within strata covariate balance, although some covariates had an ASMD>0.1 in some strata. Overall there was good covariate balance, as shown in ***Figure 4.3***. This method seems to work much better than PSSwhole in achieving balance when numbers are low, and is hence the preferred option. (126)

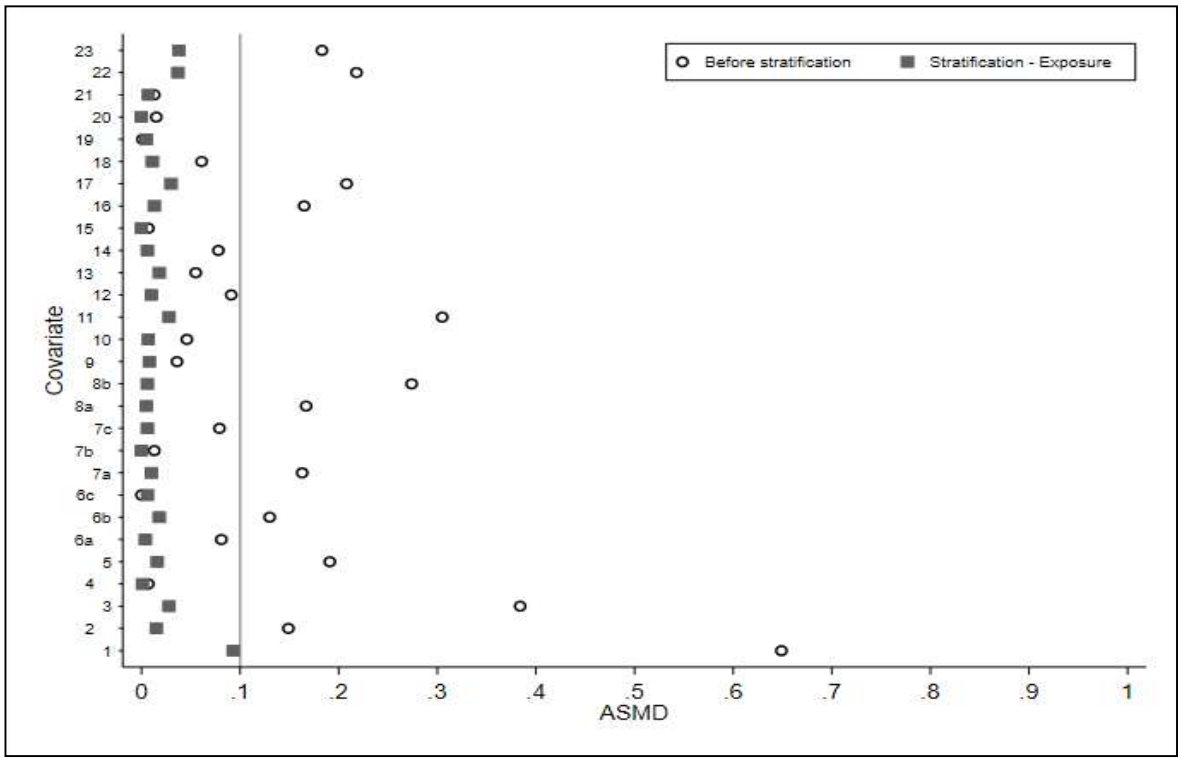
In the safety cohort, PSSexp also achieved a more evenly distributed PS **(*Appendix Figure 4.2*)** between PKR and TKR patients than the PSS_{whole} method. Similarly, as it happened to the effectiveness cohort, within-strata covariate balance was better. This method resulted in fewer variables with an ASMD>0.1 within strata. Overall covariate balance was achieved for all the covariates, with all average ASMD <0.1 **(*Figure 4.4*).**

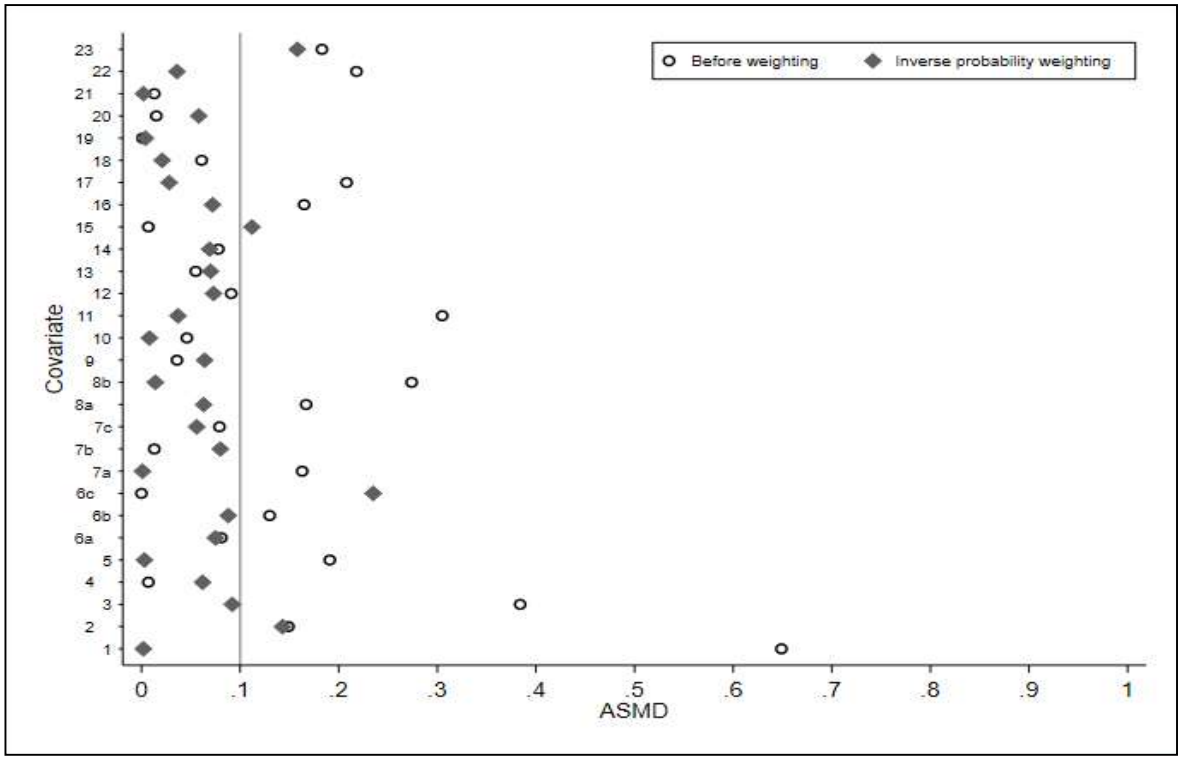
IP Weighting

Inverse Probability Weighting did not produce good balance in the effectiveness cohort. It resulted in a pseudo-population with the 145 PKR patients having a stabilised weight ranging from 0.08 to 4.45 (IQR: 0.38, 1.28). The 23,344 TKR patients had stabilised weights close to 1 (min, 25th percentile, 75th percentile, max: 0.99, 1.00, 1.00, 1.10). But covariate balance was not achieved for all covariates. Respiratory disease, sex, socio-economic deprivation, and history of spondylosis had an ASMD>0.1 after IPW (***Figure 4.3***). PKR patients pseudo population were more likely to have a history of respiratory disease (31% vs 26%), more likely to be men (53% vs 46%), resided in more deprived areas (40% vs 29%) and were less likely to have spondylosis (2% vs 4%) compared to TKR patients. These covariates were adjusted in the outcome analyses.

In the safety cohort, 2256 PKR patients had stabilised weight ranging from 0.09 to 9.45. TKR patients were given weights of around 1. IPW managed to balance all covariates with an ASMD ≤0.1 (***Figure 4.4***).

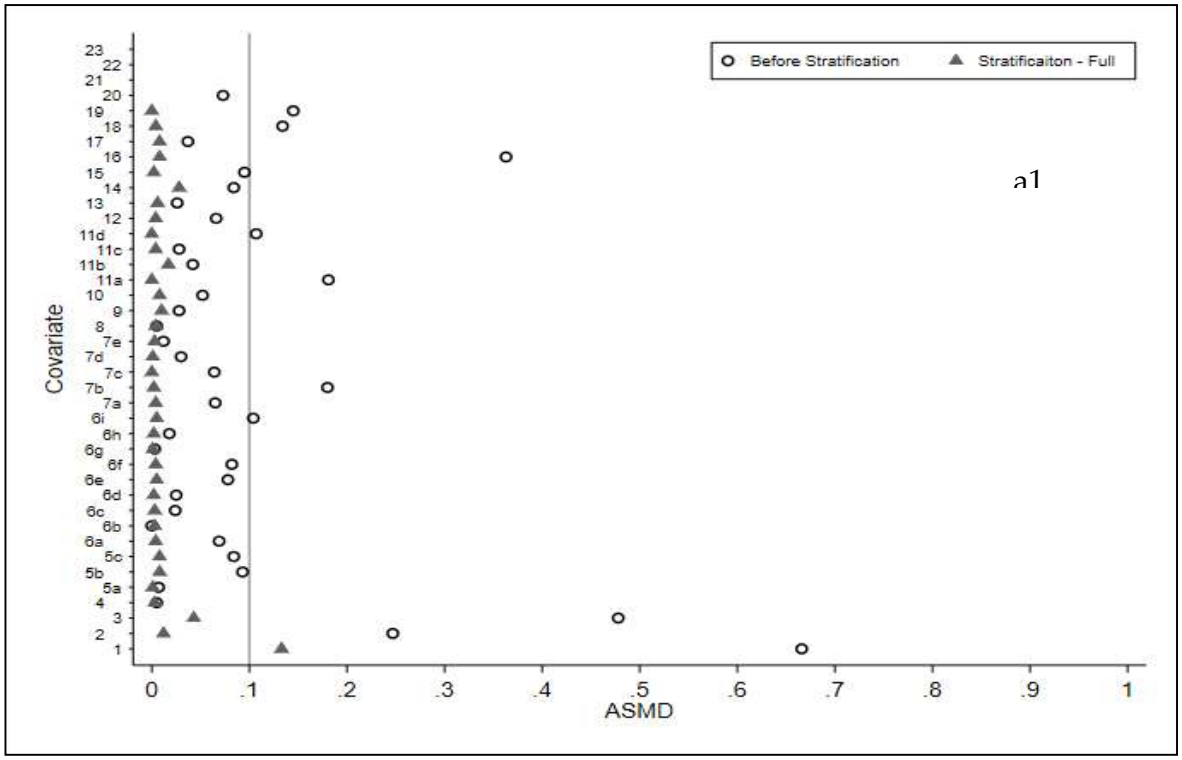

Before stratification
Stratificaiton - whole
Covariate
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8b
8a
7c
7b
7a
6c
6b
6a
5
4
3
2
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
ASMD


Before stratification
Stratification - Exposure
Covariate
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8b
8a
7c
7b
7a
6c
6b
6a
5
4
3
2
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
ASMD

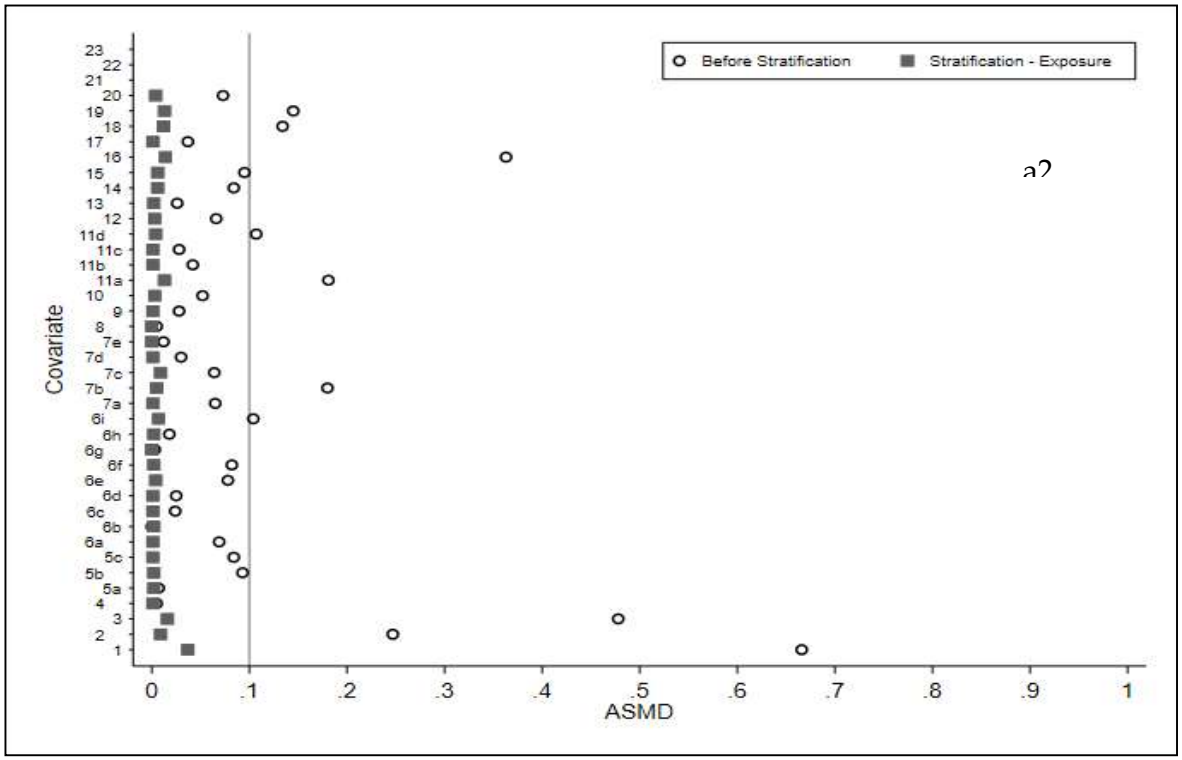


1 overall PS; 2 males; 3 age; 4 Body Mass Index; 5 Town and fringe Village/Isolated; 6a less deprived 20%-50%; 6b more deprived 10-40%; 6c more deprived 41-50%/most deprived;7a Charlson Index=1; 7b Charlson index=2; 7c Charlson index=3+; 8a General health=1; 8b General health=2; 9 ASA=4; 10 Pre-operative quality of life measure (EQ5D); 11 Pre-operative OKS; 12 Gastrointestinal Diseases; 13 Other joint problems; 14 Mental health; 15 Respiratory diseases; 16 Cardiovascular diseases; 17 Thyroid Problems; 18 Foot, hip, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other Arthrosis; 22 Polyarthrosis; 23 Spondylosis.

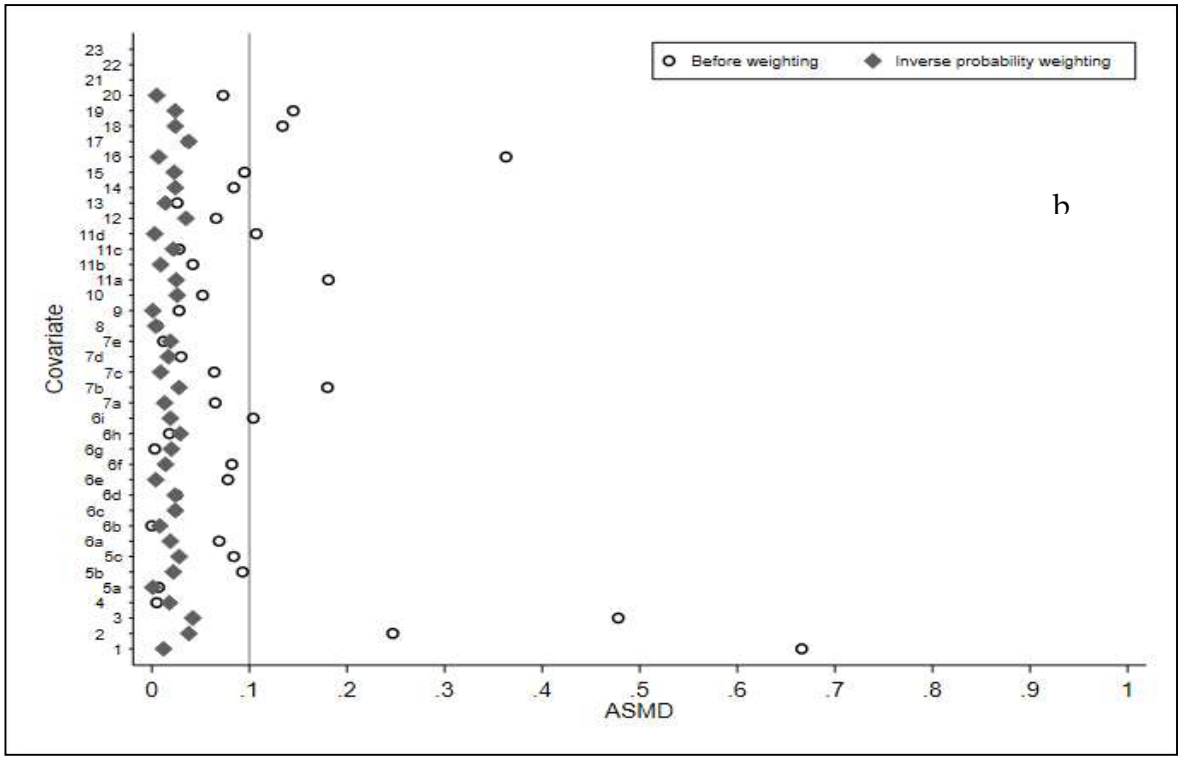
Figure 4.3 – Propensity score distribution in the Effectiveness cohort before and after PS stratification and weighting. 1st Imputation.


Before Stratification
Stratificaiton - Full
Covariate
ASMD
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

a1


Before Stratification
Stratification - Exposure
Covariate
ASMD
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1

a2



1 overall PS; 2 males; 3 age; 4 Body Mass Index; 5 Town and fringe Village/Isolated; 6a less deprived 20%-50%; 6b more deprived 10-40%; 6c more deprived 41-50%/most deprived;7a Charlson Index=1; 7b Charlson index=2; 7c Charlson index=3+; 8a General health=1; 8b General health=2; 9 ASA=4; 10 Pre-operative quality of life measure (EQ5D); 11 Pre-operative OKS; 12 Gastrointestinal Diseases; 13 Other joint problems; 14 Mental health; 15 Respiratory diseases; 16 Cardiovascular diseases; 17 Thyroid Problems; 18 Foot, hip, spinal pain; 19 Coxarthrosis; 20 Neurological disorders; 21 Other Arthrosis; 22 Polyarthrosis; 23 Spondylosis.

Figure 4.4 – Propensity score distribution in the Safety cohort before and after PS stratification and weighting. 1st Imputation.

Main Results

OKS

Table 4.3 shows pre- and post-operative OKS mean and SD for the effectiveness cohort and for the analyses validated in Chapter 3: IPW, PSSwhole, and PSSexp stratification. In the crude description, TKR patients have more than 2 points difference in pre-operative OKS. This difference reduces greatly win the IPW pseudo population, with <0.3 points difference between both (16.99 vs 17.28), but not in stratification. Postoperative OKS was also almost 2 points higher in PKR than in TKR in the crude analysis. Post-operatively, 6-8 months after the operation, OKS was almost twice the pre-operative OKS in the crude population and after weighting and stratification. This shows a massive improvement due to surgery in both TKR and PKR patients.

PSSexp and PSSwhole stratification show how PKR produced a post-operative OKS 1.83 (0.10, 3.56) points higher than TKR. IPW analyses also found an increase of OKS in the PKR patients compared to the TKR patients, but with much greater uncertainty: OKS difference of 1.00 95%CI(-1.28 to 3.27).

		TKR mean (SD)	PKR mean (SD)	Mean difference / Effect size (95% CI)
Crude	Pre-op.	16.97 (7.55)	19.44 (8.58)	-
	Post-op.	32.59 (10.24)	34.57 (10.53)	1.83 (0.15, 3.50)
IPW	Pre-op	16.99 (7.56)	17.28 (8.37)	-
	Post-op	32.60 (10.24)	33.65 (10.87)	1.00 (-1.28, 3.27)
PSS$_{whole}$	Pre-op	16.97 (7.55)	19.43 (8.55)	-
	Post-op	32.59 (10.24)	34.56 (10.53)	1.82 (0.10, 3.56)
PSS$_{exp}$	Pre-op	16.97 (7.55)	19.44 (8.55)	-
	Post-op	32.59 (10.24)	34.57 (10.53)	1.83 (0.10, 3.56)

IPW: inverse probability weighting; PS: propensity score; PSM: propensity score matching; PSS$_{whole}$: PS stratification based on the whole cohort, PSS$_{exp}$: propensity score stratification based on the exposure cohort; PS$_{Alin}$: Propensity score linear adjustment; PS$_{Anonlin}$: Propensity score non-linear adjustment; SD: standard deviation; TKR: total knee replacement; PKR: partial knee replacement.

Table 4.3. Pre- and post-operative Oxford Knee Score (OKS) in the cohort and for each PS method for the severe systemic disease cohort.

Revision

From the safety cohort, 90 (4%) PKR patients and 847 (1.5%) TKR patients had a revision surgery within 5 years of the original surgery. The cumulative risk of revision during these 5 years is shown in ***Figure 4.5***. The incidence rates of revision following PKR and TKR were 13.09 95%CI(10.64 to 16.09) and 4.88 (4.56 to 5.22) revisions per 1,000 patient-years as shown in ***Table 4.4***. This translates into an almost 3-fold increase in revision risk for PKR vs TKR (crude hazard ratio: 2.70 [2.16, 3.37]). This risk was not attenuated after adjustment using the validated methods, with resulting cause-specific HRs (95% CI) of 2.70 (2.15 to 3.38) for PS stratification methods, and of 2.60 (1.94 to 3.47) for IPW.

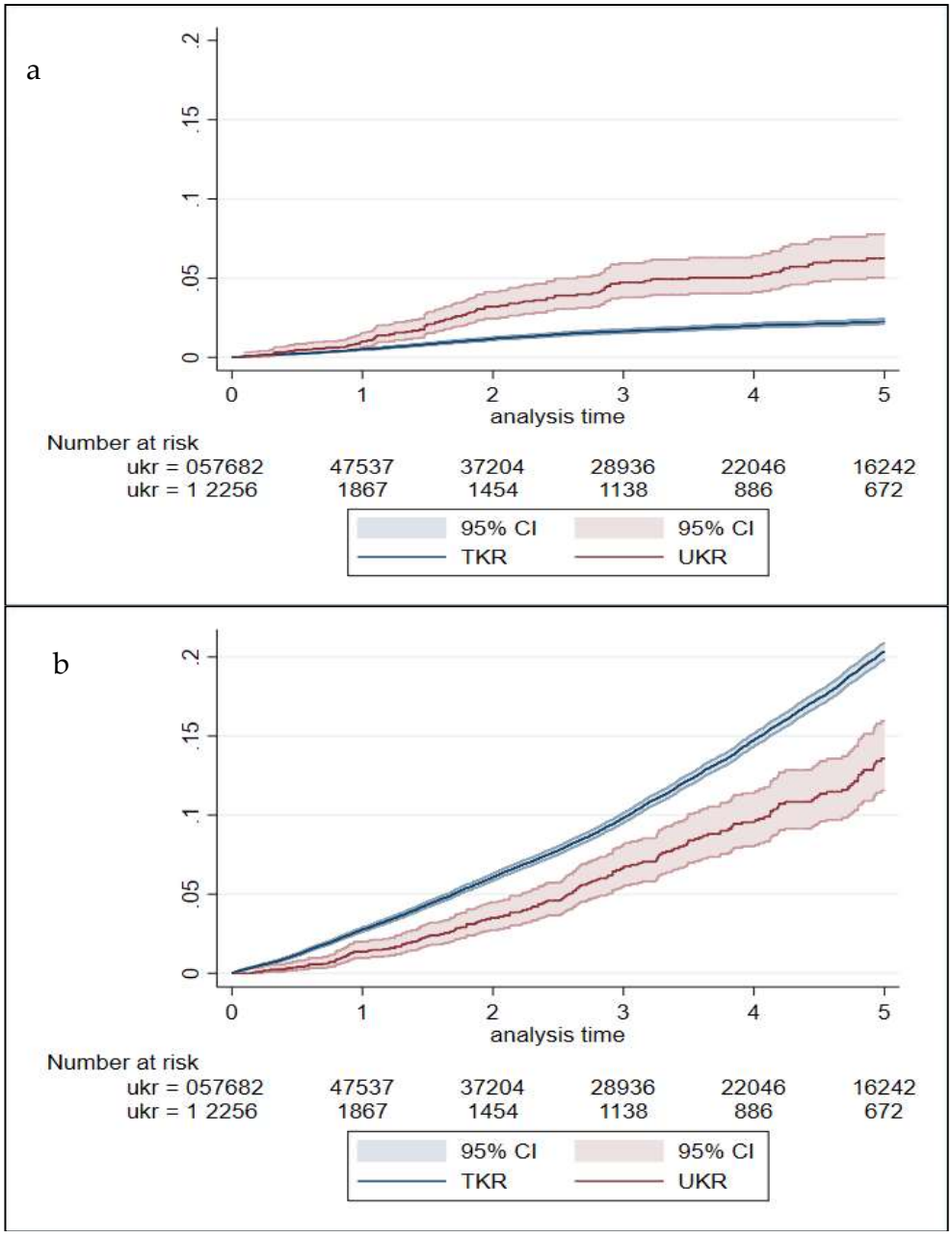
Death

Table 4.4 and *Figure 4.5* also show results for death. The number of patients who died in the 5 years after surgery is significantly larger than patients who had a revision, 6,401 (11%) in TKR and 164 (7%) in PKR. In terms of incidence rates, 36.89 (36.00 to 37.81) died per 1,000 patients-year in the TKR group and 23.85 (20.46 to 27.79) in the PKR group. This amounted to a crude mortality HR 0.64 (0.55 to 0.75) times lower for PKR patients. This decrease in mortality did not attenuate in the stratification analyses, with a csHR of 0.64 (0.55 to 0.75) for both PSSwhole and PSSexp. After using IPW for controlling confounding, the effect was attenuated, with a csHR of 0.83 (0.67, 1.03).

Group	N with event	Person-years	Incidence rates per 1,000 person-years (95% CI)	5-year cumulative incidence	Crude HR (95% CI)	PSS_{whole} HR (95% CI)	PSS_{exp} HR (95% CI)	IPW HR (95% CI)
Revision surgery								
TKR	847	1,735.01	4.88 (4.56, 5.22)	1.47 (1.37, 1.57)	REF	REF	REF	REF
PKR	90	68.77	13.09 (10.64, 16.09)	3.99 (3.21, 4.90)	2.70 (2.16, 3.37)	2.70 (2.15, 3.38)	2.70 (2.15, 3.38)	2.60 (1.94, 3.47)
All-cause mortality								
TKR	6,401	1,735.01	36.89 (36.00, 37.81)	11.10 (10.83, 11.37)	REF	REF	REF	REF
PKR	164	68.77	23.85 (20.46, 27.79)	7.27 (6.20, 8.47)	0.64 (0.55, 0.75)	0.64 (0.55, 0.75)	0.64 (0.55, 0.75)	0.83 (0.67, 1.03)

95% CI = 95% confidence interval; IPW: inverse probability weighting; N = Number of participants with the event of interest; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposed cohort; HR = hazard ratio

Table 4.4. Long-term (5-year) complications after partial knee replacement (PKR) or total knee replacement (TKR)



Number at risk: indicates numbers of TKR/PKR patients who have not experienced any of the outcomes.

Figure 4.5 Cumulative incidence functions of risks of revision (a) and mortality (b) for PKR vs and TKR over 5 years of follow-up

Complications

Table 4.5 shows the results for complications. The rates for 90-day post-operative venous thrombo-embolism (VTE) observed was lower for PKR compared to TKR participants: 2.66 [1.20 to 5.91] vs 7.96 [7.26 to 8.71] per 1,000 persons, resulting in a crude RR of 0.33 (0.15 to 0.75) less incidence of VTE in PKR. These differences were not attenuated after applying PS methods, RR of 0.33 (0.15 to 0.74) for PSSwhole and PSSexp, and a RR of 0.39 (0.16 to 0.96) for IPW.

By contrast, 90-day cumulative incidence of myocardial infraction (MI) were similar between TKR and PKR patients, 4.87 (4.34 to 5.47) and 3.55 (1.77 to 7.07) per 1,000 patients with no differences in the crude RR (0.73 [0.36 to 1.47]). This was not changed after PS stratification methods, but the point estimate turned more in favour of PKR after IPW, 0.64 (0.29, 1.45).

This is very similar to the results for prosthetic joint infection (PJI), with incidence rates of 1.92 (1.60 to 2.32) per 1,000 patients in TKR and 1.77 (0.67 to 4.71) in PKR. The crude RR was of 0.92 (0.34 to 2.50). PS stratification yielded an RR of 0.85 (0.33 to 2.19) and IPW of 0.55 (0.18 to 1.71).

	N	Cumulative Incidence per 1,000 patients (95% CI)	Crude RR (95% CI)	PSSwhole RR (95% CI)	PSSexp RR (95% CI)	IPW RR (95%CI)
Venous thrombo-embolism (VTE)						
TKR	459	7.96 (7.26, 8.71)	REF	REF	REF	REF
PKR	6	2.66 (1.20, 5.91)	0.33 (0.15, 0.75)	0.33 (0.15, 0.74)	0.33 (0.15, 0.74)	0.39 (0.16, 0.96)
Myocardial Infarction (MI)						
TKR	281	4.87 (4.34, 5.47)	REF	REF	REF	REF
PKR	8	3.55 (1.77 ,7.07)	0.73 (0.36, 1.47)	0.73 (0.36, 1.45)	0.73 (0.36, 1.45)	0.64 (0.29, 1.45)
Prosthetic Joint Infection (PJI)						
TKR	111	1.92 (1.60, 2.32)	REF	REF	REF	REF
PKR	4	1.77 (0.67,4.71)	0.92 (0.34, 2.50)	0.85 (0.33, 2.19)	0.85 (0.33, 2.19)	0.55 (0.18, 1.71)

N = Number of participants with the event of interest; Cumul. Incid. = Cumulative Incidence; 95% CI = 95% Confidence Interval; RR = Risk Ratio; PSS_{whole}: propensity score stratification based on the whole cohort; PSS_{exp}: propensity score stratification based on the exposed cohort

Table 4.5. Short term (90-day) complications following PKR vs TKR surgeries

Sensitivity Analyses

Among 57,782 TKR recipients included in the safety cohort, 51,118 (89%) were included in the high-volume surgeon cohort. As for PKR participants, reduction was higher, only including 1449 patients of 2256 amounting to 64%. Details on baseline characteristics for these sub-cohorts are reported later in this chapter in ***Appendix Table 4.2***.

The cause-specific risks of revision were attenuated after restricting the cohort. PS stratification adjusted csHR shifted from 2.70 (2.15 to 3.38) in the whole cohort to 2.17 (1.60, 2.94). Similarly IPW shifted the results from 2.60 (1.94 to 3.47) to 2.33 (1.52 to 3.57). The reduction seen in mortality on PKR patients compared to TKR was also attenuated, from 0.64 (0.55 to 0.75) to 0.89 (0.68 to 1.16) in PS stratification and for both PSSwhole and PSSexp. After using IPW for controlling confounding, the effect was attenuated, with a csHR of 0.83 (0.67 to 1.03).

Discussion

I have provided evidence on the effectiveness and safety of PKR compared of TKR in high morbidity patients. Patients with ASA>3 were not eligible for the PKR trials, and there is little chance a follow-up trial will focus on these patients. These patients represent a large proportion of patients undergoing knee replacement, approximately 16% of knee replacement surgeries.

Although the number of participants in the OKS analyses for PKR was much lower than in TOPKAT, I had a large sample for safety analyses, almost 10 times more PKR and 200 times more TKR patients. There is no RCT that included patients with ASA>3, making a validation through direct comparison as I did in **Chapter 3** impossible. I then used PS stratification and IPW that in **Chapter 1** were more likely to work for knee replacements, and I further evaluated the performance of those methods based on well stablished diagnostics.

PS stratification based on the PKR cohort was the method that achieved the best balance followed by PS stratification by the whole cohort. IPW showed imbalances in some confounders and required double adjustment in the OKS analyses but not in the safety. This is probably due to a much larger population than the patient reported outcome cohort.

Using PSSexp, I showed how PKR seems to have a very similar effectiveness (compared with TKR) to that observed in TOPKAT and in **Chapter 3** for ASA<3, with an average treatment effect of 1.83 (0.10, 3.56) OKS points in favour of PKR. After double adjustment for unbalanced variables, PSSwhole showed a very similar estimate of 1.82 (0.10 to 3.56), and IPW yielded an estimate of 1.00 (-1.28 to 3.27). This effect, although statistically significant for stratification, it is not clinically relevant. (95) This shows how effectiveness of PKR in patients with severe systemic disease and/or substantial functional limitations, as defined by an ASA of stage 3 or above, is similar to those for the general population. It is important to note, that in this case the results of the RCT were a good estimate for this population outside of the RCT. This makes PKR a useful therapy for these patients.

Safety is probably the most relevant outcome for high surgical risk patients, particularly short-term post-operative outcomes. As the two different knee replacements were no different in effectiveness, the decision to perform one or another is likely to be based on safety considerations and revision risks. All the adjusted analyses show a protective effect against venous thromboembolism (VTE) for PKR receivers: they are 3 times less likely to have a VTE compared with TKR

patients. This is a very important finding considering that VTE is the most common post-operative complication for knee replacement. It happened to 8% of TKR patients and just below 3% of the PKR patients of this cohort. Acute myocardial infarction (MI) and prosthetic joint infection (PJI) analyses did not have enough power to reach strong conclusions. In the first 90 days after surgery, 5% of TKR patients and 3.5% of PKR had an MI, and 1.9% of TKR patients and 1.8% of PKR had a PJI.

Although trials have not been powered to detect such differences, these differences have been documented in observational research. For example, in Burn et al (158) we studied over 32,000 and 250,000 PKR and TKR patients across 5 US and UK databases, and found a 50% reduction in the risk of 90-day post-operative venous thromboembolism, but no significant reduction in the risk of infection. This is also consistent with a meta-analysis,(129) which found that PKR was associated with a 60% reduction in VTE compared with TKR.

Another two important outcomes are 5-year revision surgery and mortality. For mortality I found that patients with complex health needs are more likely to die than to undergo a revision surgery. The 5-year cumulative mortality rate was 24% for PKR patients and 37% for TKR patients, and the revision rates were of 13% and

5%, respectively. This translated into an almost 3 times higher revision risk in PKR patients than in TKR patients of revision but a >30% reduction in all-cause mortality. This calls for more research on the potential mortality reduction in this population, to elucidate whether it is a real effect or due to unresolved confounding.

These analyses, although limited by the potential for residual confounding and information bias, support the use of PKR in high surgical risk patients. Although now most surgeons advocate for PKR for fit young patients, these results suggest that PKR is as good as TKR for patients with severe systemic disease. PKR had better patient-reported outcomes and a much lower rate safety events, particularly thromboembolic events, as seen in fit young patients. Despite an excess revision risk, mortality was also lower among PKR patients. This is very important information to communicate to patients, as they might prefer a safer procedure in the short term, although having a higher risk of revision in the long term.

4.2 Gender, Age, Rurality and deprivation subgroup analysis for UTMOST

Introduction

Increased or decreased risk for an outcome is one of the main sources of heterogeneity of treatment effects, as seen in the previous section with high surgical risk patients. But it is not the only one. There are reasons to believe that age and gender could be important effect modifiers in some drugs or vaccines, as explained earlier. Moreover, in surgical settings one must think of wider determinants that can affect the effectiveness and safety of a treatment. Factors like rehabilitation, access to healthcare services, having responsibilities as a care provider, or architectural reasons, e.g. having a staircase to be able to go outside to walk without help, can potentially modify the results of implantable devices.

To further explore this, in this section, I explore possible interactions between partial knee replacement (PKR) and age, gender, deprivation and rurality and its stratified effect on outcomes.

Methods

Study design, data sources, and population

For these analyses I continued to study the same cohort of patients: all patients who had undergone a first total (TKR) or partial knee replacement (PKR) between 2009 and 2016 in the UK, present in the NJR linked to HES and PROMs. For this section I analysed separately those with ASA grade 3 or 4 and ASA grade 1 or 2 at the time of surgery for consistency with the previous chapters. I explained these cohorts in **Chapters 2, 3, and 4**

Exposures

My main exposure is the use of PKR vs TKR. I studied interactions with gender and age at surgery, as recorded in the NJR. Also, I studied deprivation, as the IMD recorded in HES at the time of surgery, and Rurality index.

Outcomes

I looked at both effectiveness and safety outcomes when sample allows. Effectiveness will be measured with post-operative OKS. I further studied 5-year risk of revision, 5-year risk of death, and 90-day complications: Myocardial

Infraction (MI), Venous Thromboembolism (VTE) and Prosthetic Joint Infection (PJI). Follow-up and censoring are identical to the previous section.

Statistical methods

I described differences in the uptake of PKR by gender, age, deprivation, and rurality for each cohort. To test the interactions, I decided to use dichotomised variables. I dichotomised age and deprivation by the median for each cohort. Rurality was dichotomised by recoding to Urban vs non-Urban.

I also used the PS methods that were validated in **Chapter 2:** PS Stratification, both by the full cohort (PSSwhole) and by the exposure (PSSexp), and Inverse Probability Weighting (IPW). For this section, however, I constructed PS exclusively for each analysis-cohort, by excluding the variable of interest from the PS model. For the outcome analyses I used the same multi-level linear regressions clustered in lead surgeon for OKS and with a Poisson link for 5-year revision and death and 90-day complications. I introduced the variables of interest in these models as an interaction with treatment. For the variables of interest where p-value for the interaction was <0.1 in all three methods, I produced models stratifying for the levels of the variable to assess the effect for each category.

I generated 10 imputed datasets using MICE. Interaction analyses were conducted in the first imputed dataset. All stratified analyses were conducted in each of the imputed datasets and their estimators and their errors pooled using Rubin Rules. (128)

Results

Median age for ASA<3 cohort was 70, and for ASA>2 was 74. I dichotomised at these points. ***Table 4.6*** shows the number of TKR and PKR patients for each interaction in the revision/complication's cohort. The use of PKR differs greatly between age groups, especially in lower ASA patients, where 10.3% of patients less than 70 years old got an PKR compared to 3.9% of PKR in patients 70 years old or more. Gender, IMD, and rurality also show differences between categories and PKR uptake but not so striking.

ASA 1,2		
	PKR N	%
Age		
<70	15,985	10.3%
>70	5,774	3.9%
Gender		
F	10,372	6.1%
M	11,387	8.6%
IMD		
Less deprived	13,881	7.9%
More deprived	7,878	6.2%
Rurality		
Urban	15,108	6.7%
Rural	6,651	8.5%

ASA 3,4		
	PKR N	%
Age		
<75	1,577	5.1%
>75	699	2.4%
Gender		
F	978	3.0%
M	1,278	4.8%
IMD		
Less deprived	1,327	4.3%
More deprived	929	3.2%
Rurality		
Urban	1,629	3.6%
Rural	627	4.5%

Table 4.6. Number of patients for each proposed interaction

For OKS, significant interactions, pre-defined with a p-value<0.1, were identified with age and IMD in the ASA<3 cohort. In the same cohort, there were significant interactions for gender in 5-year revision and in 90-day VTE. For the ASA>2 cohort, there were significant interactions for gender, IMD and rurality for the revision outcome. For the 90-day complications, age had a significant interaction in PJI, and rurality in VTE. ***Table 4.7*** shows all p-values for the interactions that had enough outcome events to go ahead.

	OKS	*OKS*	*OKS*	*Revision*	*Revision*	*Revision*	*Death*	*Death*	*Death*
	IPW p-val	PSSw p-val	PSSe p-val	IPW p-val	PSSw p-val	PSSe p-val	IPW p-val	PSSw p-val	PSSe p-val
ASA<3 cohort									
Age	**<0.1**	**<0.1**	**<0.1**	0.2	0.1	0.1	0.8	0.9	0.9
Gender	0.9	0.5	0.5	**<0.1**	**<0.1**	**<0.1**	0.8	<0.1	<0.1
IMD	**<0.1**	**<0.1**	**<0.1**	0.8	0.3	0.3	0.3	0.2	0.2
Rurality	0.4	0.3	0.3	0.7	0.7	0.7	0.9	0.5	0.5
ASA>3 cohort									
Age	-	-	-	0.5	0.6	0.7	0.7	0.8	0.6
Gender	-	-	-	**<0.1**	**<0.1**	**<0.1**	0.6	0.2	0.2
IMD	-	-	-	**<0.1**	**<0.1**	**<0.1**	0.9	0.7	0.7
Rurality	-	-	-	**<0.1**	**<0.1**	**<0.1**	0.8	0.8	0.9

	MI	*MI*	*MI*	*VTE*	*VTE*	*VTE*	*PJI*	*PJI*	*PJI*
	IPW p-val	PSSw p-val	PSSe p-val	IPW p-val	PSSw p-val	PSSe p-val	IPW p-val	PSSw p-val	PSSe p-val
ASA 1,2 cohort									
Age	0.5	<0.1	<0.1	0.4	0.4	0.4	<0.1	0.3	0.3
Gender	0.5	<0.1	<0.1	**<0.1**	**<0.1**	**<0.1**	0.2	0.8	0.8
IMD	<0.1	0.2	0.2	0.9	0.8	0.8	0.4	0.2	0.2
Rurality	0.3	0.7	0.7	0.4	0.7	0.7	0.4	<0.1	<0.1
ASA 3,4 cohort									
Age	0.1	0.5	0.5	0.9	0.9	0.9	**<0.1**	**<0.1**	**<0.1**
Gender	0.9	0.9	0.8	0.4	0.3	0.4	0.3	0.5	0.5
IMD	0.5	0.5	0.5	0.5	0.2	0.2	0.4	0.4	0.4
Rurality	0.7	0.5	0.5	**<0.1**	**<0.1**	**<0.1**	-	-	-

Table 4.7. P-values of the interactions on each PS model

Means/crude rates and model results for each of the strata of the selected interactions is shown in *Table 4.8*. Regarding effectiveness, OKS increase of PKR compared to TKR seem to be much higher in patients > 70 (0.98 in IPW and 2.06 in PSS) compared to those performed in younger patients (-0,72 in IPW and 0.44 in PSS for <70). As for deprivation, the small increase of OKS using PKR seems to be

higher on patients living in the least deprived zones (1.34 in IPW and 1.45 in PSS) compared to those living in more deprived areas (-0.08 in IP and -0.21 in PSS).

As for revision, gender seems to have an important effect in both ASA<3 and ASA 3 or more cohorts. Women have an increased risk of 5-year revision, more than 3 times, of PKR compared to TKR both. Men had a less increased risk of revision, around double, with PKR compared to TKR. In the high surgical risk cohort, PKR has an increased risk of revision in more deprived areas and urban areas.

The decrease in risk of VTE in PKR is like those of men. I did not perform the stratified analyses for high surgical risk complications as there were 0 events for the selected groups.

ASA 1,2		TKR mean (SD)	PKR mean (SD)	IPW	PSSwhole	PSSexp
Oxford Knee Score						
Age	<70a	35.5 (9.8)	36.1 (10.3)	-0.72 (-1.77, 0.33)	0.44 (-0.28, 1.16)	0.44 (-0.28, 1.16)
	>70a	36.1 (8.8)	38.3 (8.2)	0.98 (-0.23, 2.20)	2.06 (1.06, 3.05)	2.06 (1.06, 3.05)
IMD	Less deprived	36.8 (8.8)	38.4 (8.4)	1.34 (0.51, 2.17)	1.45 (0.79, 2.12)	1.45 (0.79, 2.12)
	More deprived	34.3 (10.0)	34.1 (11.1)	-0.08 (-1.41, 1.22)	-0.21 (-1.22, 0.81)	-0.21 (-1.22, 0.81)

ASA 3,4		events in TKR (%)	events in PKR (%)	IPW	PSSwhole	PSSexp
Revision						
Gender	Men	426 (1.3%)	46 (4.7%)	1.84 (1.19,2.83)	2.08 (1.51,2.83)	2.08 (1.51,2.83)
	Women	421 (1.6%)	44 (3.4%)	3.44 (2.30,5.15)	3.53 (2.59,4.81)	3.53 (2.59,4.81)
IMD	Less deprived	407 (1.4%)	44 (3.1%)	1.82 (1.21,2.74)	2.20 (1.58,3.10)	2.20 (1.58,3.10)
	More deprived	440 (1.6%)	49 (5.3%)	3.11 (2.08,4.67)	3.35 (2.48,4.53)	3.35 (2.48,4.53)
Rurality	Urban	664 (1.5%)	77 (4.7%)	2.84 (2.11, 3.84)	3.10 (2.44,3.97)	3.10 (2.44,3.97)
	Rural	183 (1.4%)	13 (2.1%)	0.94 (0.49,1.83)	1.51 (0.85,2.69)	1.51 (0.85,2.69)
VTE						
Rurality	Urban	377 (0.85%)	6 (0.37%)	-	-	-
	Rural	83 (0.62%)	0 (0.0%)	-	-	-
PJI						
Age	<74	76 (0.26%)	4 (0.26%)	-	-	-
	>=74	44 (0.15%)	0 (0.0%)	-	-	-

ASA 1,2		events in TKR (%)	events in PKR (%)	IPW	PSSwhole	PSSexp
Revision						
Gender	Women	2,147 (1.4%)	433 (4.3%)	2.19 (1.95,2.46)	1.06 (0.98,1.15)	1.06 (0.98,1.15)
	Men	1,943 (1.6%)	419 (3.8%)	1.80 (1.57,2.07)	0.87 (0.76,0.99)	0.87 (0.76,0.99)
VTE						
Gender	Women	1,001 (0.6%)	22 (0.2%)	0.49 (0.35,0.68)	0.48 (0.37,0.61)	0.48 (0.37,0.61)
	Men	749 (0.6%)	40 (0.4%)	0.57 (0.39,0.84)	0.58 (0.42,0.80)	0.58 (0.42,0.80)

Table 4.8. Results for each outcome and stratified analysis

Discussion

I found some differences on the outcomes after PKR compared to TKR by age, gender and IMD and Rurality.

In patients with low surgical risk, younger patients (<70 years old) had similar OKS results from PKR to TKR while older patients (70 years old or more) had slightly better results with PKR. The benefit of PKR on OKS was also higher for patients living in less deprived areas. These differences are not clinically important but are in line with other observational research.(114) TOPKAT, however, found no effect on OKS differences after stratifying by age, baseline OKS or sex. (71)

As for Revision risks, gender seems to be an important factor both for low and high risk surgical patients, with women having twice or 3 times more risk of 5 year revision after an PKR than with a TKR. In comparison, men had a smaller increased of risk associated with PKR of less than twice. Also, women seem to favour from PKR in its reduction of 90-day VTE.

While there is evidence for a decreased revision risk for women in TKR,(180) there is little research done on the risk of revision. In the 2020 NJR report,(181) there is a crude analyses of revision rates of TKR and PKR for men and women by age groups that adds evidence to this differential effect. In these analyses, men seem to

have higher 5-year cumulative revision estimates in all age groups for TKR than women, but lower 5-year revision risk for PKR. Some authors believe that this effect might be also due to inappropriate indication of PKR. (182)

IMD and Rurality are factors known to influence PKR outcomes, related to the geographical distribution and centre and surgeon variables. (114) This seems to be particularly related to PKR volume, a relationship that I study in the next section.

This analysis is the first known in the literature to look at potential interactions on the effect of PKR vs TKR on several variables using state-of-the-art methods to control confounding, validated against a clinical trial. Power issues, unresolved confounding by indication and competing risks of death could be affecting some of these results. The most important limitation comes from the methods used to study the interactions. Probably the method I used is not the most efficient, compared to spline modelling and there is no rationale for dichotomising on median. (40) This limitations call for a more in-depth study of the risk of revision and complications in women.

4.3 The effect of surgeon volume in UTMOST

Introduction

In **Chapter 3** I have shown how restricting the cohort of patients to high volume surgeons has important effects on PKR revision. In the whole cohort, revision rates were double or triple for PKR than for TKR (4.0% vs 1.5% respectively), while restricting to surgeons who had done more than 10 surgeries of the same type in the previous year reduced the difference from 50 to 150% less (3.3% vs 1.5% respectively). This is a highly controversial point in literature. While trials and institutions highly supportive of PKR show low revision rates, analyses of huge registries show 2 to 3 times higher risk of revision. (69, 181, 183)

Some research found that a great deal of this effect is driven by the higher volume of PKR a centre has, but more importantly the volume of the surgeon performing the operation.(184) However, the research on this subject is observational, and most of it does not control for potential confounders except for age and gender. This is particularly worrying in the case of IMD and rurality, as they impact the risk of revision (as shown in the previous section), and are geographically distributed, and so is the surgeon volume variable. There is less information about the effect of volume on safety and effectiveness outcomes, with at least one paper

showing that higher volume surgeons may improve all these outcomes in patients. (185)

After validating the PS methods against TOPKAT, I am in a unique position to use methods that I know minimise confounding to explore the association of surgeon volume and revision.

As outcomes I have used post-operative OKS, death, and complications. I aimed to explore the effect of the number of surgeries the surgeon performed on the previous year in the effectiveness and safety of PKR compared to TKR.

Methods

Study design, data sources, and population

I included all patients who had undergone a first total (TKR) or partial knee replacement (PKR) between 2009 and 2016 in the UK, present in the NJR linked to HES and PROMs. I performed the analyses for ASA 1-2 and 3-4 separately. I did this to have the cleanest results, with the closest population to the TOPKAT trial. Then I checked the results were similar for patients with severe systemic disease.

Exposures

The main exposure was the use of PKR vs TKR. To further explore the effect of surgeon volume on this choice, I added surgeon volume as an interaction and as a stratifying factor (more than 10, 30 or 50 surgeries). I used surgeon volume as the number of surgeries of the same type (PKR or TKR) in the previous year.

Outcomes

I measured effectiveness using post-operative OKS. I also studied 5-year risk of revision, 5-year risk of death, and 90-day complications: Myocardial Infraction (MI), Venous Thromboembolism (VTE) and Prosthetic Joint Infection (PJI). Follow-up and censoring are identical to the previous sections.

Statistical methods

I calculated lead surgeon volume as the number of surgeries of the same type (PKR or TKR) on the previous 365 days of each patient's surgery on the whole NJR. I described the number of patients and surgeons for each cohort and stratification (more or equal to 10, 30, 50 surgeries of the same type in the previous year). I described the characteristics of the patients included in each group. I calculated OKS difference and crude incidence of outcomes for each group. 90-day

complications were not analysed further as the number of outcomes in PKR patients was very low once I restrict analyses to high-volume surgeons.

I performed the methods validated in **Chapter 2:** PS Stratification, both by the full cohort (PSSwhole) and by the exposure (PSSexp), and Inverse Probability Weighting (IPW). I constructed the PS inside each of the groups. I used multi-level linear regression clustered in lead surgeon for OKS and Poisson regression for 5-year revision, and death. I performed a first analysis with the full cohorts introducing lead surgeon volume as an interaction. I did subsequent analyses for each stratification if the interaction p-value was >0.1. I continued to use the same imputation strategy using MICE and 10 imputed datasets.

Results

Restricting the analysis to high-volume surgeons greatly decreased the numbers of surgeons and patients eligible, especially in the PKR group. As shown in ***Table 4.9***, **t**he number of surgeons who performed the techniques reduced even more with increasing volume: only around 30% of surgeons who had performed PKR on these patients had done more than 10 in the previous year. For TKR, more than 60% of surgeons had a volume of more than 10 TKR on the previous year.

Table 4.9 shows that restricting to patients operated by surgeons who had done over 10 surgeries of the same type in the previous year reduces TKR patients by around 10% in all cohorts, and almost halve the number PKR patients. Subsequent reductions bring the numbers of patients down further, for example in the ASA 1-2 cohort, restricting to surgeons with more than 30 surgeries further reduces TKR to 72% and PKR to 26%; and restricting to surgeons with more than 50 surgeries in the previous year halved the TKR population, but that was only the case for 10% of the TKR patients. Details on baseline characteristics for these sub-cohorts are reported in *Appendix Tables 4.2*. There seems to be little to no difference between the main cohorts and the ones whose surgeries were performed by high volume surgeons.

ASA 1-2	OKS cohort				Full cohort			
	Patients		Surgeons		Patients		Surgeons	
	TKR	PKR	TKR	PKR	TKR	PKR	TKR	PKR
All	125,834	1,197	3,895	452	273,530	21,026	4,597	1,462
≥10 surgeries previous year	114,871 *91.3%*	602 *50.3%*	2,625 *67.4%*	164 *36.3%*	248,785 *91.0%*	13,334 *63.4%*	3,001 *65.3%*	474 *32.4%*
≥30 surgeries previous year	91,504 *72.7%*	217 *18.1%*	1,556 *39.9%*	43 *9.5%*	195,898 *71.6%*	5,555 *26.4%*	1,730 *37.6%*	128 *8.8%*
≥50 surgeries previous year	66,166 *52.6%*	83 *6.9%*	996 *25.6%*	17 *3.8%*	139,396 *51.0%*	2,550 *12.1%*	1,109 *24.1%*	51 *3.5%*
ASA 3-4								
All	23,344	145	2,916	106	57,682	2,256	4,606	1,469
≥10 surgeries previous year	20,736 *88.8%*	90 *62.1%*	2,120 *72.7%*	56 *52.8%*	51,118 *88.6%*	1,449 *64.7%*	2,583 *56.1%*	294 *20.0%*
≥30 surgeries previous year	15,939 *68.3%*	36 *24.8%*	1,348 *46.2%*	19 *17.9%*	38,321 *66.4%*	610 *27.0%*	1,590 *34.5%*	92 *6.3%*
≥50 surgeries previous year	11,071 *10.3%*	15 *10.3%*	881 *30.2%*	10 *9.4%*	25,944 *45.0%*	242 *10.7%*	1,035 *22.5%*	38 *2.6%*
TKR: total knee replacement; PKR: partial knee replacement.								

Table 4.9. Number of patients and surgeons in the effectiveness and safety cohort according to surgeon volume of the index procedure in the previous year.

In ***Table 4.10*** I show the crude outcomes for each cohort. For patients with ASA 1-2, restricting to higher volume surgeons made the post-operative OKS get slightly higher, 1 point in OKS more in those patients operated by the highest volume surgeons compared to the general PKR cohort, and 0.4 points in TKR. However, the baseline (pre-surgery) OKS also went up by similar increases. For patients with ASA 3-4 it seems that patients operated by more experienced surgeons had slightly lower OKS both pre-operatively and post-operatively. All these differences are not clinically significant.

As for revision, there were major differences between surgeon volume groups, particularly for PKR. TKR patients had a 1.5% 5-year rate of revision overall both in ASA 1-2 and in patients with severe systemic disease, that went down to 1.3% for ASA 1-2 and 1.2% for ASA 3-4 in patients who had their surgeries performed by high volume surgeons. Strikingly, for ASA 1-2 the overall 4.1% revision rate for PKR patients went down to 1.9% for those operated by experienced surgeons, similar to those with TKR. For ASA 3-4 there was less decrease, from 4.0% 5y revision overall to 2.9% in patients operated by surgeons who had performed 50 or more surgeries of the same type.

Death rates were also reduced, but less steadily. From 2.4% and 5.1% in the main PKR and TKR cohorts, respectively, to 1.7% and 4.6% in the high-volume cohort. The reduction was similar for the severe systemic disease cohort, except for PKR, where the reduction was from 7.3% to 5.0%.

As for complications, although the number of cases in PKR patients was very low when I restricted the cohort to higher volume surgeons, there did not seem to be an impact of volume on 90-day complications. These rates are shown in ***Appendix table 4.3***.

	PKR		TKR	
	Pre-surgery OKS Mean (SD)	Post-surgery OKS Mean (SD)	Pre-surgery OKS Mean (SD)	Post-surgery OKS Mean (SD)
ASA 1-2				
Main	21.5 (7.7)	36.7 (9.8)	19.2 (7.7)	35.8 (9.3)
10+ surgeries	21.7 (7.7)	37.6 (9.0)	19.3 (7.7)	35.9 (9.3)
30+ surgeries	21.6 (7.8)	37.5 (9.1)	19.4 (7.7)	36.0 (9.2)
50+ surgeries	22.0 (7.5)	37.8 (8.8)	19.5 (7.7)	36.2 (9.2)
ASA 3-4				
Main	19.0 (8.5)	34.6 (10.6)	16.4 (7.6)	32.6 (10.2)
10+ surgeries	18.8 (8.5)	35.5 (10.2)	16.4 (7.6)	32.7 (10.2)
30+ surgeries	18.1 (8.6)	35.9 (10.5)	16.5 (7.6)	32.8 (10.2)
50+ surgeries	18.8 (8.2)	34.0 (9.9)	16.5 (7.7)	32.4 (10.3)
	5-year revision		**5-year mortality**	
	N of PKR patients with 5-year revision /total PKR (%)	**N of TKR patients with 5-year revision /total TKR (%)**	**N of PKR patients with 5-year revision /total PKR (%)**	**N of TKR patients with 5-year revision /total TKR (%)**
ASA 1-2				
Main	852/21,026 (4.1%)	4090/273,530 (1.5%)	496/21,026 (2.4%)	14004/273,530 (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	137/5,555 (2.5%)	2670/195,898 (1.4%)	122/5,555 (2.2%)	9472/195,898 (4.8%)
50+ surgeries	48/2,550 (1.9%)	1791/139,396 (1.3%)	43/2,550 (1.7%)	6403/139,396 (4.6%)
ASA 3-4				
Main	90/2,256 (4.0%)	847/57,682 (1.5%)	164/2,256 (7.3%)	6,401/57,682 (11.1%)
10+ surgeries	45/1,449 (3.1%)	746/51,118 (1.5%)	107/1,449 (7.4%)	5,601/51,118 (11.0%)
30+ surgeries	14/610 (2.3%)	519/38,321 (1.4%)	34/610 (5.6%)	4,055/38,321 (10.6%)
50+ surgeries	7/242 (2.9%)	315/25,944 (1.2%)	12/242 (5.0%)	2,624/25,944 (10.1%)

Table 5.10. Pre and Post surgery OKS, 5-year revision, and 5-year mortality of participants operated on by surgeons who had performed 10+, 30+ and 50+ surgeries of the same type as the index surgery in the year before the index surgery

The p-value for the interaction term between lead surgeon volume and PKR was less than 0.1 for 5-year revision in patients ASA 1-2 both for IPW and PS Stratification. All other analyses, ASA 3-4 and death and OKS in ASA 1-2, had p-values higher than 0.1.

Figure 4.6 shows interaction graphs for each studied outcome and ASA cohort with IPW analyses. The ASA 1-2 5-year revision graph suggests that the revision rate is more than double in PKR patients than in TKR patients for those operated by surgeons with <10 surgeries of the same type in the previous year. However, this relation decreases until the revision rates became similar or even lower in high volume surgeons, who had done 50 or more surgeries of the same type in the previous year.

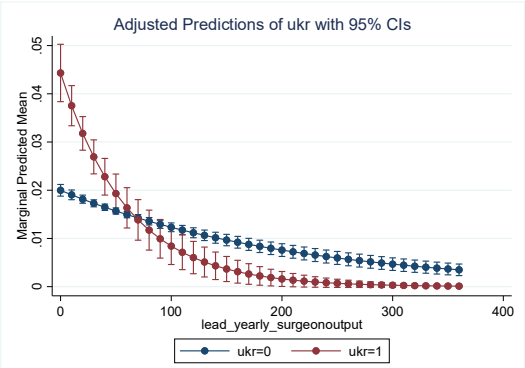
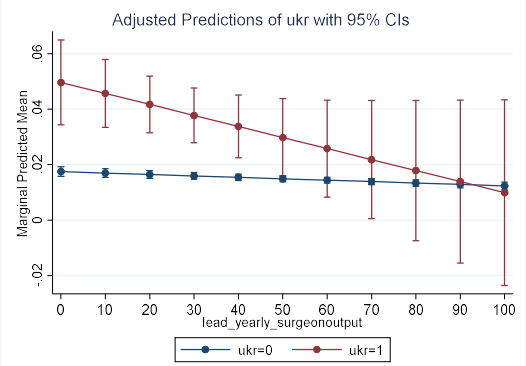
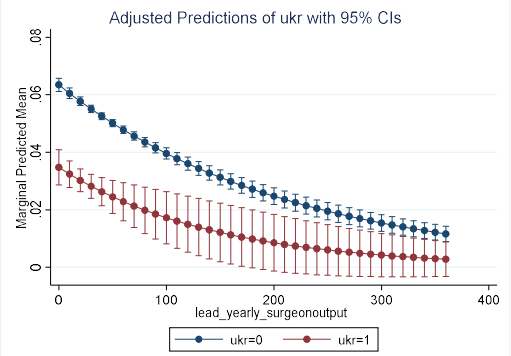
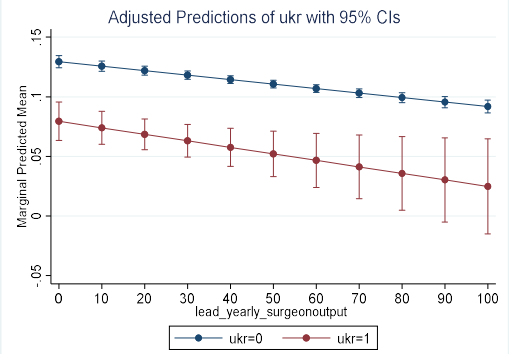
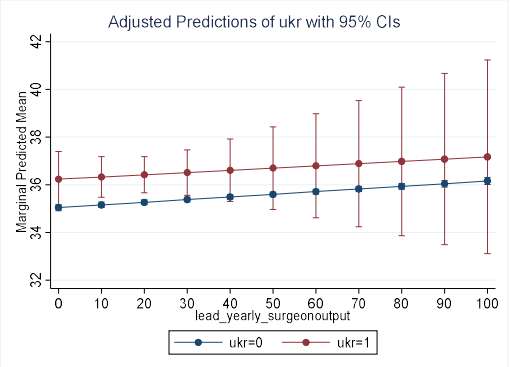
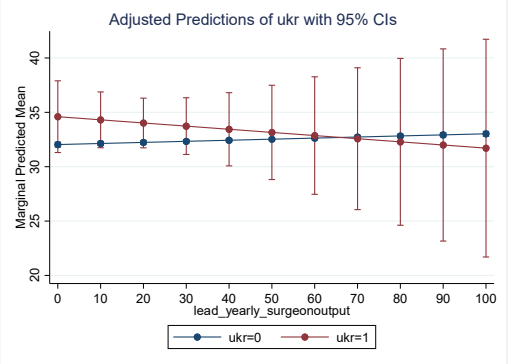
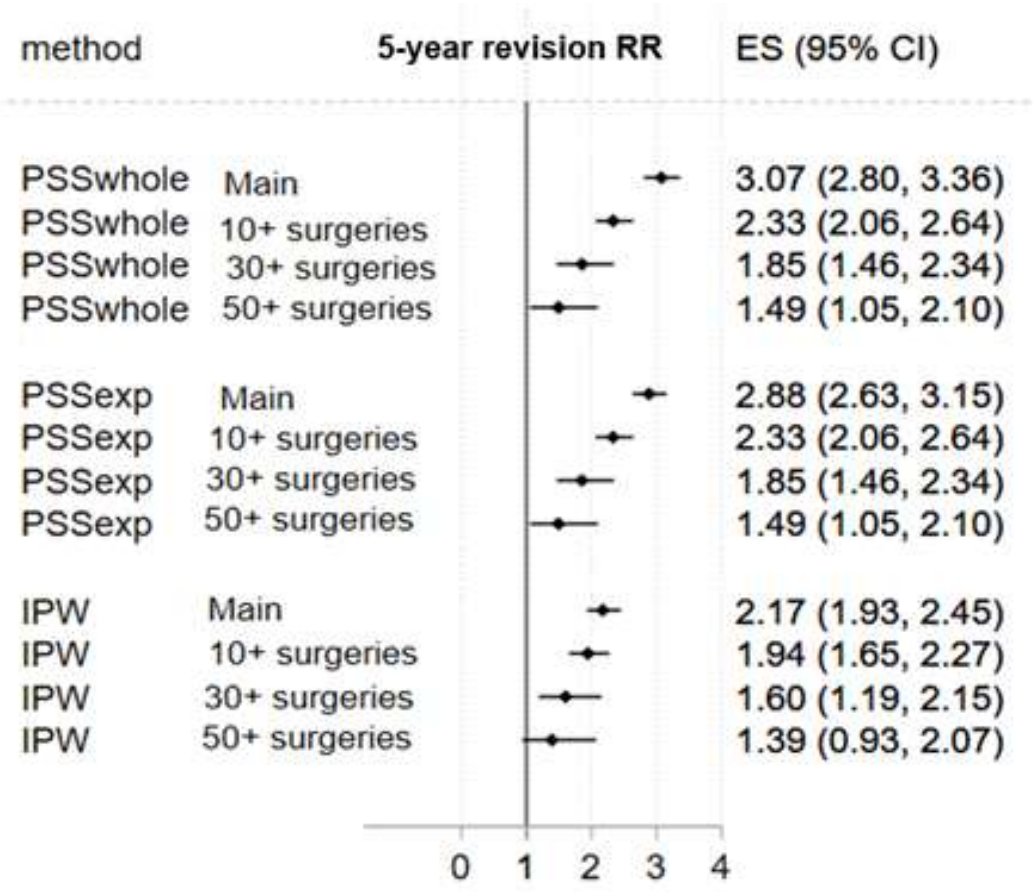
ASA 1-2	ASA 3-4
Revision	
	
Death	
	
OKS	
	

Figure 4.6. Marginal estimation after IPW of the incidence of 5-y revision and 5-year death, and 1-year Oxford Knee score after a PKR or TKR surgery by lead surgeon volume in the previous year. ASA 1-2 and 3-4 separately.

Figure 4.7 shows the estimates for 5y revision for each stratification category (all, 10+, 30+, 50+ surgeries of the same type in the previous year). This figures confirms the results seen in the interaction graphs, for all propensity score methods, confirming that there is a reduction in the risk of 5-year revision in PKR compared to TKR with an increase of number of surgeries the lead surgeon had performed on the previous year.



Note: IPW: inverse probability weighting; PSS_{exp}: propensity score stratification based on the exposure cohort; PSS_{whole}: propensity score stratification based on the whole cohort.

Figure 4.7. Forest plot of the relative risk of revision surgery with 5 years for each of the validated methods in the main cohort 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)

Discussion

In this section I explored the effect of lead surgeon volume on knee replacement outcomes. I have found that half of the PKR surgeries are performed by surgeons who had done less than 10 PKR surgeries in the previous year. This low volume is in clear contrast with TKR surgeries; here more than 90% patients are operated by surgeons who had done that volume. Surgeon volume does not seem to have a strong impact on OKS or mortality but has a striking effect on PKR 5-year revision rates.

The percentage of PKR surgeries performed by high PKR volume surgeons is much lower than for TKR: 63% surgeries with surgeons with more than 10 PKR in the previous year vs 91% surgeries performed by surgeons that had done more than 10 surgeries in the previous year. These differences are accentuated even more when restricting to 30 and 50 surgeries of the same type in the previous year. This coincides with the numbers reported in Liddle et al., where surgeons who performed PKR had done a mean of 7.6PKR surgeries in the previous year. (96)

I also found that surgery volume seems to have a dramatic effect for PKR: The rates of revision get greatly reduced as the surgeon volume increases, even overtaking the revision risk of TKR. This is in line with previous literature, (184)

but I show how this effect seems to be still present after controlling by patient confounders. My analyses also seem to point that the volume of surgeries needed to achieve comparable results to TKR may be around 50 per year per surgeon, similar to previous studies. (184, 186)

I did not find a differential effect of surgeon volume in terms of OKS, death, and complications after surgery. The effect on complications could not be interpretable due to low power to perform these analyses, even there were no differences in crude analyses. However, although this effect might not be present in the primary surgery, avoiding a revision surgery may greatly reduce the risk of complications, that are increased in this type of surgery. (187, 188)

These results point to the need of concentrating the assessment for eligibility of PKR and its surgery to high volume surgeons, as they are more likely to get revision rates similar than those of TKR. However, they may be other variables that play a part that were not explored in this work, as the relationship between learning and volume, capacity issues, and how the care of the centres where high volume surgeons work may differ from other centres.

Chapter 5

Methods - Using simulations to assess the validity of IV based on the PKR vs TKR studies and wider multilevel MD studies, and different multilevel modelling strategies of PS in the comparative effectiveness and safety of PKR vs TKR

In this Chapter I aim to explore methodological questions that arose from the work done in the previous chapters. First, I tried to understand why instrumental variables failed to replicate the findings seen in TOPKAT in **the trial emulation study in Chapter 3** by using a Plasmode analysis, a type of re-sampling simulation that keeps the relationship between variables by injecting new synthetic variables generated based on the real data, therefore providing settings closer to the real ones, (189) as detailed in **section 5.1**. This kind of simulation allows me to specify the effects of the IV, exposure and covariates while maintaining the relationships between them. Second, I explored the impact of a multilevel structure, where patients are grouped among surgeons, on preference-based instrumental variable

analyses. For this I used a parametric simulation, that allows greater control over the scenarios, totally pre-specified and not dependant on the original data, to test and mimic the multilevel structure of the dataset. Finally, I aimed to find the best way of including surgeon variables in PS models using the real data from the previous Chapters. I explored this using random effects models on the PS and the outcome models and its impact on the results of propensity score analyses.

5.1 Simulation study: Instrumental Variables and bias

Introduction

True instrumental variables (IVs) are very difficult to find, as they should be strongly related to the exposure and related to the outcome only through the exposure.(99) In **Chapter 2 Section 2.1.3**, I selected some of the IVs previously used in pharmacoepidemiology (physician prescription preference, geographical location, and calendar time), and tested them. Out of all the tested IVs, only surgeon preference (equivalent to physician prescription preference) was deemed strong enough (highly related to the exposure). Furthermore, it seemed to achieve good balance on known confounders. However, surgeon preference IV yielded unplausible treatment effect estimates of Oxford Knee Score that were far from the TOPKAT trial estimate.

This failure to replicate the trial results for effectiveness analyses could be due to several factors. The most clinically plausible is that surgeon's preference influences the postoperative OKS through other variables. This would mean that either unmeasured confounders exist, or that the balance method I used to exclude known confounding failed to detect it.

In this **Section 5.1** I aim to explore sources of bias and error related to the two assumptions I tested for in **Sections 2.1 and 3.1:** strength of the IV and known confounders. To do so, I propose to first explore how the selected IVs perform in terms of coverage, bias and error in a setting similar to the ones in **Chapters 2 and 3**. I compared the best performing IV to a synthetic IV, both with the same set of strength settings, and explored how strength and known confounders impact performance metrics. I additionally explored how introducing one confounding variable at a time impacted the same metrics.

Methods

Data Generation

Plasmode Simulation

To keep the confounding structure present in the data, I used the Plasmode simulation method. This is a re-sampling method with replacement. (189) This method re-samples from the original dataset covariates without any modification. Exposure and outcome are generated based on these re-sampled covariates and on pre-specified true treatment effects.

In contrast with a parametric simulation, plasmode simulation retains the relationships among covariates creating also a more "realistic" scenario. Retaining these complex relationships among covariates is a great tool to explore confounding present naturally in real data. This confounding could be mediated through several covariates, which would be difficult to replicate in a parametric simulation. In addition, since I am interested in discovering whether I failed to find a known confounder, it is important to keep the relationships between variables and the structure of the original dataset.

The original method I build on was developed to evaluate confounder adjustment via propensity score. I have made several revisions to the original method, rewriting a substantial part of the existing R code: I modified the package "Plasmode"(189) to add the ability to model IVs, as well as the possibility of breaking the IV assumptions. Second, I added the possibility of creating non-normally distributed outcomes that can also be bounded and rounded, similar to the **Chapter 2** study outcome, OKS. Finally, I have debugged the code and changed the implementation to improve performance and reliability. The revised code can be found ***Appendix Text 5.1***.

Original datasets

I used the datasets from **Chapter 2**, for the *continuous* outcome (post-operatory OKS), and **Chapter 3**, for the *binary* outcome (5-year revision), to resample observations, to extract the prevalence of treatment and the betas of the relationship of covariates and treatment and outcomes.

Number of observations and datasets generated

I generated the same number of observations per dataset as those in **Chapter 2** (1,197 PKR and 125,834 TKR) and in **Chapter 3** (21,026 PKR and 273,530 TKR). I generated 1,000 simulated datasets per scenario.

Instrumental Variable

I compared two different IVs. First, the original surgeon preference from **Chapters 2 and 3** (% of surgeries in the 30 previous surgeries of the same type PKR or TKR performed by the same surgeon), that I will refer to as *real-world IV*. This represents the operating surgeon's real preference for PKR (with a mean prevalence of PKR of 7% among the simulated datasets).

I also generated a so called *synthetic IV* trying to mimic the distribution of the *real-world IV* but unrelated to any variable other than being directly related to the treatment. This *synthetic IV* was defined as gamma distributed with k=0.16, θ=1 and truncated at 0 and 1. The resulting distributions are shown in ***Appendix Figure 5.1***.

The *real-world IV* is the original surgeon preference in the data and would maintain the relationships between it and potential confounders. The *synthetic IV* is a

variable with the same overall shape as the *real-world IV* but generated independently from any other variable. This makes the *synthetic IV* a true IV. This allows me to explore if there was any violation of the unconfoundedness assumption, at least through known confounders. These potential confounders may be correlated to the *real-world IV* (thanks to the resampling of the Plasmode simulation) but they shouldn't affect the *synthetic IV* (as it is totally exogenous and not related to any covariate in the dataset).

Covariates

All covariate values were re-sampled from the original dataset. These included age, gender, tenths of IMD, Rural-Urban Index, ASA grade, BMI, EQ5D general health scale, EQ5D index, OKS pre-surgery, Charlson Co-morbidity Index, and a list of pre-specified relevant comorbidities: cardiovascular disease, coxarthrosis, foot, hip and back pain, gastrointestinal diseases, neurological disorders, other arthrosis, polyarthrosis, respiratory diseases, spondylosis, thyroid problems, other joint problems, mental health diseases, and surgeon volume. These are known confounders of the relationship between knee replacement treatment and the outcomes of interest, and they were clinically identified.

Exposure

The exposure is the type of treatment, 0 being TKR and 1 PKR. Treatment (t_i) was a function of:

$$t_i \sim Bernoulli\ (\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\ \pi_i}\right) = \alpha_0 + \ \alpha_1 \times x_1 + \cdots + \ \alpha_n \times x_n + \ Strength_{IV} \times IV$$

Where:

- α_1 to α_n are the coefficients of each covariate $x_1,\ldots, x_n$ identified in a logistic regression on the treatment in the original dataset;
- strength IV was generated as log (1, 1.1, 2, 5, 10, 25, 50, 100);
- α_0 is calculated as the value necessary to reach a prevalence of treatment of 7%, to mirror the original dataset, by one-dimensional root finding;

Outcome

Continuous Outcome

The continuous outcome (Y_i) was designed to be similar to the OKS outcome on **Chapter 2.** It was a function of the covariates, and the true treatment effect of 2:

$$Y_i = \beta_0 + \ \beta_1 \times x_1 + \cdots + \ \beta_{1_n} \times x_n + \beta_{treat} \times Treatment$$

Where:

- β_1 to β_n are the coefficients of each covariate obtained β using a linear regression of all variable and the treatment on the outcome OKS in the original **Chapter 2** dataset;
- β_0 is normally distributed with μ =0, and σ equal to the SD of the errors on the regression of the variables and treatment on the outcome in the original **Chapter 2** dataset;
- β_{treat} is a selected true treatment effect: 2;

And Y_i is truncated at 0 and 48.

Binary Outcome

The binary outcome (Y_i) was designed to be similar to the 5-year revision outcome on **Chapter 3.** The binary outcome (Y_i) was a function of the covariates and the true treatment effect:

$t_i \sim Bernoulli\ (\pi_i)$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times x_1 + + \beta_{1\ldots n} \times x_{\ldots n} + \beta_{treat} \times Treatment$$

Where:

- β_1 to β_n are the coefficients of each covariate obtained using a logistic regression of all variable and the treatment on the outcome 5 year revision in the original dataset;
- β treat is a selected true treatment effect: log (2)

Covariate testing

To test if there is a violation of the IV assumptions, where the IV is related to the outcome through one of the known confounders, I repeated the data generation using each confounding variable with their beta at a time, both in the outcome and exposure generation, plus the IV and the treatment effects. I repeated the data generation with only the IV, treatment and outcome, without any confounder. For these analyses, I used the strength beta of log(50), a very high strength, to ensure I did not have spurious findings due to low IV strength.

Instrumental Variable diagnostics

Here, I relied on the same diagnostics used in **Chapters 2 and 3, sections 2.1.3 and 3.1.3.** For instrument strength I used the odds ratio (OR) between the IV and the exposure and the F statistic. I pre-specified a minimum OR of 2 to deem the IV strong enough as based on the previous analyses.(110) For the F statistic, I used a threshold of >10. Finally, to evaluate if the IV was related to any known

confounder I assessed balance with the absolute standardised mean difference, where I set a threshold of less or equal to 0.1, following previous literature. (107)

Methods and Estimand

I looked at the treatment effect yielded by the same method used in **Chapters 2 and 3**, namely two-stage least squares with a median dichotomised IV. I evaluated the beta for PKR in a linear two-stage least squares IV regression for the continuous outcome and in a Poisson IV regression for the binary outcome.

Performance measure(s)

As performance measures, I used several measures: (190)

- Absolute bias, defined as the average difference between the true treatment effect and the one estimated using the IV regression.

$$Absolute\ bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \widehat{\theta_i} - \theta$$

- Empirical standard error, defined as the standard error of the point estimates.

$$EmpSE = \sqrt{\frac{1}{n_{sim} - 1} \sum_{i=1}^{n_{sim}} (\widehat{\theta_i} - \theta)^2}$$

- Coverage, as the proportion of 95% confidence intervals obtained that included the true treatment effect.

Coverage = $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{low,i} \leqq \theta \leqq \hat{\theta}_{upp,i})$

Results

Original dataset and regression coefficients

The original dataset used for the analysis of the continuous outcome included 125,834 TKR patients and 1,197 PKR patients. After excluding those for whom IV could not be computed, the first 30 patients operated by each surgeon (see **Chapter 2** for detail), I kept 100,916 TKR and 965 PKR. In parallel, the dataset used for the analysis of binary outcomes included 273,530 TKR and 21,026 PKR and kept 199,118 TKR and 16,607 PKR after restricting to those whose IV was calculated.

I performed a logistic regression to estimate the coefficients of relationships between the IVs and the selected covariates and the exposure for each model: *real-world IV* and *synthetic IV* for continuous and binary datasets. I got the coefficients for the outcome regressions using a similar strategy for the outcome variables, with a logistic regression for the binary outcome and a linear regression for the continuous variable. The betas found are shown in ***Table 5.1***. The betas were similar between *real-world* and *synthetic* IVs for each dataset, except for the deprivation betas, that were lower in the *real-world IV* regression. I used the *real-world IV* betas for the one confounder at a time analysis.

	Binary Outcome: 5y Revision				Continuous Outcome: OKS			
	Synthetic IV		Real-world IV		Synthetic IV		Real-world IV	
	Exposure	Outcome	Exposure	Outcome	Exposure	Outcome	Exposure	Outcome
Gender: Male	0.14	0.205	0.169	0.204	0.15	0.38	0.15	0.383
Age	-0.074	-0.035	-0.07	-0.035	-0.078	0.02	-0.074	0.02
BMI	-0.037	0	-0.037	0	-0.054	-0.067	-0.057	-0.067
General Health: 2-4	0.366	0.213	0.356	0.213	0.372	-0.539	0.348	-0.538
EQ-5D Health Scale	0	0	0	0	0.002	0.053	0.001	0.053
pre-operative OKS	0.044	-0.02	0.044	-0.02	0.039	0.322	0.034	0.321
CHARLSON >1	0.031	-0.022	0.04	-0.022	-0.086	-0.258	-0.098	-0.259
Non-Urban	0.144	-0.014	0.145	-0.014	0.071	0.439	0.047	0.439
Less deprived 10-20%	-0.108	0	-0.052	-0.001	-0.187	-0.048	-0.14	-0.045
Less deprived 20-30%	-0.218	0.058	-0.115	0.056	-0.143	-0.362	-0.012	-0.355
Less deprived 30-40%	-0.338	-0.046	-0.203	-0.048	-0.329	-0.449	-0.145	-0.441
Less deprived 40-50%	-0.295	0.03	-0.156	0.027	0.029	-0.655	0.213	-0.646
More deprived 10-20%	-0.705	0.035	-0.533	0.032	-0.132	-2.123	0.104	-2.11
More deprived 20-30%	-0.565	-0.012	-0.398	-0.015	-0.353	-1.38	-0.183	-1.367
More deprived 30-40%	-0.424	0.032	-0.293	0.029	-0.22	-1.07	-0.005	-1.06
More deprived 40-50%	-0.395	-0.012	-0.255	-0.014	-0.504	-0.725	-0.377	-0.714
Most deprived 10%	-0.813	-0.063	-0.624	-0.067	-0.596	-2.936	-0.436	-2.922
ASA 2	-0.22	-0.053	-0.215	-0.054	-0.035	-0.419	-0.036	-0.419
Cardiovascular diseases	-0.08	0.012	-0.068	0.011	-0.219	-0.178	-0.158	-0.176
Coxarthrosis	-0.393	-0.049	-0.369	-0.05	-0.432	-0.106	-0.453	-0.101
Foot, hip, spinal pain	0.009	0.221	-0.029	0.222	0.292	-2.04	0.232	-2.045
Gastrointestinal disease	0.084	0.033	0.088	0.033	-0.05	-0.909	-0.073	-0.91
Neurological disorders	0.098	0.108	0.084	0.108	0.212	-0.76	0.172	-0.761
Other arthrosis	-0.045	0.12	-0.019	0.12	-0.231	-0.558	-0.208	-0.557
Polyarthrosis	-0.497	-0.219	-0.457	-0.22	-0.912	-0.082	-0.987	-0.08
Respiratory diseases	0.029	0.01	0.008	0.01	0.132	-0.353	0.156	-0.351
Spondylosis	-0.164	0.004	-0.127	0.003	-0.453	-1.195	-0.413	-1.191
Thyroid problems	-0.043	-0.017	-0.066	-0.017	0.106	-0.426	0.068	-0.426
Other Joint Problems	-0.04	0.177	-0.057	0.177	0.053	-0.59	0.081	-0.593
Mental health	0.2	-0.034	0.253	-0.034	-0.076	-1.286	-0.018	-1.284
Lead Surgeon Output	-0.064	-0.002	-0.063	-0.002	-0.114	0.004	-0.115	0.004

Table 5.1 Betas used in the data generation

I generated 1,000 datasets for each scenario. Scenarios included 8 strengths per IV type (real-world vs synthetic) per type of variable (continuous vs binary). In addition, I also generated a scenario per each variable tested (23 selected confounders) plus one with no confounder variable, for continuous and for binary outcome. This amounts to 32 IV testing scenarios and 48 variable testing scenarios and amounts to 80,000 simulated datasets.

IV strength

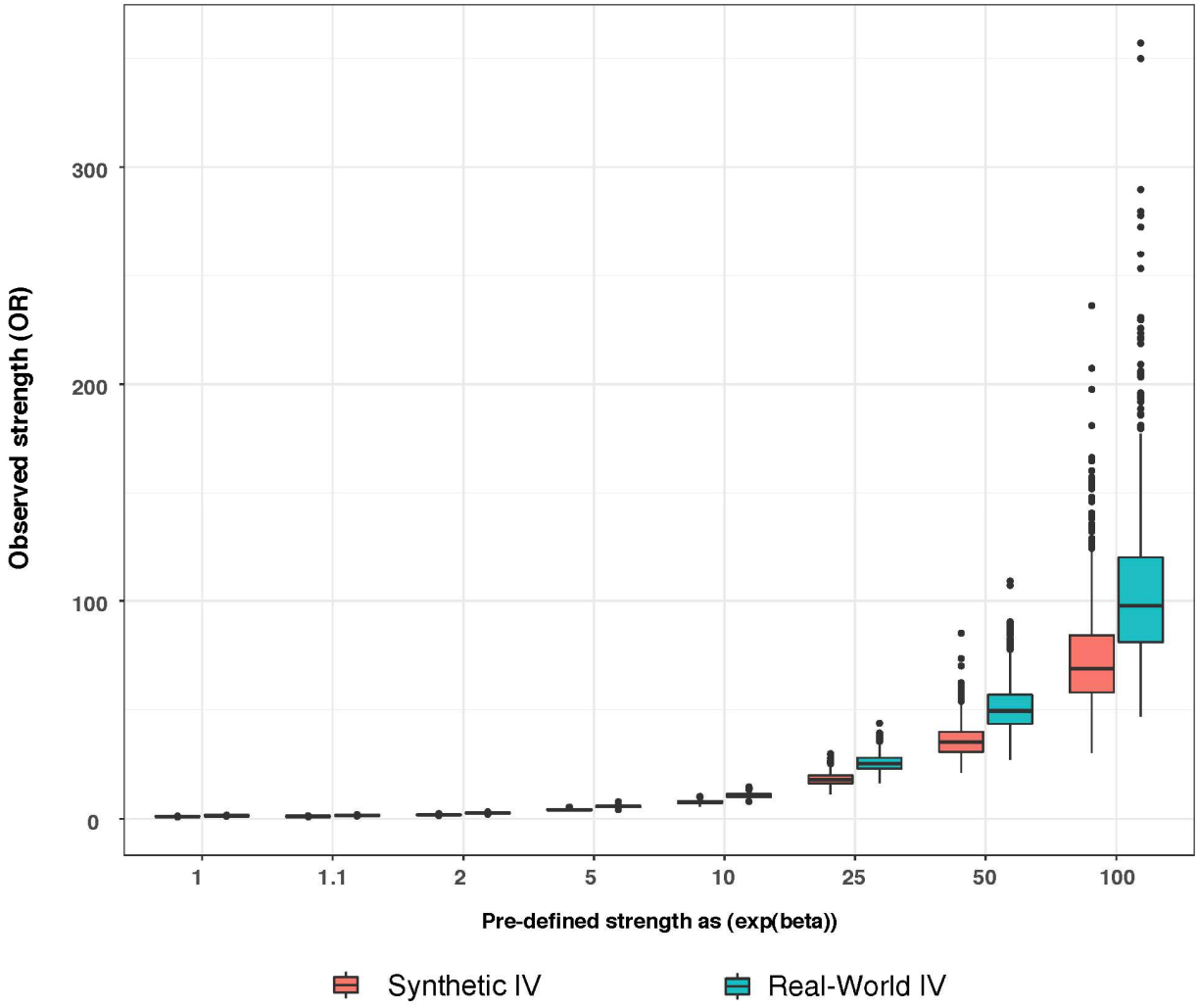
For both the continuous and binary outcome datasets, IV strength increased exponentially with higher beta values in all scenarios. Strengths started to have an OR higher than 2 with a beta of log(2) in the data generation, but all OR for all simulated datasets were higher than 2 with a beta of log(5) or more. Mean OR for a beta of log(5) ranged from 3.3 to 6.7 across all scenarios and IVs. Minimum F-statistic was higher than 10 for scenarios with a beta higher than 2. These results are shown in ***Table 5.2*** and ***Figure 5.1***. The mean strength achieved was consistently lower when using a synthetic IV, compared to the *real-world IV*. In addition, on the binary outcome dataset the OR for the *synthetic IV* went up more slowly when increasing the beta than the OR for the *real-world IV*.

		Odds Ratio (OR)							
		Synthetic IV				Real-world IV			
Database	**Real strength**	**mean**	**sd**	**min**	**max**	**mean**	**sd**	**min**	**max**
Binary	**1**	1.0	0.0	1.0	1.1	1.9	0.0	1.8	2.0
	1.1	1.1	0.0	1.0	1.1	2.1	0.0	2	2.2
	2	1.7	0.0	1.6	1.7	3.3	0.1	3.1	3.5
	5	3.3	0.1	3.1	3.4	6.7	0.1	6.3	7.2
	10	5.5	0.1	5.2	5.9	12.0	0.3	11.1	13.4
	25	11.8	0.3	10.9	12.8	27.0	1.0	24.4	30.1
	50	21.7	0.8	19.5	24.4	51.4	2.5	43.6	60.7
	100	40.9	1.9	36.2	48.3	99.8	6.5	80.4	121.8
Continuous	**1**	1.0	0.1	0.8	1.2	1.4	0.1	1.2	1.7
	1.1	1.1	0.1	0.9	1.4	1.5	0.1	1.2	1.9
	2	1.8	0.1	1.5	2.4	2.6	0.2	2.1	3.2
	5	4.1	0.3	3.3	5.4	5.7	0.5	4.2	7.8
	10	7.7	0.8	5.6	10.2	10.8	1.1	7.9	14.6
	25	18.3	2.7	11.4	29.9	25.7	3.9	16.3	43.8
	50	36.0	7.4	21.2	85.2	51.3	11.2	27.3	109.1
	100	73.7	23.4	30.3	236.1	105.6	35.3	47	356.8

		F-statistic							
		Synthetic IV				Real-world IV			
Database	**Real strength**	**mean**	**SD**	**min**	**max**	**mean**	**SD**	**min**	**max**
Binary	**1**	1	1	0	12	1629	81	1379	1908
	1.1	20	8	1	50	1994	84	1766	2248
	2	936	60	726	1121	4962	133	4533	5401
	5	4570	129	4214	4947	10837	185	10281	11397
	10	8254	154	7742	8850	15142	216	14508	15966
	25	12900	173	12357	13462	19432	211	18774	20118
	50	15557	181	15093	16101	21448	212	20793	22082
	100	17328	176	16813	17927	22657	213	22040	23294
Continuous	**1**	1	2	0	11	30	11	6	69
	1.1	4	4	0	22	45	13	12	102
	2	84	19	33	156	210	28	126	312
	5	356	32	273	452	595	45	419	751
	10	573	36	454	679	877	46	742	1020
	25	782	37	664	896	1140	46	984	1282
	50	874	34	766	986	1251	47	1115	1406
	100	926	33	835	1054	1319	47	1184	1464

Table 5.2: Mean, SD, median, min and max achieved strength (in Odds Ratio) compared to the exponentiated beta of the IV in the exposure generation formula (Expected OR) per each method (Synthetic vs Real IV)

Continuous Outcome datasets



Binary Outcome datasets

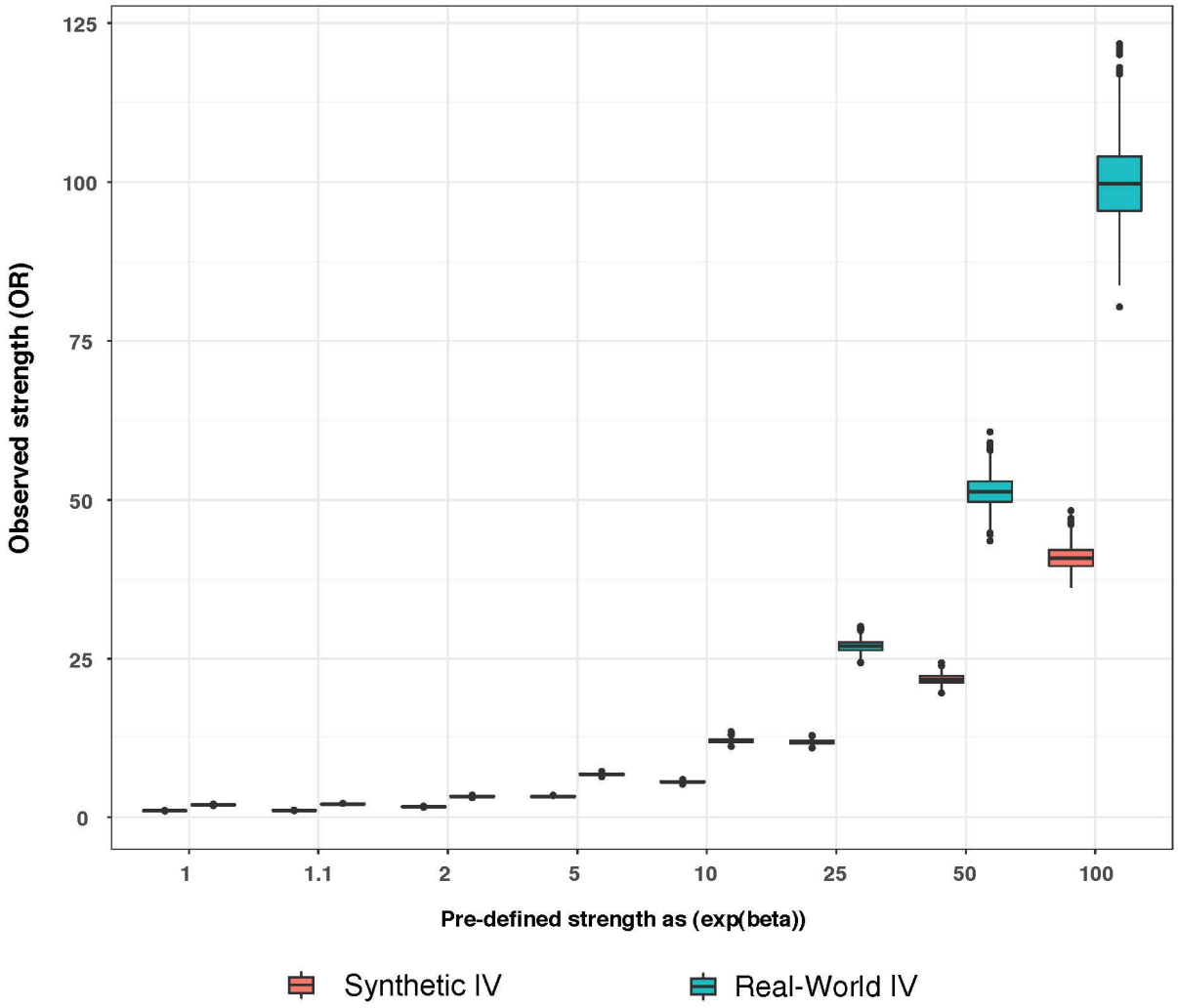


Figure 5.1 Box Plot of the achieved strength (in Odds Ratio) compared to the exponentiated beta of the IV in the exposure generation formula (Expected OR) per each method (Synthetic vs Real IV)

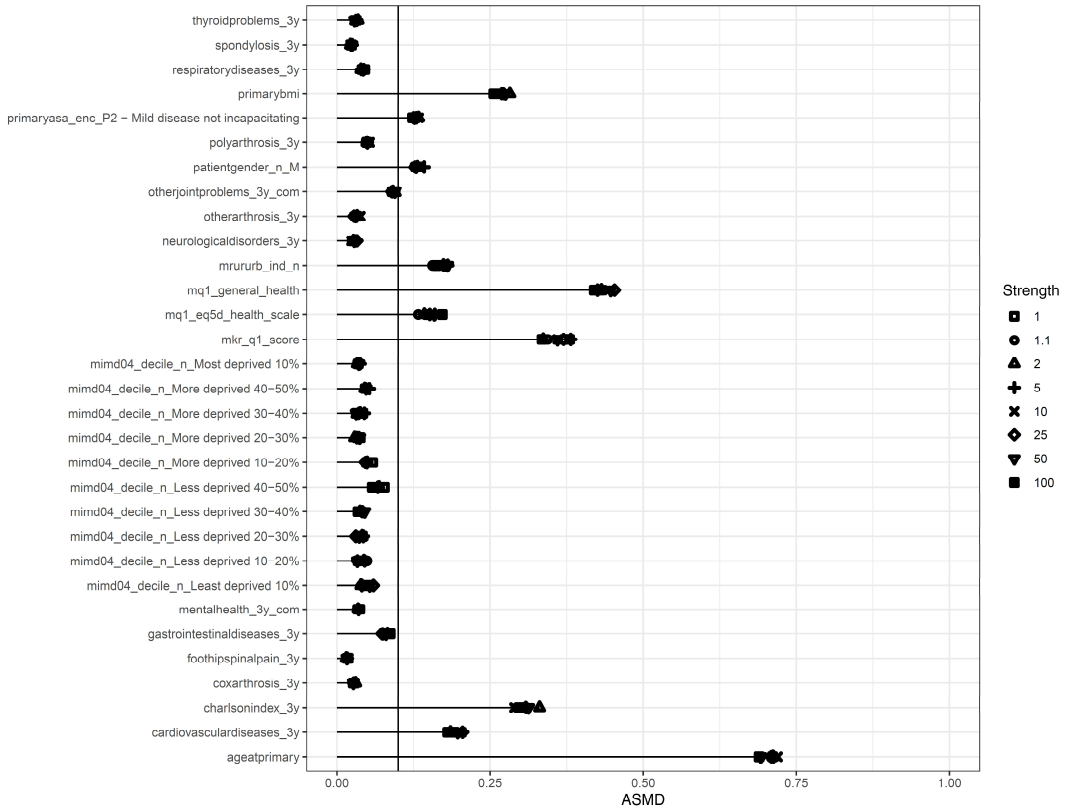
The second pre-defined selection criterion for IVs was the ability to balance known confounders between IV=1 and IV=0. Maximum ASMD for each variable among all simulations for both datasets before and after applying the IVs is shown in ***Figure 5.2*****.** Before applying IV for confounding adjustment, there were significant imbalances in many covariates between treated and untreated, showing with 10 covariates with an ASMD higher than 0.1 both in the continuous and binary outcomes (***Figure 5.2***). The higher imbalances were seen in age at primary (mean of all scenarios ASMD= 0.60, max ASMD= 0.73), general health scale (mean ASMD=0.35, max ASMD=0.45), and pre-operatory OKS (mean ASMD= 0.26 , max ASMD= 0.38).

After applying the IVs, the balance improved greatly in all simulated datasets and variables, as seen in ***Figure 5.2***. The synthetic IVs, as expected, almost perfectly balanced all confounders with a max ASMD of 0.03 and a mean ASMD of 0.003 in both the binary and continuous datasets, well below the 0.1 cut-off. The *real-world IV* performs well in reducing the balance on the continuous outcome datasets, with none of the variables in any of the datasets showing an imbalance higher than 0.1. This was true for all 8 different strengths of the IV. The maximum imbalances across all these datasets were seen in pre-operatory OKS (mean ASMD=0.05, max

ASMD=0.07), and age at primary (mean ASMD=0.05, max ASMD=0.07), all with an ASMD below the proposed threshold of 0.1. The *real-world IV* performed worse in the binary outcome datasets, although it maintained the ASMD below 0.1 for most variables across datasets and IV strengths. However, the *real-world IV* did not manage to balance patient age at time of the surgery, in none of the strengths, with a mean SMD of 0.13 and a maximum SMD of 0.15. None of the different strengths balanced this variable in any of the simulated datasets, but it was balanced in the original dataset.

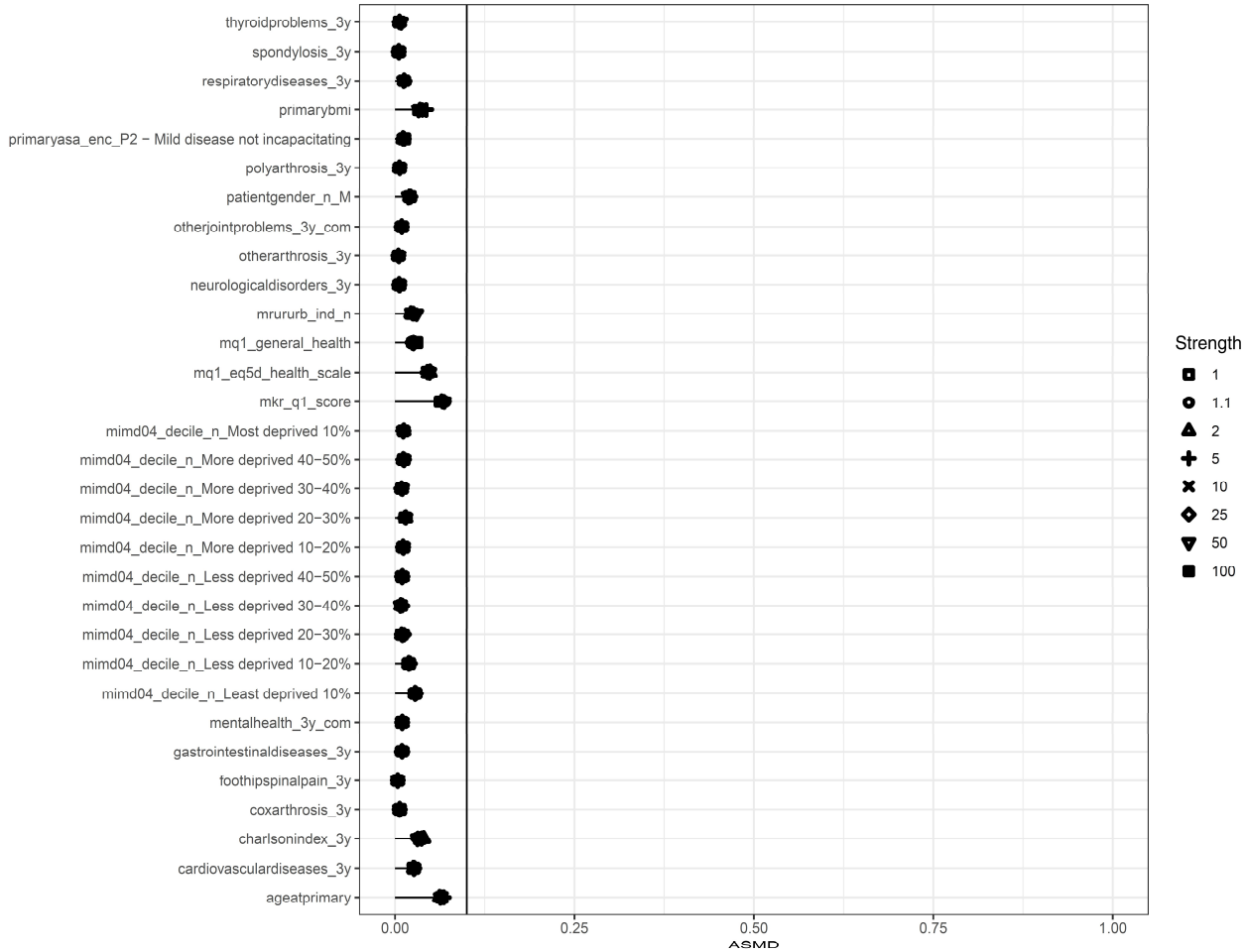
Continuous outcome variable

ASMD before IV



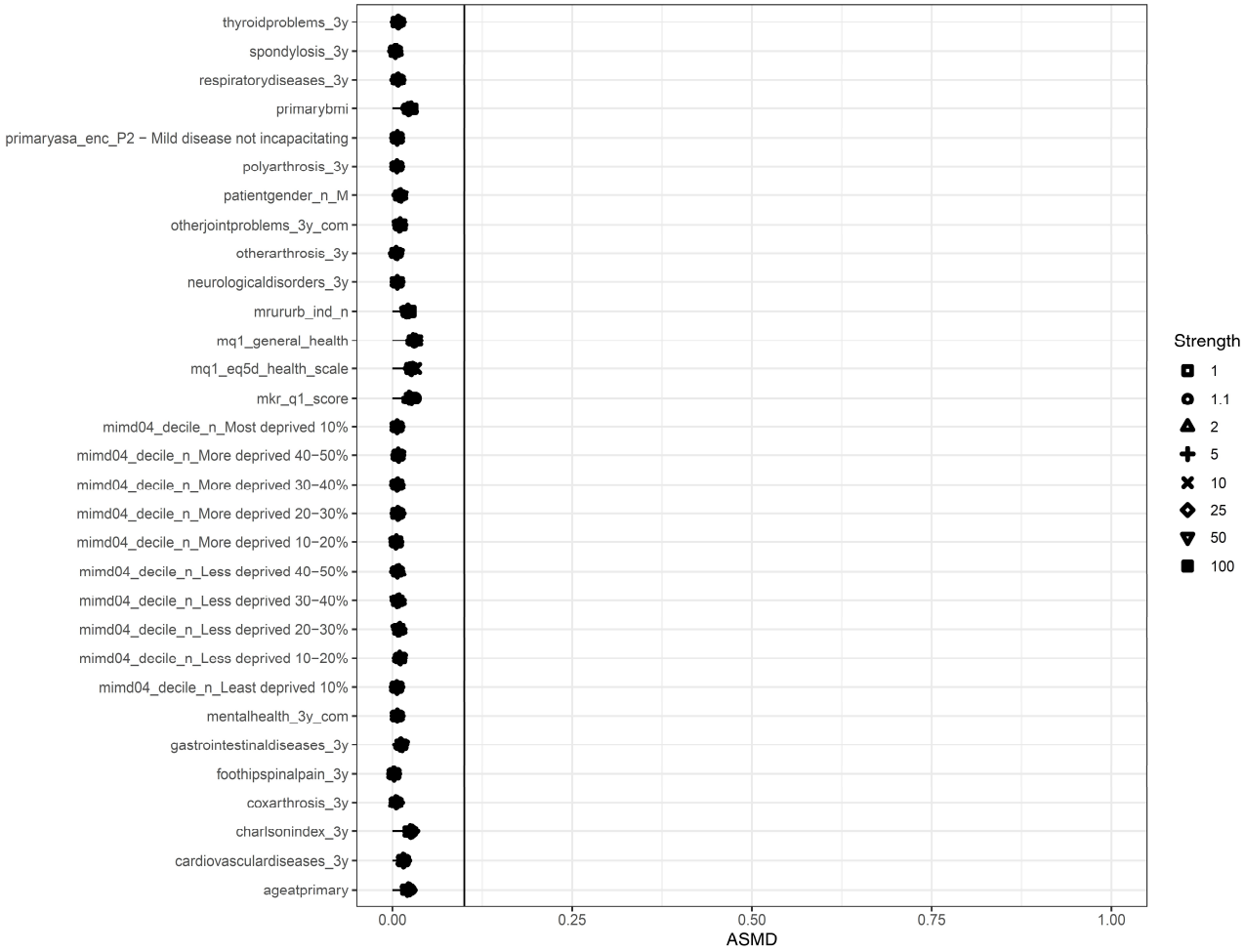
Continuous outcome variable

ASMD after real-world IV



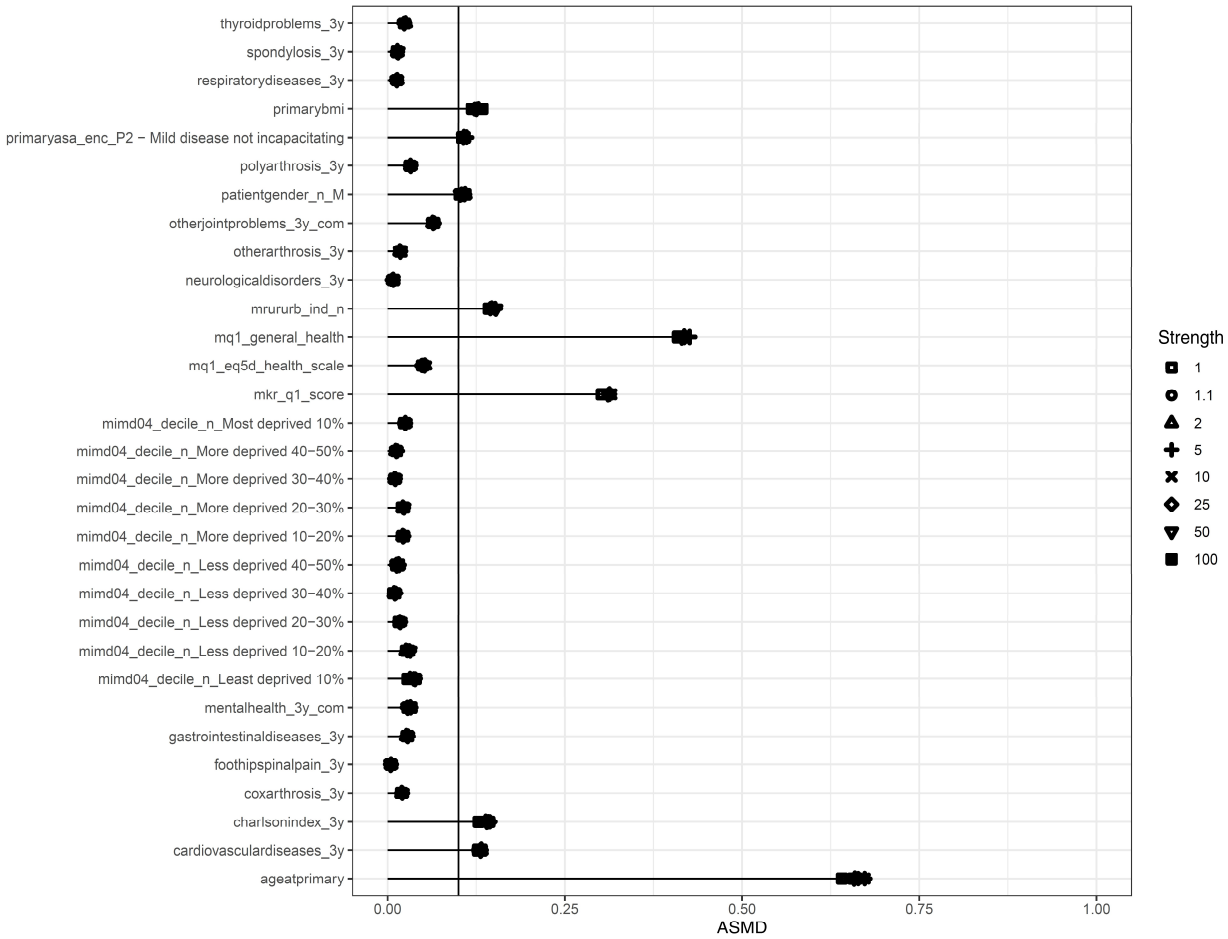
Continuous outcome variable

ASMD after synthetic IV



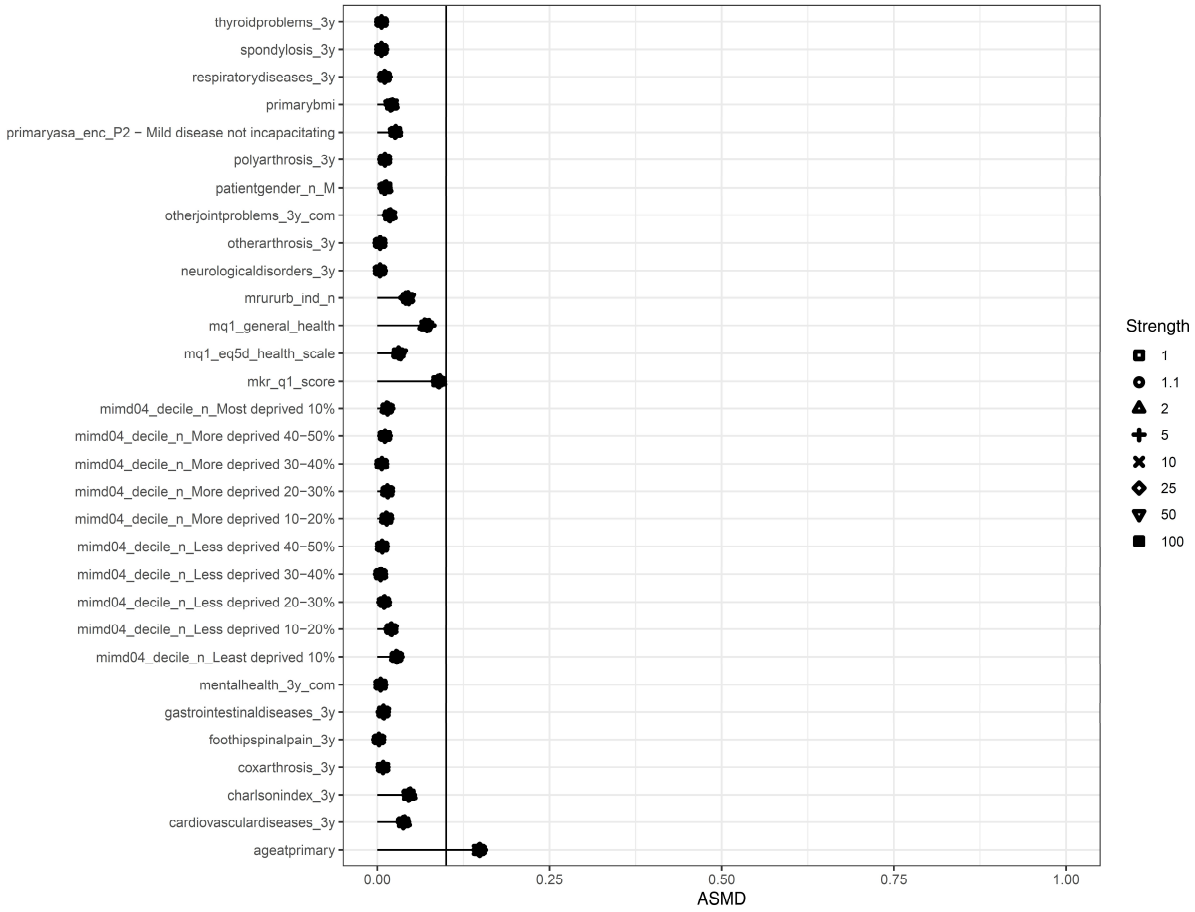
Binary outcome variable

ASMD before IV



Binary outcome variable

ASMD after real-world IV



Binary outcome variable

ASMD after synthetic IV

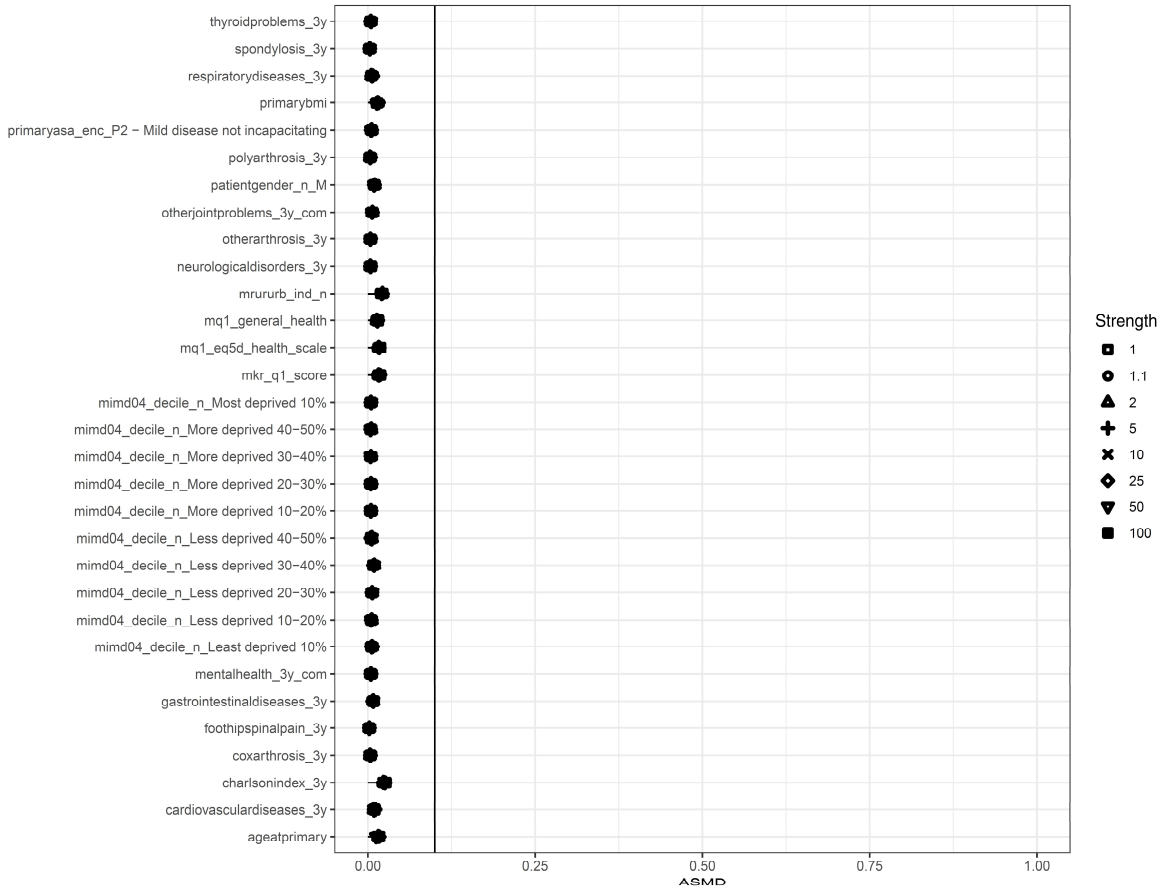


Figure 5.2 Maximum absolute standardised mean differences for each considered confounder for each cohort dataset (continuous/binary outcome variable) and instrumental variable (IV) method.

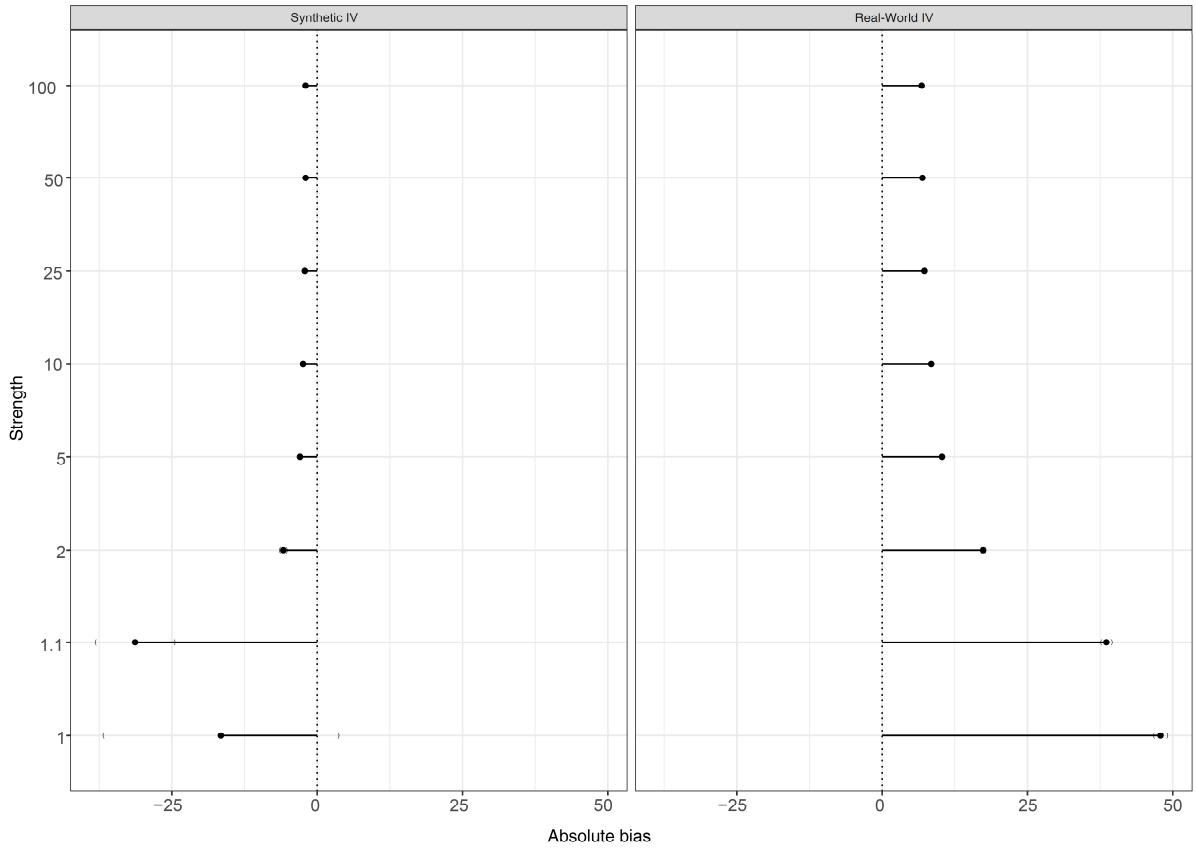
Bias, Error, and Coverage

Appendix Table 5.1 shows measures for all the proposed metrics (simulations that converged, estimates with SE, bias, errors, coverage and power) with their respective Monte Carlo errors.

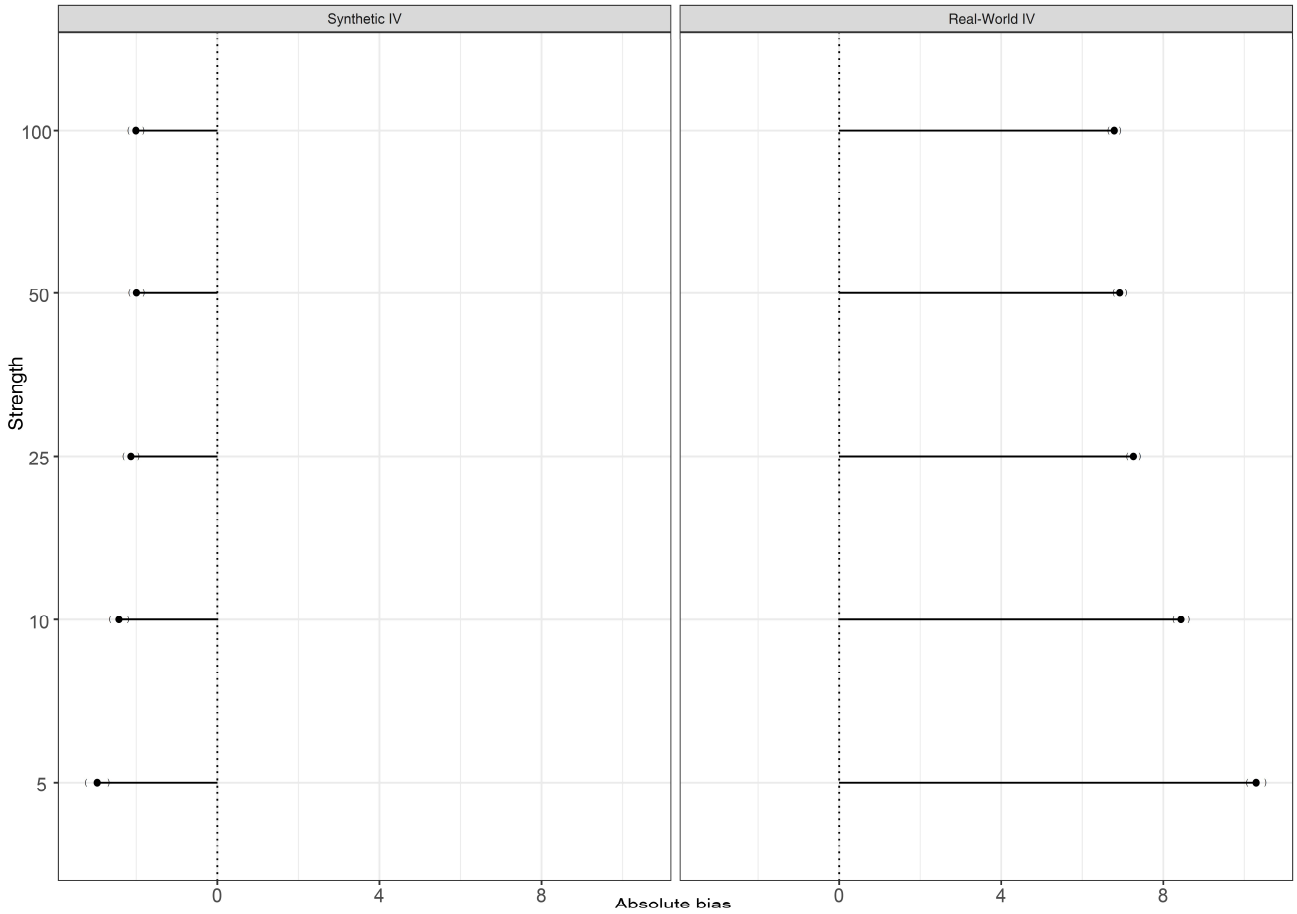
As for bias, in the continuous outcome dataset, all methods had substantial bias, even for the strongest instruments. Bias for the Synthetic IV with a strength of 100 is -2.01 95%CI (-2.19 to -1.83) , a 100% relative bias, and 6.79 CI (6.65 to 6.93) for the real IV, with a striking 340% relative bias. This bias seemed to get higher with weakest IV, specially, from instruments with an OR of 2 or lower. For example, with a strength of 2, the absolute bias was 17.4 (17.0 to 17.8). Weakest IVs also had the higher bias for the binary outcome, with absolute biases of 0.47 (0.45 to 0.49), a relative bias of 70% for the Synthetic IV with a strength OR=100, and of 1.02 (1.01 to 1.04) for the real IV, with a relative bias of 145%. Graphs of these biases and their Monte Carlo 95% confidence intervals are shown in *Figure 5.3.*

Continuous outcome

All

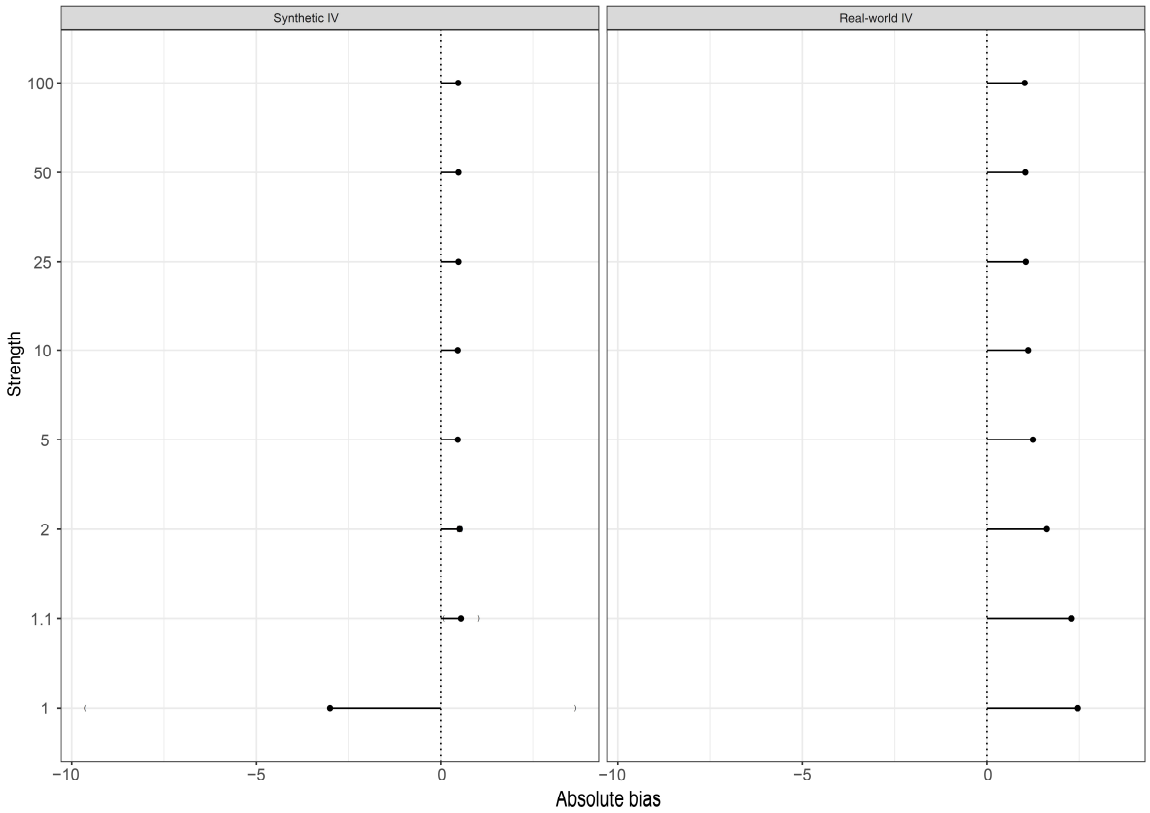


Only OR>2


Synthetic IV
Real-World IV
Strength
100
50
25
10
5
0
4
8
Absolute bias

Binary outcome

All



Only OR>2

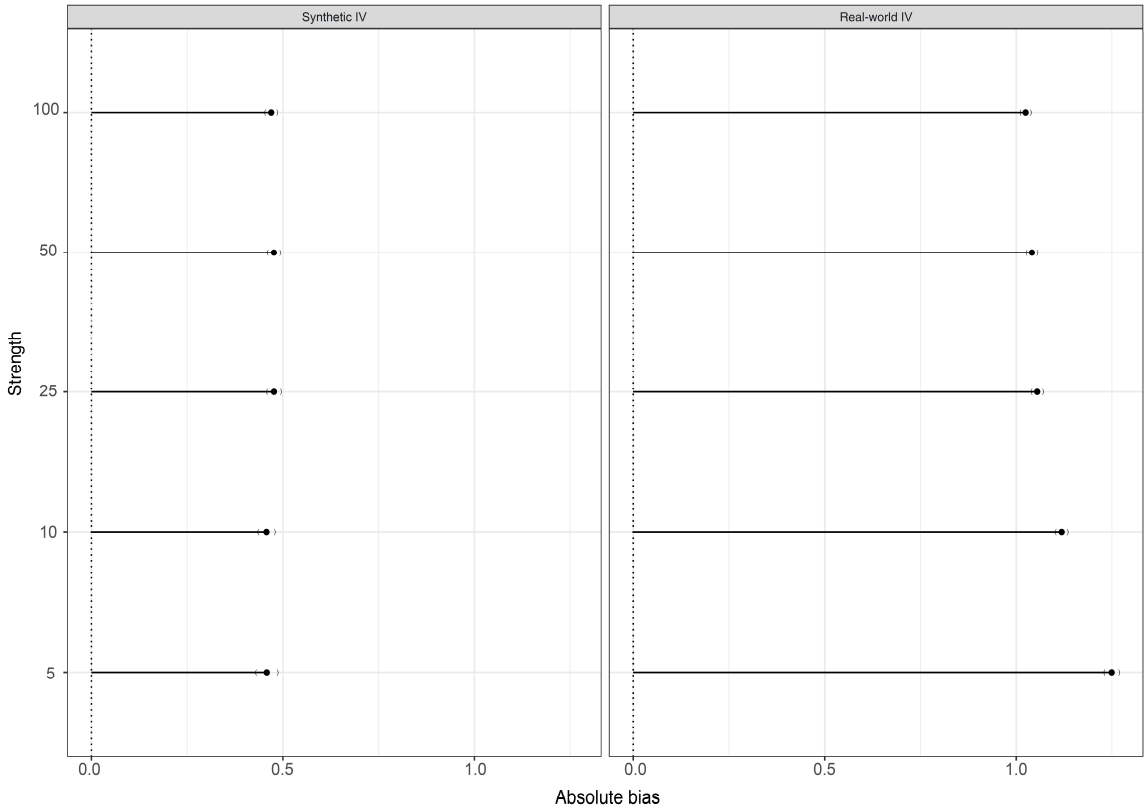
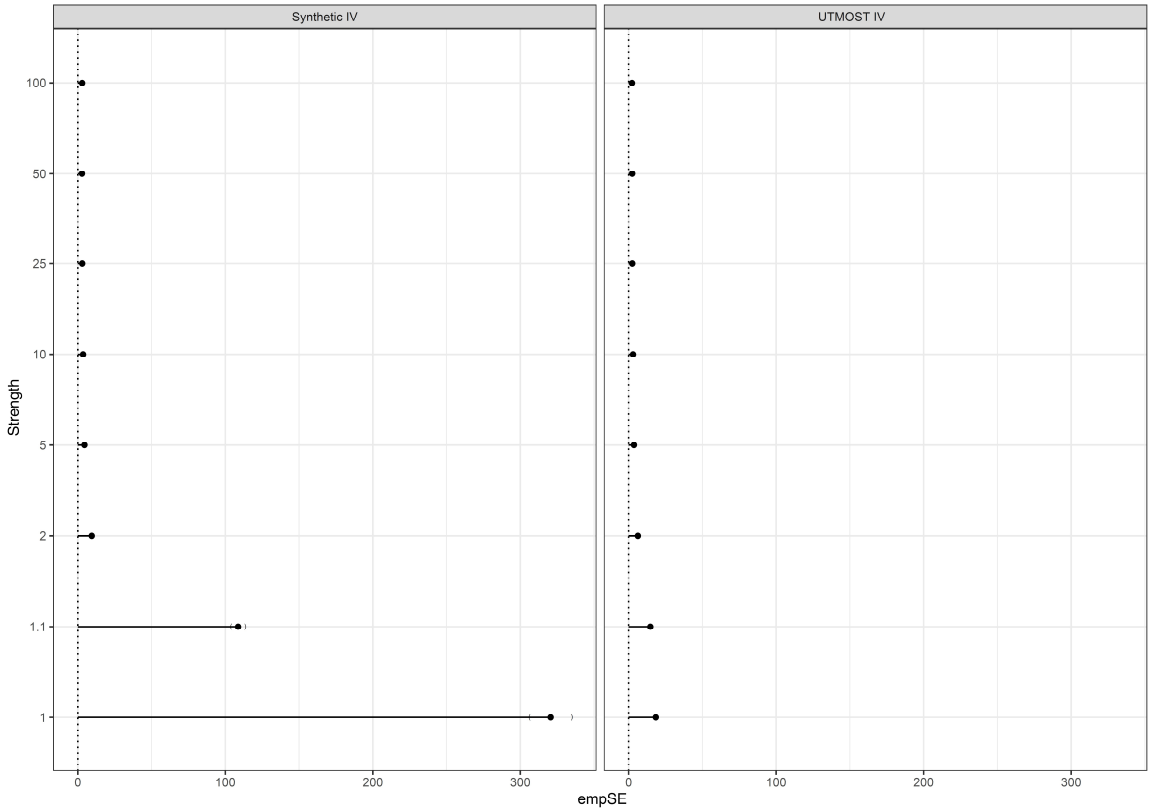


Figure 5.3 Absolute bias for each method and IV strength for each cohort dataset (continuous/binary outcome variable) and instrumental variable (IV) method

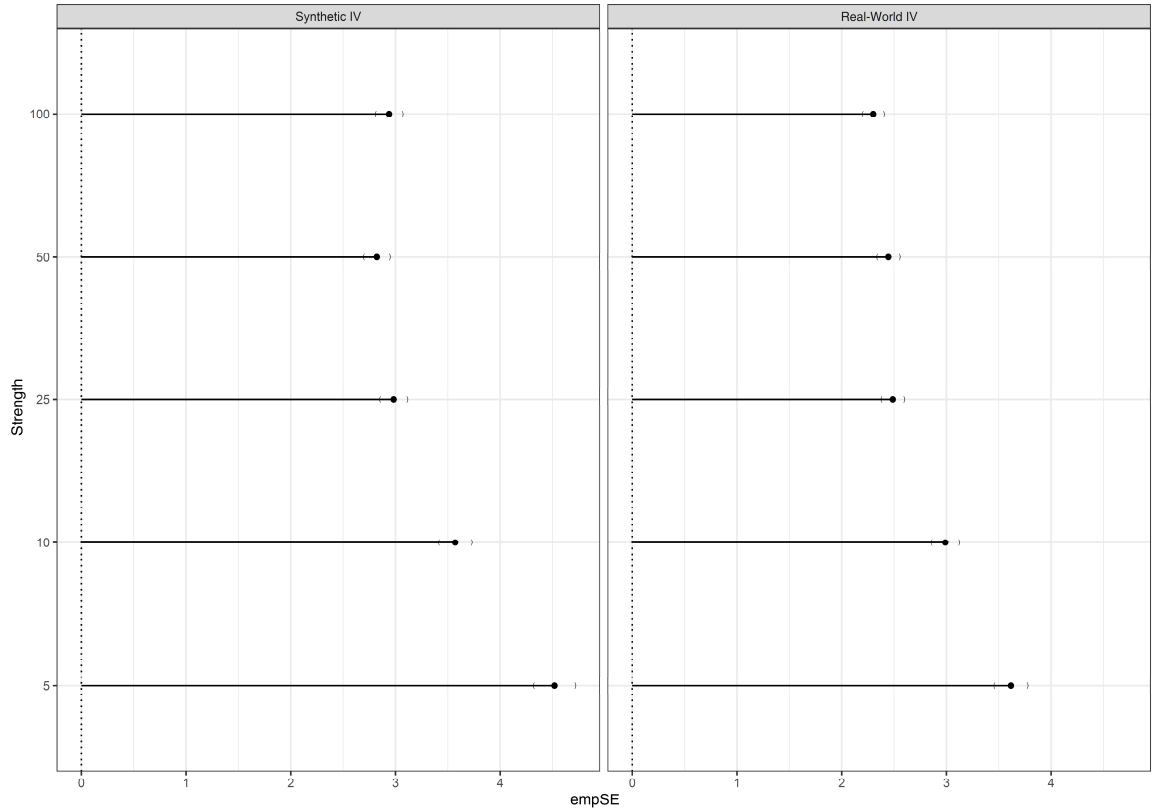
The empirical standard error seemed to get lower when IV strength increased. It ranged from 2.94 (2.81 to 3.07) for the Synthetic IV and 2.30 (2.20 to 2.40) in the real-world IV with strength OR=100, to 320 (306 - 334) with the Synthetic IV and 18 (17 to 19) with the real-world IV in the IVs with strength OR=1. The errors were very similar for all strengths between the synthetic IV and the real IV, except for the weak IVs (OR=1, and OR=1.1) where it increased much more in the synthetic IV. Similarly, standard errors among all simulated datasets with the binary outcome were small, shown in ***Figure 5.4.*** For example, the scenario with a strength of OR=100 for both synthetic and real-world IV had a standard error of 0.25 95%CI(0.24 - 0.26) and 0.21 95%CI(0.20 - 0.22), respectively. By contrast, standard error of 1.03 CI(0.98 - 1.07) in the synthetic IV and 0.45 CI(0.43 - 0.47) in the real-world IV were observed with a weakest strength of OR=2.

Continuous outcome

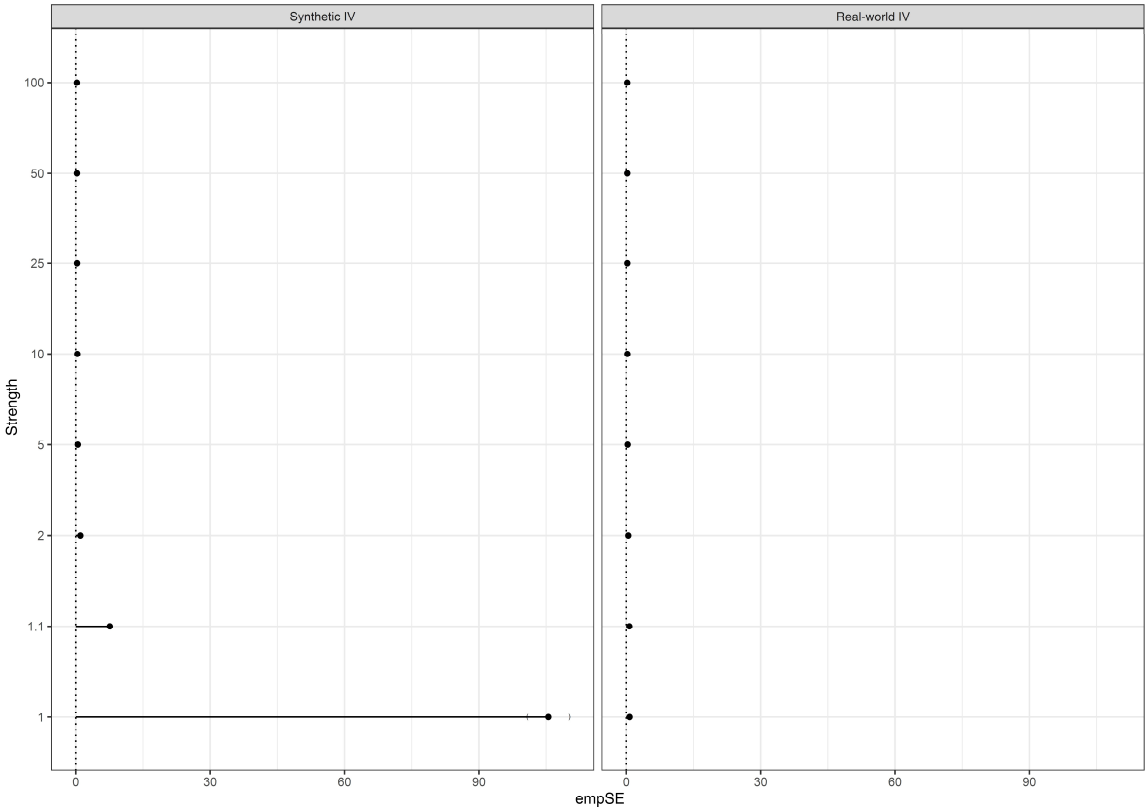
All



Only OR>2


Synthetic IV
Real-World IV
Strength
100
50
25
10
5
0
1
2
3
4
empSE

Binary outcome
All



Only OR>2

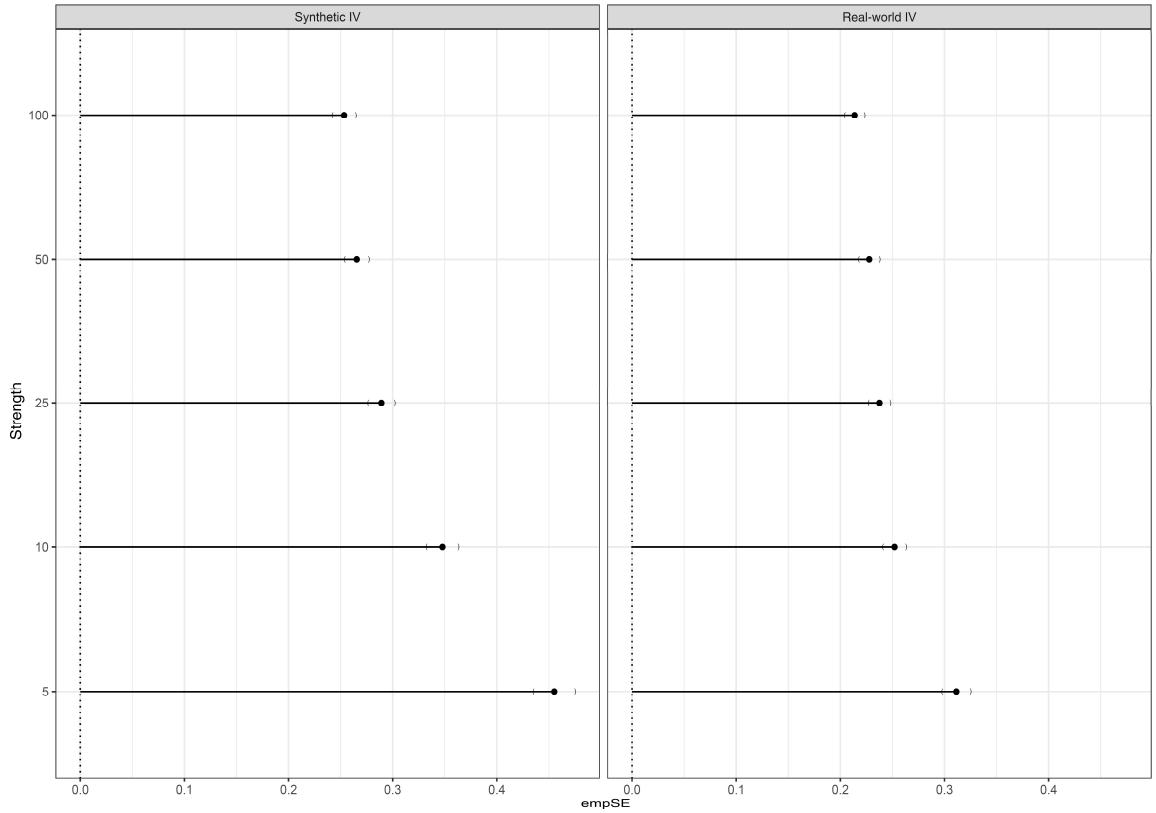
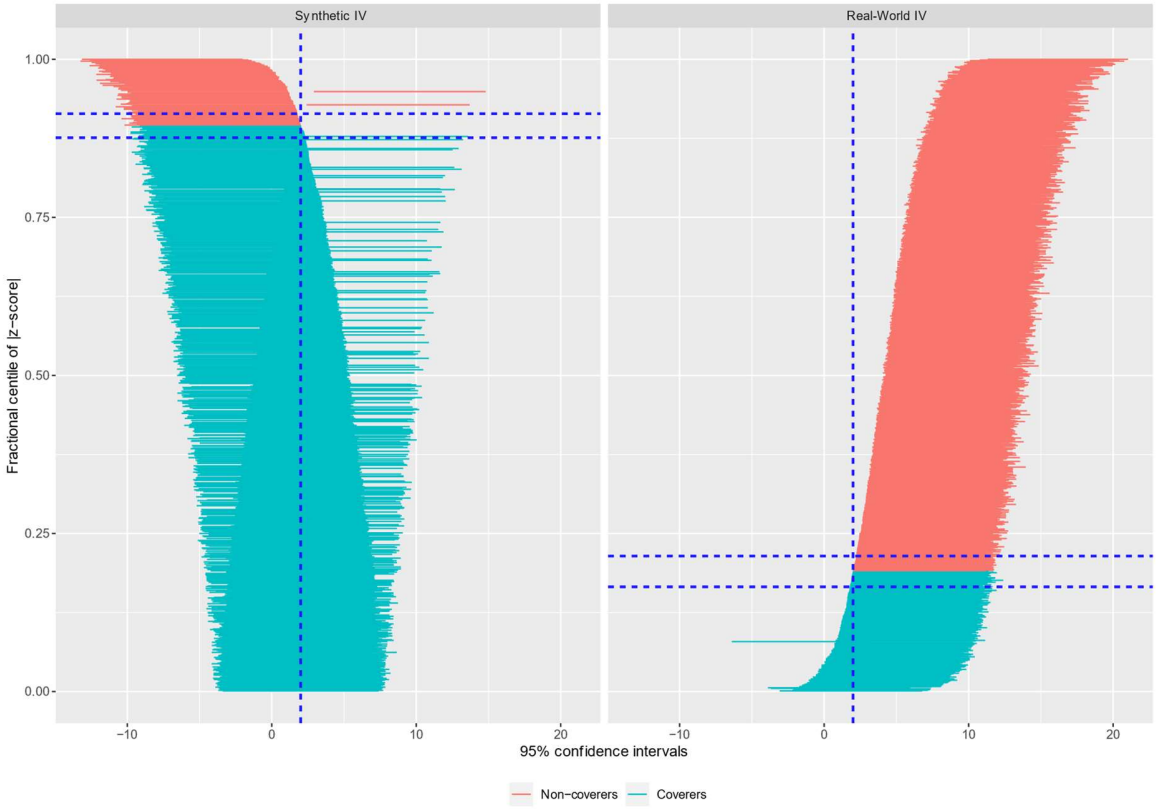


Figure 5.4 - Empirical standard Errors for each method and IV strength for each cohort dataset (continuous/binary outcome variable) and instrumental variable (IV) method

Coverage, the proportion of simulated estimators whose confidence intervals comprise the real effect, was significantly different between the Synthetic IV and the real-world IV. While the Synthetic IV coverage was very good for all strengths of the continuous variables, with a minimum of 86% (84% - 88%) in the IV strength of 100, the coverage of the real-world IV was poor for all strengths, with 17% (15% - 20%) for the same OR=100 IV strength. The coverage for the binary variable outcome was lower for very strong IVs in the synthetic IV, 53% (50%-56%) for the IV with strength 100 but increased with lower strength, 84% (81%-86%) for a strength of 5. However, the real-world IV coverage was near 0% for all strengths except for the weak IVs 1 and 1.1 where it was 10% (8%-11%). ***Figure 5.5*** shows the zip plots for Synthetic and real-world IV for both types of outcome with a strength of 50 as an illustration.

Continuous outcome



Binary Outcome

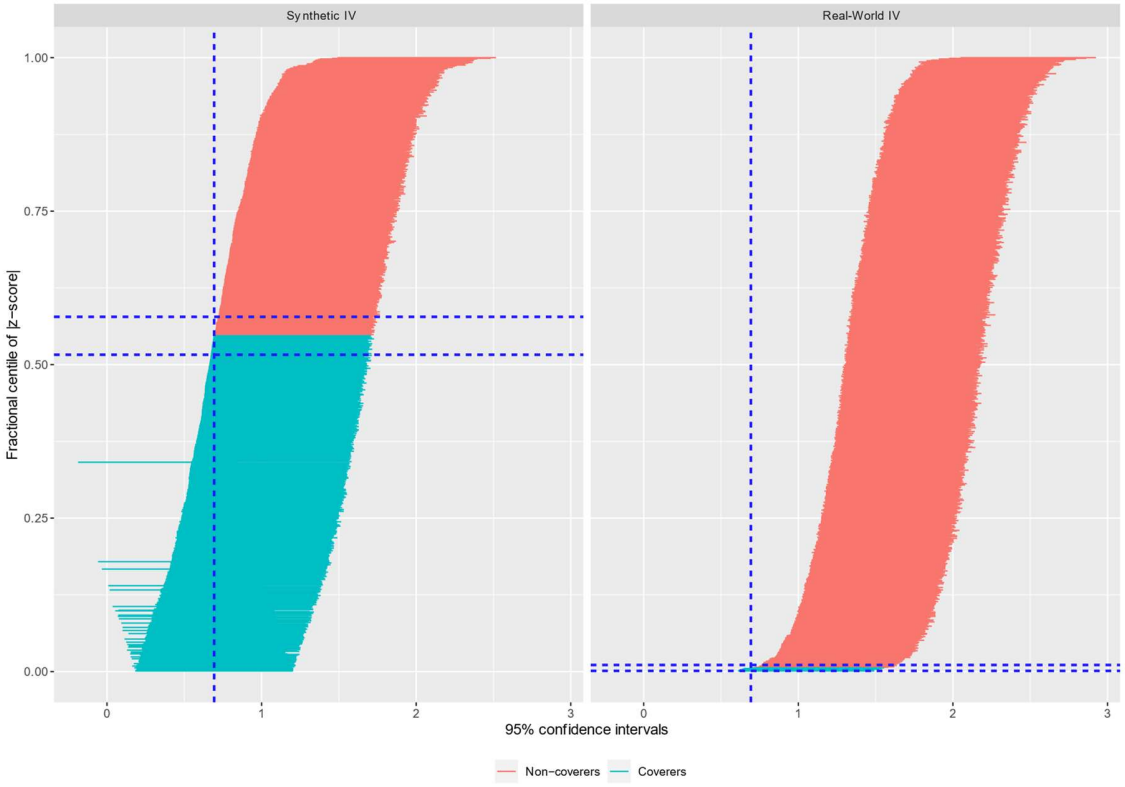


Figure 5.5 - Zip Plots sowing coverage for each simulated dataset and instrumental variable (IV) method on the a) continuous outcome and b) binary outcome with an IV strength of OR=50

Covariate testing

To investigate if any of the known confounders were driving the bias on the real IV, I repeated the strength OR=50 IV scenario by studying both in the exposure and outcome generation one confounder at a time. For this, I performed a simulation without any confouder (just IV, exposure and outcome) and a simulation per each confounder (IV, exposure, confounder, and outcome). Mean calculated strength ORs for the IV was consistent with means of around 50 in all variable scenarios. Some variables introduced massive absolute bias in the continuous outcome, in descending order: pre-operative OKS 4.42 95%CI(4.27 to 4.56), IMD 2.43 (2.27 to 2.59), EQ5D health scale 1.08 (0.92 to 1.24) and lead surgeon yearly output -0.66 (-0.81 to -0.51). The highest introducers of bias on the binary outcome analyses were age, which introduced a bias of 0.67 (0.65 to 0.68), lead surgeon yearly output with a bias of 0.43 (0.42 to 0.44) and EQ5D general health index 0.40 (0.38 to 0.41). There were no differences in empirical standard error. The coverage of the confidence intervals of the treatment effect estimate is shown in ***Figures 5.6 and 5.7***. All coverages for the continuous outcome were higher than 80% except for the scenario with pre-operative OKS score as a confounder, with a coverage of 54% (51 to 57%). For the binary outcome, the variables that decrease coverage below 80%

were age, with a coverage of 16% (13 to 18%), yearly surgeon volume 51% (48% to 54%) and EQ5D general health index 57% (54 to 60%).

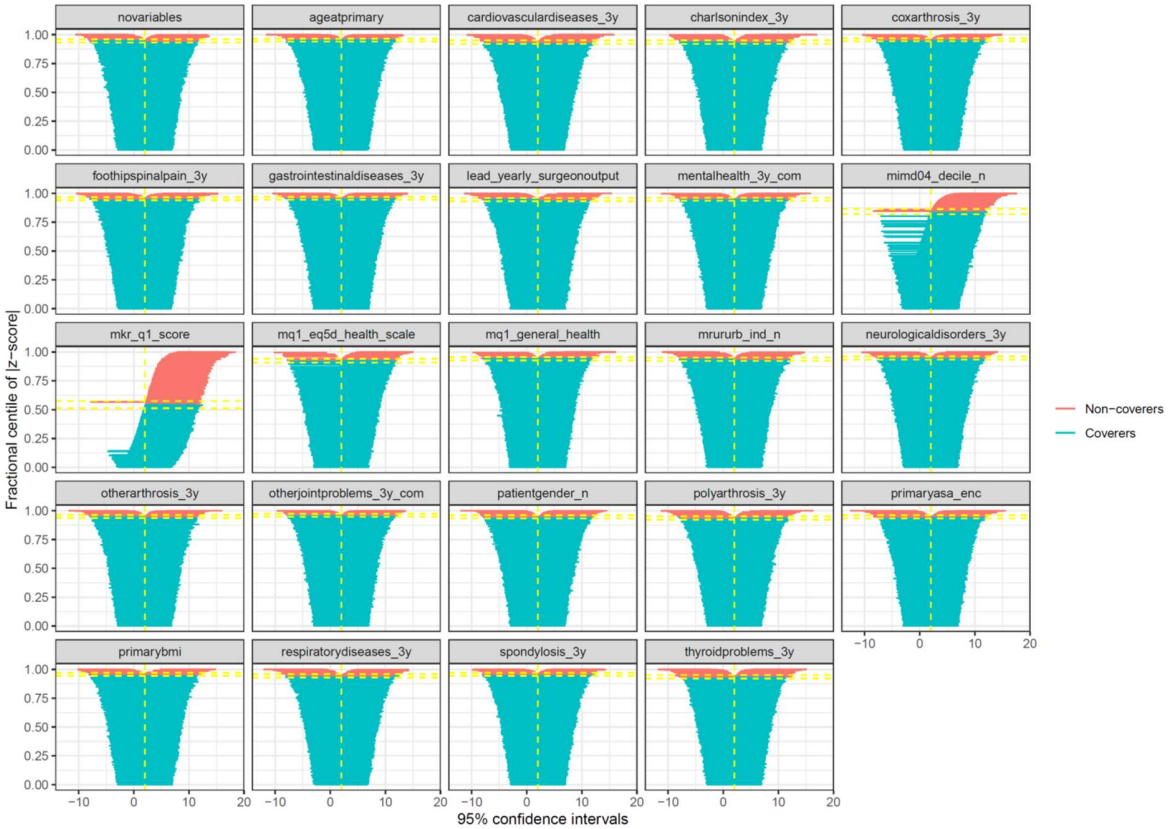


Figure 5.6 - Zip Plots sowing coverage for each simulated dataset using one confounder at a time with the Real-world IV on the continuous outcome with an IV strength of OR=50.

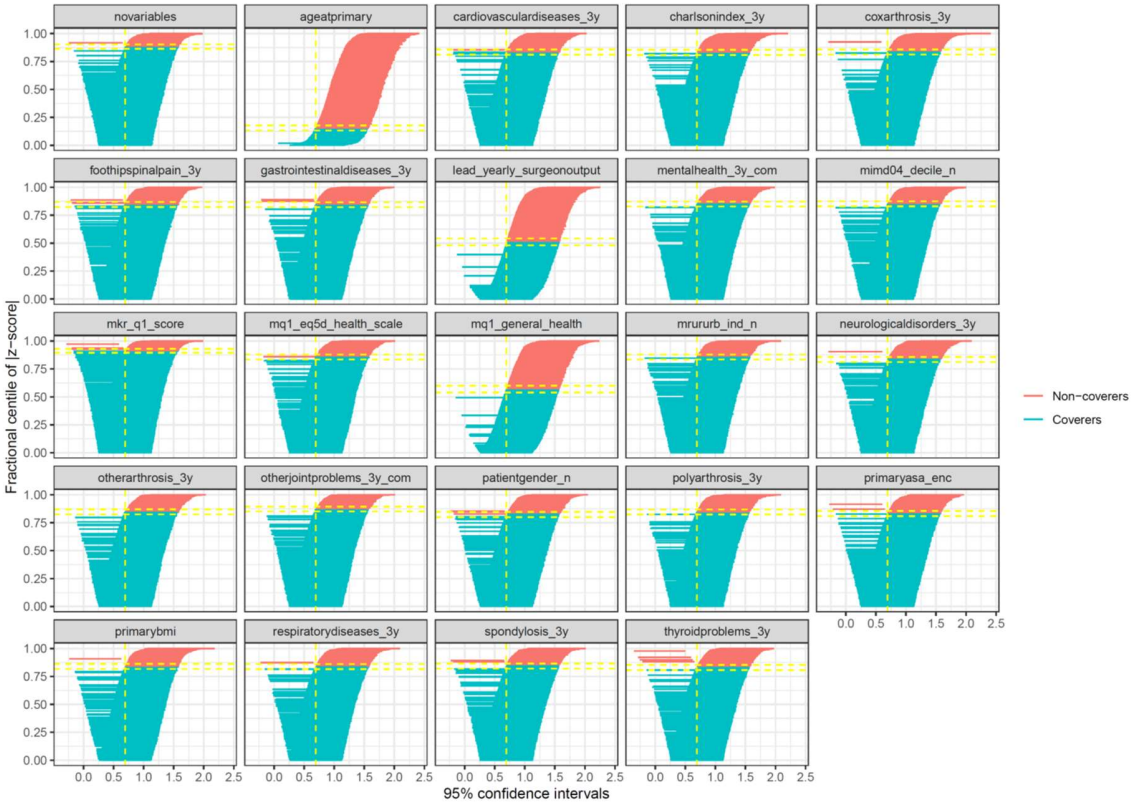


Figure 5.7 - Zip Plots sowing coverage for each simulated dataset using one confounder at a time with the Real-world IV on the binary outcome with an IV strength of OR=50.

Discussion

The mean strength of the IV is consistently inferior when using a Synthetic IV, totally exogenous per definition, compared to using the real-world IV used in the previous chapters. This could point to the relationship of the real-world IV with some other variable that predicts treatment, with the danger of this variable being a confounder, measured or unmeasured.

Balance for all variables was measured with the ASMD and evaluated using as a threshold the commonly accepted 0.1. (107) On the continuous outcome datasets, there was no imbalance, with ASMDs well below the threshold and a maximum of 0.07. In the binary outcome datasets, patient age at surgery was imbalanced.

There seems to be a 100% bias even for the high strength synthetic instrument in the continuous outcome datasets. This bias is highly increased to 340% with the IV used. The high standard error, even for strong instruments, includes the real effect in most of the synthetic database results, but not for the real-world IV, with coverages of around 17%. The real-world IV did not cover the real effect in any case for the binary variable, where the synthetic IV performed variably.

The high amount of bias points to the real-world IV used in previous chapters, preference of the surgeon, not being fit for purpose, and probably being related to

the outcome through variables other than the treatment. After investigating if this confounding was due to a known confounder, I found out that pre-operative OKS introduced bias, with a coverage of only 67%. Surgeon volume, and EQ5D introduce bias in the binary outcome. In the case of the continuous variable, we would have deemed the IV appropriate and went on with the analyses in all simulated datasets, as the problematic variable pre-operative OKS, had a maximum ASMD of 0.007 across datasets: over 10 times lower than the threshold. This raises concerns about the adequacy of the use of SMD as a suitable diagnostic to evaluate the exogeneity of IVs, as in this case it failed completely to identify relevant confounders.

These results are consistent with the odd results yielded previously by IVs, and how unreliable they seem to be in this context. The mechanism by which this high bias occurs seems to have two origins: First, a mix of known and unknown confounding was creating highly biased estimates with poor coverage, even though no variables were flagged as having imbalance issues; second, the bias of the IV method itself, even in the synthetic IV it produced more 100% bias. However, this high bias seems to be countered by the high SE, bringing coverage to acceptable levels. Having a point estimate way off the real effect and very wide

confidences intervals seems of little use for estimating even the biggest effects expected in MD effectiveness and safety.

This simulation study has several limitations. First, the fact that all simulated relationships between exposure and outcome were added as linear predictors. Some of these variables may have other kinds of relationships as a quadratic relationship or may be interacting with other variables, and this can bias the results. Furthermore, as the relationship between the exposure, IV, covariates and outcome is pre-defined, I was not able to explore unobserved confounders effect.

There is a chance, however, that this bias and high uncertainty are related to the structure of the real data itself. To explore that, I use a parametric simulation in a multi-level setting, such as those seen in surgical/MD epidemiology, in the next section.

5.2 Simulation study: Instrumental Variables and bias in multilevel surgical settings

Introduction

In the previous Section, I have shown how the use of Instrumental Variables (IVs) to produce unconfounded estimates may be strongly biased if the 3rd assumption of IVs, which states that the IV should be only related to the outcome through the exposure, is broken. I also have shown how this violation of assumption was not noticeable with the commonly used method of evaluating balance with SMDs. This contrasts with previous research suggesting that IVs yield correct unbiased estimates in cohort studies if all assumptions are fulfilled and IVs are strong enough. (110)

There are several potential explanations for this discrepancy. First, bias in my study could be caused by the non-normal distribution of the continuous outcome. OKS, like many functional and health related quality of life scores, had a non-normal distribution in the sample, bounded to 0–48. In the previous section, I used a normal distribution but truncated at 0 and 48. This could mean that the magnitude of the estimated effect could have been infra-estimated when the predicted post-surgery OKS plus the effect surpassed 48 in most patients. This

potential source of bias, using continuous non-normal and bounded outcomes, is crucial to be understood when the outcome results from a questionnaire or a scale, as it is often used in MD and surgical effectiveness research.

A second possibility is that not considering the multilevel structure is affecting the IV estimation. The selected IV is based on the observed preference. Although the true preference could be, in theory, a valid IV, the multilevel structure could be also affecting the observed preference. This could happen through the interactions with other surgeon variables, such as surgical volume, which is known to impact outcomes in surgical settings. Furthermore, the ratio patient-surgeon could have an effect on the error of the estimate. Another source of bias could be the low prevalence of the exposure (around 7%).

In this section, I looked at the effect of the multilevel structure on the bias and error of the IV estimation, exploring: 1) different surgeon-patient ratios, 2) different generation models for the outcome, 3) different effects of the surgeon volume on the outcome, and 4) different exposure prevalence. I tested these scenarios with different strengths and true effects on bias, error, and coverage.

Methods

Data Generation

In this section, I created simulated datasets using a Monte Carlo parametric simulation. I created 130,000 observations per dataset. I generated 240 simulated datasets per each scenario. A graphical depiction of the whole data generation process can be seen in *Figure 5.8*.

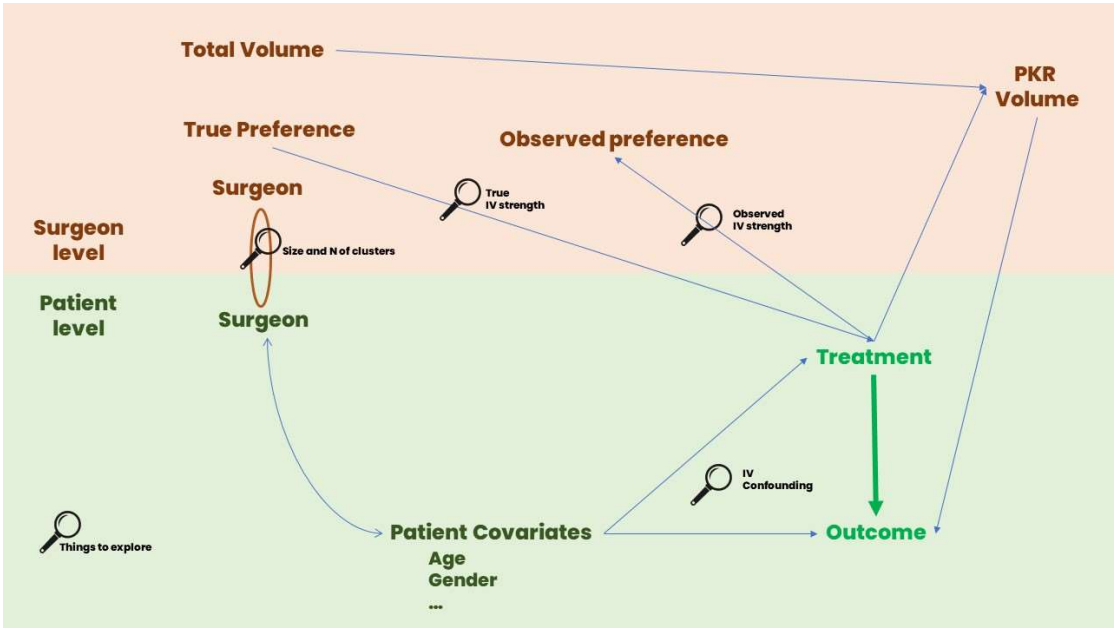


Figure 5.8 Theoretical model for the data generation.

Patient-level covariates

The theoretical capacity of IVs for handling patient level confounding has already been demonstrated.(40) For simplicity, I only used 2 patient level variables, inspired by gender and age. The binary variable (x1) created as a binomial process (0,1) with 50% chance and the continuous one (x2) created as a normal distribution with mean 70 and SD 8.9.

Surgeon-level covariates

The 130,000 patients were assigned to different number of surgeons, depending on the scenario. I created datasets with 1 (where there was no multilevel structure in the data), 5, 10, 50, 100, 300, 500 and 1000 (where there was complex multilevel structure in the data) surgeons. Patients were distributed evenly between surgeons, creating a fix patient-surgeon ratio. I used 2 independently generated surgeon level variables: the *true preference* as gamma distributed among surgeons with shape 0.1 scale 3 and truncated at 0 and 1, and deprivation index as normally distributed with mean 100 and SD 8.9. These distributions tried to mimic the real variables seen in previous chapters.

Exposure

The exposure here is the type of treatment, 0 being TKR and 1 PKR. Treatment (ti) was a function:

$t_i \sim Bernoulli\ (\pi_i)$

$$\log\left(\frac{\pi_i}{1-\ \pi_i}\right) = \alpha_0 + \ \alpha_1 \times x_1 + \ \alpha_2 \times x_2 \ + \ Strength_{IV} \times Real\_Preference$$

Where:

- α_1 and α_2 are the coefficients of each covariate: α_1=-0.073 α_2=0.30;
- strength IV could be 0, 1.5, 5, 10, 50, 100;
- α_0 is calculated as the necessary to reach a prevalence equal to 0.05, 0.25, and 0.5 by one dimensional root finding;

Observed Instrumental Variable and Volume Confounding

The *observed preference* IV means to mimic the way surgeon preference IVs are estimated in the real-world data. This variable is supposed to be a proxy of the *true preference*, a real instrumental variable, but differs in the way that the *observed preference* is calculated from the observed treatment in the database. To do so I used the same method from **Chapters 2 and 3** (% of surgeries in the 30 previous surgeries of the same type of PKR or TKR performed by the same surgeon). I also

generated a volume variable, as the number of PKR performed previously by the surgeon.

Outcome

Continuous Outcome

The continuous outcome (Y_i) was modelled as a function of two patient covariates, the true treatment effect and the surgeon volume confounding as an interaction:

$$Y_i = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_{treat} \times Treatment + \beta_{vol\_conf} \times Vol_{PKR} \times Treatment$$

Where:

- β_1 and β_2 are the coefficients of each covariate: β_1=0.062 β_2=2.4;
- β_0 is generateβ in 3 different ways;
    - Normally distributed with μ =30, and σ = 8;
    - Normally distributed with μ =30, and σ = 8 and truncated at 0 and 48.
    - Gamma distributed with shape=4.2, and scale = 8, truncated at 0 and 1 and multiplied per 48.
- β treat is the true treatment effect: 0,1, or 3;
- βvol_conf is the magnitude of the interaction of the volume and the treatment: 0, 0.1, or 1;

Following these settings, I generated three different outcomes: one normal without any range limit, where there should be no bias due to an outcome regression misspecification, and two with a range from 0 to 48, one modelled with Gamma and the other with Normal distributions, where this bias may occur.

Binary Outcome

The binary outcome (Y_i) was also modelled in the same fashion:

$$t_i \sim Bernoulli\ (\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_{treat} \times Treatment$$

$$+ \beta_{vol\_conf} \times Vol_{UKR} \times Treatment$$

Where:

- β_1 and β_2 are the coefficients of each covariate: β_1=-0.07 β_2=0.3;
- β_0 is calculated as the necessary to reach a prevalence equal to 0.05 by one dimensional root finding;
- β treat is the true treatment effect: 0,1, or 3;
- βvol_conf is the magnitude of the interaction of the volume and the treatment: 0, 0.1, or 1.

Methods and Estimand

I used the same method as in **Section 5.1**, 2 stage least squares with a median dichotomised IV as a method to estimate the treatment effects, as it is the most widely used in pharmacoepidemiology studies. I evaluated the beta for PKR in a linear IV regression for the continuous variable and in a Poisson IV regression for the binary variable.

Scenarios and rationale

Table 5.3 summarises the different values that change for each scenario. I explored the combination of 6 IV strengths, 3 exposure prevalences, 8 patient:surgeon ratios, 3 volume confounding effects, 3 effect betas, and 4 types of outcomes. In total, 5,184 scenarios, with 240 repetitions each. The choice of IV strength came from the need of having both week IVs and very strong IVs, similar to what we would see between treatment allocation and treatment on a trial. Exposure prevalence tries to mimic a low exposure prevalence, as seen in previous chapters, a middle scenario of 25% and a 50/50 scenario. The election of patient per surgeon tries to explore the difference of how having a different number of clusters affects the estimation, from 1 to 1000 clusters. True effects and outcomes are similar to what I have seen in the previous chapters.

Variable	Values
IV strength	0, 1.5, 5, 10, 50, 100
Exposure prevalence	5%, 25%, 50%
Patients per surgeon	130, 260, 433, 1300, 2600, 13000, 26000, 130000
True effects beta	0, 1, 3
Volume Confounding	0, 0.1, 1
Outcomes	Binary (5% outcome prevalence)
	Continuous: normal, truncated normal, truncated gamma

Table 5.3: Summary of values that change in each simulated scenario

Performance measure(s)

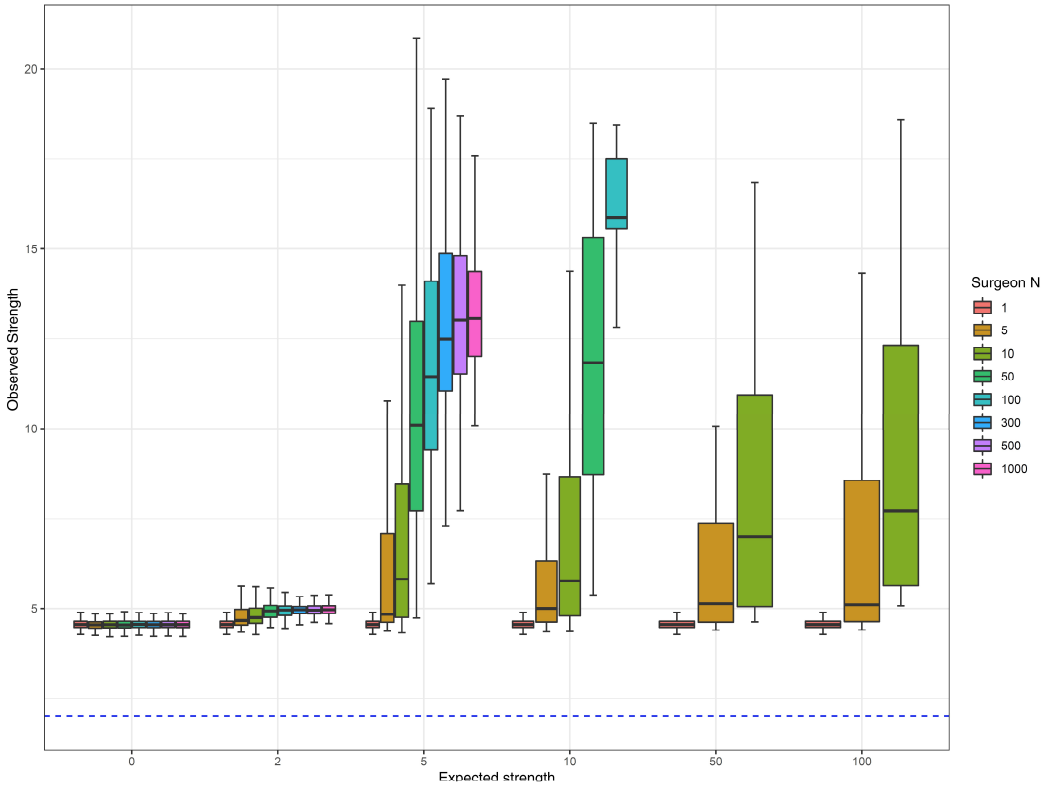
I followed the same strategy as in **Section 5.1**. I quantified absolute bias, defined as the average difference between the true treatment effect and the one yielded by the 2sls regression; empirical standard error, defined as the standard error of the point estimates. I also calculated the root mean square error (RMSE). I calculated the average model-based SE, as the average of the SE of each analysis. I calculated coverage, as the proportion of 95% confidence intervals obtained that included the true treatment effect.

Results

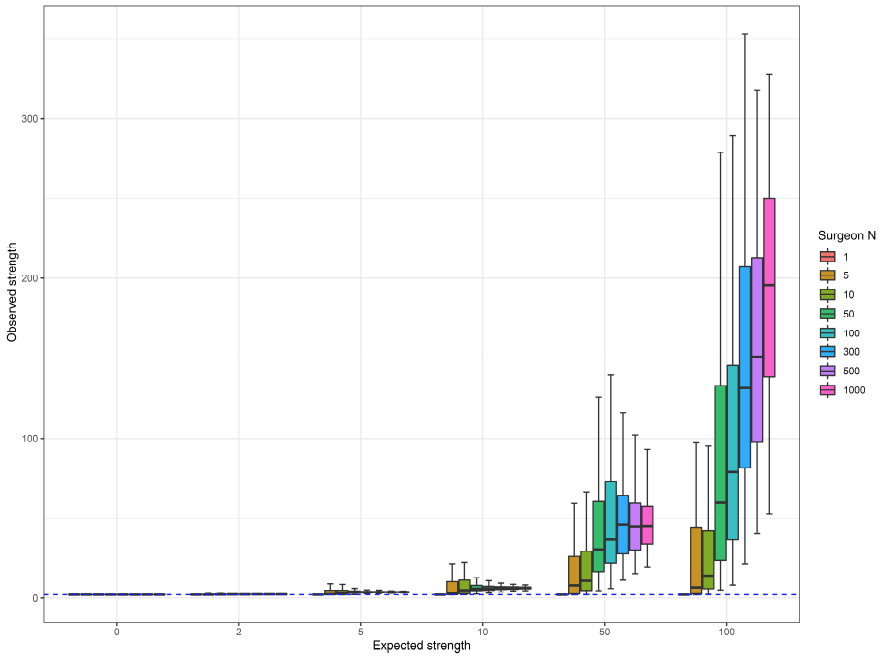
Strength

Table 5.4 and *Figure 5.9* show the strength achieved by the observed preference vs the theoretical strength (beta) of the *real preference*, by prevalence of exposure and number of surgeons. As shown, the strength achieved depended greatly on the prevalence of the exposure.

Exposure prevalence 5%



Exposure prevalence 25%



Exposure prevalence 50%

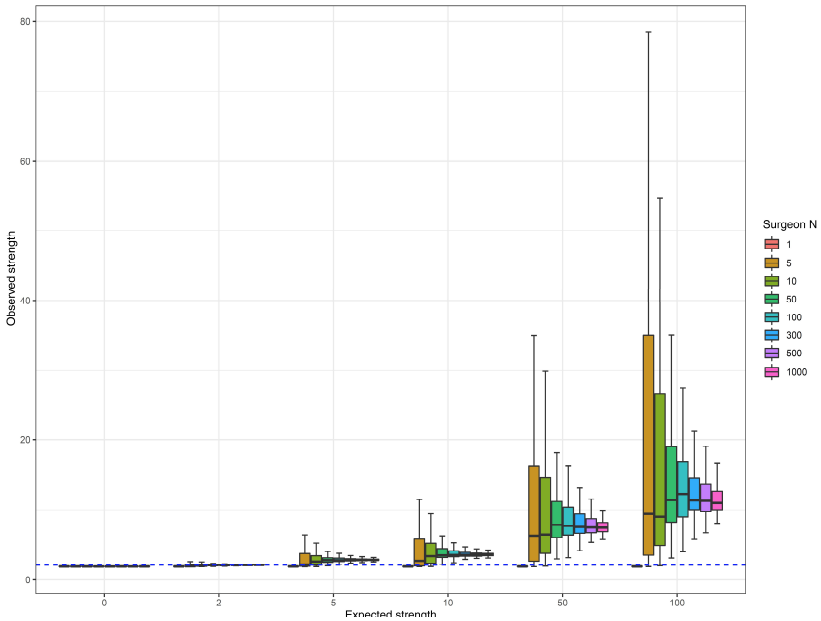


Figure 5.9 Box Plot of the achieved strength (observed Odds Ratio) of the observed preference IV compared to the exponentiated beta of the IV in the exposure generation formula (OR strength of the true preference IV)

				Prevalence								
				5%			25%			50%		
True IV - Exposure beta	OR	surgeons	patients per surgeon	p50	Min	Max	p50	Min	Max	p50	Min	Max
0	1	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
0	1	5	26000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
0	1	10	13000	4.6	4.2	5	2	1.9	2.1	1.8	1.8	1.9
0	1	50	2600	4.6	4.2	4.9	2	1.9	2.1	1.8	1.8	1.9
0	1	100	1300	4.6	4.2	4.9	2	1.9	2.1	1.8	1.8	1.9
0	1	300	433	4.6	4.2	4.9	2	1.9	2.1	1.8	1.8	1.9
0	1	500	260	4.6	4.2	5	2	1.9	2.1	1.8	1.8	1.9
0	1	1000	130	4.6	4.2	5	2	1.9	2.1	1.8	1.8	1.9
1.5	4	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
1.5	4	5	26000	4.7	4.3	6.6	2	1.9	3.4	1.9	1.8	3
1.5	4	10	13000	4.7	4.3	6.5	2.1	1.9	3.2	1.9	1.8	3.1
1.5	4	50	2600	4.9	4.5	5.7	2.2	2	2.6	2	1.8	2.4
1.5	4	100	1300	5	4.5	5.7	2.2	2	2.6	2	1.9	2.3
1.5	4	300	433	5	4.5	5.4	2.2	2.1	2.4	2	1.9	2.2
1.5	4	500	260	5	4.6	5.4	2.2	2.1	2.3	2	1.9	2.1
1.5	4	1000	130	5	4.6	5.4	2.2	2.1	2.3	2	1.9	2.1
5	148	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
5	148	5	26000	5.1	4.4	Inf	2.3	2	39.5	2.1	1.8	16.1
5	148	10	13000	7.2	4.3	Inf	3.1	1.9	34.5	2.5	1.8	25.4
5	148	50	2600	11.6	4.7	Inf	3.5	2.1	8.6	2.7	1.9	5.4
5	148	100	1300	12	5.7	Inf	3.5	2.6	7	2.7	2.1	4.6
5	148	300	433	12.6	7.3	Inf	3.6	2.7	5	2.8	2.3	3.4
5	148	500	260	13	7.7	Inf	3.6	2.9	4.7	2.8	2.4	3.3
5	148	1000	130	13.1	10.1	19.4	3.6	3	4.4	2.8	2.5	3.2
10	22026	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
10	22026	5	26000	6.6	4.4	Inf	3	2	Inf	2.6	1.8	149.7
10	22026	10	13000	Inf	4.4	Inf	4.8	2	Inf	3.3	1.8	239.1
10	22026	50	2600	Inf	5.4	Inf	5.7	2.4	46.7	3.4	2.1	13.9
10	22026	100	1300	Inf	12.5	Inf	5.8	3	22.4	3.5	2.3	7.4
10	22026	300	433	Inf	Inf	Inf	6.1	3.7	12.6	3.5	2.8	5
10	22026	500	260	Inf	Inf	Inf	6.1	4.3	8.4	3.5	3	4.9
10	22026	1000	130	Inf	Inf	Inf	6.1	4.5	8	3.5	3	4.3
50	5.18E+21	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
50	5.18E+21	5	26000	Inf	4.4	Inf	17.7	2	Inf	6.3	1.8	13798.6
50	5.18E+21	10	13000	Inf	4.6	Inf	25	2.1	Inf	6.4	1.9	7914.5
50	5.18E+21	50	2600	Inf	Inf	Inf	42.4	4.5	Inf	7.8	2.9	377.6
50	5.18E+21	100	1300	Inf	Inf	Inf	44.7	5.8	Inf	7.7	3.1	27.2
50	5.18E+21	300	433	Inf	Inf	Inf	46.1	11	Inf	7.5	4.1	14
50	5.18E+21	500	260	Inf	Inf	Inf	44.3	15.2	294.8	7.5	5.3	12
50	5.18E+21	1000	130	Inf	Inf	Inf	44.6	19.3	116	7.4	5.7	11.5
100	2.69E+43	1	130000	4.6	4.3	5	2	1.9	2.1	1.8	1.8	1.9
100	2.69E+43	5	26000	Inf	4.4	Inf	69.6	2	Inf	9.4	1.8	10828.7
100	2.69E+43	10	13000	Inf	5.1	Inf	92	2.2	Inf	9	1.9	16627.7
100	2.69E+43	50	2600	Inf	Inf	Inf	258.9	4.9	Inf	11.4	3	5032.3
100	2.69E+43	100	1300	Inf	Inf	Inf	Inf	7.9	Inf	12.3	3.9	66
100	2.69E+43	300	433	Inf	Inf	Inf	Inf	21.8	Inf	11.4	5.7	23.3
100	2.69E+43	500	260	Inf	Inf	Inf	Inf	40.2	Inf	11.3	6.7	20.3
100	2.69E+43	1000	130	Inf	Inf	Inf	Inf	52.7	Inf	11	8	17.5

Table 5.4: Median, min and max achieved strength (in Odds Ratio) compared to the beta of the IV in the exposure generation formula per each scenario

Prevalence of exposure limited the IV strength achieved, especially in the lower limit as seen in ***Table 5.4***. When real preference was unrelated to the exposure, with an IV strength of 1, the observed preference had a mean OR strength of 4.5 for 5% prevalence, 2 for 25% and 1.8 for 50%.

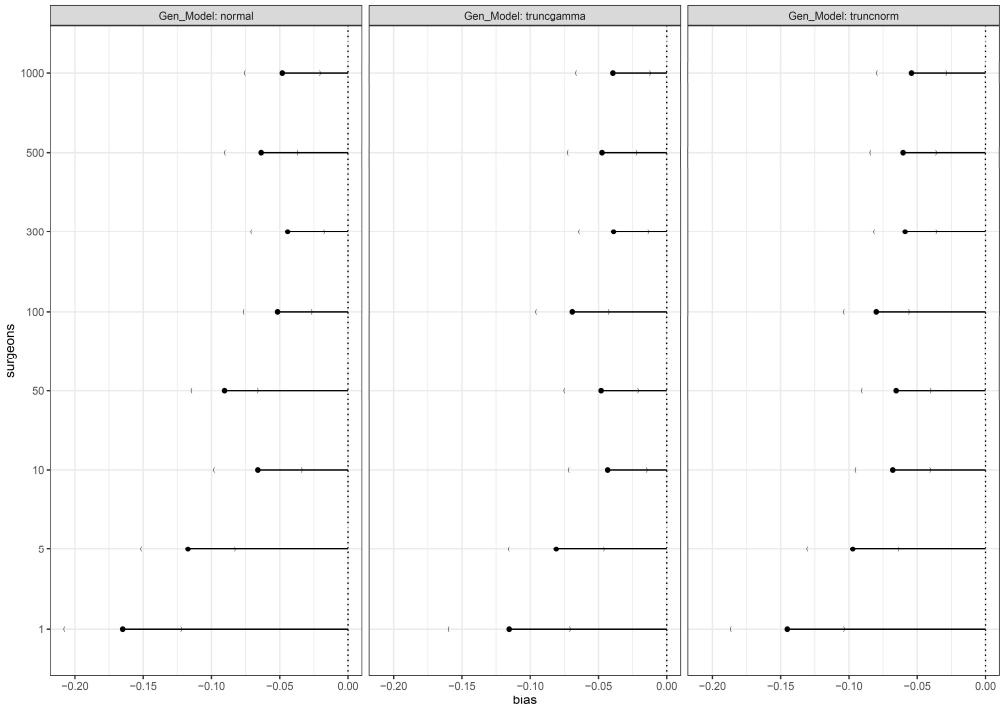
In the low exposure prevalence setting (5%), the observed strength stayed the same (mean OR of 4.5) with an increase in the real-world IV strength, up to very strong IVs, where the observed strength tended to go to infinity. In the medium prevalence setting (25%), observed strength increased with real strength from a median OR of 2 in the low strength setting to OR of over 200 in high strength settings. It also tended to infinity in some high strength settings. The high prevalence scenario (50%) increased the observed strength with an increase of the real strength, but in a more limited range, from median observed IV of 1.8 in a low strength setting to over 11 in a high strength setting.

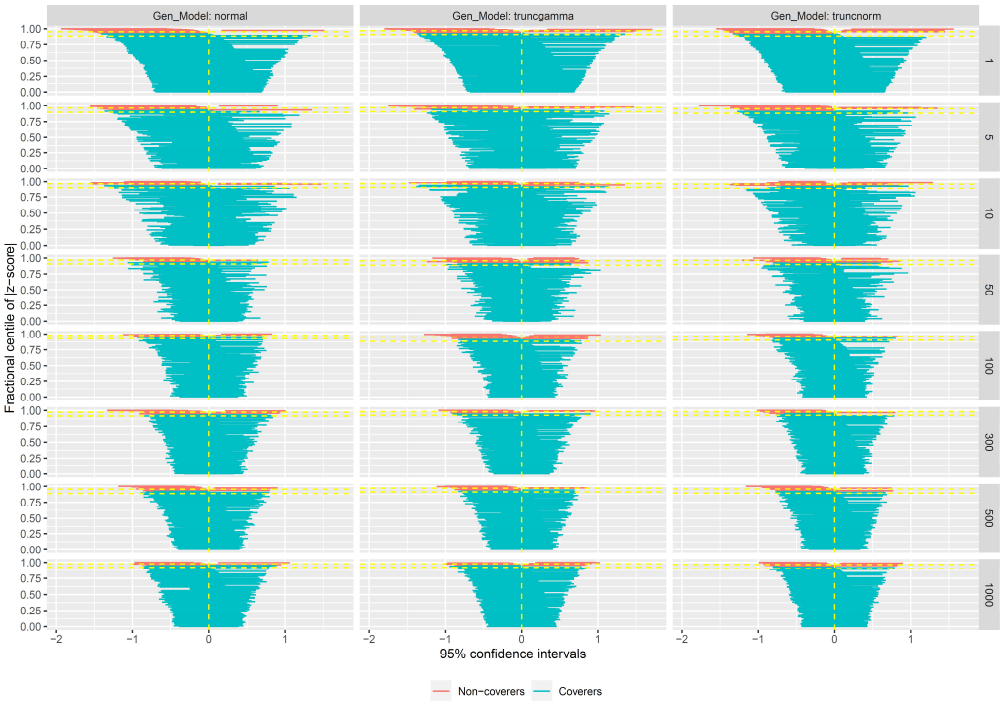
The third factor that influences the strength is the number of surgeons (or the ratio surgeon:patient). A higher number of surgeons seemed to increase the mean and median observed OR for most scenarios. The high number of surgeons increased the minimum OR achieved among simulations, but the highest maximum OR was seen with a relatively low number of surgeons (5 or 10).

Generation model

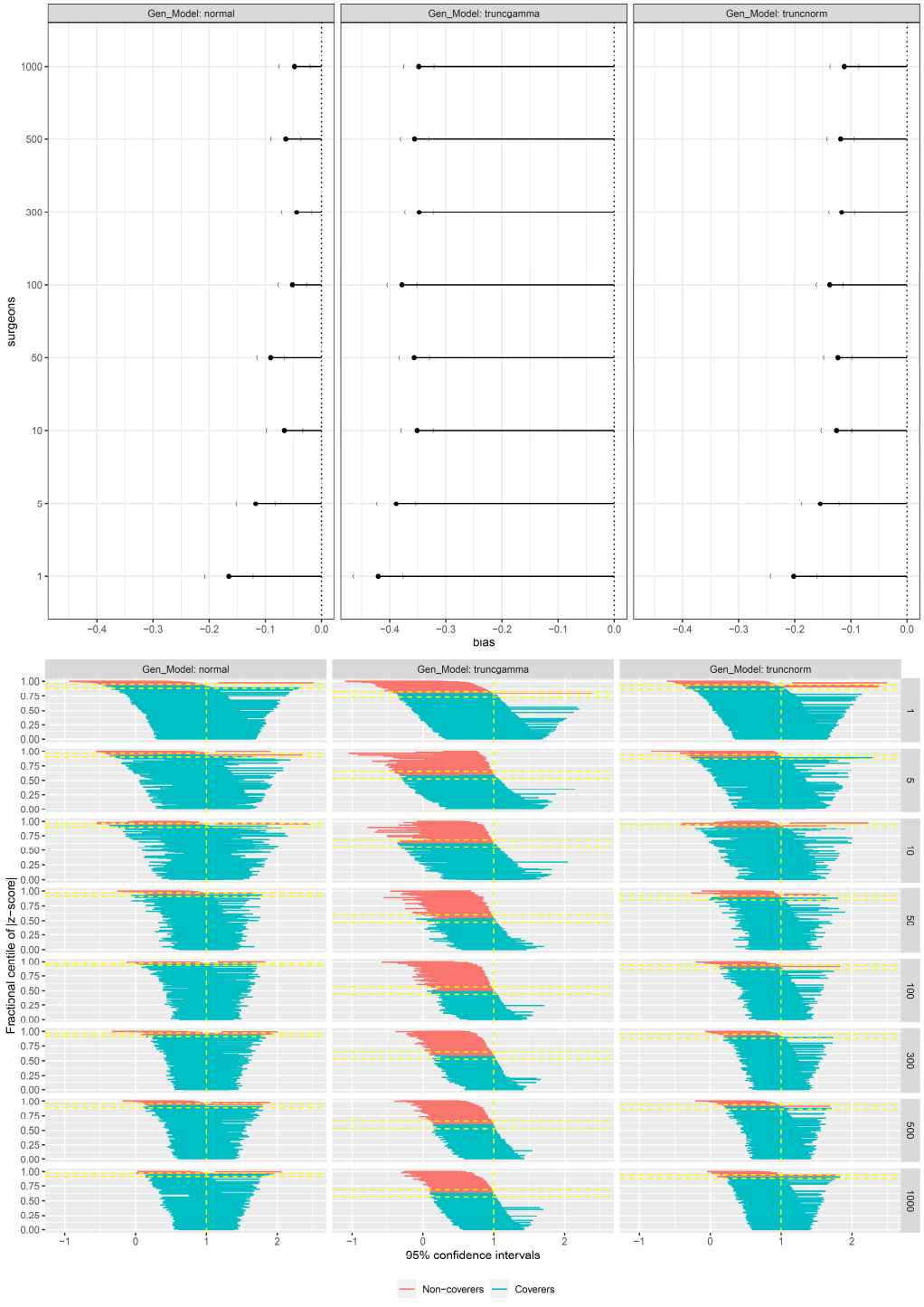
For the analysis of the bias in the continuous outcome caused by the generation model I compared a normal outcome vs a normal bounded outcome vs a gamma bounded outcome. I focused on the scenarios with a medium prevalence of exposure (25%), no volume confounding and an IV strength beta of 5. ***Figure 5.10*** shows absolute bias and coverage plots for all number of surgeons. I will refer to the 1000 surgeons in text, as results vary little across surgeon number, except for 1 and 5 surgeons where bias is slightly increased.

True Effect = 0




Gen_Model: normal
Gen_Model: truncgamma
Gen_Model: truncnorm
1
5
10
50
100
300
500
1000
Fractional centile of |z−score|
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
95% confidence intervals
Non−coverers
Coverers

True Effect = 1



True Effect = 3

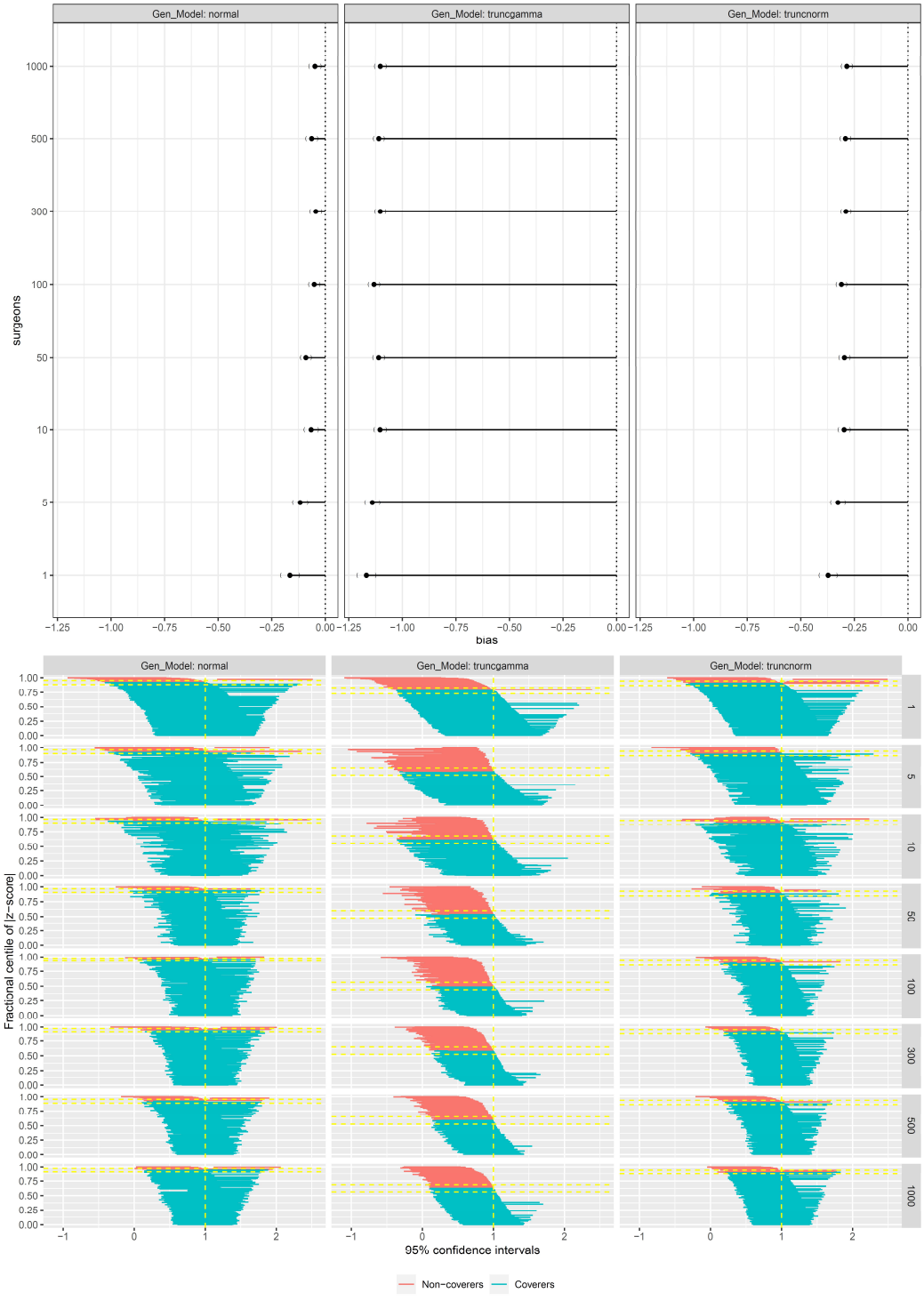


Figure 5.10 Absolute bias and Coverage per generation model, normal, normal bounded, and non-normal bounded(normal, truncnorm, and truncgamma in the figure), number of surgeons and True effect.

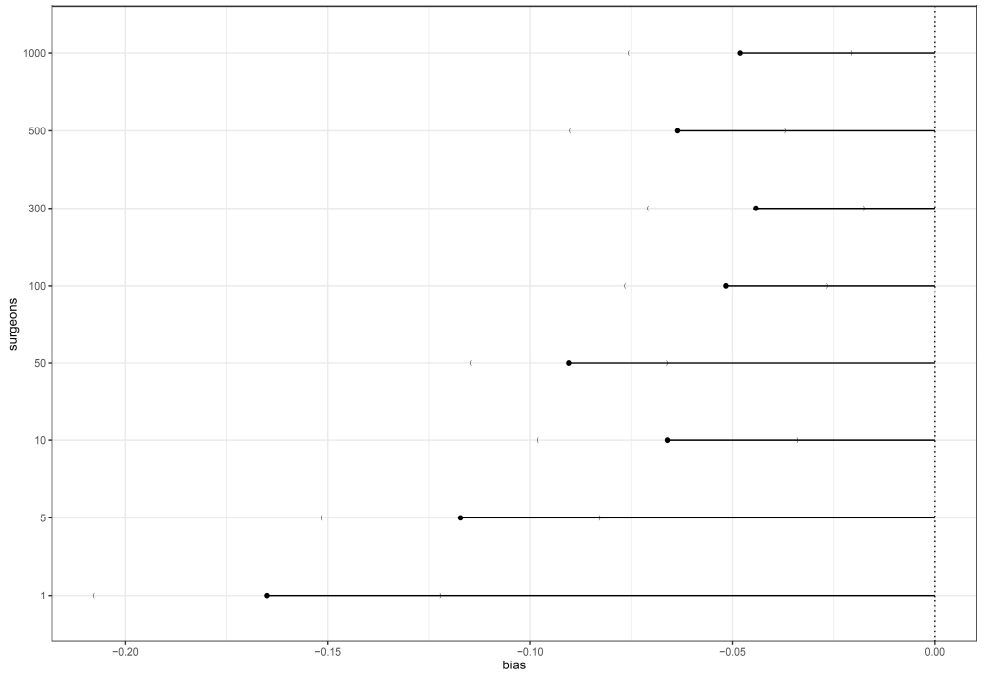
The 2sls IV method performed well with the normally distributed outcome, with minimal bias: an absolute bias of -0.05 95%CI(-0.08 to -0.02) for all effect sizes (0,1, and 3). Bias increased when truncating the possible outcome as the effect increases: Truncated normal showed bias of -0.05 (-0.08 to -0.03) with effect 0, -0.11 (-0.14 to -0.09) with effect 1, -0.28 (-0.31 to -0.26) with effect 3. This increase in bias with the effect was even greater when the underlying distribution is not normal, with truncated gamma bias is -0.04 (-0.07 to -0.01) with effect 0, -0.35 (-0.37 to -0.32) with effect 1, and -1.10 (-1.13 to -1.09) with effect 3. Coverage was good (>80%) in all cases with effect 0, and in the normal and truncated normal with effect 1. In cases with effect size 3, only the normal distribution had a good coverage of 95%, as coverage for truncated gamma is 0% and 70% for truncated normal.

Volume confounding vs bias

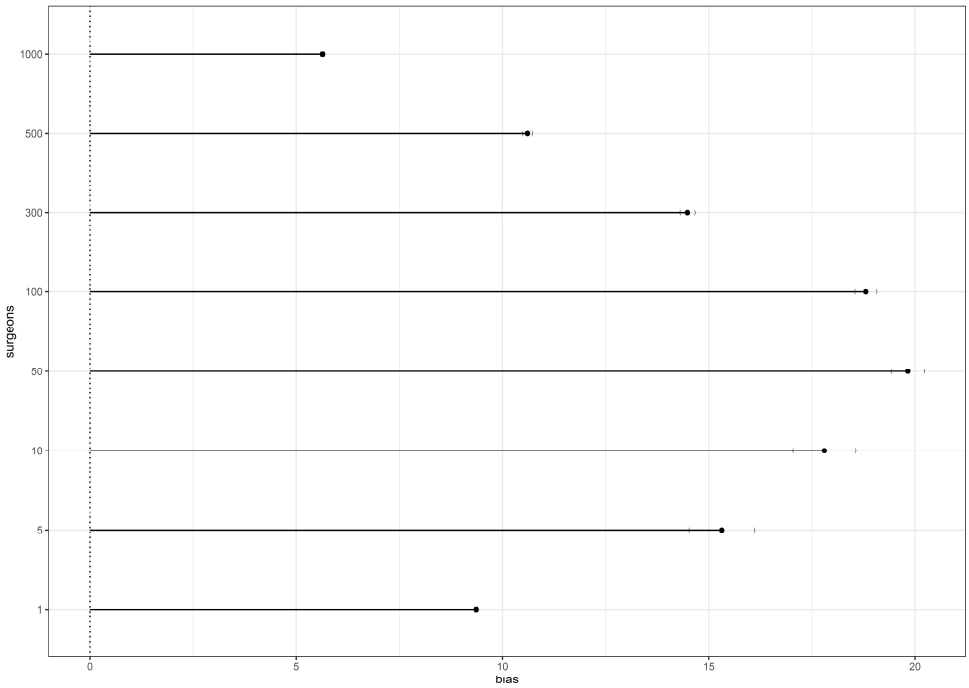
Figure 5.11 shows the effect of different volume confounding betas 0 , 0.1, 1 on the IV bias, for a True effect of 3, an IV strength of 5 and a prevalence exposure of 25%. The bias was similar in other scenarios. When the beta of volume confounding is 0, meaning no effect of volume on the outcome, the bias was minimal, of 0.05 95%CI (-0.08 to -0.02) with 1000 surgeons. Relative biases were between 1.6% and 5% and coverage exceeded 90% in all surgeon: patient ratios. This quickly changed when

minimal confounding, demonstrated by a beta of 0.1, was introduced. Coverage dropped to 0% in all cases, and absolute bias ranged from 5.6 (5.6 to 5.7) with 100 surgeons to 17.8 (17.0 to 18.6) with 10 surgeons, representing a relative bias of between 187% and 593%. A larger beta of 1 added one order of magnitude to the bias, with absolute bias between 56.8 (56.4 to 57.2) with 100 surgeons and 199 (195 to 203) with 50 surgeons.

beta=0



beta=0.1



beta=1

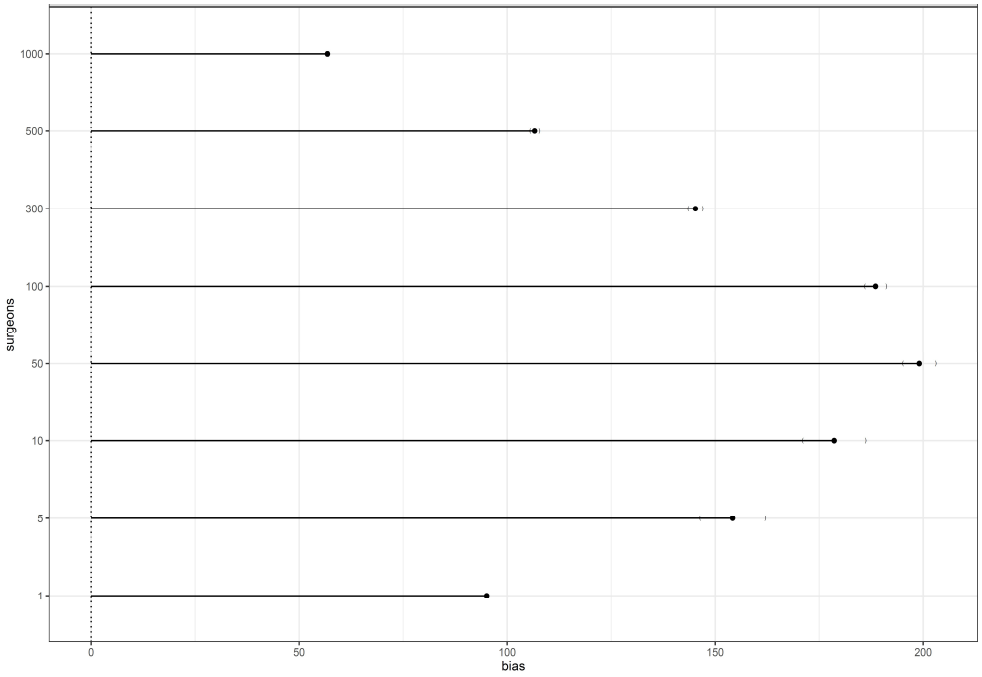
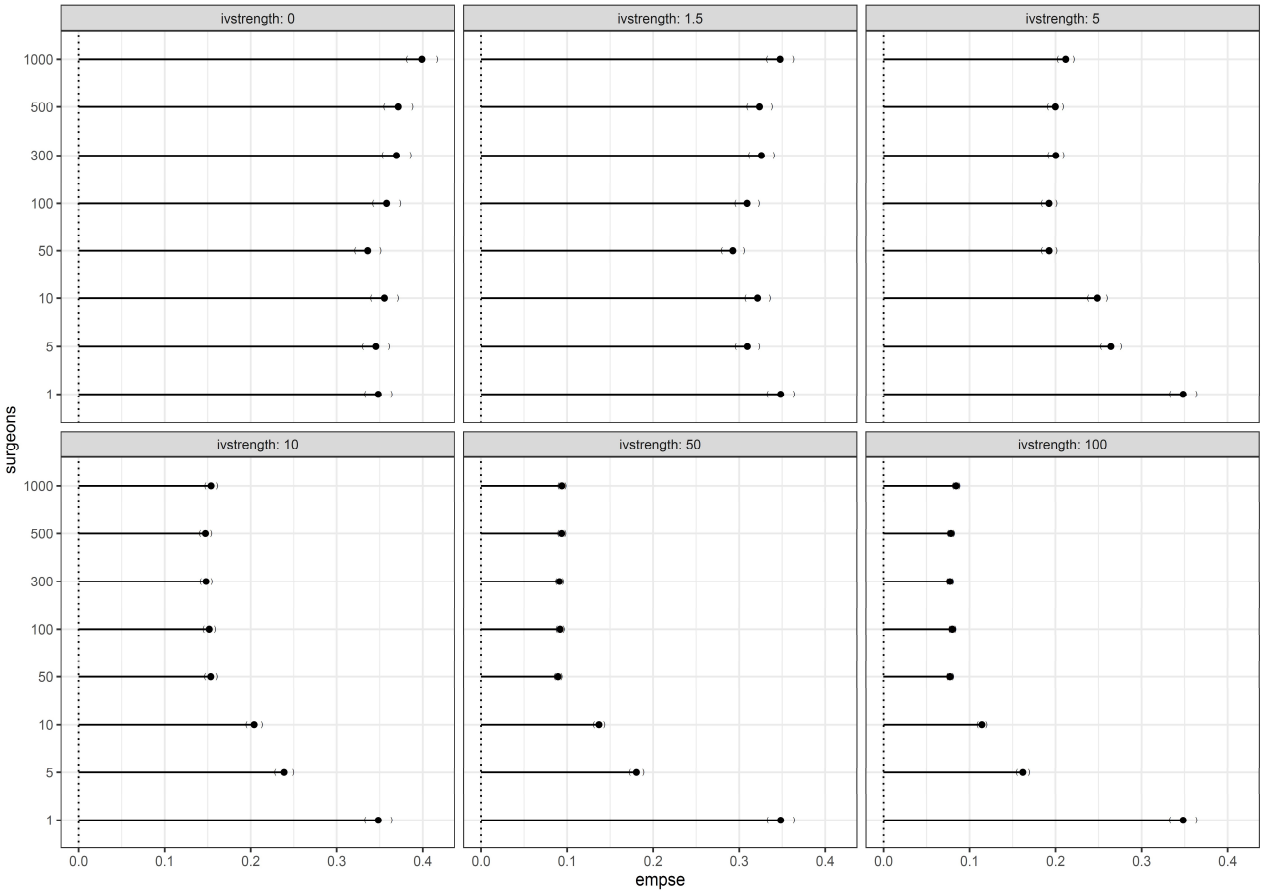


Figure 5.11 Absolute bias per different volume confounding betas, for a True effect of 3, normally distributed outcome, an IV strength of 5 and a prevalence exposure of 25%.

Number of surgeons effect on error and bias

Figure 5.12 shows how empirical standard error (empSE) is modified for different IV strength (0, 1.5, 5, 10, 50, 100) and different number of surgeons (1, 5, 10, 50, 100, 300, 500, 1000) with a True effect of 3, and a prevalence exposure of 25%. In 0 strength IVs the empSE was consistently high for all number of surgeons: from 0.35 (0.33 to 0.36) with 5 surgeons to and 0.40 (0.38 to 0.42) with 1000 surgeons. In the scenarios of large strength, 10 to 100, the empSE was still high for small number of surgeons, for example, 0.35 (0.33 to 0.36) for 100 surgeons, but it quickly reduced and plateaued at around 50 surgeons, with an empSE of 0.07 (0.07 to 0.08). Absolute bias followed a similar trend: bias decreased with an increasing number of surgeons across different scenario with the highest IV strengths (of 5 or more).

EmpSE


ivstrength: 0
ivstrength: 1.5
ivstrength: 5
ivstrength: 10
ivstrength: 50
ivstrength: 100
surgeons
1000
500
300
100
50
10
5
1
0.0
0.1
0.2
0.3
0.4
empse

Bias

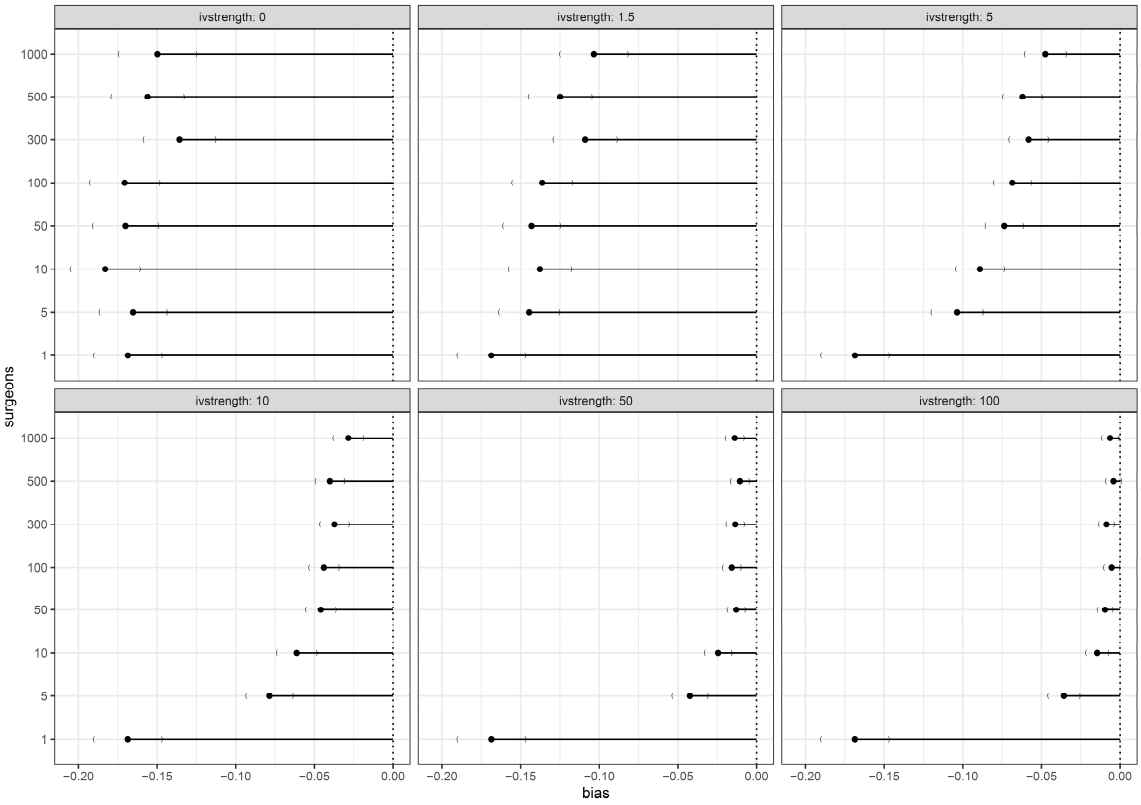


Figure 5.12 Empirical Standard Error and bias by IV strength and Surgeon number, for a True effect of 3, normally distributed outcome and a prevalence exposure of 25%.

Discussion

Preference IVs may not be fit for purpose when analysing some surgical interventions and MDs.

Here, I define the *real preference* of the surgeon as a theoretical variable that affects the treatment allocation, which fulfils all the criteria to be a valid and useful IV, and the *observed preference* as a variable calculated from the data as percentage of patients operated by the same surgeon in n previous surgeries that receive the exposure under study.

The strength of this observed preference variable is affected by different phenomena. First, the prevalence of the exposure limits its strength, most worryingly in its lower limit: when real preference is totally unrelated to the exposure, the observed preference has a mean strength of 4.5 for 5% prevalence. This could mean that I mistakenly took preference as a strong IV when it was actually a measure of prevalence, and likely to break the 3rd IV assumption.

The upper strength limit achieved by the observed preference variable was also capped on low prevalence and high prevalence exposures, no matter what the real preference strength was. In medium exposure prevalence settings (25%) the achieved strength of the observed variable is more similar to the real strength.

Another factor affecting the observed preference strength was the number of surgeons. An increased number of surgeons seemed to consistently increase the mean observed strength and reduces the SD of it. In scenarios with less than 50 surgeons, the high SD could mean that although we had a very strong real preference IV (OR=148), the observed preference would be deemed not strong per the OR>2 criteria.

In surgical epidemiology, the number of surgeries a surgeon performs per year, the volume, has sometimes an impact on the outcomes. (191) To explore the effect this would have on an IV analysis, I added an effect of the observed volume on the outcome. With a medium beta of 0.1, the bias for a true effect of 3 ranges from 6 to 18, estimates between 9 and 7 times what would be expected. This adds to the problem of using the observed preference, based on the treatment, as an IV.

This method is also sensitive to the shape of the outcome variable. Under perfect conditions, the method works well with normally distributed outcomes. However, when the outcome is truncated or not normally distributed, as many patients reported outcome scales are, this leads to increased bias.

Another important consideration is how the number of surgeons impacts the estimates. I found that when using a low number of surgeons, the error and the

bias is high for all IVs, even the stronger ones granting low coverages. When the number of surgeons is more than 50 the bias and error reduce in strong IV, and the coverage increases.

This simulation study also has limitations. First, all simulated relationships between variables are prespecified. The covariate and confounding structure in the real world are probably much more complex, and this could mean the results I arrive to are biased or won't be generalisable to a real study. This study, in addition to representing only a very specific case, uses only two stage least-squares regression to estimate the effect. It is possible that with other methods and parameters the issues seen in this simulation can be resolved.

To conclude, preference-based IVs seem to be dangerous tools for surgical epidemiology. Basing the IV on the very thing we want to predict seems to make it prone to several sources of bias, especially volume confounding, and the observed strength could vastly misrepresent the real strength the surgeon preference has. This is even more worrying considering that, as I showed in **Section 5.1**, the usual falsification techniques were unable to identify this confounding.

5.3 Propensity Scores in multilevel surgical settings

Introduction

In **Chapters 3, 4, and 5** I also used Propensity Score methods (PS) as an analytical strategy for calculating treatment effects minimising confounding. These methods were able to give results quite similar to those on the trial and had reasonable results in other outcomes. However, I also showed how covariates affecting the surgeon, such as volume, can strongly impact some outcomes. This led me to think that the patients receiving surgery and their outcomes may not be independent from each other, and they may be clustered on the surgeon performing the surgery.

This clustered structure may be the case for most of the MD epidemiology. Propensity scores are widely used in pharmacoepidemiology, where the effect of the prescriber could be negligible in most cases. But as we have seen, that is not the case for MDs and surgeries. New studies have proposed and extended PS-methods for clustered data (192-196), but only a few have tested different modelling strategies in real data and evaluated if they impact the results.

I aim to apply different modelling strategies to take into account multilevel structure for propensity score methods in the study of the PKR vs TKR

effectiveness and safety and to compare them to the results of the trial and of the previous **Chapters**.

Methods

Study design, data sources and population

The study and cohorts used have been extensively described in **Chapters 2 and 3**. I used both the effectiveness and safety cohorts of these chapters: patients who underwent a first primary PKR or TKR from 2009 until December 2016 with ASA 1 or 2. I repeated the analyses using the cohort of patients operated by surgeons with a volume of more than 10 surgeries of the same type on the previous year, as to mimic as closely the inclusion criteria of the trial.

Outcomes

I also explored the well powered outcomes from the previous chapters. For this section I used OKS as the main outcome, as has a clear comparison to the TOPKAT trial. This outcome is explained in **Section 2.1**. I also used 5y revision, which although does not have a clear gold standard, can be compared to my results in previous chapters in terms of effect estimate and error. This outcome is explained in **Section 3.1.**

Propensity Score calculation

I tested 4 different strategies of PS calculation. The first was the logistic regression with patient-level covariates, from now on *logistic PS*, as used in previous chapters and shown in ***Appendix Table 3.2***. As a second strategy I added PKR and TKR volume to the logistic regression, *logistic + volumes PS*. For the third and fourth strategies, I used the covariates of the previous strategies in a mixed-effect logistic regression with lead surgeon as cluster variable, *mixed effects PS* and *mixed effects + volume PS*.

Treatment effect estimation and outcome model

As in the previous chapters, I modelled the outcome using a lead surgeon clustered random effect regression, linear for OKS, and Poisson for revision and death. The main PS strategies were PS weighting with stabilised inverse probability weights and stratification (by the whole cohort and by the exposure) derived from each PS specification strategy. I double adjusted for the imbalanced variables after weighting or stratification.

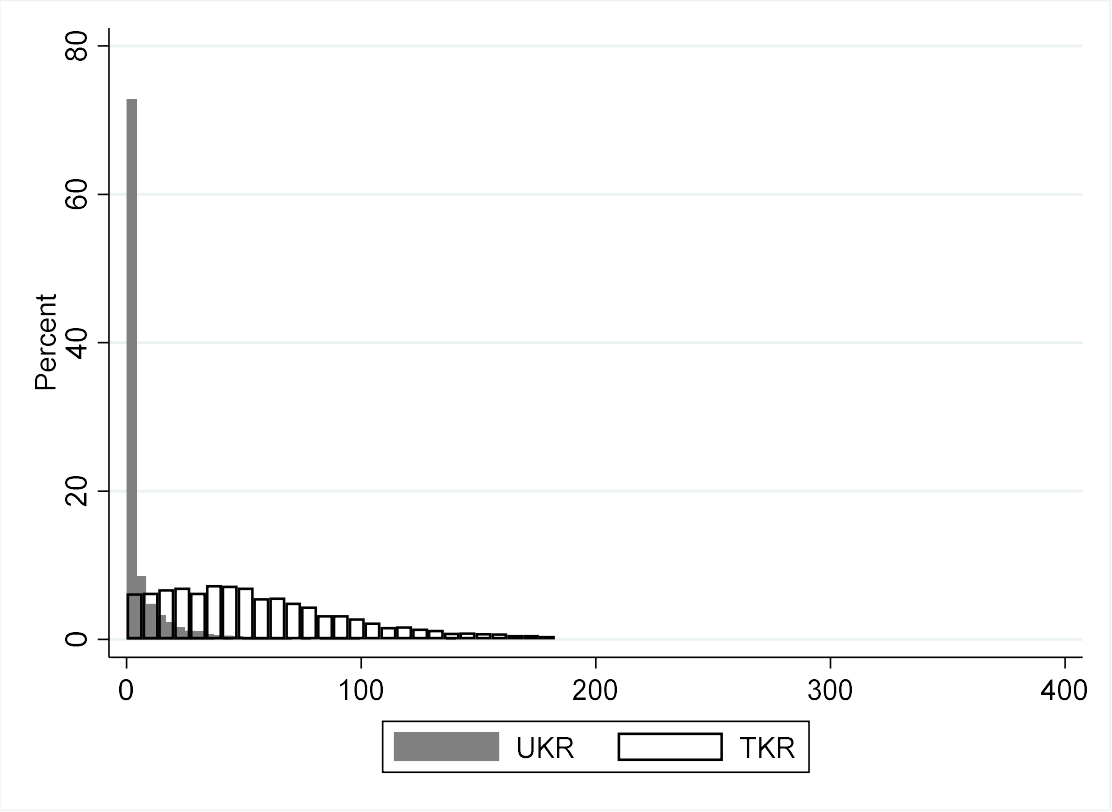
Comparison to TOPKAT

I only formally compared the estimates for the OKS effect to the trial. I used the same 5 agreement criteria as in **Section 2.1** to assess whether a method could be close to the results from the TOPKAT RCT. These criteria were: 1) coverage, 2) whether the treatment effect estimate is inside the 95% CI of the TOPKAT estimate, 3) the chi-square for heterogeneity test with a p-val>=0.05; 4) an I2 for heterogeneity below 40%; having a small between method variance (Tau2); 5) and statistical significance agreement, defining as equivalent results those with the same direction and significance as TOPKAT.

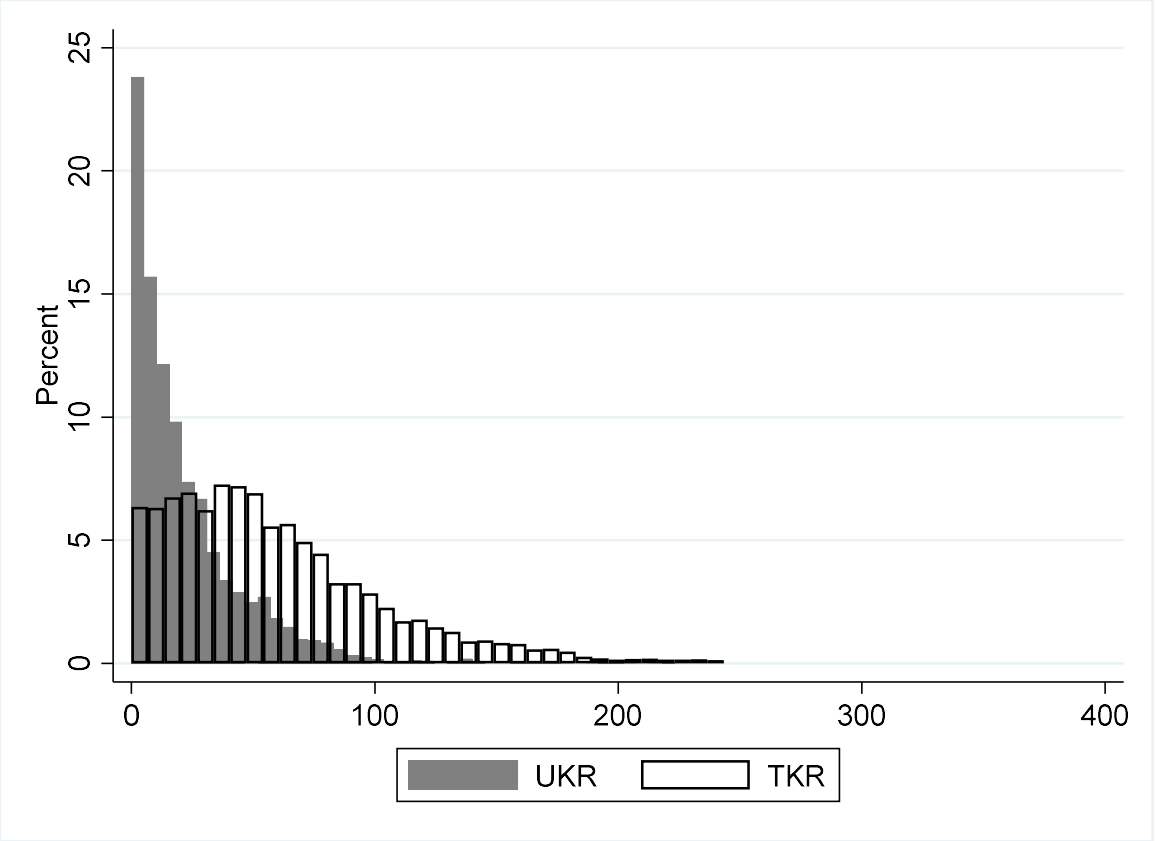
Results

Surgeon characteristics

The variables of surgeon volume of PKR and TKR have been described in detail in **Chapter 4**. *Figure 5.13* shows the distribution of these two variables, for all patients, and for patients who received the same surgery. It shows how a great number of patients are operated by surgeons that have done no PKR in the previous year (>60%), also for those receiving an PKR (>20%).



a) For all patients



b) Surgery Specific

Figure 5.13 – Proportion of patients who were operated by surgeons that had done a certain volume of surgeries in the previous year. For all patients and surgery-specific (for those patients with PKR, surgeon volume of PKR and surgeon volume of TKR for those with a TKR).

As shown in ***Table 5.5***, the ratio surgeon-patient is highly variable. Of all surgeons 50% had done 16 or less surgeries on the whole period, with ratios ranging from 1 patient per surgeon to 1,154 patients per surgeon. When we look at PKR, the

number of surgeons that had perform any is very low: of 6,420 surgeons in the whole cohort, 68.4% did not perform any PKR in the whole period. And in the effectiveness cohort that number increases to 88.4%.

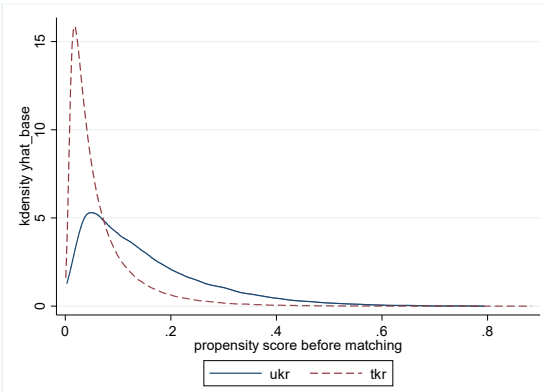
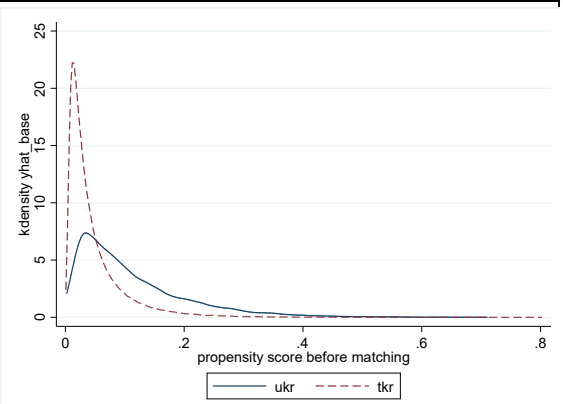
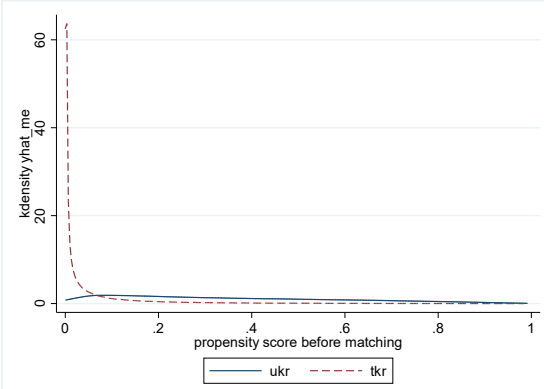
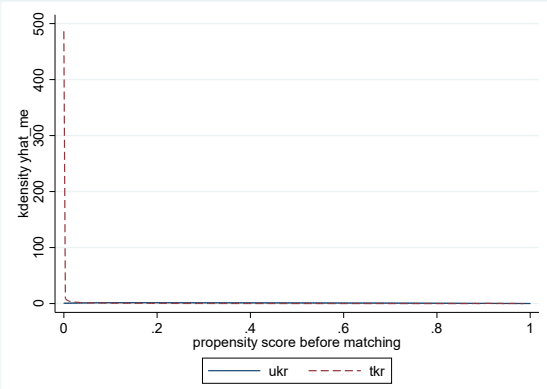
Cohort (n of surgeons)		mean	sd	p50	p10	p25	p75	p90	min	max
Revision (n= 4,620)	Total	63.8	106.2	16	2	5	78	194	1	1154
	PKR	4.6	16.5	0	0	0	1	10	0	271
	TKR	59.2	98.5	16	2	4	73	179	0	1137
Effectiveness (n= 6,420)	Total	32.6	52.6	11	1	3	40	95	1	612
	PKR	0.3	1.3	0	0	0	0	1	0	35
	TKR	32.3	52.1	10	1	3	40	94	0	612
Revision high volume (n= 3,009)	Total	87.1	120.5	37	2	8	124	239	1	1148
	PKR	4.4	18.6	0	0	0	0	7	0	265
	TKR	82.7	113.4	36	2	8	116	226	0	1137
Effectiveness high volume (n= 2,625)	Total	44.0	59.8	21	2	5	61	115	1	612
	PKR	0.2	1.4	0	0	0	0	0	0	35
	TKR	43.8	59.4	21	2	5	61	115	1	612

Table 5.5 Number of each type of surgery (Total or Partial Knee Replacement) performed by each surgeon on the whole study period.

Propensity Score calculation and Diagnostics

The betas from the propensity score calculation can be found in *Appendix Table 5.2*.

Figure 5.14 and Figure 5.15 shows the distribution of PS for each cohort and method.

Full cohort	High volume cohort
	
Logistic	
	
Mixed Effects	

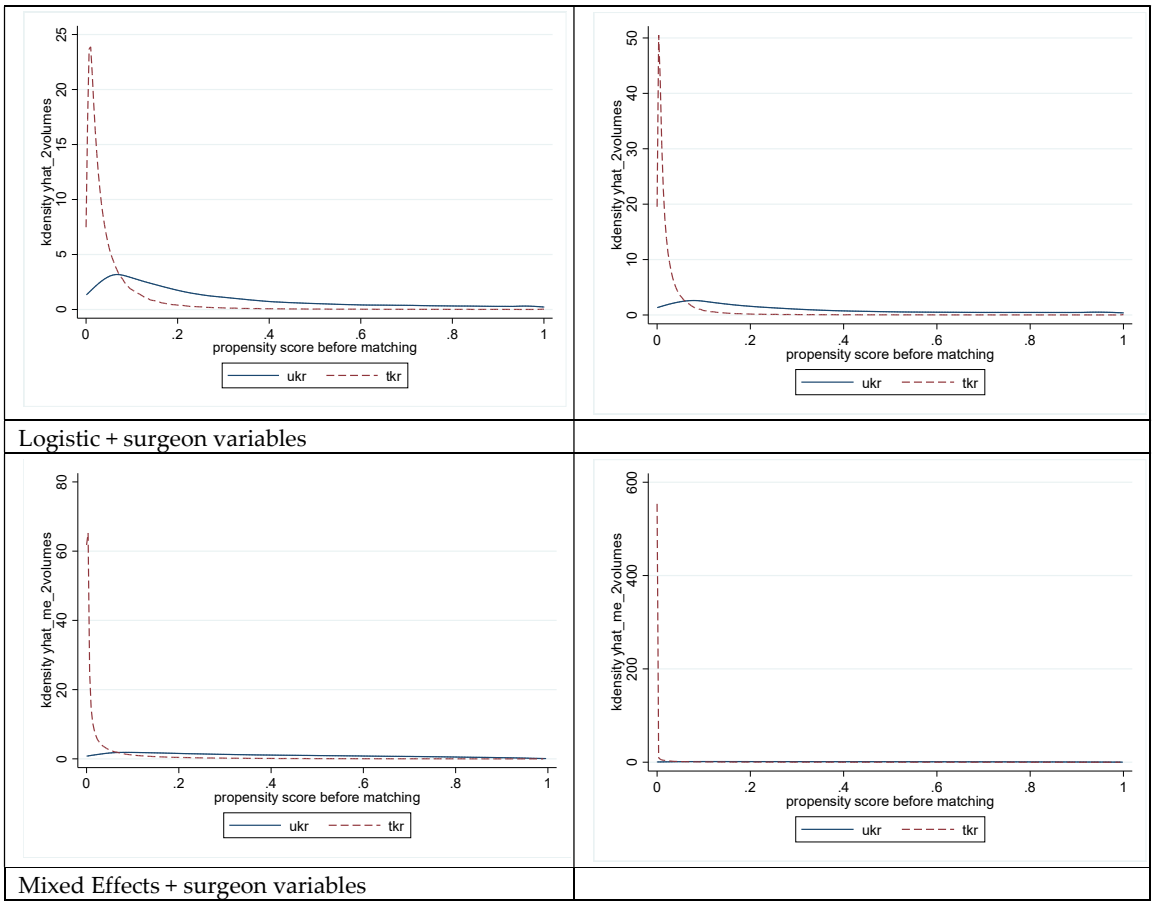
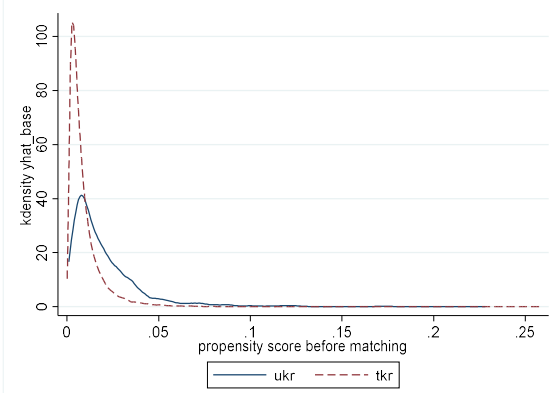
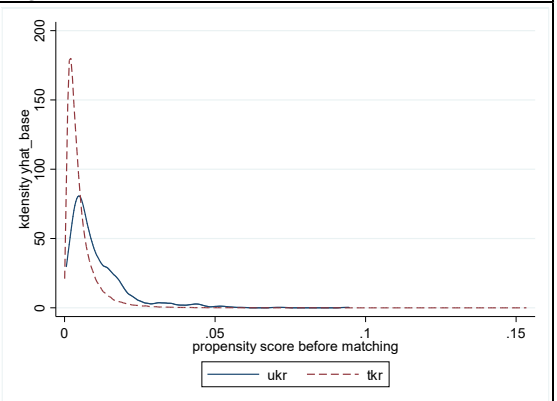
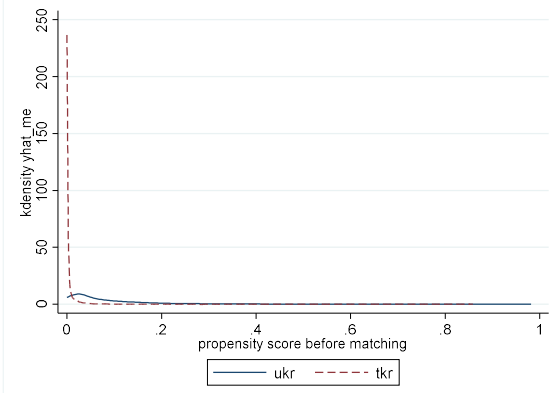
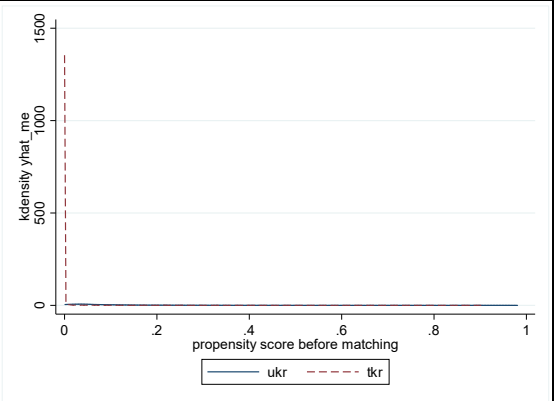


Figure 5.14 – Propensity score distribution (density) in the Revision cohort. 1st Imputation.

Full cohort	High volume cohort
	
Logistic	
	
Mixed Effects	

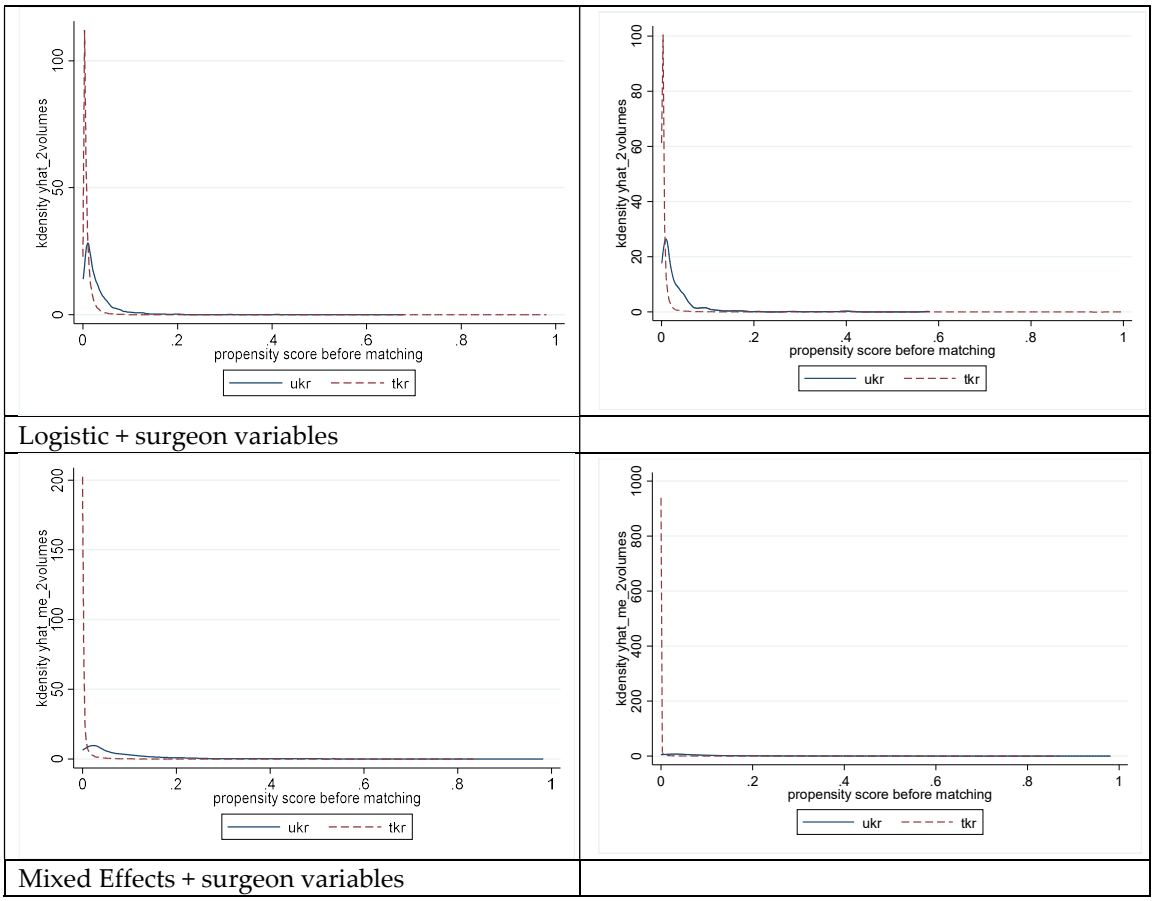


Figure 4.15 – Propensity score distribution in the OKS cohort. 1st Imputation.

The addition of the volume variables on the propensity score flattens the PKR propensity score curve and reduces the overlap between those who received TKR and PKR. Mixed effects reduce the overlap even more to being almost non-existent in both cohorts.

Using a random effect for the lead surgeon seem to better capture geographical variability and reduces the beta of deprivation on the prediction of treatment, from betas ranging 0.5 to 0.8 for each deprivation tenth in the fixed effects to betas

ranging 0.8 to 0.99 in the random effects model. When introducing lead surgeon volume in the previous year, both PKR and TKR volume are strong predictors of PKR treatment (with betas between 1.03 to 1.09 for PKR and of 0.99 for TKR). Similar results were produced in the high-volume surgeon cohort.

Including surgeon level covariates to the propensity score changed radically the weights assigned to the PKR and TKR patients in both effectiveness and safety cohorts. This is shown in ***Table 5.6***. Adding volumes in the logistic regression increased disproportionately the maximum weights for TKR (up to 119,159). Using mixed effects increased both maximum weights, for TKR and PKR, but not as much as using the PS generated by the logistic regression with surgeon volumes.

	High Volume Cohort				Full Cohort			
	TKR	n=114,871	PKR	n=602	TKR	n=125,834	PKR	n=1,197
OKS Cohort	Min	Max	Min	Max	Min	Max	Min	Max
Logistic	0.99	1.18	0.06	9.22	0.99	1.33	0.04	9.94
Mixed effects	1.00	10.14	0.02	21.43	0.99	7.02	0.02	81.77
Logistic + volume	0.99	268.15	0.02	14.41	0.99	50.55	0.02	16.22
Mixed effects + volume	1.00	7.62	0.02	39.47	0.99	5.97	0.02	78.30
Revision	TKR	n=248,785	PKR	n=13,334	TKR	n=273,530	PKR	n=21,026
Logistic	0.95	4.81	0.07	20.88	0.93	7.93	0.09	23.19
Mixed effects	0.95	65.36	0.14	381.27	0.93	57.89	0.12	118.65
Logistic + volume	0.03	119159	0.09	83.96	0.26	77991	0.09	68.40
Mixed effects + volume	0.94	136.88	0.15	314.60	0.93	94.32	0.12	127.35

Table 5.6- Minimum and maximum patient assigned weights after stabilised IPW

The ASMDs for each method and cohort are shown in ***Appendix Table 5.3.*** Adding volumes to the PS logistic regression resulted in imbalance in almost all variables when using IPW in the revision cohort, for example in PS (3.4), deprivation (0.7), or pre-operative OKS (0.8). The OKS cohort only had small imbalance in PS (0.1).

Modelling the PS with lead surgeon as a random effect resulted in some imbalance when using IPW, and stratification by the full cohort, in most cases on the PS itself, and on age (0.3) and on cardiovascular disease (0.12). Similar imbalances were

seen in the mixed effects + volumes for all cohorts in IPW and PS stratification by the whole cohort.

By contrast, stratification by the exposed perfectly balanced all the variables for all methods and cohorts.

Outcome Estimation

Effectiveness

Figure 5.16 show the results of each method for the whole cohort. It also shows the metanalysis metrics used in **Chapter 3** to validate methods against TOPKAT. In the main cohort, including volumes in the propensity score calculation seemed to have little effect for IPW and stratification by the whole cohort. The estimates changed from 0.67 95%CI(-0.10 to 1.44) in the logistic PS to 0.61 (-0.22 to 1.45) in the logistic + volume PS for IPW and from 0.18 (-0.43 to 0.80) in the logistic PS to 0.42 (-0.19 to 1.04) in the logistic + volume PS for PS stratification by the whole cohort. Using a mixed regression to calculate PS had little effect on the PS stratification by the full cohort. However, it increased the error on IPW and moved the point estimate out of the confidence interval from the trial, 1.91 (0.20 to 3.62), moving to 0.01 (-1.13 to

1.16). Stratification by the exposed continued to have the closest point estimate to the one from TOPKAT and it is totally unaffected by the PS calculation method.

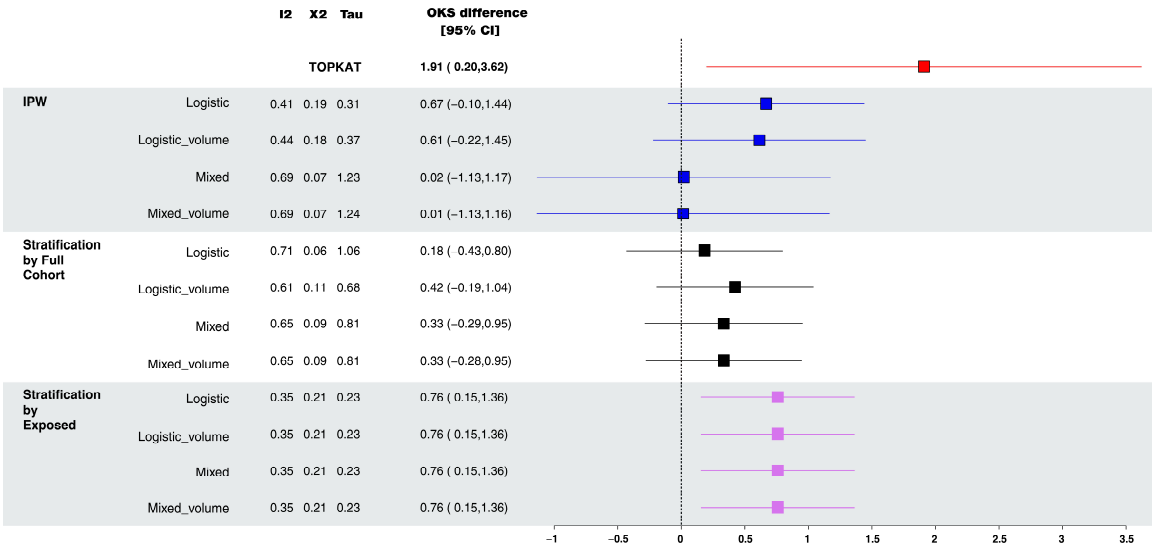


Figure 5.16 – Results of each PS strategy and methods for the whole cohort on effectiveness for Partial vs Total Knee replacement

Figure 5.17 shows the same results and metrics for patients whose surgery was performed by high volume surgeons. In this setting, adding surgeon volume variables to the propensity score calculation moved the estimate closer to the one from TOPKAT without increasing error, both in IPW and stratification by the full cohort. Using a multilevel regression to calculate the propensity score seemed to have a small effect on PS stratification by the whole cohort. This effect was much bigger for IPW, where estimates moved closer to the ones from the trial, even more

when adding the surgeon volume to the model, but doubling the amount of error. Stratification by the exposed stays consistent between PS modelling strategies.

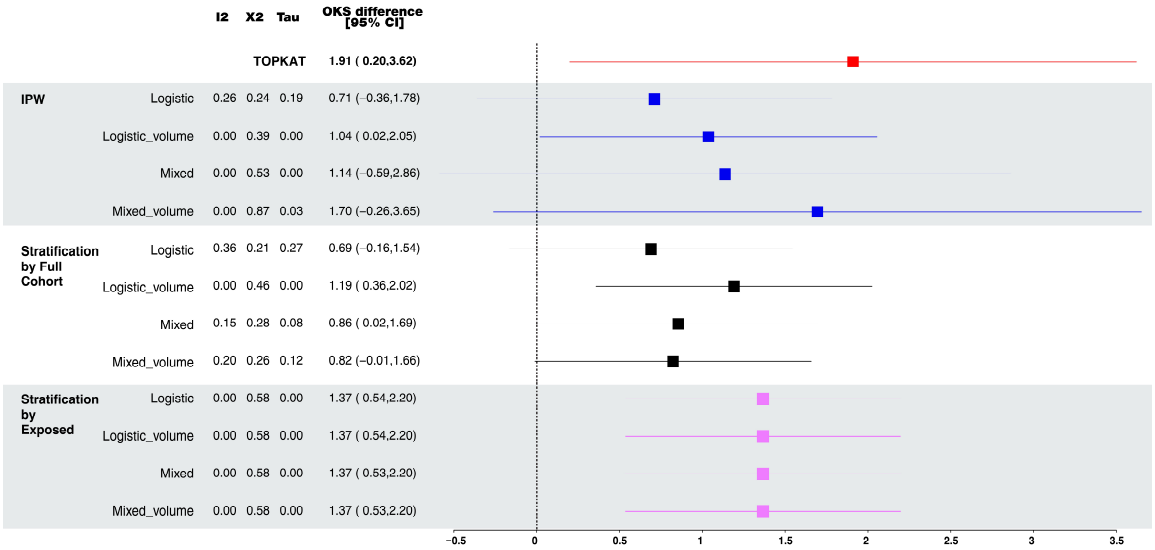


Figure 5.17 – Results of each PS strategy and methods for the high surgeon volume cohort on effectiveness for Partial vs Total Knee replacement

As for the pre-specified agreement criteria in IPW, all the different modelling techniques seem to increase heterogeneity (in I2 and Tau) and using surgeon as a random effect gets the point estimate outside of the trial confidence intervals, although confidence intervals overlap. In stratification has the opposite effect, decreasing heterogeneity and moving the point estimate inside the RCT confidence intervals. In the high-volume cohort, adding the surgeon volume or/and surgeon

as random intercept in the PS model improves all agreement measures with the trial both in IPW and stratification by the full cohort.

Revision

Figures 5.18 and 5.19 show the results for 5-y revision for the full cohort and the high-volume cohort respectively. On the full cohort, adding surgeon volumes to the propensity score calculation results in an increase of the estimator, both in IPW and stratification by the full cohort. Using a mixed effect model further increases the calculated risk in IPW, with little differences in stratification. Again, stratification by the exposed remains unaffected by the modelling strategy. In the high-volume surgeon cohort, the effect of introducing volumes or the surgeon as a random effect on the final estimate is similar to the one it has in the full cohort.

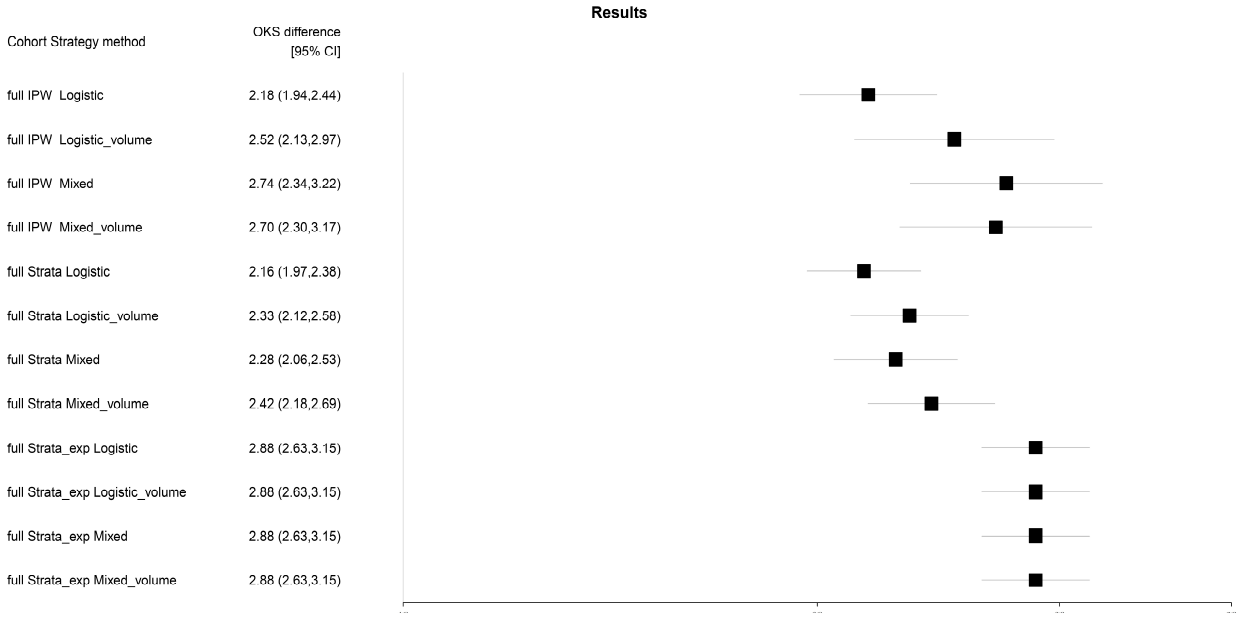


Figure 5.16 – Results of each PS strategy and methods for the whole cohort on revision for Partial vs Total Knee replacement

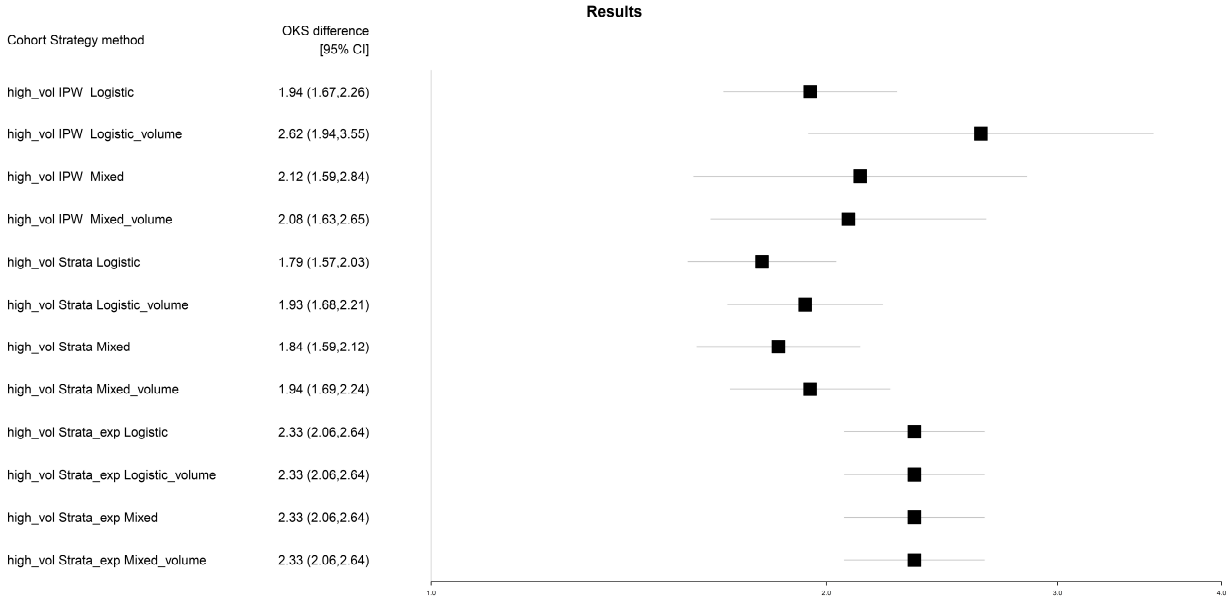


Figure 5.17 – Results of each PS strategy and methods for the high surgeon volume cohort on revision for Partial vs Total Knee replacement

Discussion

Previous methodological research on propensity scores for multilevel confounding focused on time varying confounding or matching. (193, 195) This research suggests that considering the multilevel structure has an important effect on bias. In this case, ignoring the multilevel structure, thus using a fixed effects logistic regression with only patient level confounders, seems to be a good enough strategy, but not the best in all cases. The propensity score resulting has very little predictive power. This results in the PS distributions of TKR and PKR overlapping a lot in lower propensity score values. This strategy, ignoring the cluster level, seems to create some imbalance and has high potential for unmeasured confounding.

Using surgeon variables in the propensity score calculation changed considerably the distribution of the predicted propensity score, reducing the overlap of PS between those who received a TKR and those with PKR. It seems to flatten the PKR PS distribution, probably getting a much better prediction of patients who would receive a PKR. It also increased the imbalance on patient balance for IPW and PS stratification by the whole cohort.

Considering the multilevel structure, using mixed effects both with and without surgeon volume, turns the PS more predictive, and decreases the overlap to perform PS techniques. As with adding surgeon volumes, it increases imbalanced variables in both IPW and stratification by the full cohort, possibly suggesting that the potential for residual confounding if the structure is not considered is high.

When applying this PS calculation strategies to IPW and PS stratifications results differ slightly. IPW seems to perform poorly when the surgeon structure or confounders are taking into account. Although I used stable weights (77) to calculate the weighting, some patients get assigned astronomical weights. That could mean that those patients can be driving the whole effect with high potential for spurious results. When using mixed effects (with or without volume confounders) to model PS, IPW estimation yields a lot of uncertainty doubling the standard error than the one achieved with logistic regression. This could be due to the high sensitiveness of IPW to large surgeon-patient ratio differences with multilevel modelling, where most surgeon clusters are modelled using only a handful of patients, and most without any PKR. In contrast stratification by the full cohort SE seems unaffected by the PS calculation strategy. Finally, stratification by the exposure seems totally unaffected by the PS calculation strategy, possibly being

a good consistent method for this kind of research, where there is a low prevalence of patients receiving one of the treatments.

As shown in the results, adding the surgeon volume in the propensity score did not improve the estimates from the full cohort. However, on the high-volume surgery cohort, considering the surgeon in the PS calculation improved the estimates. This points to a possible effect modification that not even taking into account the surgeon structure and volume could amend. Again, this warns of the importance of designing studies and choosing populations carefully, for example, by thinking in a trial emulation framework.

This section is clearly limited by the specificity of its analysis and cannot be easily generalisable as I only tested the methods in one study. I only used, for the sake of computing time, one of the imputed datasets. This led to slightly different results, although very similar in the final estimate of treatment effect, and different imbalanced variables in the patient level propensity score methods. However, the existence of an RCT gave me an unvaluable opportunity to test methods with a known true effect, a luxury that is usually reserved for simulation studies. The

findings will need to be confirmed in more comprehensive methodological research.

Chapter 6

Discussion and Conclusions

In this chapter I summarise the main results of the dissertation and contextualise their implications. I explain the clinical implications that could impact standard practice for knee replacement and tracheotomy, and I explore the implications for MD and surgical epidemiology methods. I reflect on the strengths and limitations of the research done and propose future avenues of research in MD and surgical epidemiological methods.

6.1 Summary of Main Results

The aim of this thesis was to find the best methods for conducting observational research of surgical implants. This search for the optimal methodology to tackle some of the specific caveats of MD and surgical epidemiology has been performed using routinely collected observational health data and simulated datasets. I explored the performance of multiple designs and methods including instrumental variables (IVs), propensity score (PS) methods and inverse probability weighting (IPW), self-controlled case series (SCCS), and the trial emulation framework. I tested these methods and answered clinical questions such as the effectiveness and safety of Partial Knee Replacement (PKR) and Total Knee Replacement (TKR) in patient with and without severe systemic disease (and higher surgical risk), the best timing of tracheotomy in COVID-19 patients, and the effect of surgical volume on PKR and TKR effectiveness and safety.

Chapter 2 focused on the study of effectiveness of MD. In **Section 2.1** I emulate TOPKAT,(70, 71) an RCT looking at the comparative effectiveness of PKR and TKR, using registry data linked to hospital records and patient reported outcomes. I tested different methods, IV and PS, to see if they could yield similar results to the TOPKAT trial. I found that only 5 of the 17 potential instrumental variables

tested passed diagnostics, and none of these 5 managed to replicate the results of the trial. PS methods got closer estimates to the trial, with PS stratification (both by the whole cohort and by the exposed) and IPW achieving the best results (according to pre-specified agreement criteria). The results of all methods were even better and obtained results closer to those of the trial once I restricted the analysis to those operated by high volume surgeons, matching the inclusion criteria of the RCT. I found that PKR is slightly superior to TKR in improving Oxford Knee Score (a measure of pain and function) but that difference was not clinically relevant. This shows how difficult is to study MD in observational epidemiology, as their effect cannot be isolated from the effect of the procedures and the surgeons performing them.

In **Section 2.2**, I explored the best timing for tracheotomy in COVID-19 patients in need of invasive mechanical ventilation. To do so I used the target trial emulation framework (43) to design the study using a COVID-19 tracheotomy registry of several hospitals in Spain. This helped me understand the possible caveats and limitations of performing this study, such as immortal time bias,(42) and how to minimise them by carefully thinking about surgery and surgery allocation timings. I found that performing a tracheotomy early (on day 7-10 after initiation of

mechanical ventilation) compared to late (after day 10) did not pose increased risks of death or increased the chance of weaning (successfully stopping mechanical ventilation).

Chapter 3 studied the safety of knee replacements. **Section 3.1** applied the same methods used in **Section 2.1** to compare the 5-year revision and death and the 90-day complications of PKR vs TKR. The cohort analysed in this section was also selected using the TOPKAT trial's inclusion criteria; however, the sample was larger as it did not require linked PROMS information. Unlike contradictory findings between IV estimation and al PS based analyses found in Chapter 2, results for this study were similar. Both showed a double risk of 5-year revision with PKR compared to TKR, but a reduction in mortality and complications, especially for venous thromboembolism.

Section 3.2 focused on describing the short-term risks of knee replacement using a SCCS design. In this design, widely used in vaccine safety studies, the chance of having the outcome after the exposure is compared to the period before the exposure.(164, 165) This reduces the chance of time-fixed confounding, as patients are compared to themselves. These analyses showed that a knee replacement is associated with a 22% increase in the risk of having an MI and 800% increase in the

risk of having a VTE. Both increases in risk were lower for PKR, with a 400% increase in risk in VTE, and no increase in the risk of MI.

Chapter 4 investigated different subgroups and variables that could potentially modify treatment effects. First, in **section 4.1** I applied the methods that were able to replicate the RCT in **section 2.1** to the cohort of patients with high surgical risk. These patients were excluded from the TOPKAT trial, but they could benefit from the lower risk of complications of PKR compared to TKR. These analyses showed that the effectiveness of PKR is similar to TKR, in line with the findings from **Chapter 2** and from the trial. (70, 71) In terms of safety, PKR related to much lower rates of 90-day VTE, almost a 3-fold reduction, with no differences in terms of MI and PJI. **Section 4.2** explored differences on the effect of PKR vs TKR according to different subgroups. PKR performed better, in terms of effectiveness, when patients were older or lived in less deprived areas. In terms of revision, women had a higher risk with PKR, but a lower risk of 90-day VTE.

Section 4.3 explored the modifying effect of surgeon volume, defined as the number of surgeries of the same type performed by the same surgeon in the previous year, on effectiveness and safety of PKR vs TKR. I found that half of the PKR surgeries were performed by surgeons who had done less than 10 PKR

surgeries in the previous year. The volume did not seem to have an impact on effectiveness or post-surgical complications but had a great impact on 5-year revision: patients who were operated by surgeons that have performed more than 50 PKR in the previous year had similar revision rates to those who received a TKR, and not 3 times more as with patients operated by low volume surgeons.

In **Chapter 5** I explored some methodological questions on the performance of different statistical methods, IV and PS, that arose from the previous chapters. **Section 5.1** explored some hypotheses for why the results from the preference based IV in **Chapter 2** were extremely different to those in the RCT, using a Plasmode simulation.(189) I found that the IV that I used had poor coverage (17% in a very strong IV) compared to a real synthetic IV (86% on an IV of the same strength) despite passing our falsification tests and related diagnostics. This increased bias seemed to be mediated, at least partly, by known confounders: pre-operative OKS, deprivation, and EQ5D for the continuous outcome and age, surgeon volume, and EQ5D for the binary outcome. However, this was not detected when using the SMD to check for known confounder imbalances.

Section 5.2 explored the impact of a multilevel structure, such as the one that arose when patients were clustered in surgeons, on preference based IV analysis using a

parametric simulation. The results highlighted that using preference as an IV calculated from the available data can mistakenly show a strong preference-treatment relationship when no such preference really exists. Furthermore, if the outcome was influenced by surgeon volume, as was the case with revision in PKR, the use of a preference-based IV strongly biased the results.

Section 5.3 looked at the use of different techniques to calculate PS in clustered data, with or without random effects, and with and without including surgeon volume. I compared these results to those from the previous chapters and the TOPKAT trial. Adding the surgeon volume or using a random effects model in the PS calculation worsened the estimates compared to logistic regression, in terms of agreement with the trial. However, the same techniques got estimates closer to the trial, when compared to PS based only on patient covariates estimated using logistic regression, once the cohort was restricted to high volume surgeons. These effects were only seen when using IPW but made little or no difference for PS Stratification.

6.2 Implications of the research

Methodological Implications

Chapter 2 highlights the need of careful thinking when designing an observational study using routinely collected data on MDs and surgery. One of the best tools to design a comparative effectiveness or safety study is the Target Trial Framework as explained in **Chapter 1**. (43) One key step to take, especially for surgical epidemiology, is to think about how to define the different timings in the study. A critical timing is the "decision date", or the moment when a patient would be randomised in a target trial. Getting this right is crucial to minimise and/or to understand and acknowledge certain limitations such as immortal time bias.(42) This bias arises when there is a period where death or an event cannot occur, such as the time between treatment assignment to actual treatment, if we only study patients who have been treated. This also happens if an event that may contraindicate the assigned treatment occurs at this time. These biases can greatly change the results of the study and should be discussed as limitations and minimised when possible. Another key step on defining a target trial is the "inclusion criteria". Deciding which patients to include, making sure treatment contraindications and other factors are considered, is easier when thinking of a

target trial. As we have seen in **Chapters 2, 3, and 4** not only patient factors should be considered in this decision, but other factors such as surgeon or hospital features can be critical to getting unbiased estimates. **Section 4.3** shows the importance of surgeon volume on revision outcomes for PKR, and how a failure to take that into account could lead to different conclusions and implications.

As seen throughout the thesis, another major caveat for MD epidemiology is the complexity of disentangling the effect of the MD by itself from the effect on the surgical procedure, the surgeon, the setting. Looking at devices of different kind would always be affected by the differences in procedures. Careful design and the capacity to differentiate MDs implanted (like the registry of device IDs) with the same procedure would help make possible a more purely MD comparison study.

The comparison to the TOPKAT trial in **Section 2.1** and the application of the validated methods in **Section 4.1** show how some PS based methods can be useful for MD epidemiology. PS stratification (122, 123) and Inverse Probability Weighting (121, 125, 127) worked better in this setting to replicate the trial. PS stratification by the exposed (126) achieved better balance and similar results when having a low number of treated patients, as shown in **Section 4.1.** This could be of use as some MDs, like partial knee replacement, have a low prevalence of use

compared to more widely established alternatives. The tested PS strategies in **Section 5.3** show how considering the multilevel structure of the data and surgeon variables can change the results when calculating PS and using them to estimate effectiveness and safety. Further research is needed to understand which are the best strategies for modelling PS in multilevel studies.

As for using IVs (102, 103) to compare effectiveness and safety of MD, it seems that finding real instruments is a tough task. Most of the instruments proposed in **Section 2.1**, taken from the pharmacoepidemiology literature,(100, 101) failed to fulfil the prespecified criteria and testable assumptions. (103-105) Worryingly, even IVs that passed such diagnostic criteria failed to produce results similar to the trial. **Section 5.1** shows how the criteria used to clear an instrument for use, balancing known confounders with an SMD<0.1, failed to detect the effect of known confounders. This calls for treating this widely used criterion with caution, and for trying to find an alternative to this test.

Another call for caution is on the use of preference-based instruments. In **Section 5.2** I have shown how using a preference-based instrument, estimated from the observed preference seems to cause many problems in IV estimation. If we consider the real IV a true preference that affects the exposure, the observed

preference seems to misrepresent the true strength when prevalence is low. This kind of IV seems also very sensitive to volume confounding, generating biases doubling the true effect when volume impacts outcomes. It is likely that instrumental variables based on preference have very limited use in MD epidemiology as they are subject to many potential biases.

A method that has proven useful for the study of MD has been the self-controlled case series (SCCS). (164, 165) With a thoughtful definition of the times at risk, a long enough wash-out period, and the performance of sensitivity analyses, SCCS can be a useful tool for the study of MD safety. This is exemplified in **section 3.2**.

Clinical Implications

This work also has clinical implications. I have generated evidence of the comparative effectiveness and safety of PKR vs TKR for patients with multi-morbidity (ASA 3 or above). Patients with ASA>2 would not have been eligible for the TOPKAT RCT, (70, 71) and it is unlikely that there will be a follow-up trial to include them. This population, however, encompasses 15% to 20% of patients who undergo knee replacement surgery in the NHS.

The effectiveness of PKR compared with TKR was similar to that seen in TOPKAT and patients ASA<3, with an average treatment effect of 1.83 95%CI (0.10, 3.56)

OKS points in favour of PKR for stratification and 1.00 95%CI (-1.28 to 3.27) for IPW. PKR therefore provided benefits equivalent to those from TKR amongst patients with severe systemic disease and/or substantial functional limitations, just as in the overall population.

In terms of safety, all analyses suggested a strongly protective effect against post-operative venous thromboembolism in these patients, with a 60% to 67% relative reduction in risk compared with TKR. Acute myocardial infarction and prosthetic joint infection showed no differences in occurrence on the PS analyses. However, in the first 90 days after surgery, almost 5% of TKR patients and just over 3.5% of PKR patients experienced an MI, and 1.9% of TKR patients and 1.8% of PKR patients a prosthetic joint infection. These results are consistent with previous literature, that finds that PKR has a 50-60% reduction compared to TKR in post-operative VTE. (129, 156) This is particularly important for patients with multi-morbidity, where VTE is one of the most common complications and could lead to worse outcomes and longer hospital stays.

Using a SCCS analysis, which inherently controls for confounding, I found that having a knee replacement is associated with a 22% increased risk of having an MI and an 8-fold increase in risk of a VTE in the first 90 days post-surgery. This is

consistent with previous observational research. (168-171) Furthermore, when looking at the risk of VTE after TKR or PKR the increase in risk for PKR was lower than for TKR, consistent with PS analyses. (172, 174)

As for long-term consequences, **Chapter 5** showed that patients with high surgical risk are more likely to die than to have revision surgery in the 5 years following surgery. The mortality rate was 24% for PKR and 37% for TKR, compared with 13% and 5% respectively for revision. PS analyses showed a threefold higher revision risk for PKR vs TKR. However, they also showed a >30% reduction in all-cause mortality with PKR vs TKR. These results should be taken with caution, as they could be driven by unresolved residual confounding and/or information bias.

There seem to be differences in comparative effectiveness and safety by sex and deprivation. The increase on revision rates seems to be substantially higher for women with PKR, but they also have a decreased VTE risks. Deprivation also conditioned outcomes, where patients from less deprived areas had more effectiveness benefit from PKR and a lower increase in the risk of revision.

A possible explanation for the differential risk of revision is differences in the surgeons performing the operation. I have shown the effect that surgeon volume has on the revision risks of PKR. Using the three validated methods, I show how

the excess risk observed among patients undergoing PKR decreases dramatically when PKR is performed by surgeons with higher PKR volume. PKR patients have as average more than double the risk of 5-year revision surgery of TKR patients. This increased risk drops to around 40% to 50% when the analysis is restricted to patients operated on by surgeons who had performed 50 or more surgeries of the same type in the previous year. However, the number of patients who had a surgery performed by a high-volume surgery is low, especially for PKR. While 50% of patients had their TKR performed by a surgeon who had done more than 50 TKR in the previous year, only 12% of the PKR were performed by surgeons with the same volume of PKR, approximately one a week. A possible way to improve outcomes on these patients is to centralise PKR in specialised treatment centres where these surgeries can be provided by specialist surgeons.

As for the COVID emergency study, I showed how an early tracheotomy has similar or higher rates of weaning than a late tracheotomy, without detecting differences on complications or death. There was no clear consensus of when to perform this procedure,(142) but the results from this section speak in favour of performing an early tracheotomy, especially when in a critical situation, as it could potentially liberate much needed ICU beds.

6.3 Strengths and Limitations

This Thesis focuses on observational research using routinely collected data and uses knee replacement and COVID related tracheotomy as case uses.

With regards to knee replacement, this implant is one of the most used in the world and has RCTs that have been used to confirm the results of this thesis.(197) Additionally, the granularity of the data, with reliable registry data linked to hospital and patient reported outcomes, allowed us to study both effectiveness and safety. However, there are some pitfalls. First, the differential response in the PROMs database may mean some results could be due to selection bias. (97, 114) I tried to minimise this by using the trial results as "gold standard" but the potential for bias remains. Furthermore, I haven't delved deep into the types of MDs, or how to use device IDs, as I have studied PKR and TKR as groups and hence looked at "class effects", mixing also device and procedure effects. This could have led to biases due to differences in uptake of model of implants of the same group regionally or changes in time. (198, 199)

In Section 3.1, I replicate a surgical trial using the same outcomes and methods as the trial, trying to follow the trial protocol. Having information of patient reported outcomes is a rare occurrence in observational research using routinely collected

data. In addition, all the analyses were designed and performed before knowing the results of the trial (i.e., blindly), making it impossible to change methods to confirm hypothesis.

However, there are also has limitations to the studies on the knee replacement cohorts. I was not able to implement one of the most important inclusion criteria in the trial: the presence of medial compartment osteoarthritis with exposed bone on both femur and tibia.(200) This was due to insufficient information on the NJR and HES databases for the indication of surgery. This could mean that the population studied in this thesis might be different from the trial, impacting which methods were deemed valid. Similarly, as for the lack of data to assess indication, some residual confounding may be left in due to unmeasured confounders, as HES records only hospital occurrences and can lack some information about chronic and milder disease. Nevertheless, PS stratification and IPW were able to replicate TOPKAT with the available information. Another limitation of this trial replication is on the outcome timing. TOPKAT had a post-operative OKS collection at 1-year post randomisation and my Thesis at 6-12 months post-surgery, which could also contribute to some of the differences.

There are further limitations on the analyses of the cohorts with severe systemic disease. The first is the much fewer number of participants (especially for the effectiveness analyses where there were only 145 patients receiving an PKR). However, the revision and safety analyses included a much higher number of patients. The second, the higher potential for confounding, as the population receiving a PKR may be very different to TKR and, in this case, there is no 'gold standard' results for comparison.

In **Section 3.2**, where I examine the best timing for tracheotomy on patients in ICU with COVID-19. This study applies the trial emulation framework to a situation where there is no trial protocol to follow or results to be compared to.(59) It focuses on a different clinical question than the rest of the clinical applications of this thesis, as, after the emergence of COVID it was important to help resolve these questions. Unfortunately, I was not able to contribute to setting the inclusion criteria. The fact that the registry included only patients who had received a tracheotomy, instead of all patients potentially eligible at day 7, prevented me from analysing the causal effects of tracheotomy timing on total days of IMV. This could mean that the differences in duration of IMV can be artificially inflated by immortal time bias.(42) In this study, there is a strong chance of confounding by

indication, where the tracheotomy could have been postponed in critically ill patients on the decision day. The results could also be limited by the presence of residual confounding. Although it was not part of the initial protocol, I applied IPW as one of the validated methods I learned in previous sections, and it did not change the results. As strengths, following the trial emulation framework, with "randomisation" at day 7 following clinical advice, makes for robust results. The large number of participants and the recruitment in multiple centres with different treatment protocols and no standardized weaning criteria, sedation agents, or type of tracheotomy increases external validity. Another strength is the granularity of information available on day 7 that give a wider picture of the patient's prognosis.

As for the studies looking at the performance of IVs for MD epidemiology, there are several strengths and limitations. Strengths include the use of the real variables and conservation of the relations among them on the plasmode simulation as well as the use of similar variables to the real ones on the parametric simulation. This helps focus the research question on scenarios occurring naturally and distances from extreme scenarios. However, there are limitations. The low number of variables used on the multilevel simulation, and the lack of testing of non-linear relationships may limit the simulation validity and more testing considering these

may be necessary. The choice of a low number of repetitions in the parametric simulation, due to being computationally intensive, gives an increased uncertainty for some scenarios. However, the amount of bias seen is likely to be true and would still be unacceptable with lower uncertainty. In the parametric simulation the test of the volume confounding and of the different strengths of IVs have a large range, and could be improved by adding more scenarios, thereby being more granular on the amount of bias for smaller changes in magnitude.

6.4 Future Research

This thesis has shown how observational data can provide insightful information to help plan and deliver implants in healthcare systems. Although it seems that PKR may be the preferable option for patients with high surgical risk, this should be properly confirmed. Furthermore, the hypothesis that highly specialised, high volume centres delivering PKR may have better revision outcomes could be tested in a pragmatic RCT and would provide invaluable information on how to organise surgical care. The mechanisms behind these improved outcomes in high volume surgeons are also uncertain. More work scrutinising whether this effect is due to more appropriate patient selection, better care or improved surgical expertise would also provide hints on how to minimise revision risk on PKR.

As for the tracheotomy timing, the study performed was a rapid evaluation due to the emergency of the pandemic. After these preliminary findings, an RCT trial with an intention-to-treat analysis would be preferable to establish the effect of early tracheotomy on the total duration of mechanical ventilation for patients admitted to ICU. This would help tackle the potential of immortal time bias and help explore the interactions with pronation and other clinical factors. Performing

this in a lower pressure pandemic stage could help inform the best course of action for optimal resource allocation during higher pressure stages.

I have found that PS stratification and IPW were useful for minimising confounding when evaluating the comparative effectiveness and risks of partial vs total knee replacement and could potentially be a good tool for other MDs and surgical procedures. Further research on why PSM and PS adjustment failed is needed, and whether the difference between ATE and ATT could be playing a role. I used the most used technique for propensity score estimation, logistic regression with clinically selected variables. Future research could explore alternative methods for variable selection such as machine learning, or large-scale propensity score. Also, alternatives methods of PS matching, with different algorithms, and other implementations of IPW could be explored. More research on how to treat multilevel when applying PS based methods is needed to have better guidance on how to include the multi-level structure of surgical scenarios and surgeon and provider-based confounders to the estimations, such as multilevel propensity models or intra-group matching techniques.

Although preference-based IVs have been deemed unfit for my study cases, further work to determine which IVs could be useful in MD epidemiology is

needed. Instruments derived from temporal changes in uptake, which did not exist in this case, may be useful. The falsification of the IV assumption that the IV used only has an effect on the outcome through the treatment should be improved. The criteria used nowadays, an SMD<0.1, to move forward with an IV as unconfounded at least for known confounders, seems insufficient. More research on whether a more stringent SMD could work is needed, but it does not seem to be the case in my study. An alternative could be the performance of a plasmode simulation testing this lack of confounding, but further testing is needed to assess its usefulness.

Observational studies using routinely collected data will continue to flourish in the absence of RCTs and with the new European and global regulations. Clear specific guidelines highlighting the differences with pharmacoepidemiology studies are needed. A key learning from this thesis is that MD studies should consider operator variables and trial emulations should think about eligibility criteria for patient and surgeons. There is still a need for more methodological research to inform guidelines on which methods are better suited for different situations and settings for the post-marketing surveillance of MDs and surgical epidemiology.

6.5 Conclusion

This thesis has important implications for both the clinical delivery of knee replacements and the way observational MD and surgical research is performed. Clinically, although most PKR surgeries are currently performed in fit young patients, the clinical findings of this thesis suggest that PKR has similar benefits over TKR for patients with severe systemic disease and a lower rate of short-term complications. These benefits and risks are also modulated by gender and deprivation, and very importantly by surgeon volume. There is a need to establish clear guidelines on whom to offer PKR, and to inform patients about risks and benefits of both procedures. Another recommendation arising from this thesis is to continue research to disentangle surgeon volume and other variables related to it, and to investigate if PKR done by specialist surgeons in specialised treatment centres, where enough volume of surgeries is performed, would improve patient selection, recovery, and reduce the risk of revision.

This thesis also finds that, with a careful design, some methods, such as IPW and PS stratification, can get to similar results as surgical RCTs. This is of fundamental importance for the compliance with the new regulation of MDs, which will require comprehensive observational post-marketing surveillance. This kind of research

should aim to consider the variables of both patients and surgeons and the optimal timings and methods to minimise biases. PS stratification and inverse probability weighting are useful for minimising confounding when evaluating comparative effectiveness and risks of alternative MDs and surgical procedures. This should be confirmed with further research on other scenarios and circumstances.

Finally, IVs based on surgeon/operator preference were inaccurate in my use cases and simulations, and its use in surgical and MD epidemiology should be further investigated. Surgeon preference is frequently affected by structural and geographical variables, which could impact outcomes. Moreover, surgeon preference calculated from the data is intimately linked to volume, which sometimes affects outcomes directly or through interaction. Using a measure of balance of known confounders, such as SMD, seems to fail to detect confounders in some cases. More research on appropriate methods to detect this confounding and on which real instruments can be used in MD epidemiology is needed if IV estimation are to be used safely.

Bibliography

1. Ring ME. Dentistry: An Illustrated History. Abrams HN, editor1985.

2. Virginia R, Visitors of the University o. Surgical Instruments from Ancient Rome 2007 [Available from: http://exhibits.hsl.virginia.edu/romansurgical/.

3. Digital N. Hospital Admitted Patient Care Activity 2020-21. 2021.

4. Digital N. Sexual and Reproductive Health Services, England (Contraception) 2020/21. 2021.

5. SM B, JM W, RE F. Systemic toxicity related to metal hip prostheses. Clinical toxicology (Philadelphia, Pa). 2014;52(8).

6. Devlin H. 'I had pain all the time': health issues after Essure implants. The Guardian. 2018.

7. Devlin H. Hernia mesh implants cost top British athlete five years of his career. The Guardian. 2018.

8. Dick K. The Bleeding Edge. 2018.

9. Torrres F. 'Implant Files' Investigation Sheds Light on Dark Side of Medical Devices: International Center for Journalists; 2021 [Available from:

https://www.icfj.org/news/implant-files-investigation-sheds-light-dark-side-medical-devices.

10. Hwang TJ, Sokolov E, Franklin JM, Kesselheim AS. Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. Bmj. 2016;353:i3323.

11. Jin J. JAMA patient page. FDA authorization of medical devices. JAMA. 2014;311(4):435.

12. Kramer DB, Xu S, Kesselheim AS. Regulation of medical devices in the United States and European Union. The New England journal of medicine. 2012;366(9):848-55.

13. Classification Of Medical Devices And Their Routes To CE Marking 2021 [Available from: https://support.ce-check.eu/hc/en-us/articles/360008712879-Classification-Of-Medical-Devices-And-Their-Routes-To-CE-Marking.

14. Chai JY. Medical device regulation in the United States and the European Union: a comparative study. Food Drug Law J. 2000;55(1):57-80.

15. COUNCIL DIRECTIVE 93/42/EEC of 14 June 1993 concerning medical devices, (1993).

16. Cohen D. How a fake hip showed up failings in European device regulation. Bmj. 2012.

17. S.510 - 94th Congress (1975-1976): An Act to amend the Federal Food, Drug, and Cosmetic Act to provide for the safety and effectiveness of medical devices intended for human use, and for other purposes., (1976).

18. Rathi VK, Krumholz HM, Masoudi FA, Ross JS. Characteristics of Clinical Studies Conducted Over the Total Product Life Cycle of High-Risk Therapeutic Medical Devices Receiving FDA Premarket Approval in 2010 and 2011. JAMA. 2015;314(6):604-12.

19. Dhruva SS, Bero LA, Redberg RF. Strength of study evidence examined by the FDA in premarket approval of cardiovascular devices. JAMA. 2009;302(24):2679-85.

20. Zheng SY, Dhruva SS, Redberg RF. Characteristics of Clinical Studies Used for US Food and Drug Administration Approval of High-Risk Medical Device Supplements. JAMA. 2017;318(7):619-25.

21. Zuckerman DM, Brown P, Nissen SE. Medical device recalls and the FDA approval process. Arch Intern Med. 2011;171(11):1006-11.

22. News A. FDA's Fast-Track Medical Device Approval Process Under Fire. 2011.

23. Administration FaD. The Least Burdensome Provisions: Concept and Principles 1 Guidance for Industry and Food and Drug Administration Sta. 2019.

24. Hines JZ, Lurie P, Yu E, Wolfe S. Left to their own devices: breakdowns in United States medical device premarket review. PLoS Med. 2010;7(7):e1000280.

25. Medicine Io. Medical devices and the public's health: the FDA 510(k) clearance process at 35 years. Washington, DC; 2011.

26. Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on latest steps to strengthen FDA's 510(k) program for premarket review of medical devices [press release]. 2019.

27. Rathi VK, Ross JS. Modernizing the FDA's 510(k) Pathway. The New England journal of medicine. 2019;381(20):1891-3.

28. EU. REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and

repealing Council Directives 90/385/EEC and 93/42/EEC 2017 [Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02017R0745-20170505&from=EN.

29. REGULATION (EU) 2020/561 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2020 2020 [Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32020R0561.

30. Medicines and Medical Devices Act 2021 - Parliamentary Bills - UK Parliament, (2021).

31. EUDAMED database - EUDAMED. 2021.

32. Fraser AG, Byrne RA, Kautzner J, Butchart EG, Szymanski P, Leggeri I, et al. Implementing the new European Regulations on medical devices-clinical responsibilities for evidence-based practice: a report from the Regulatory Affairs Committee of the European Society of Cardiology. Eur Heart J. 2020;41(27):2589-96.

33. Regulation on EMA's extended mandate becomes applicable [press release]. 2022.

34. Lubbeke A, Smith JA, Prieto-Alhambra D, Carr AJ. The case for an academic discipline of medical device science. EFORT Open Rev. 2021;6(3):160-3.

35. Sedrakyan A, Campbell B, Merino JG, Kuntz R, Hirst A, McCulloch P. IDEAL-D: a rational framework for evaluating and regulating the use of medical devices. Bmj. 2016;353:i2372.

36. Fleetcroft C, McCulloch P, Campbell B. IDEAL as a guide to designing clinical device studies consistent with the new European Medical Device Regulation. BMJ Surgery, Interventions, & Health Technologies. 2021;3(1):e000066.

37. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci. 2011;13(2):217-24.

38. Lubbeke A, Silman AJ, Barea C, Prieto-Alhambra D, Carr AJ. Mapping existing hip and knee replacement registries in Europe. Health Policy. 2018;122(5):548-57.

39. Lubbeke A, Silman AJ, Prieto-Alhambra D, Adler AI, Barea C, Carr AJ. The role of national registries in improving patient safety for hip and knee replacements. BMC musculoskeletal disorders. 2017;18(1):414.

40. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology. . European Medicines Agency; 2020.

41. VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology (Cambridge, Mass). 2009;20(6):863-71.

42. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167:492-99.

43. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183:758-64.

44. OECD. Health Care Utilisation : Surgical procedures 2021 [Available from: https://stats.oecd.org/#.

45. Singh JA, Yu S, Chen L, Cleveland JD. Rates of Total Joint Replacement in the United States: Future Projections to 2020-2040 Using the National Inpatient Sample. J Rheumatol. 2019;46(9):1134-40.

46. UK N. Knee replacement - NHS: NHS UK; 2019 [updated 24 Oct 201. Available from: https://www.nhs.uk/conditions/knee-replacement/.

47. Ethgen O, Bruyere O, Richy F, Dardennes C, Reginster JY. Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. J Bone Joint Surg Am. 2004;86(5):963-74.

48. Kane RL, Saleh KJ, Wilt TJ, Bershadsky B. The functional outcomes of total knee arthroplasty. J Bone Joint Surg Am. 2005;87(8):1719-24.

49. Beard D, Holt M, Mullins M, Malek S, Massa E, Price A. Decision making for knee replacement: variation in treatment choice for late stage medial compartment osteoarthritis. The Knee. 2012;19(6):886-9.

50. Board TNE. 15th Annual Report 2018. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Surgical data to 31st December 2017. 2018.

51. arthroplasty, knee replacement[MeSH Terms] - Search Results - PubMed 2021 [Available from: https://www.ncbi.nlm.nih.gov/pubmed/.

52. UK N. Tracheostomy: NHS UK; 2019 [updated 20 Oct 201. Available from: https://www.nhs.uk/conditions/tracheostomy/.

53. Wei-jie G, Zheng-yi N, Yu H, et al. Clinical characteristics of coronavirus disease 2019 in China. The New England journal of medicine. 2020;382:1708-20.

54. Esteban A AAAI, et al. How is mechanical ventilation employed in the intensive care unit: an international utilization review. Am J Respir Crit Care Med. 2000;161:1450-58.

55. Mehta AB, Cooke CR, Wiener RS, Walkey AJ. Hospital Variation in Early Tracheostomy in the United States: A Population-Based Study. Crit Care Med. 2016;44(8):1506-14.

56. Wartolowska K, Judge A, Hopewell S, Collins GS, Dean BJ, Rombach I, et al. Use of placebo controls in the evaluation of surgery: systematic review. Bmj. 2014;348:g3253.

57. London AJ, Kadane JB. Placebos that harm: sham surgery controls in clinical trials. Statistical methods in medical research. 2002;11(5):413-27.

58. Macklin R. The ethical problems with sham surgery in clinical research. The New England journal of medicine. 1999;341(13):992-6.

59. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American Journal of Epidemiology. 2019;183(8):758-64.

60. 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):1-126.

61. Bennell KL, Hunter DJ, Hinman RS. Management of osteoarthritis of the knee. Bmj. 2012;345:e4934.

62. Litwic A, Edwards MH, Dennison EM, Cooper C. Epidemiology and burden of osteoarthritis. Br Med Bull. 2013;105:185-99.

63. Joint replacement (primary): hip, knee and shoulder. NICE; 2020.

64. CA W-O, K B, H A, M M, JP C. Unicondylar knee arthroplasty in the UK National Health Service: an analysis of candidacy, outcome and cost efficacy. The Knee. 2009;16(6).

65. Pearse AJ, Hooper GJ, Rothwell A, Frampton C. Survival and functional outcome after revision of a unicompartmental to a total knee replacement: the New Zealand National Joint Registry. J Bone Joint Surg Br. 2010;92(4):508-12.

66. Price AJ, Webb J, Topf H, Dodd CA, Goodfellow JW, Murray DW, et al. Rapid recovery after oxford unicompartmental arthroplasty through a short incision. J Arthroplasty. 2001;16(8):970-6.

67. Brown NM, Sheth NP, Davis K, Berend ME, Lombardi AV, Berend KR, et al. Total knee arthroplasty has higher postoperative morbidity than

unicompartmental knee arthroplasty: a multicenter analysis. J Arthroplasty. 2012;27(8 Suppl):86-90.

68. Hassaballa MA, Porteous AJ, Newman JH. Observed kneeling ability after total, unicompartmental and patellofemoral knee arthroplasty: perception versus reality. Knee Surg Sports Traumatol Arthrosc. 2004;12(2):136-9.

69. 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):1-98.

70. 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. The Lancet. 2019;394(10200):746-56.

71. 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. 2020;24:20.

72. Board TNE. 12th Annual Report 2015. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Surgical data to 31 December 2014. 2015.

73. 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(9952):1437-45.

74. 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. The bone & joint journal. 2015;97-b(6):793-801.

75. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiology and Drug Safety. 2010;19(6):537-54.

76. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. Journal of Epidemiology & Community Health. 2006;60(7):578-86.

77. Hernán MA, Robins JM. Causal Inference: What If. . Boca Raton:: Chapman & Hall/CRC; 2020.

78. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. [Lancaster, Pa.]: [Lancaster Press]; 1984.

79. Partial knee replacement 'could be first choice' for suitable patients with osteoarthritis 2019 [Available from: https://discover.dc.nihr.ac.uk/content/signal-000824/partial-knee-replacement-could-be-first-choice-in-some-patients.

80. 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:dc190069.

81. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. Circulation. 2021;143(10):1002-13.

82. Grose E, Wilson S, Barkun J, Bertens K, Martel G, Balaa F, et al. Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature. J Am Coll Surg. 2020;230(1):101-12 e2.

83. First guidance on new rules for certain medical devices 2019 [Available from: https://www.ema.europa.eu/en/news/first-guidance-new-rules-certain-medical-devices.

84. Sabah S, Henckel J, Cook E, Whittaker R, Hothi H, Pappas Y, et al. Validation of primary metal-on-metal hip arthroplasties on the National Joint Registry for England, Wales and Northern Ireland using data from the London Implant Retrieval Centre: a study using the NJR dataset. The bone & joint journal. 2015;97(1):10-8.

85. Sabah S, Henckel J, Koutsouris S, Rajani R, Hothi H, Skinner J, et al. Are all metal-on-metal hip revision operations contributing to the National Joint Registry implant survival curves? A study comparing the London Implant Retrieval Centre and National Joint Registry datasets. The bone & joint journal. 2016;98(1):33-9.

86. National Joint Registry. Patient characteristics for primary knee replacement procedures. 2017.

87. 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 Services and Delivery Research. Southampton (UK)2016.

88. Hawley S, Cordtz R, Dreyer L, Edwards CJ, Arden NK, Delmestri A, et al., editors. 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. Seminars in arthritis and rheumatism; 2018: Elsevier.

89. Devlin N, Appleby J. Getting the Most Out of PROMS: Putting Health Outcomes at the Heart of NHS Decision-Making. The Kings Fund, 2010.

90. 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 Open Quality. 2016;5(1):u204088. w3767.

91. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. The Journal of bone and joint surgery British volume. 1996;78(2):185-90.

92. Murray D, Fitzpatrick R, Rogers K, Pandit H, Beard D, Carr A, et al. The use of the Oxford hip and knee scores. The Journal of bone and joint surgery British volume. 2007;89(8):1010-4.

93. Brooks R, Group E. EuroQol: the current state of play. Health policy. 1996;37(1):53-72.

94. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br. 1998;80(1):63-9.

95. Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol. 2015;68(1):73-9.

96. AD L, H P, A J, DW M. Optimal usage of unicompartmental knee arthroplasty: a study of 41,986 cases from the National Joint Registry for England and Wales. The bone & joint journal. 2015;97-B(11).

97. Patient Reported Outcome Measures (PROMs) in England - 2012-13, Special Topic: Patient engagement with PROMs by demographic characteristics, procedure type and self-reported pre-operative health - NHS Digital 2015 [Available from:

https://webarchive.nationalarchives.gov.uk/20180307194233/http://digital.nhs.uk/catalogue/PUB16482.

98. 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.

99. Uddin MJ, Groenwold RH, Ali MS, de Boer A, Roes KC, Chowdhury MA, et al. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. International journal of clinical pharmacy. 2016;38(3):714-23.

100. 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(3):278-85.

101. Gokhale M, Buse JB, DeFilippo Mack C, Jonsson Funk M, Lund J, Simpson RJ, et al. Calendar time as an instrumental variable in assessing the risk of heart failure with antihyperglycemic drugs. Pharmacoepidemiology and drug safety. 2018;27(8):857-66.

102. Angrist J, Krueger AB. Instrumental variables and the search for identification: From supply and demand to natural experiments. National Bureau of Economic Research; 2001.

103. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology (Cambridge, Mass). 2006;17(3):260-7.

104. Ionescu-Ittu R, Abrahamowicz M, Pilote L. Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. J Clin Epidemiol. 2012;65(2):155-62.

105. Uddin M, Groenwold RH, Boer A, Belitser SV, Roes KC, Hoes AW, et al. Performance of instrumental variable methods in cohort and nested case–control studies: a simulation study. Pharmacoepidemiology and drug safety. 2014;23(2):165-77.

106. Ertefaie A, Small DS, Flory JH, Hennessy S. A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017;26(4):357-67.

107. Ali MS, Uddin MJ, Groenwold RH, Pestman WR, Belitser SV, Hoes AW, et al. Quantitative falsification of instrumental variables assumption using balance measures. Epidemiology (Cambridge, Mass). 2014;25(5):770-2.

108. 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(1):9.

109. StataCorp. Stata Statistical Software: Release 15. . College Station, TX: : StataCorp LLC.; 2017.

110. Uddin MJ, Groenwold RH, de Boer A, Belitser SV, Roes KC, Hoes AW, et al. Performance of instrumental variable methods in cohort and nested case-control studies: a simulation study. Pharmacoepidemiol Drug Saf. 2014;23(2):165-77.

111. 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(5):770-2.

112. 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. The Journal of arthroplasty. 2011;26(8):1343-9.

113. 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.

114. 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 network open. 2019;2(10):e1914325-e.

115. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

116. Royston P, Sauerbrei W. Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley-Blackwell; 2008. xiii, 303 pages p.

117. Sauerbrei W, Royston P. Fractional Polynomials [Available from: http://mfp.imbi.uni-freiburg.de/fp.

118. Rubin DB. On principles for modeling propensity scores in medical research. Pharmacoepidemiology and drug safety. 2004;13(12):855-7.

119. Nguyen T-L, Collins GS, Spence J, Daurès J-P, Devereaux P, Landais P, et al. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC medical research methodology. 2017;17(1):78.

120. 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 (Cambridge, Mass). 2009;20(4):512-22.

121. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Statistical methods in medical research. 2012;21(3):273-93.

122. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine. 2004;23(19):2937-60.

123. Pitblado J, editor Survey data analysis in Stata. Canadian Stata Users Group Meeting; 2009.

124. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and statistics. 2004;86(1):4-29.

125. Ali MS, Prieto-Alhambra D, Lopes L, Ramos D, Bispo N, Ichihara MY, et al. Propensity score methods in health technology assessment: principles, extended applications, and recent advances. Frontiers in Pharmacology. 2019;10:973.

126. 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 (Cambridge, Mass). 2017;28(2):249-57.

127. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011;46(3):399-424.

128. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.

129. 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.

130. Strauss V, Prats-Uribe A, Kolovos S, Judge A, Arden N, Wilkinson M, et al. NIH HTA Protocol: Risk-benefit and costs of unicompartmental (compared to

total) knee replacement for patients with multiple co-morbidities: a non-randomised study, and different novel approaches to minimise confounding.; 2019.

131. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661-79.

132. Ryan CP, Hripcsak G. Proving reliable real-world evidence: Replicating RCTs using LEGEND 2019 [Available from: https://www.ohdsi.org/wp-content/uploads/2019/09/4-Plenary-3-Replicating-LEGEND.pdf.

133. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60.

134. Jonathan J Deeks JPH, Douglas G Altman. Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins J, Thomas J, Chandler J, Cumpston MLT, Page M, Welch V, editors. Cochrane Handbook for Systematic Reviews of Interventions version 62: Cochrane; 2021.

135. Andriolo BN, Andriolo RB, Saconato H, Atallah AN, Valente O. Early versus late tracheotomy for critically ill patients. Cochrane Database Syst Rev. 2015;1.

136. Huang H, Li Y, Ariani F, Chen X, Lin J. Timing of tracheotomy in critically ill patients: a meta-analysis. PLoS One. 2014;9.

137. Smith D, Montagne J, Raices M, et al. Tracheostomy in the intensive care unit: Guidelines during COVID-19 worldwide pandemic. Am J Otolaryngol. 2020;41(10257):8.

138. Shiba T, Ghazizadeh S, Chhetri D, St. John M, Long J. Tracheostomy considerations during the COVID-19 pandemic. OTO Open. 2020;4:24739.

139. Bernal-Sprekelsen M, Aviles-Jurado FX. Consensus document of the Spanish Society of Intensive and Critical Care Medicine and Coronary Units (SEMICYUC), the Spanish Society of Otorhinolaryngology and Head and Neck Surgery (SEORL-CCC) and the Spanish Society of Anesthesiology and Resuscitation (SEDAR) on tracheotomy in patients with COVID-

2020;19:386-92.

140. McGrath BA, Brenner MJ, Warrillow SJ, et al. Tracheostomy in the COVID-19 era: global and multidisciplinary guidance. Lancet Respir Med. 2020;8:717-25.

141. Takhar A, Walker A, Tricklebank S, et al. Recommendation of a practical guideline for safe tracheotomy during the COVID-19 pandemic. Eur Arch Otorhinolaryngol. 2020;277:2173-84.

142. Tay JK, Khoo ML, Loh WS. Surgical considerations for tracheostomy during the COVID-19 pandemic: lessons learned from the severe acute respiratory syndrome outbreak. JAMA Otolaryngol Head Neck Surg. 2020;146:517-18.

143. Aviles-Jurado FX, Prieto-Alhambra D, Gonzlez-Snchez N, et al. Timing, complications, and safety of tracheotomy in critically ill patients with COVID-19. JAMA Otolaryngol Head Neck Surg. 2020;8.

144. Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. Clin Cancer Res. 2012;18(8):2301-8.

145. Jason PF, Robert JG. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.

146. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. 2014;2014:14.

147. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC medical research methodology. 2014;14(1):75.

148. Terragni PP, Antonelli M, Fumagalli R, et al. Early vs late tracheostomy for prevention of pneumonia in mechanically ventilated adult ICU: a randomized controlled trial. Jama. 2010;303:1483-89.

149. Young D, Harrison D, Cuthbertson B, et al. Early vs late tracheostomy for prevention of pneumonia in mechanically ventilated adult ICU. The tracman randomized trial. 2013;309:2121-29.

150. Oliver ER, Gist A, Gillespie MB. Percutaneous versus surgical tracheotomy: an updated meta-analysis. Laryngoscope. 2007;117:1570-75.

151. Halum SL, Ting JY, Plowman EK, et al. A multi-institutional analysis of tracheotomy complications. Laryngoscope. 2012;122:38-45.

152. Botti C, Lusetti F, Neri T, et al. Comparison of percutaneous dilatational tracheotomy versus open surgical technique in severe COVID-19: Complication rates, relative risks and benefits. Auris Nasus Larynx. 2020;8146(20):30296-0.

153. Picetti E, Fornaciari A, Taccone FS, et al. Safety of bedside surgical tracheostomy during COVID-19 pandemic: A retrospective observational study. PLoS One. 2020;15.

154. Bier-Laning C, Cramer JD, Roy S, et al. Tracheostomy during the COVID-19 pandemic: comparison of international perioperative care protocols and practices in 26 countries. Otolaryngol Head Neck Surg. 2020;3(19459):9820961985.

155. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernan MA. A causal framework for classical statistical estimands in failure-time settings with competing events. Stat Med. 2020;39(8):1199-236.

156. 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. The Lancet Rheumatology. 2019.

157. 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. Clinical epidemiology. 2018;10:697.

158. 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. The Lancet Rheumatology. 2019;1(4):e229-e36.

159. Burn E, Liddle AD, Hamilton TW, Judge A, Pandit HG, Murray DW, et al. 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(4):e020977.

160. Partial knee replacement 'could be first choice' for suitable patients with osteoarthritis - Informative and accessible health and care research. NIHR Evidence. 2010.

161. Farrington P, Pugh S, Colville A, Flower A, Nash J, Morgan-Capner P, et al. A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. Lancet. 1995;345(8949):567-9.

162. Miller E, Goldacre M, Pugh S, Colville A, Farrington P, Flower A, et al. Risk of aseptic meningitis after measles, mumps, and rubella vaccine in UK children. Lancet. 1993;341(8851):979-82.

163. Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs. Bmj. 2016;354:i4515.

164. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. Stat Med. 2006;25(10):1768-97.

165. Whitaker HJ, Ghebremichael-Weldeselassie Y, Douglas IJ, Smeeth L, Farrington CP. Investigating the assumptions of the self-controlled case series method. Stat Med. 2018;37(4):643-58.

166. Cadarette SM, Maclure M, Delaney JAC, Whitaker HJ, Hayes KN, Wang SV, et al. Control yourself: ISPE-endorsed guidance in the application of self-controlled study designs in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2021;30(6):671-84.

167. Weldeselassie YG, Whitaker HJ, Farrington CP. Use of the self-controlled case-series method in vaccine safety studies: review and recommendations for best practice. Epidemiol Infect. 2011;139(12):1805-17.

168. Mantilla CB, Horlocker TT, Schroeder DR, Berry DJ, Brown DL. Frequency of myocardial infarction, pulmonary embolism, deep venous thrombosis, and

death following primary hip or knee arthroplasty. Anesthesiology. 2002;96(5):1140-6.

169. Pulido L, Parvizi J, Macgibeny M, Sharkey PF, Purtill JJ, Rothman RH, et al. In hospital complications after total joint arthroplasty. J Arthroplasty. 2008;23(6 Suppl 1):139-45.

170. Mahomed NN, Barrett J, Katz JN, Baron JA, Wright J, Losina E. Epidemiology of total knee replacement in the United States Medicare population. J Bone Joint Surg Am. 2005;87(6):1222-8.

171. Lalmohamed A, Vestergaard P, Klop C, Grove EL, de Boer A, Leufkens HG, et al. Timing of acute myocardial infarction in patients undergoing total hip or knee replacement: a nationwide cohort study. Arch Intern Med. 2012;172(16):1229-35.

172. Warwick D. Prevention of venous thromboembolism in total knee and hip replacement. Circulation. 2012;125(17):2151-5.

173. Keller K, Hobohm L, Barco S, Schmidtmann I, Munzel T, Engelhardt M, et al. Venous thromboembolism in patients hospitalized for knee joint replacement surgery. Sci Rep. 2020;10(1):22440.

174. Schneider AM, Schmitt DR, Brown NM. Unicompartmental knee arthroplasty and revision total knee arthroplasty have a lower risk of venous thromboembolism disease at 30 days than primary total knee arthroplasty. Knee Surgery & Related Research. 2020;32(1):59.

175. Kravitz RL, Duan N, Braslow J. Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages. The Milbank Quarterly. 2004;82(4):661-87.

176. Carr A, Smith JA, Camaradou J, Prieto-Alhambra D. Growing backlog of planned surgery due to covid-19. Bmj. 2021;372:n339.

177. Santoso MB, Wu L. Unicompartmental knee arthroplasty, is it superior to high tibial osteotomy in treating unicompartmental osteoarthritis? A meta-analysis and systemic review. J Orthop Surg Res. 2017;12(1):50.

178. 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-.

179. Digital N. Patient Reported Outcome Measures (PROMs) in England - 2012-13, Special Topic: Patient engagement with PROMs by demographic

characteristics, procedure type and self-reported pre-operative health 2015 [Available from: https://webarchive.nationalarchives.gov.uk/20180307194233/http://digital.nhs.uk/catalogue/PUB16482.

180. Singh JA, Kwoh CK, Richardson D, Chen W, Ibrahim SA. Sex and surgical outcomes and mortality after primary total knee arthroplasty: a risk-adjusted analysis. Arthritis Care Res (Hoboken). 2013;65(7):1095-102.

181. Board TNE. 17th Annual Report 2020. National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Surgical data to 31st December 2019. 2020.

182. Kennedy JA, Burn E, Mohammad HR, Mellon SJ, Judge A, Murray DW. Lifetime revision risk for medial unicompartmental knee replacement is lower than expected. Knee Surg Sports Traumatol Arthrosc. 2020;28(12):3935-41.

183. Labek G, Sekyra K, Pawelka W, Janda W, Stockl B. Outcome and reproducibility of data concerning the Oxford unicompartmental knee arthroplasty: a structured literature review including arthroplasty registry data. Acta Orthop. 2011;82(2):131-5.

184. Baker P, Jameson S, Critchley R, Reed M, Gregg P, Deehan D. Center and surgeon volume influence the revision rate following unicondylar knee replacement: an analysis of 23,400 medial cemented unicondylar knee replacements. J Bone Joint Surg Am. 2013;95(8):702-9.

185. Garriga C, Leal J, Sanchez-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(10):e1914325.

186. Katz JN, Losina E, Barrett J, Phillips CB, Mahomed NN, Lew RA, et al. Association between hospital and surgeon procedure volume and outcomes of total hip replacement in the United States medicare population. J Bone Joint Surg Am. 2001;83(11):1622-9.

187. Sabah SA, Alvand A, Knight R, Beard DJ, Price AJ. Patient-Reported Function and Quality of Life After Revision Total Knee Arthroplasty: An Analysis of 10,727 Patients from the NHS PROMs Program. J Arthroplasty. 2021.

188. Lee DH, Lee SH, Song EK, Seon JK, Lim HA, Yang HY. Causes and Clinical Outcomes of Revision Total Knee Arthroplasty. Knee Surg Relat Res. 2017;29(2):104-9.

189. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219-26.

190. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-102.

191. Du M, Ali S, Strauss VY, Prieto-Alhambra D. Random effects modelling vs logistic regression for the inclusion of surgeon covariates in propensity scores for medical device epidemiology: A simulation study. Pharmacoepidemiology and Drug Safety. 2020;29(S3).

192. Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. Computational Statistics & Data Analysis. 2011;55(4):1770-80.

193. Cafri G, Austin PC. Propensity score methods for time-dependent cluster confounding. Biom J. 2020;62(6):1443-62.

194. Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. Stat Med. 2013;32(19):3373-87.

195. Schuler MS, Chu W, Coffman D. Propensity score weighting for a continuous exposure with multilevel data. Health Serv Outcomes Res Methodol. 2016;16(4):271-92.

196. Thoemmes FJ, West SG. The Use of Propensity Scores for Nonrandomized Designs With Clustered Data. Multivariate Behav Res. 2011;46(3):514-43.

197. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet. 2018;392(10158):1672-82.

198. Jalbert JJ, Ritchey ME, Mi X, Chen CY, Hammill BG, Curtis LH, et al. Methodological considerations in observational comparative effectiveness research for implantable medical devices: an epidemiologic perspective. Am J Epidemiol. 2014;180(9):949-58.

199. Sedrakyan A, Marinac-Dabic D, Normand SL, Mushlin A, Gross T. A framework for evidence evaluation and methodological issues in implantable device studies. Med Care. 2010;48(6 Suppl):S121-8.

200. 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(1):292.

Appendix A

Appendix: Tables and Figures

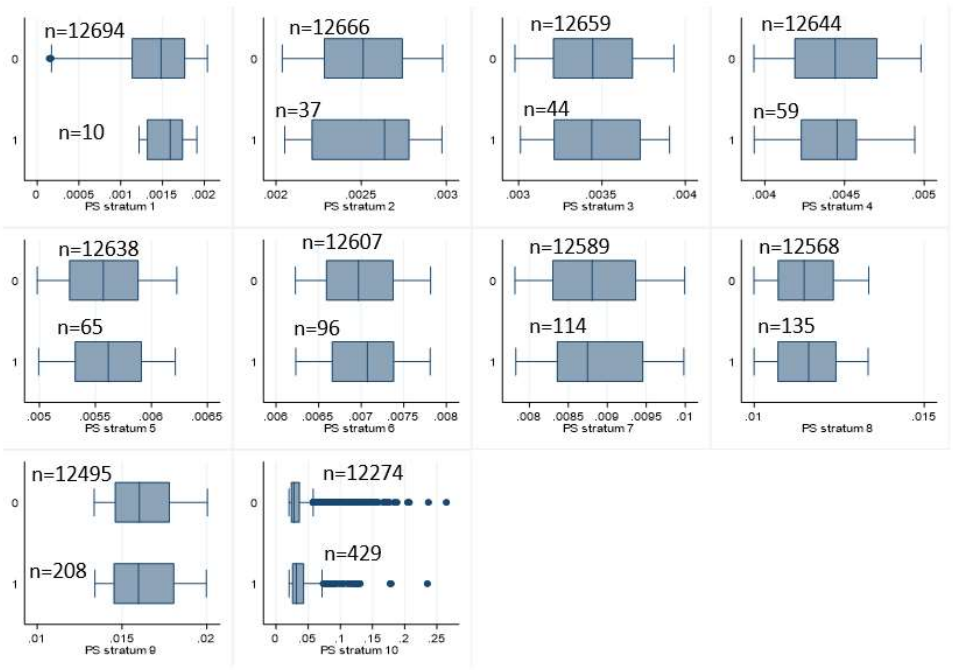
2. Comparative Effectiveness

Appendix Material

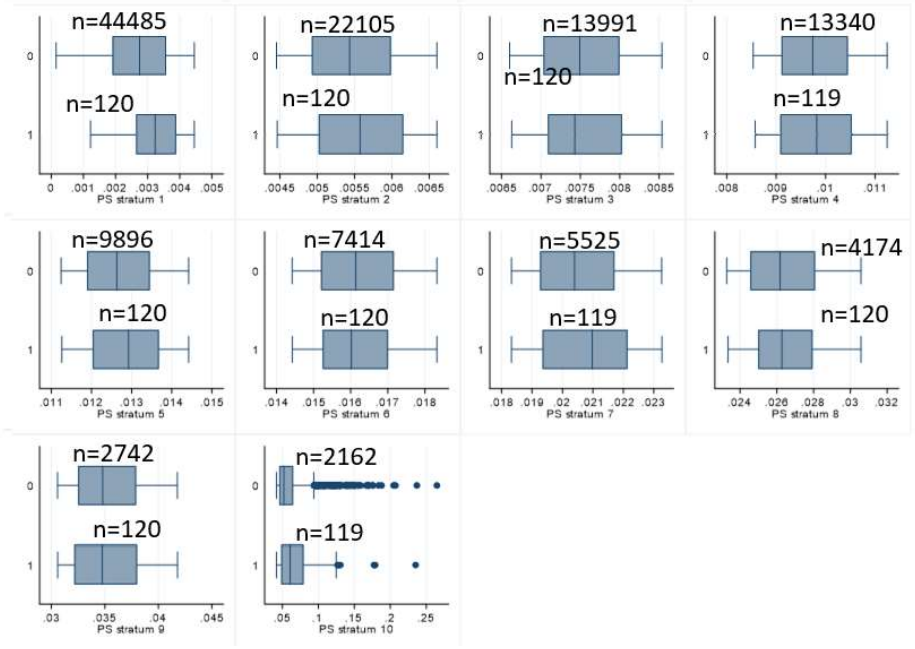
Figures

Appendix Figure 2.1 – Box plots of propensity score distribution for TKR(0) and PKR (1) in each stratum of the OKS cohort based on the PSSwhole (a) and PSSexp (b) method. PS: propensity score.

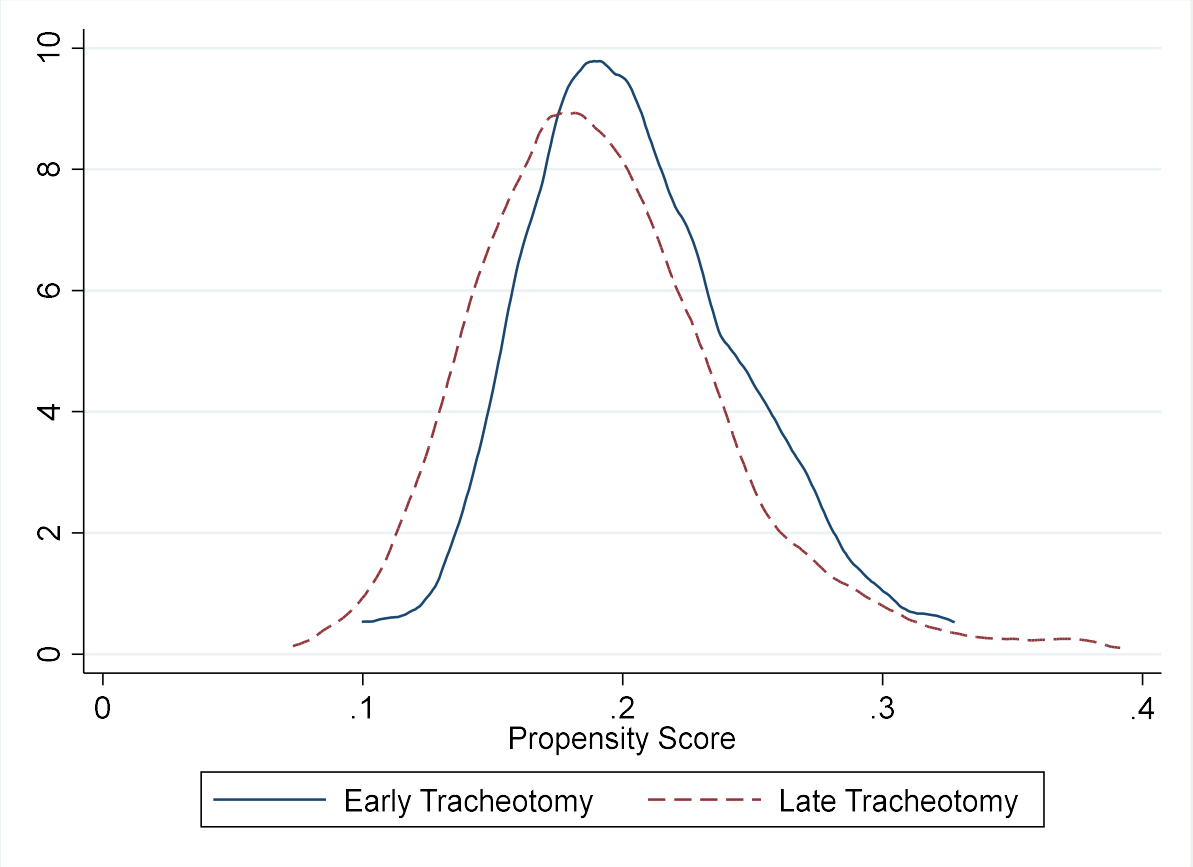
a)



b)


n=44485
n=120
PS stratum 1
n=22105
n=120
PS stratum 2
n=13991
n=120
PS stratum 3
n=13340
n=119
PS stratum 4
n=9896
n=120
PS stratum 5
n=7414
n=120
PS stratum 6
n=5525
n=119
PS stratum 7
n=4174
n=120
PS stratum 8
n=2742
n=120
PS stratum 9
n=2162
n=119
PS stratum 10

Appendix Figure 2.2 – Propensity score of an early tracheotomy distribution in early vs late tracheotomy



Tables

Appendix Table 2.1 - Codes used to identify HES based exclusion criteria in UTMOST stage 1.

Cruciate ligament injury. ICD10 Codes.

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.

Rheumatoid arthritis or other inflammatory disorders. ICD10 codes

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

Foot, hip and spinal pain. ICD10 codes

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	Unespecified Dorsalgia, Cervicothoracic region
M5494	Unespecified Dorsalgia, Thoracic region
M5495	Unespecified Dorsalgia, Thoracolumbar region
M5496	Unespecified Dorsalgia, Lumbar region
M5497	Unespecified Dorsalgia, Lumbosacral region
M5498	Unespecified Dorsalgia, Sacral and sacrococcygeal region
M5499	Unespecified 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
M0127	Arthritis in Lyme 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 thig
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

Foot, hip, or spinal pain. OPCS4 codes.

Code	**Description**
U503	Delivery of rehabilitation for joint replacement
W05	Prostethic 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
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
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

Knee surgery. OPCS4 codes.

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	Capsullorrhapy 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

Septic arthritis. ICD10 codes

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 disease, pelvic region and thigh
M0126	Arthritis in Lyme 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

Patellofemoral Damage or Varus Deformity. ICD 10 codes

Code	**Description**
M22	Disorder of patella
M2116	Varus deformity, not elsewhere classified, knee

Appendix Table 2.2 – Patient-level characteristics for the OKS and whole cohorts.

Stage 1 N (%) or mean (SD)	Whole cohort				OKS cohort			
	TKR (n=273,530)		PKR (n=21,026)		TKR (N=125,834)		PKR (n=1,197)	
Gender								
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								
Urban	203,938	74	14,607	70	92,052	73	844	71
Town and fringe	32,573	12	2,698	13	15,730	13	164	14
Village	26,012	10	2,596	12	12,637	10	138	12
Isolated	11,007	4	1,125	5	5,415	4	51	4
IMD								
Least deprived 10%	29,339	11	2,917	14	14,168	11	149	12
Less deprived 10-20%	31,518	12	2,871	14	15,194	12	137	11
Less deprived 20-30%	31,946	12	2,669	13	15,435	12	142	12
Less deprived 30-40%	32,593	12	2,480	12	15,405	12	138	12
Less deprived 40-50%	31,209	11	2,456	12	14,611	12	164	14
More deprived 10-20%	20,502	7	1,224	6	8,628	7	102	9
More deprived 20-30%	23,357	9	1,415	7	10,110	8	84	7
More deprived 30-40%	26,174	10	1,917	9	11,621	9	123	10
More deprived 40-50%	29,479	11	2,156	10	13,557	11	106	9
Most deprived 10%	17,413	6	921	4	7,105	6	52	4
ASA								
P1 - Fit and healthy	30,224	11	4,394	21	13,849	11	242	20
P2 - Mild disease not incapacitating	243,306	89	16,632	79	111,985	89	955	80
Charlson comorbidity								
0	187,509	69	15,408	73	86,474	69	915	76
1	58,781	21	4,134	20	26,733	21	224	19
2	17,834	7	996	5	8,357	7	41	3
3+	9,406	3	488	2	4,270	3	17	1
Age*	70.2	8.9	64.3	9.5	70.4	8.6	64.9	9.4
BMI*	30.5	5.1	30.0	4.9	30.4	5.0	29.6	4.7
PROM pre-operative OKS*	19.3	6.8	21.3	6.2	19.7	7.6	21.9	7.5
PROM EQ-5D*	69.2	19.4	69.7	19.2	70.0	19.2	71.1	19.0
PROM General Health								
Excellent	161,904	59	6,546	31	88,778	71	604	50

Stage 1 N (%) or mean (SD)	Whole cohort				OKS cohort			
	TKR (n=273,530)		PKR (n=21,026)		TKR (N=125,834)		PKR (n=1,197)	
1	43,913	16	6,643	32	1,433	1	33	3
2	30,058	11	4,400	21	10,398	8	181	15
3	26,008	9	2,217	10	17,504	14	271	23
4	10,024	4	834	4	6,886	5	94	8
Poor	1,623	1	386	2	835	1	14	1
Gastrointestinal disease	52,029	19	3,621	17	25,142	20	174	15
Osteoarthritis and other joint problems	49,941	18	2,696	13	23,578	19	149	12
Mental health	25,823	9	2,380	11	11,421	9	101	8
Respiratory diseases	37,754	14	2,827	13	17,078	14	147	12
Cardiovascular diseases	157,504	58	9,592	46	73,382	58	515	43
Thyroid problems	20,724	8	1,249	6	9,742	8	80	7
Foot, hip, spinal pain	3,096	1	205	1	1,519	1	15	1
Coxarthrosis	8,966	3	381	2	4,395	3	25	2
Neurological disorders	16,435	6	1,208	6	7,491	6	67	6
Other arthrosis	12,818	5	708	3	5,930	5	41	3
Polyarthrosis	15,935	6	675	3	7,520	6	29	2
Spondylosis	7,378	3	349	2	3,501	3	17	1

Note: SD: standard deviation; ASA: American Society of Anaesthesiologists physical status classification system; BMI: body mass index; IMD: index of multiple deprivation; OKS: Oxford Knee Score; PROM: patient-reported outcome measure.
* mean (SD) is presented.

Appendix Table 2.3– Propensity Score covariates included in PS Model codes.

Charlson Index- AIDS. ICD 10 codes

Code	Description
B20	Human immunodeficiency virus
B21	HIV disease resulting in Kaposi sarcoma
B22	Human immunodeficiency virus [HIV] disease resulting in other specified diseases
B24	Unspecified human immunodeficiency virus [HIV] disease

Charlson Index- Metastatic. ICD 10 codes

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

Charlson Index- Moderate to Severe Liver diseases. ICD 10 codes

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

Charlson Index- Cancer. ICD 10 codes

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 esophagus
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
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 lymphoma
C82	Follicular lymphoma
C83	Non-follicular lymphoma
C84	Mature T/NK-cell lymphomas
C85	Other specified and unspecified types of non-Hodgkin lymphoma
C88	Malignant immunoproliferative diseases and certain other B-cell lymphomas
C90	Multiple myeloma and malignant plasma cell neoplasms
C91	Lymphoid leukemia
C92	Myeloid leukemia
C93	Monocytic leukemia
C94	Other leukemias of specified cell type
C95	Leukemia of unspecified cell type
C96	Other and unspecified malignant neoplasms of lymphoid, hematopoietic and related tissue
C97	Malignant neoplasms of independent (primary) multiple sites

Charlson Index- Renal diseasea. ICD 10 codes

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

Charlson Index- Paraplegia. ICD 10 codes

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

Charlson Index- Diabetes complications. ICD 10 codes

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
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

Charlson Index- Diabetes without complications. ICD 10 codes

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
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

Charlson Index- Liver disease. ICD 10 codes

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 hemorrhagic necrosis of liver
K763	Infarction of liver
K764	Peliosis hepatis
K768	Other specified diseases of liver
K769	Liver disease, unspecified
Z944	Liver transplant status

Charlson Index- Peptic ulcer. ICD 10 codes

Code	**Description**
K25	Gastric ulcer
K26	Duodenal ulcer
K27	Peptic ulcer, site unspecified
K28	Gastrojejunal ulcer

Charlson Index- Connective tissue disorder. ICD 10 codes

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

Charlson Index- Pulmonary disease. ICD 10 codes

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 fibers
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 vapors
J701	Chronic and other pulmonary manifestations due to radiation
J703	Chronic Pulmonary Disease

Charlson Index- Dementia. ICD 10 codes

Code	**Description**
F00	Dementia in Alzheimer 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 1. Charlson Index- Cerebral vascular Disease. ICD 10 codes

Code	Description
G45	Transient cerebral ischemic attacks and related syndromes
G46	Vascular syndromes of brain in cerebrovascular diseases
I60	Nontraumatic subarachnoid hemorrhage
I61	Nontraumatic intracerebral hemorrhage
I62	Other and unspecified nontraumatic intracranial hemorrhage
I63	Cerebral infarction
I64	Stroke, not specified
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

Charlson Index- Peripheral vascular disease. ICD 10 codes

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

Charlson Index- Congestive Heart Failure. ICD 10 codes

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	Ischemic 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

Charlson Index- Acute myocardial infarction. ICD 10 codes

Code	**Description**
I21	ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction
I22	STEMI and NSTEMI myocardial infarction
I252	Old myocardial infarction

Osteoartritis and other joint Problems. ICD 10 codes

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
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 thigh
M2597	Unspecified joint disorder lower leg
M2598	Joint disorder NOS ankle and foot
M2599	Unspecified joint disorder other site

Appendix Table 2.4 – Coefficient table for the 1st Imputed dataset PS logistic regression model.

Variable	Odds Ratio	95% Confidence Interval	p,val
Gender: Male	1.15	(1.02 , 1.30)	0.02
Age	0.93	(0.92 , 0.94)	<0.01
BMI	0.95	(0.94 , 0.96)	<0.01
Rural/Urban: Urban	1		
Town and fringe	1.13	(0.95 , 1.34)	0.16
Village	1.15	(0.96 , 1.39)	0.14
Isolated	0.93	(0.70 , 1.24)	0.62
IMD: Least deprived 10%	1		
Less deprived 10,20%	0.86	(0.68 , 1.08)	0.20
Less deprived 20,30%	0.88	(0.69 , 1.11)	0.27
Less deprived 30,40%	0.86	(0.68 , 1.09)	0.23
Less deprived 40,50%	1.09	(0.87 , 1.37)	0.46
More deprived 10,20%	1.15	(0.89 , 1.50)	0.28
More deprived 20,30%	0.80	(0.61 , 1.05)	0.11
More deprived 30,40%	1.02	(0.80 , 1.30)	0.89
More deprived 40,50%	0.75	(0.58 , 0.97)	0.03
Most deprived 10%	0.72	(0.52 , 0.99)	0.04
PROM General Health: Excellent	1		
1	2.54	(1.76 , 3.65)	<0.01
2	2.23	(1.87 , 2.65)	<0.01
3	2.28	(1.97 , 2.64)	<0.01
4	2.30	(1.83 , 2.89)	<0.01
Poor	2.83	(1.63 , 4.91)	<0.01
PROM EQ,5D Health Scale	1.00	(1.00 , 1.00)	0.49
PROM pre,operative OKS	1.04	(1.03 , 1.04)	<0.01
ASA 1	1		
ASA 2	0.88	(0.75 , 1.02)	0.09
CHARLSON 0	1		
1	0.95	(0.78 , 1.16)	0.60
2	0.71	(0.51 , 0.99)	0.04
3	0.80	(0.45 , 1.40)	0.44
4	0.50	(0.19 , 1.36)	0.18
Gastrointestinal disease	0.88	(0.75 , 1.04)	0.14
Osteoarthritis and other joint problems	1.01	(0.76 , 1.33)	0.97
Mental health	0.95	(0.77 , 1.18)	0.64

Variable	Odds Ratio	95% Confidence Interval	p,val
Respiratory diseases	1.11	(0.88 , 1.40)	0.38
Cardiovascular diseases	0.84	(0.74 , 0.95)	0.01
Thyroid problems	1.15	(0.91 , 1.45)	0.25
Foot, hip, spinal pain	1.42	(0.84 , 2.41)	0.19
Coxarthrosis	0.77	(0.48 , 1.21)	0.26
Neurological disorders	1.11	(0.86 , 1.43)	0.41
Other arthrosis	0.99	(0.67 , 1.47)	0.98
Polyarthrosis	0.48	(0.31 , 0.75)	<0.01
Spondylosis	0.70	(0.41 , 1.19)	0.19
Intercept	3.56	(1.67 , 7.59)	<0.01

Appendix Table 2.5 – Baseline patient-level characteristics for patients who received TKR and PKR surgeries in the OKS cohort before and after PS-Matching. Imputed dataset 1.

Stage 1 N (%) or mean (SD)	Before PS matching				After PS matching			
	TKR (N=125,834)		PKR (n=1,197)		TKR (N=5,652)		PKR (n=1,197)	
Gender								
Female	70671	56	576	48	2691	48	576	48
Male	55163	44	621	52	2961	52	621	52
Rural index								
Urban	92052	73	844	71	4038	71	844	71
Town and fringe	15730	13	164	14	737	13	164	14
Village	12637	10	138	12	638	11	138	12
Isolated	5415	4	51	4	239	4	51	4
IMD								
Least deprived 10%	14168	11	149	12	693	12	149	12
Less deprived 10-20%	15194	12	137	11	664	12	137	11
Less deprived 20-30%	15435	12	142	12	702	12	142	12
Less deprived 30-40%	15405	12	138	12	672	12	138	12
Less deprived 40-50%	14611	12	164	14	714	13	164	14
More deprived 10-20%	8628	7	102	9	453	8	102	9
More deprived 20-30%	10110	8	84	7	407	7	84	7
More deprived 30-40%	11621	9	123	10	578	10	123	10
More deprived 40-50%	13557	11	106	9	515	9	106	9
Most deprived 10%	7105	6	52	4	254	4	52	4
ASA								
P1 - Fit and healthy	13849	11	242	20	1100	19	242	20
P2 - Mild disease not incapacitating	111985	89	955	80	4552	81	955	80
Charlson comorbidity								
0	86474	69	915	76	4319	76	915	76
1	26733	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
Age*	3234	1	180	1	21	0	4	0
BMI*	70.4	8.6	64.9	9.4	65.3	9.0	64.9	9.4
PROM pre-operative OKS*	30.4	5.0	29.6	4.7	29.6	4.9	29.6	4.7
PROM EQ-5D*	19.7	7.6	21.9	7.5	21.82	7.73	21.88	7.52

Stage 1 N (%) or mean (SD)	Before PS matching				After PS matching			
	TKR (N=125,834)		PKR (n=1,197)		TKR (N=5,652)		PKR (n=1,197)	
PROM General Health	70.0	19.2	71.1	19.0	71.09	19.22	71.13	18.97
Excellent								
1	88778	71	604	50	2918	52	604	50
2	1433	1	33	3	138	2	33	3
3	10398	8	181	15	815	14	181	15
4	17504	14	271	23	1273	23	271	23
Poor	6886	5	94	8	442	8	94	8
Gastrointestinal disease	835	1	14	1	66	1	14	1
Osteoarthritis and other joint problems	25142	20	174	15	798	14	174	15
Mental health	23578	19	149	12	751	13	149	12
Respiratory diseases	11421	9	101	8	469	8	101	8
Cardiovascular diseases	17078	14	147	12	686	12	147	12
Thyroid problems	73382	58	515	43	2465	44	515	43
Foot, hip, spinal pain	9742	8	80	7	327	6	80	7
Coxarthrosis	1519	1	15	1	79	1	15	1
Neurological disorders	4395	3	25	2	121	2	25	2
Other arthrosis	7491	6	67	6	306	5	67	6
Polyarthrosis	5930	5	41	3	210	4	41	3
Spondylosis	7520	6	29	2	140	2	29	2

Note: SD: standard deviation; ASA: American Society of Anaesthesiologists physical status classification system; BMI: body mass index; IMD: index of multiple deprivation; OKS: Oxford Knee Score; PROM: patient-reported outcome measure.
* mean (SD) is presented.

Appendix Table 2.6 – Baseline patient-level characteristics for patients who received TKR and PKR surgeries. Full OKS cohort vs those whose surgeons had done >10 surgeries of the same type on the previous year.

Stage 1 N (%) or mean (SD)	*OKS* cohort				*OKS* sensitivity cohort			
	TKR (N=125,834)		PKR (n=1,197)		TKR (N=125,834)		PKR (n=1,197)	
Gender								
Female	70,671	56	576	48	64,468	56	287	48
Male	55,163	44	621	52	50,403	44	315	52
Rural index								
Urban	92,052	73	844	71	83,810	73	396	66
Town and fringe	15,730	13	164	14	14,446	13	97	16
Village	12,637	10	138	12	11,587	10	79	13
Isolated	5,415	4	51	4	5,028	4	30	5
IMD								
Least deprived 10%	14,168	11	149	12	12,981	11	75	12
Less deprived 10-20%	15,194	12	137	11	13,992	12	72	12
Less deprived 20-30%	15,435	12	142	12	14,159	12	82	14
Less deprived 30-40%	15,405	12	138	12	14,140	12	65	11
Less deprived 40-50%	14,611	12	164	14	13,371	12	94	16
More deprived 10-20%	8,628	7	102	9	7,731	7	42	7
More deprived 20-30%	10,110	8	84	7	9,178	8	39	6
More deprived 30-40%	11,621	9	123	10	10,551	9	59	10
More deprived 40-50%	13,557	11	106	9	12,333	11	52	9
Most deprived 10%	7,105	6	52	4	6,435	6	22	4
ASA								
P1 - Fit and healthy	13,849	11	242	20	12,748	11	118	20
P2 - Mild disease not incapacitating	111,985	89	955	80	102,123	89	484	80
Charlson comorbidity								
0	86,474	69	915	76	79,157	69	447	74
1	26,733	21	224	19	24,269	21	121	20
2	8,357	7	41	3	7,582	7	23	4
3+	4,270	3	17	1	2,579	2	8	1
Age*	70.4	8.6	64.9	9.4	1,284	1	3	0

Stage 1 N (%) or mean (SD)	*OKS* cohort				*OKS* sensitivity cohort			
	TKR (N=125,834)		PKR (n=1,197)		TKR (N=125,834)		PKR (n=1,197)	
BMI*	30.4	5.0	29.6	4.7	70.3	8.6	65.6	9.3
PROM pre-operative OKS*	19.7	7.6	21.9	7.5	30.4	5.0	29.5	4.6
PROM EQ-5D*	70.0	19.2	71.1	19.0	19.7	7.6	22.1	7.6
PROM General Health					70.0	19.2	71.3	18.8
Excellent	88,778	71	604	50	81,617	71	306	51
1	1,433	1	33	3	1,311	1	14	2
2	10,398	8	181	15	9,395	8	100	17
3	17,504	14	271	23	1,5652	14	128	21
4	6,886	5	94	8	6,148	5	48	8
Poor	835	1	14	1	748	1	6	1
Gastrointestinal disease	25,142	20	174	15	22,766	20	93	15
Osteoarthritis and other joint problems	23,578	19	149	12	21,434	19	71	12
Mental health	11,421	9	101	8	10,528	9	46	8
Respiratory diseases	17,078	14	147	12	15,503	13	79	13
Cardiovascular diseases	73,382	58	515	43	66,546	58	272	45
Thyroid problems	9,742	8	80	7	8,868	8	39	6
Foot, hip, spinal pain	1,519	1	15	1	1,408	1	7	1
Coxarthrosis	4,395	3	25	2	4,000	3	14	2
Neurological disorders	7,491	6	67	6	6,794	6	38	6
Other arthrosis	5,930	5	41	3	5,340	5	15	2
Polyarthrosis	7,520	6	29	2	6,877	6	9	1
Spondylosis	3,501	3	17	1	3,196	3	7	1

Note: SD: standard deviation; ASA: American Society of Anaesthesiologists physical status classification system; BMI: body mass index; IMD: index of multiple deprivation; OKS: Oxford Knee Score; PROM: patient-reported outcome measure.
* mean (SD) is presented.

Appendix Table 2.7 - Baseline characteristics in tracheotomised patients stratified by exclusion

	Total	Included	General Exclusions	Tracheostomy <7 d
	N=794	N=754	N=40	N=58
Sex				
Female	29.8%	30.4%	20.0%	24.1%
Male	70.2%	69.6%	80.0%	75.9%
Age (years)	62.9 (10.2)	62.9 (10.2)	62.8 (9.7)	61.3 (11.1)
missing	1.6%	0.0%	32.5%	0.0%
Tobacco consumption				
Never	74.9%	74.4%	85.0%	72.4%
Smoker	16.0%	16.4%	7.5%	20.7%
Missing	9.1%	9.2%	7.5%	6.9%
Smoking Index (pack/year)	3.5 (12.8)	3.5 (12.7)	4.1 (15.3)	4.9 (13.7)
Missing _	16.0%	16.4%	7.5%	12.1%
Weigth (Kg)	83.1 (15.6)	83.2 (15.6)	81.4 (15.6)	85.1 (17.0)
Missing	20.0%	19.6%	27.5%	19.0%
Heigth	168.8 (9.2)	168.8 (9.2)	168.6 (9.4)	17..5 (10.5)
Missing	22.2%	21.9%	27.5%	22.4%
BMI	29.3 (5.5)	29.3 (5.5)	28.8 (4.9)	29.0 (5.9)
Missing	24.2%	23.7%	32.5%	24.1%
High Blood Pressure				
No	53.8%	53.4%	60.0%	56.9%
Yes	45.8%	46.2%	40.0%	43.1%
Missing	0.4%	0.4%	0.0%	0.0%
Immunodepression				
No	92.8%	93.0%	90.0%	98.3%
Yes	6.8%	6.6%	10.0%	1.7%
Missing	0.4%	0.4%	0.0%	0.0%
Cardiac insuficiency				
No	96.0%	96.3%	90.0%	96.6%
Yes	3.8%	3.4%	10.0%	3.4%
Missing	0.3%	0.3%	0.0%	0.0%
Autoimmune disease				
No	94.1%	93.8%	100.0%	91.4%
Yes	5.7%	6.0%	0.0%	8.6%

	Total	Included	General Exclusions	Tracheostomy <7 d
Missing	0.3%	0.3%	0.0%	0.0%
COPD				
No	9234%	92.3%	95.0%	91.4%
Yes	7.2%	7.3%	5.0%	8.6%
Missing	0.4%	0.4%	0.0%	0.0%
Pregnancy				
No	99.4%	99.5%	97.5%	100.0%
Yes	0.5%	0.4%	2.5%	0.0%
Missing	0.1%	0.1%	0.0%	0.0%
Diabetes Mellitus				
No	78.8%	78.6%	82.5%	84.5%
Yes	20.9%	21.1%	17.5%	15.5%
Missing	0.3%	0.3%	0.0%	0.0%
Neuromuscular disease				
No	98.5%	98.4%	100.0%	100.0%
Yes	1.3%	1.3%	0.0%	0.0%
Missing	0.3%	0.3%	0.0%	0.0%
Ischemic cardiopathy				
No	88.9%	89.1%	85.0%	74.1%
Yes	10.8%	10.6%	15.0%	25.9%
Missing	0.3%	0.3%	0.0%	0.0%
APACHE	15.0 (6.6)	15.0 (6.6)	15.1 (7.0)	14.8 (6.9)
Missing	18.8%	19.5%	5.0%	34.5%
SOFA	6.3 (3.8)	6.3 (3.8)	6.2 (3.3)	8.4 (4.8)
Missing	22.7%	21.9%	37.5%	22.4%
INR tracheotomy	1.5 (2.1)	1.6 (2.1)	1.4 (1.7)	1.5 (2.0)
Missing	16.8%	17.0%	12.5%	6.9%
PAFI (O2/ FIO2)	140.6 (69.8)	141.4 (69.9)	123.5 (65.5)	160.1 (77.6)
Missing	11.2%	10.7%	20.0%	10.3%
PAFI (O2/ FIO2) 7 days	183.8 (74.8)	184.3 (74.8)	173.3 (74.8)	203.6 (88.3)
Missing	14.4%	14.1%	20.0%	24.1%
PAFI (O2/ FIO2) trach	193.7 (70.4)	193.6 (70.2)	194.4 (74.8)	204.7 (80.3)
Missing	7.8%	7.4%	15.0%	5.2%
PEEP intubation	12.6 (5.0)	12.6 (5.0)	12.7 (3.0)	13.2 (3.8)
Missing	11.0%	10.6%	17.5%	8.6%
PEEP 7 days	11.1 (7.4)	11.1 (7.6)	10.4 (3.8)	11.3 (3.7)
Missing	14.7%	14.5%	20.0%	20.7%
PEEP tracheotomy	9.8 (3.1)	9.8 (3.1)	9.7 (3.0)	11.5 (3.9)
Missing	4.2%	3.8%	10.0%	0.0%
Ventilator problems				
No	85.5%	85.4%	87.5%	91.4%

	Total	Included	General Exclusions	Tracheostomy <7 d
Yes	13.5%	13.5%	12.5%	8.6%
Missing	1.0%	1.1%	0.0%	0.0%
Anticoagulant drug				
No	40.2%	40.1%	42.5%	34.5%
Yes	48.5%	48.5%	47.5%	48.3%
Missing	11.3%	11.4%	10.0%	17.2%
Pronation				
No	31.5%	32.1%	20.0%	60.3%
Yes	68.5%	67.9%	80.0%	39.7%
Pronation days	5.8 (7.9)	5.6 (7.7)	8.3 (10.5)	1.8 (3.6)
Missing	10.1%	10.5%	2.5%	5.2%
Pronation days before trach	9.3 (9.2)	9.1 (9.0)	13.2 (12.2)	1.6 (2.4)
Missing	0.0%	0.0%	0.0%	0.0%
Pronation before_7d				
No	36.3%	37.0%	22.5%	53.4%
Yes	63.7%	63.0%	77.5%	46.6%
Pronation days before 7d	3.9 (3.8)	3.9 (3.8)	4.2 (2.9)	3.1 (4.1)
Missing	0.3%	0.0%	5.0%	0.0%
Pronation before trach				
No	35.3%	35.5%	30.0%	67.2%
Yes	64.7%	64.5%	70.0%	32.8%
Last pronation before tracheotomy				
No	55.5%	56.1%	45.0%	82.8%
Yes	44.5%	43.9%	55.0%	17.2%
Last pronation after tracheotomy				
No	80.5%	80.6%	77.5%	81.0%
Yes	19.5%	19.4%	22.5%	19.0%
Vasoactive drugs tracheostomy				
No	47.4%	46.8%	57.5%	46.6%
Yes	40.6%	40.8%	35.0%	46.6%
Missing	12.1%	12.3%	7.5%	6.9%
Vasoactive drugs OTI				
No	42.9%	42.2%	57.5%	34.5%
Yes	52.8%	53.3%	42.5%	65.5%
Missing	4.3%	4.5%	0.0%	0.0%
Secretions problems				
No	74.4%	74.4%	75.0%	82.8%

	Total	Included	General Exclusions	Tracheostomy <7 d
Increase pressure	12.3%	12.2%	15.0%	6.9%
Obstruction	3.8%	3.8%	2.5%	3.4%
Missing	9.4%	9.5%	7.5%	6.9%
Indication tracheotomy				
Prolonged mechanical ventilation	81.5%	81.2%	87.5%	75.9%
Secretions management	10.6%	10.7%	7.5%	17.2%
Other	7.8%	8.0%	5.0%	6.9%
Missing	0.1%	0.1%	0.0%	0.0%
Total linfocites	5238.0 (26329.0)	5155.4 (26054.9)	6808.5 (31439.1)	3395.0 (12670.6)
missing	1.8%	1.7%	2.5%	0.0%
INR	1.6 (2.2)	1.6 (2.3)	1.2 (0.2)	1.6 (2.2)
missing	7.8%	8.1%	2.5%	3.4%
D-Dimer	1499.8 (1708.0)	1509.5 (1719.2)	1299.6 (1468.7)	1424.2 (1479.6)
missing	23.8%	23.5%	30.0%	39.7%
Ferritine	1357.7 (1306.6)	1345.5 (1292.6)	1607.5 (1573.6)	1114.0 (1043.2)
missing	24.3%	24.0%	30.0%	13.8%
LDH	645.5 (837.1)	651.0 (858.4)	547.2 (221.3)	1360.0 (1880.3)
missing	14.4%	14.6%	10.0%	24.1%
Leukocites	4445.0 (9343.6)	4448.9 (9494.0)	4370.0 (5854.7)	3396.3 (6162.1)
missing	1.5%	1.5%	2.5%	0.0%
Linfocites	62.2 (183.5)	61.2 (178.8)	81.3 (259.0)	52.2 (163.9)
missing	1.4%	1.3%	2.5%	0.0%
CRP	20.0 (22.6)	20.2 (22.7)	17.3 (21.4)	13.3 (21.1)
missing	42.6%	42.8%	37.5%	27.6%

COPD: Chronic obstructive pulmonary disease, DM: diabetes mellitus, INR: International normalised ratio, PAFI (PaO2/FiO2), PEEP: positive end-expiratory pressure, OTI: orotracheal intubation, LDH: lactate dehidrogenase, CRP C-reactive protein.

Appendix Table 2.8 - Pronation and pronation days before tracheotomy in tracheostomised patients stratified by early or late weaning.

	Total	Late (>10d after IOT)	Early (10d after IOT)
	N=696	N=554	N=142
Pronation anytime	70·3%	72·0%	63·4%
Days of pronation (if pronated)	9·0 (8·1)	9·5 (8·1)	6·8 (7·5)
Missing	10·9%	11·0%	10·6%
Pronation on 7 days post-OTI	64·4%	65·2%	61·3%
Pronation days on 7 days post-OTI	4·9 (2·9)	5·0 (2.5)	4·6 (4·1)
Pronation before tracheotomy	67·1%	68·8%	60·6%
Days of pronation before tracheotomy	8·3 (6·1)	9·0 (6·2)	5·3 (4·6)
Last pronation before tracheotomy	46·1%	49·5%	33·1%
Pronation continued after tracheotomy	19·4%	18·1%	24·6%

OTI: orotracheal intubation

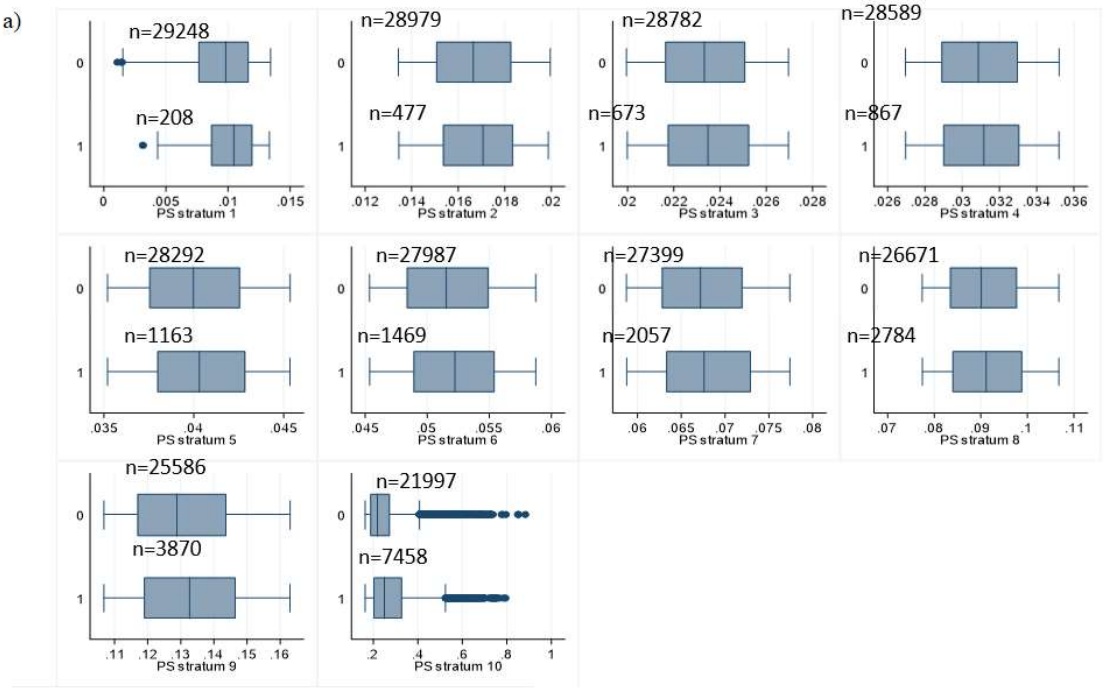
3. Comparative Safety

Appendix Material

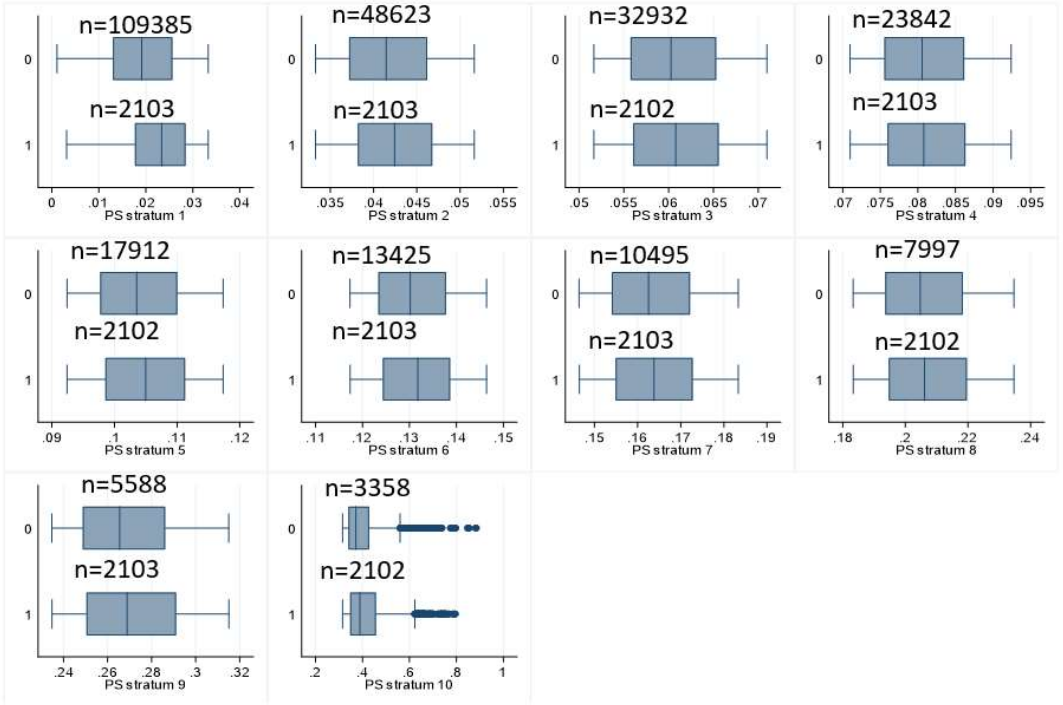
Figures

Appendix Figure 3.1 – Box plots of propensity score distribution for TKR(0) and PKR (1) in each stratum of the OKS cohort based on the PSSwhole (a) and PSSexp (b) method. PS: propensity score.

a)


n=29248
n=208
PS stratum 1
n=28979
n=477
PS stratum 2
n=28782
n=673
PS stratum 3
n=28589
n=867
PS stratum 4
n=28292
n=1163
PS stratum 5
n=27987
n=1469
PS stratum 6
n=27399
n=2057
PS stratum 7
n=26671
n=2784
PS stratum 8
n=25586
n=3870
PS stratum 9
n=21997
n=7458
PS stratum 10

b)


n=109385
n=2103
PS stratum 1
n=48623
n=2103
PS stratum 2
n=32932
n=2102
PS stratum 3
n=23842
n=2103
PS stratum 4
n=17912
n=2102
PS stratum 5
n=13425
n=2103
PS stratum 6
n=10495
n=2103
PS stratum 7
n=7997
n=2102
PS stratum 8
n=5588
n=2103
PS stratum 9
n=3358
n=2102
PS stratum 10

Tables

Appendix Table 3.1 – Codes used to identify complications in UTMOST.

Myocardial Infarction. ICD10 Codes.

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

Venous Thromboembolism. ICD10 Codes.

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

Prosthetic joint infection. ICD-10 and OPCS codes.

Code	Description
Infection	
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 orthopedic 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 codes	
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)	

Appendix Table 3.2 – Coefficient table for the 1st Imputed dataset PS logistic regression model.

Variable	Odds Ratio	95% Confidence Interval	p-val
Gender: Male	1.16	(1.12 - 1.19)	<0.01
Age	0.93	(0.93 - 0.93)	<0.01
BMI	0.97	(0.96 - 0.97)	<0.01
Rural/Urban: Urban	1		
Town and fringe	1.09	(1.04 - 1.14)	<0.01
Village	1.30	(1.24 - 1.36)	<0.01
Isolated	1.29	(1.20 - 1.38)	<0.01
IMD: Least deprived 10%	1		
Less deprived 10-20%	0.89	(0.84 - 0.94)	<0.01
Less deprived 20-30%	0.82	(0.77 - 0.86)	<0.01
Less deprived 30-40%	0.74	(0.70 - 0.78)	<0.01
Less deprived 40-50%	0.76	(0.72 - 0.81)	<0.01
More deprived 10-20%	0.58	(0.54 - 0.62)	<0.01
More deprived 20-30%	0.59	(0.55 - 0.64)	<0.01
More deprived 30-40%	0.71	(0.67 - 0.76)	<0.01
More deprived 40-50%	0.71	(0.67 - 0.75)	<0.01
Most deprived 10%	0.51	(0.47 - 0.55)	<0.01
PROM General Health: Excellent	1		
1	3.93	(3.78 - 4.07)	<0.01
2	3.74	(3.59 - 3.91)	<0.01
3	2.26	(2.15 - 2.39)	<0.01
4	2.48	(2.29 - 2.68)	<0.01
Poor	7.44	(6.59 - 8.41)	<0.01
PROM EQ-5D Health Scale	1.00	(1.00 - 1.00)	<0.01
PROM pre-operative OKS	1.05	(1.04 - 1.05)	<0.01
ASA 1	1		
ASA 2	0.78	(0.75 - 0.81)	<0.01
CHARLSON 0	1		
1	1.03	(0.98 - 1.08)	0.26
2	1.05	(0.98 - 1.13)	0.19
3	1.09	(0.96 - 1.23)	0.17
4	1.26	(1.08 - 1.48)	<0.01
Gastrointestinal disease	1.16	(1.11 - 1.21)	<0.01
Osteoarthritis and other joint problems	0.99	(0.92 - 1.06)	0.71
Mental health	1.40	(1.33 - 1.47)	<0.01

Variable	Odds Ratio	95% Confidence Interval	p-val
Respiratory diseases	1.08	(1.02 - 1.14)	0.01
Cardiovascular diseases	0.97	(0.94 - 1.00)	0.07
Thyroid problems	1.03	(0.97 - 1.10)	0.35
Foot, hip, spinal pain	1.16	(1.00 - 1.35)	0.06
Coxarthrosis	0.71	(0.63 - 0.81)	<0.01
Neurological disorders	1.07	(1.00 - 1.14)	0.05
Other arthrosis	0.96	(0.87 - 1.06)	0.43
Polyarthrosis	0.67	(0.61 - 0.74)	<0.01
Spondylosis	0.84	(0.74 - 0.95)	0.01
Intercept	12.30	(10.10 - 14.98)	<0.01

Appendix Table 3.3 – Baseline patient-level characteristics for patients who received TKR and PKR surgeries in the OKS cohort before and after PS-Matching. Imputed dataset 1.

Stage 1 N (%) or mean (SD)	Before PS matching				After PS matching			
	TKR (N=273,530)		PKR (n=21,026)		TKR (N=92,071)		PKR (n=21,026)	
Gender								
Female	155267	57	10016	48	35300	50	10016	48
Male	118263	43	11010	52	35745	50	11010	52
Rural index								
Urban	203938	74	14607	70	50141	71	14607	69
Town and fringe	32573	12	2698	13	9035	13	2698	13
Village	26012	10	2596	12	8288	12	2596	12
Isolated	11007	4	1125	5	3581	5	1125	5
IMD								
Least deprived 10%	29339	11	2917	14	9315	13	2917	14
Less deprived 10-20%	31518	12	2871	14	9325	13	2871	14
Less deprived 20-30%	31946	12	2669	13	8899	13	2669	13
Less deprived 30-40%	32593	12	2480	12	8422	12	2480	12
Less deprived 40-50%	31209	11	2456	12	8276	12	2456	12
More deprived 10-20%	20502	7	1224	6	4394	6	1224	6
More deprived 20-30%	23357	9	1415	7	5087	7	1415	7
More deprived 30-40%	26174	10	1917	9	6570	9	1917	9
More deprived 40-50%	29479	11	2156	10	7339	10	2156	10
Most deprived 10%	17413	6	921	4	3418	5	921	4
ASA								
P1 - Fit and healthy	30224	11	4394	21	12213	17	4394	21
P2 - Mild disease not incapacitating	243306	89	16632	79	58832	83	16632	79
Charlson comorbidity								
0	187509	69	15408	73	51019	72	15408	73
1	58781	21	4134	20	14467	20	4134	20
2	17834	7	996	5	3759	5	996	5
3+	4270	3	17	1	1800	3	488	2
Age*	70.2	8.9	64.3	9.5	66.1	9.1	64.3	9.5

Stage 1 N (%) or mean (SD)	Before PS matching				After PS matching			
	TKR (N=273,530)		PKR (n=21,026)		TKR (N=92,071)		PKR (n=21,026)	
BMI*	30.5	5.1	30.0	4.9	30.2	5.1	30.0	4.9
PROM pre-operative OKS*	19.3	6.8	21.3	6.2	1.20	1.25	1.30	1.23
PROM EQ-5D*	69.2	19.4	69.7	19.2	69.66	19.43	69.70	19.17
PROM General Health								
Excellent	161904	59	6546	31	26651	38	6546	31
1	43913	16	6643	32	19224	28	6643	32
2	30058	11	4400	21	13233	18	4400	21
3	26008	9	2217	10	7838	11	2217	10
4	10024	4	834	4	2926	4	834	4
Poor	1623	1	386	2	852	1	386	2
Gastrointestinal disease	52029	19	3621	17	12701	18	3621	17
Osteoarthritis and other joint problems	49941	18	2696	13	9998	14	2696	13
Mental health	25823	9	2380	11	7645	11	2380	11
Respiratory diseases	37754	14	2827	13	9636	14	2827	13
Cardiovascular diseases	157504	58	9592	46	35015	49	9592	46
Thyroid problems	20724	8	1249	6	4568	6	1249	6
Foot, hip, spinal pain	3096	1	205	1	731	1	205	1
Coxarthrosis	8966	3	381	2	1461	2	381	2
Neurological disorders	16435	6	1208	6	4201	6	1208	6
Other arthrosis	12818	5	708	3	2511	4	708	3
Polyarthrosis	15935	6	675	3	2665	4	675	3
Spondylosis	7378	3	349	2	1349	2	349	2

Note: SD: standard deviation; ASA: American Society of Anaesthesiologists physical status classification system; BMI: body mass index; IMD: index of multiple deprivation; OKS: Oxford Knee Score; PROM: patient-reported outcome measure.
* mean (SD) is presented.

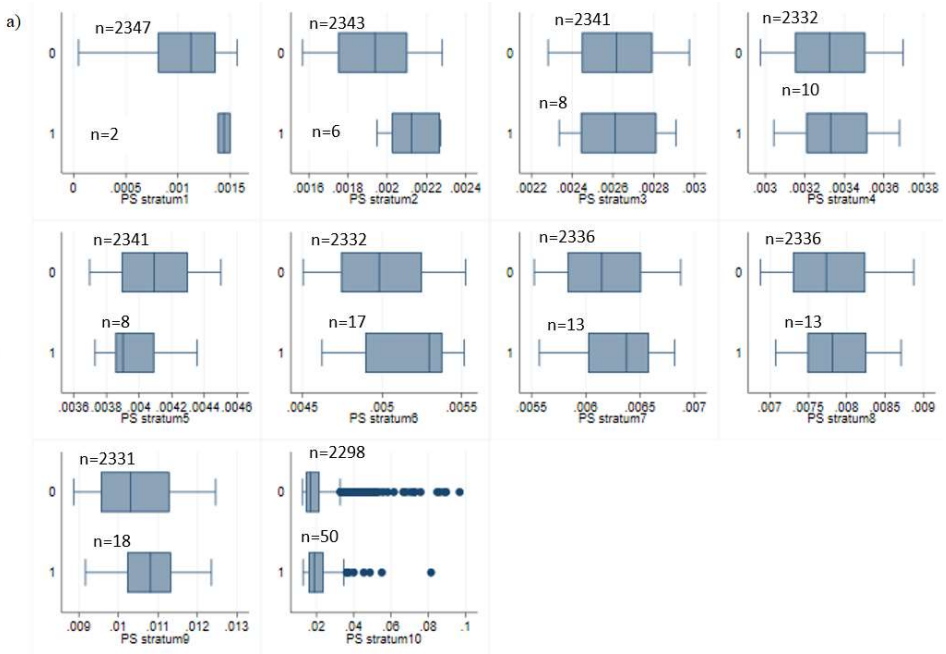
4. Treatment Heterogeneity

Appendix Material

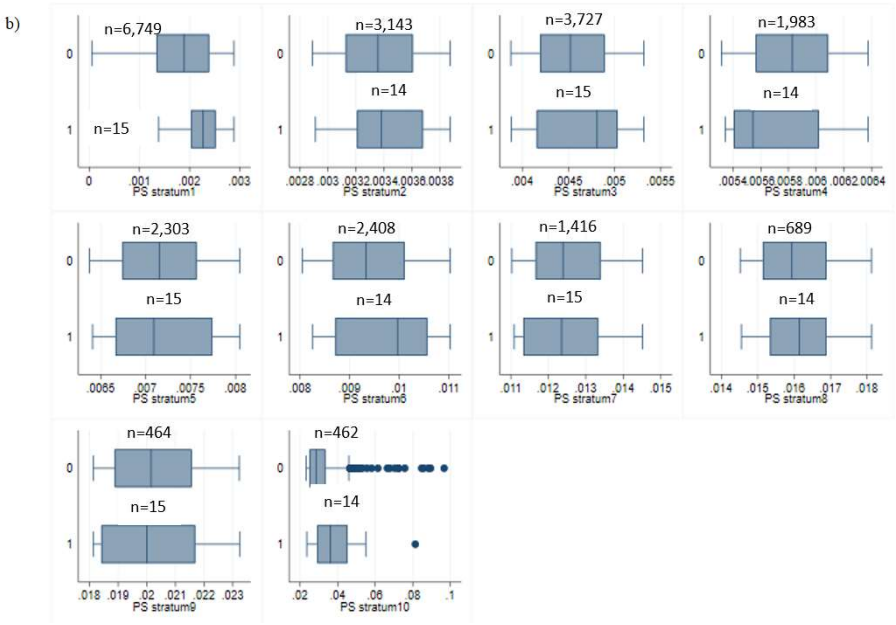
Figures

Appendix Figure 4.1 – Box plot of propensity score distribution for TKR(0) and PKR (1) in each stratum of the effectiveness cohort based on the PSSwhole (a) and PSSexp (b) method

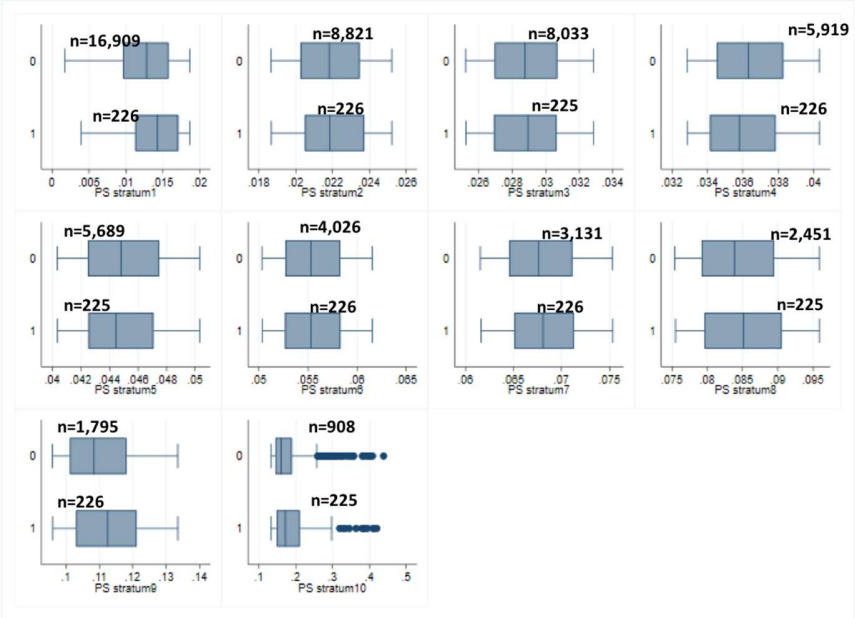
a: PSS_{whole}



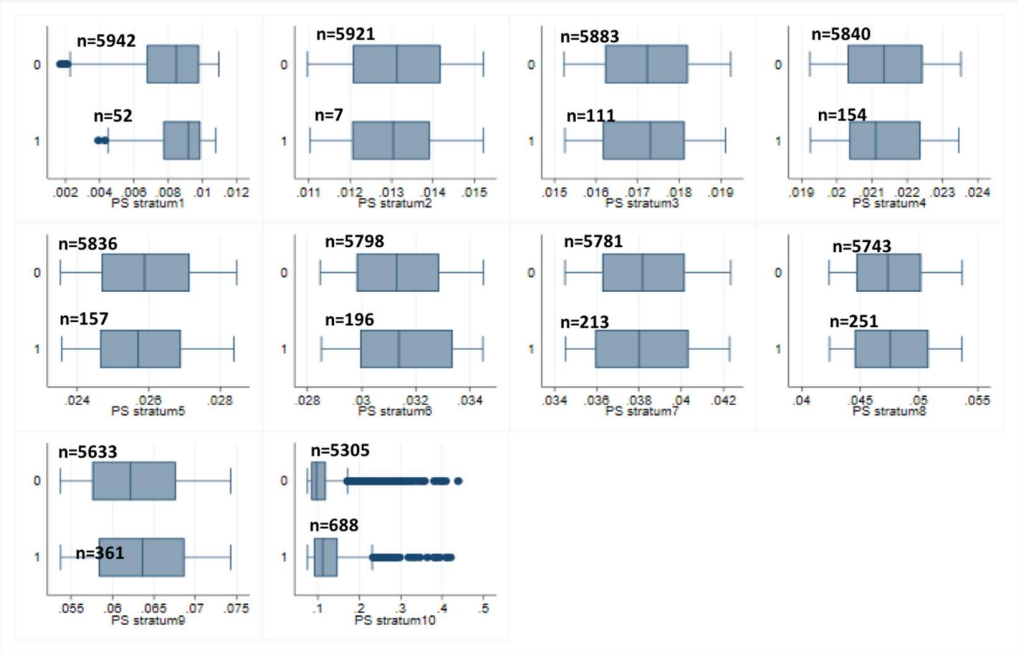
b: PSS_{exp}


b)
n=6,749
n=15
PS stratum1
n=3,143
n=14
PS stratum2
n=3,727
n=15
PS stratum3
n=1,983
n=14
PS stratum4
n=2,303
n=15
PS stratum5
n=2,408
n=14
PS stratum6
n=1,416
n=15
PS stratum7
n=689
n=14
PS stratum8
n=464
n=15
PS stratum9
n=462
n=14
PS stratum10

Appendix Figure 4.2 – Box plot of propensity score distribution for TKR(0) and PKR (1) in each stratum of the safety cohort based on the PSSwhole (a) and PSSexp (b) method



a: PSS_{whole}



b: PSS_{exp}

Tables

Appendix Table 4.1 – Coefficient table for the 1st Imputed dataset PS logistic regression model.

Effectiveness cohort

	Odds Ratio	95% Conf.	Interval	p-val
patientgender_n	1.04	(0.73 -	1.48)	0.83
ageatprimary	0.95	(0.93 -	0.97)	<0.01
primarybmi	0.98	(0.95 -	1.01)	0.21
_Inew_mruru_2	1.39	(0.97 -	2.00)	0.07
_Inew_imd_2	1.07	(0.70 -	1.66)	0.75
_Inew_imd_3	0.74	(0.43 -	1.29)	0.29
_Inew_imd_4	1.00	(0.65 -	1.54)	0.98
_Imq1_gener_1	1.69	(1.12 -	2.57)	0.01
_Imq1_gener_2	2.57	(1.67 -	3.97)	<0.01
mq1_eq5d_health_scale	1.00	(0.99 -	1.01)	0.92
mkr_q1_score	1.04	(1.02 -	1.07)	<0.01
primaryasa_enc	1.36	(0.43 -	4.35)	0.60
_Inew_charl_1	1.63	(1.07 -	2.49)	0.02
_Inew_charl_2	1.48	(0.85 -	2.59)	0.17
_Inew_charl_3	1.39	(0.75 -	2.58)	0.30
gastrointestinaldiseases_3y	0.93	(0.63 -	1.37)	0.71
otherjointproblems_3y_com	1.56	(0.87 -	2.81)	0.14
mentalhealth_3y_com	0.83	(0.47 -	1.46)	0.52
respiratorydiseases_3y	0.88	(0.58 -	1.34)	0.56
cardiovasculardiseases_3y	0.76	(0.50 -	1.14)	0.19
thyroidproblems_3y	0.53	(0.26 -	1.10)	0.09
foothipspinalpain_3y	0.76	(0.27 -	2.10)	0.60
coxarthrosis_3y	0.92	(0.36 -	2.36)	0.86
neurologicaldisorders_3y	0.96	(0.58 -	1.59)	0.86
otherarthrosis_3y	0.96	(0.45 -	2.02)	0.91
polyarthrosis_3y	0.31	(0.11 -	0.92)	0.04
spondylosis_3y	0.30	(0.07 -	1.27)	0.10
_cons	0.07	(<0.01 -	4.46)	0.21

Safety cohort

	Odds Ratio	95% Conf.	Interval	p-val
patientgender_n	1.33	(1.21 -	1.46)	<0.01
ageatprimary	0.94	(0.93 -	0.94)	<0.01
primarybmi	0.98	(0.97 -	0.99)	<0.01
mrururb_ind_n	1.00			
2	0.98	(0.85 -	1.12)	0.74
3	1.27	(1.11 -	1.47)	<0.01
4	1.45	(1.17 -	1.78)	<0.01
mimd04_decile_n	1.00			
Less deprived 10-20%	0.73	(0.61 -	0.86)	<0.01
Less deprived 20-30%	0.59	(0.50 -	0.71)	<0.01
Less deprived 30-40%	0.53	(0.44 -	0.63)	<0.01
Less deprived 40-50%	0.62	(0.53 -	0.74)	<0.01
More deprived 10-20%	0.44	(0.36 -	0.53)	<0.01
More deprived 20-30%	0.44	(0.36 -	0.54)	<0.01
More deprived 30-40%	0.60	(0.50 -	0.72)	<0.01
More deprived 40-50%	0.57	(0.47 -	0.68)	<0.01
Most deprived 10%	0.39	(0.31 -	0.48)	<0.01
mq1_general_health	1.00			
1	1.29	(1.15 -	1.45)	<0.01
2	1.87	(1.67 -	2.11)	<0.01
mq1_eq5d_health_scale	1.00	(1.00 -	1.00)	0.12
mkr_q1_score	1.04	(1.03 -	1.05)	<0.01
primaryasa_enc	0.60	(0.40 -	0.91)	0.02
charlsonindex_3y	1.00			
1	1.09	(0.98 -	1.22)	0.12
2	1.21	(1.05 -	1.39)	0.01
3	1.29	(1.08 -	1.55)	0.01
4	1.24	(1.01 -	1.53)	0.04
gastrointestinaldiseases_3y	1.00	(0.90 -	1.11)	1.00
otherjointproblems_3y_com	0.91	(0.77 -	1.08)	0.28
mentalhealth_3y_com	1.09	(0.96 -	1.24)	0.17
respiratorydiseases_3y	1.07	(0.96 -	1.20)	0.20
cardiovasculardiseases_3y	0.95	(0.85 -	1.06)	0.39
thyroidproblems_3y	1.01	(0.87 -	1.17)	0.92
foothipspinalpain_3y	1.08	(0.85 -	1.38)	0.53
coxarthrosis_3y	0.82	(0.61 -	1.10)	0.19
varusdeformity_3y	0.94	(0.67 -	1.34)	0.75
neurologicaldisorders_3y	1.08	(0.95 -	1.23)	0.22
otherarthrosis_3y	0.75	(0.59 -	0.95)	0.02
polyarthrosis_3y	0.70	(0.55 -	0.89)	<0.01
spondylosis_3y	0.92	(0.70 -	1.21)	0.55
_cons	14.32	(3.66 -	56.03)	<0.01

Appendix Table 4.2a – Baseline characteristics of study participants receiving PKR vs TKR in the safety cohort with ASA 1-2. Patients operated by surgeons who had performed at least >10, >30 and >50 surgeries of the same type in the previous year

	Full cohort 10 +				Full cohort 30 +				Full cohort 50+			
N(%) or mean (SD)	**TKR (n=248,785)**		**PKR (n=13,334)**		**TKR (n=195,898)**		**PKR (n=5,555)**		**TKR (n=139,396)**		**PKR (n=2,550)**	
Gender												
F	141124	57	6401	48	110807	57	2636	47	78641	56	1242	49
M	107661	43	6933	52	85091	43	2919	53	60755	44	1308	51
Rural Index												
1	185028	74	8984	67	144874	74	3513	63	102350	73	1550	61
2	29793	12	1810	14	23913	12	815	15	17349	12	379	15
3	23784	10	1790	13	18964	10	881	16	13711	10	443	17
4	10180	4	750	6	8147	4	346	6	5986	4	178	7
IMD												
Least deprived 10%	26808	11	1936	15	21504	11	823	15	15215	11	386	15
Less deprived 10-20%	28936	12	1908	14	23211	12	827	15	16656	12	366	14
Less deprived 20-30%	29178	12	1711	13	23177	12	775	14	16663	12	383	15
Less deprived 30-40%	29751	12	1572	12	23744	12	674	12	17103	12	292	11
Less deprived 40-50%	28532	11	1605	12	22448	11	716	13	16100	12	347	14
More deprived 10-20%	18313	7	678	5	14045	7	212	4	9735	7	91	4
More deprived 20-30%	21123	8	873	7	16385	8	323	6	11642	8	135	5
More deprived 30-40%	23669	10	1169	9	18462	9	445	8	13084	9	200	8
More deprived 40-50%	26781	11	1371	10	20990	11	582	10	14945	11	287	11
Most deprived 10%	15694	6	511	4	11932	6	178	3	8253	6	63	2
ASA												
P1 - Fit and healthy	27829	11	2707	20	22227	11	1104	20	15725	11	539	21

	Full cohort 10 +				Full cohort 30 +				Full cohort 50+			
N(%) or mean (SD)	**TKR (n=248,785)**		**PKR (n=13,334)**		**TKR (n=195,898)**		**PKR (n=5,555)**		**TKR (n=139,396)**		**PKR (n=2,550)**	
P2 - Mild disease not incapacitating	220956	89	10627	80	173671	89	4451	80	123671	89	2011	79
Charlson Comorbidity												
0	170990	69	9694	73	134945	69	3980	72	95884	69	1862	73
1	53212	21	2645	20	41686	21	1124	20	29628	21	481	19
2	16101	6	652	5	12654	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*	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*	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 pre-operative OKS *	19.3	6.8	21.4	6.2	19.4	6.8	21.6	6.2	19.4	6.9	21.7	6.1
PROMS EQ5D*	69.3	19.4	69.9	19.2	69.4	19.4	70.3	19.1	69.5	19.4	70.3	19.2
PROMS General Health												
0	147872	59	4099	31	118240	60	1743	31	85617	61	817	32
1	40068	16	4324	32	31166	16	1846	33	21766	16	860	34
2	27233	11	2819	21	21043	11	1121	20	14603	10	502	20
3	23188	9	1357	10	17617	9	559	10	12116	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
Gastrointestinal Disease	46976	19	2346	18	36986	19	1025	18	26435	19	449	18
Osteoarthritis and Other Joint Problems	45193	18	1655	12	35589	18	677	12	25192	18	296	12
Mental Health	23773	10	1487	11	19113	10	630	11	13867	10	286	11
Respiratory Diseases	34160	14	1793	13	26882	14	750	14	19222	14	306	12
Cardiovascular Diseases	142322	57	6275	47	111485	57	2604	47	79174	57	1167	46
Thyroid Problems	18786	8	794	6	14744	8	322	6	10479	8	151	6
Foot, hip, spinal pain	2831	1	127	1	2220	1	50	1	1574	1	28	1

	Full cohort 10 +				Full cohort 30 +				Full cohort 50+			
N(%) or mean (SD)	**TKR (n=248,785)**		**PKR (n=13,334)**		**TKR (n=195,898)**		**PKR (n=5,555)**		**TKR (n=139,396)**		**PKR (n=2,550)**	
Coxarthrosis	8158	3	245	2	6454	3	106	2	4518	3	42	2
Neurological Disorders	14848	6	796	6	11684	6	335	6	8409	6	144	6
Other Arthrosis	11518	5	449	3	9029	5	204	4	6399	5	94	4
Polyarthrosis	14466	6	371	3	11306	6	140	3	7912	6	64	3
Spondylosis	6677	3	215	2	5344	3	76	1	3812	3	25	1
Note: * mean (SD) is presented.												

Appendix Table 4.2b - Baseline characteristics of study participants receiving PKR vs TKR in the effectiveness cohort with ASA 1-2. Patients operated by surgeons who had performed at least >10 of the same type in the previous year

	OKS cohort>10			
N(%) or mean (SD)	**TKR (n=114,871)**		**PKR (n=602)**	
Gender				
F	64468	56	287	48
M	50403	44	315	52
Rural Index				
1	83810	73	396	66
2	14446	13	97	16
3	11587	10	79	13
4	5028	4	30	5
IMD				
Least deprived 10%	12981	11	75	12
Less deprived 10-20%	13992	12	72	12
Less deprived 20-30%	14159	12	82	14
Less deprived 30-40%	14140	12	65	11
Less deprived 40-50%	13371	12	94	16
More deprived 10-20%	7731	7	42	7
More deprived 20-30%	9178	8	39	6
More deprived 30-40%	10551	9	59	10
More deprived 40-50%	12333	11	52	9
Most deprived 10%	6435	6	22	4
ASA				
P1 - Fit and healthy	12748	11	118	20
P2 - Mild disease not incapacitating	102123	89	484	80
Charlson Comorbidity				
0	79157	69	447	74
1	24269	21	121	20
2	7582	7	23	4
3	2579	2	8	1
4	1284	1	3	0
Age*	70.3	8.6	65.6	9.3
BMI*	30.4	5.0	29.5	4.6

	OKS cohort>10			
N(%) or mean (SD)	**TKR (n=114,871)**		**PKR (n=602)**	
PROMS pre-operative OKS *	19.7	7.6	22.1	7.6
PROMS EQ5D*	70.0	19.2	71.3	18.8
PROMS General Health				
0	81617	71	306	51
1	1311	1	14	2
2	9395	8	100	17
3	15652	14	128	21
4	6148	5	48	8
5	748	1	6	1
Gastrointestinal Disease	22766	20	93	15
Osteoarthritis and Other Joint Problems	21434	19	71	12
Mental Health	10528	9	46	8
Respiratory Diseases	15503	13	79	13
Cardiovascular Diseases	66546	58	272	45
Thyroid Problems	8868	8	39	6
Foot, hip, 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
Note: * mean (SD) is presented.				

Appendix Table 4.2c. Baseline characteristics of study participants receiving PKR vs TKR in the safety cohort with ASA 3-4. Patients operated by surgeons who had performed at least >10, >30 and >50 surgeries of the same type in the previous year

	Full cohort 10 +				Full cohort 30 +				Full cohort 50+			
N(%) or mean (SD)	**TKR (n=51,118)**		**PKR (n=1,449)**		**TKR (n=38,321)**		**PKR (n=610)**		**TKR (n=25,944)**		**PKR (n=242)**	
Gender												
F	28470	56	627	43	21296	56	280	46	14310	28470	56	627
M	22648	44	822	57	17025	44	330	54	11634	22648	44	822
Rural Index												
1	39185	77	1026	71	29239	76	404	66	19669	39185	77	1026
2	6069	12	172	12	4661	12	78	13	3192	6069	12	172
3	4319	8	181	12	3266	9	92	15	2259	4319	8	181
4	1545	3	70	5	1155	3	36	6	824	1545	3	70
IMD												
Least deprived 10%	4275	8	220	15	3287	9	101	17	2277	4275	8	220
Less deprived 10-20%	5146	10	194	13	3925	10	99	16	2708	5146	10	194
Less deprived 20-30%	5606	11	145	10	4246	11	49	8	2938	5606	11	145
Less deprived 30-40%	5613	11	136	9	4237	11	56	9	2951	5613	11	136
Less deprived 40-50%	5705	11	178	12	4252	11	74	12	2890	5705	11	178
More deprived 10-20%	4752	9	97	7	3534	9	41	7	2300	4752	9	97
More deprived 20-30%	4883	10	112	8	3624	9	49	8	2436	4883	10	112
More deprived 30-40%	5137	10	147	10	3782	10	49	8	2517	5137	10	147
More deprived 40-50%	5505	11	151	10	4118	11	64	10	2786	5505	11	151
Most deprived 10%	4496	9	69	5	3316	9	28	5	2141	4496	9	69
ASA												
P3 - Incapacitating systemic disease	50171	98	1432	99	37637	98	608	100	25508	98	242	100
P4 - Life threatening disease	947	2	17	1	684	2	2	0	436	2	0	0
Charlson Comorbidity												

	Full cohort 10 +				Full cohort 30 +				Full cohort 50+			
N(%) or mean (SD)	**TKR (n=51,118)**		**PKR (n=1,449)**		**TKR (n=38,321)**		**PKR (n=610)**		**TKR (n=25,944)**		**PKR (n=242)**	
0	20126	39	538	37	14972	39	235	39	10082	39	100	41
1	16304	32	480	33	12207	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*	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*	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 pre-operative OKS *	16.4	7.6	19.2	7.8	16.4	7.6	19.0	7.9	16.5	7.7	18.8	8.2
PROMS EQ5D*	61.8	20.5	64.2	20.4	61.8	20.5	64.5	20.0	61.7	20.5	64.0	19.9
PROMS General Health												
0	36509	71	909	63	27727	72	367	60	19045	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
Gastrointestinal Disease	14360	28	360	25	10786	28	159	26	7371	28	70	29
Osteoarthritis and Other Joint Problems	13378	26	269	19	10042	26	112	18	6891	27	42	17
Mental Health	6705	13	205	14	5151	13	99	16	3551	14	42	17
Respiratory Diseases	13379	26	383	26	10016	26	146	24	6730	26	52	21
Cardiovascular Diseases	41694	82	1142	79	31223	81	476	78	21048	81	190	79
Thyroid Problems	5597	11	140	10	4179	11	61	10	2853	11	19	8
Foot, hip, 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
Note: * mean (SD) is presented.												

Appendix Table 4.3. Number (%) of participants having a complication in the 90 days after index surgery operated on by surgeons who had performed 10+, 30+ and 50+ surgeries of the same type as the index surgery in the year before the index surgery

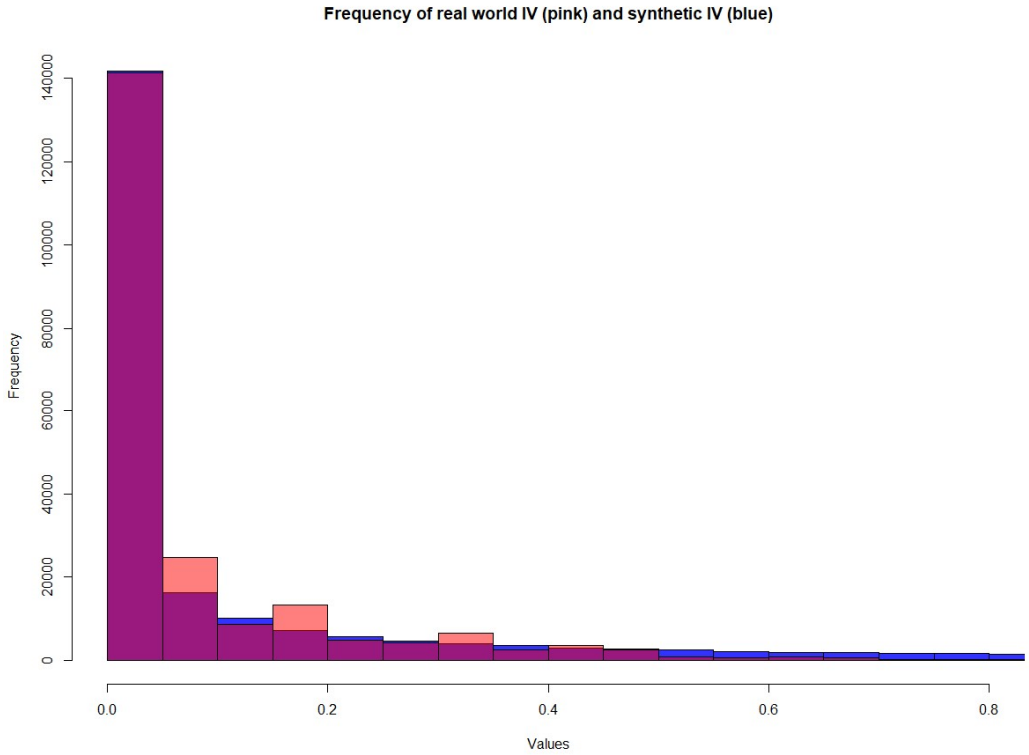
	90-day MI (%)		90-day VTE (%)		90-day PJI (%)	
	PKR	**TKR**	**PKR**	**TKR**	**PKR**	**TKR**
ASA 1-2						
Main	26 (0.12%)	572 (0.21%)	62 (0.29%)	1,750 (0.64%)	17 (0.08%)	338 (0.12%)
10+ surgeries	20 (0.15%)	517 (0.21%)	33 (0.25%)	1,587 (9.64%)	13 (0.10%)	295 (0.12%)
30+ surgeries	8 (0.14%)	401 (0.20%)	11 (0.20%)	1,244 (0.64%)	<5	225 (0.11%)
50+ surgeries	<5	227 (0.20%)	<5	861 (0.62%)	<5	156 (0.11%)
ASA 3-4						
Main	8 (0.35%)	282 (0.49%)	6 (0.27%)	460 (0.80%)	<5	120 (0.21%)
10+ surgeries	6 (0.41%)	282 (0.49%)	5 (0.35%)	411 (0.80%)	<5	109 (0.21%)
TKR: total knee replacement; PKR: partial knee replacement.						

5. Methods Research

Appendix Material

Figures

Appendix Figure 5.1 – Distribution of the real-world IV (pink) and the synthetic IV (blue). The overlap is shown in purple.



Tables

Appendix Table 5.1 – Plasmode Simulation complete results

		Continuous Variable								Binary Variable							
		Synthetic IV				UTMOST IV				Synthetic IV				UTMOST IV			
IV_str_real	stat	est	mcse	lower	upper	est	mcse	lower	upper	est	mcse	lower	upper	est	mcse	lower	upper
1	nsim	967.00				1000.00				972.00	NA	NA	NA	1000.00	NA	NA	NA
1	thetamean	-14.57				49.88				-2.31	NA	NA	NA	3.15	NA	NA	NA
1	thetamedian	-14.68				48.82				-0.30	NA	NA	NA	3.14	NA	NA	NA
1	se2mean	1.73E+07				403.06				1898040	NA	NA	NA	0.60	NA	NA	NA
1	se2median	27131.70				318.37				4281.78	NA	NA	NA	0.60	NA	NA	NA
1	bias	-16.57	10.31	(-36.77	, 3.63)	47.88	0.58	(46.74	, 49.03)	-3.00	3.38	(-9.64	, 3.63)	2.46	0.02	(2.41	, 2.50)
1	empse	320.56	7.29	(306.26	, 334.85)	18.41	0.41	(17.61	, 19.22)	105.50	2.39	(100.81	, 110.20)	0.74	0.02	(0.71	, 0.77)
1	mse	102924.58	10817.84	(81,721.99	, 124,127)	2631.60	62.76	(2,508.60	, 2,754.60)	11128.78	1070.89	(9,029.88	, 13,227.7)	6.59	0.12	(6.36	, 6.82)
1	relprec	0.00	0.00	(-0.00	, 0.00)	30204.12	1915.56	(26,449.69	, 33,958.5)	0.00	0.00	(0.00	, 0.00)	2022003.6	127812.45	(1,771,496	, 2,272,511)
1	modelse	4154.59	458.14	(3,256.66	, 5,052.53)	20.08	0.23	(19.62	, 20.53)	1377.69	177.42	(1,029.96	, 1,725.42)	0.77	0.00	(0.77	, 0.77)
1	relerror	1196.06	145.93	(910.04	, 1,482.07)	9.03	2.74	(3.65	, 14.40)	1205.81	170.75	(871.15	, 1,540.48)	3.98	2.33	(-0.58	, 8.54)
1	cover	1.03				0.21	0.01	(0.19	, 0.24)	1.03	NA	NA	NA	0.10	0.01	(0.08	, 0.11)
1	becover	1.03				0.95	0.01	(0.94	, 0.96)	1.03	NA	NA	NA	0.96	0.01	(0.95	, 0.97)
1	power	0.00	0.00	(0.00	, 0.00)	0.83	0.01	(0.81	, 0.86)	0.00	0.00	(0.00	, 0.00)	0.99	0.00	(0.98	, 0.99)
1.1	nsim	981.00				1000.00				1000.00	NA	NA	NA	1000.00	NA	NA	NA
1.1	thetamean	-29.33				40.57				1.24	NA	NA	NA	2.98	NA	NA	NA
1.1	thetamedian	-23.22				39.74				1.30	NA	NA	NA	2.98	NA	NA	NA
1.1	se2mean	242309.97				226.34				67.97	NA	NA	NA	0.49	NA	NA	NA
1.1	se2median	3134.60				202.63				53.13	NA	NA	NA	0.49	NA	NA	NA
1.1	bias	-31.33	3.47	(-38.13	, -24.53)	38.57	0.47	(37.65	, 39.48)	0.55	0.24	(0.08	, 1.02)	2.29	0.02	(2.24	, 2.33)
1.1	empse	108.67	2.45	(103.86	, 113.48)	14.73	0.33	(14.08	, 15.38)	7.60	0.17	(7.27	, 7.94)	0.69	0.02	(0.66	, 0.72)
1.1	mse	12778.49	1337.26	(10,157.52	#########	1704.11	38.59	(1,628.48	, 1,779.74)	58.05	3.01	(52.15	, 63.95)	5.71	0.10	(5.51	, 5.91)
1.1	relprec	0.00	0.00	(0.00	, 0.00)	5342.20	344.36	(4,667.26	, 6,017.14)	0.00	0.00	(0.00	, 0.00)	11994.36	765.26	(10,494.48	, 13,494.2)
1.1	modelse	492.25	58.13	(378.31	, 606.19)	15.04	0.11	(14.83	, 15.26)	8.24	0.13	(7.99	, 8.50)	0.70	0.00	(0.70	, 0.70)
1.1	relerror	352.99	54.47	(246.24	, 459.74)	2.13	2.40	(-2.57	, 6.84)	8.43	2.97	(2.62	, 14.24)	0.89	2.26	(-3.54	, 5.31)
1.1	cover	1.02				0.19	0.01	(0.17	, 0.21)	0.97	0.01	(0.96	, 0.98)	0.10	0.01	(0.08	, 0.11)
1.1	becover	1.00	0.00	(1.00	, 1.00)	0.95	0.01	(0.93	, 0.96)	0.97	0.01	(0.96	, 0.98)	0.94	0.01	(0.93	, 0.96)
1.1	power	0.00	0.00	(-0.00	, 0.00)	0.85	0.01	(0.83	, 0.87)	0.03	0.01	(0.02	, 0.04)	0.99	0.00	(0.98	, 0.99)
2	nsim	1000.00				1000.00				1000.00	NA	NA	NA	1000.00	NA	NA	NA
2	thetamean	-3.83				19.37				1.20	NA	NA	NA	2.31	NA	NA	NA
2	thetamedian	-3.40				19.05				1.21	NA	NA	NA	2.30	NA	NA	NA
2	se2mean	93.24				37.40				1.04	NA	NA	NA	0.20	NA	NA	NA
2	se2median	89.37				36.71				1.04	NA	NA	NA	0.20	NA	NA	NA
2	bias	-5.83	0.30	(-6.42	, -5.25)	17.37	0.20	(16.98	, 17.76)	0.51	0.03	(0.45	, 0.57)	1.62	0.01	(1.59	, 1.65)
2	empse	9.44	0.21	(9.03	, 9.86)	6.28	0.14	(6.01	, 6.56)	1.03	0.02	(0.98	, 1.07)	0.45	0.01	(0.43	, 0.47)
2	mse	123.06	5.80	(111.70	, 134.43)	341.01	7.44	(326.42	, 355.60)	1.32	0.06	(1.20	, 1.43)	2.82	0.05	(2.73	, 2.92)
2	relprec	0.00	0.00	(-0.00	, 0.00)	125.94	14.29	(97.93	, 153.96)	0.00	0.00	(0.00	, 0.00)	414.19	32.54	(350.42	, 477.96)
2	modelse	9.66	0.04	(9.58	, 9.73)	6.12	0.02	(6.08	, 6.15)	1.02	0.00	(1.02	, 1.02)	0.44	0.00	(0.44	, 0.44)
2	relerror	2.28	2.32	(-2.28	, 6.83)	-2.64	2.19	(-6.94	, 1.65)	-0.83	2.22	(-5.18	, 3.52)	-2.14	2.19	(-6.43	, 2.15)
2	cover	0.91	0.01	(0.89	, 0.93)	0.20	0.01	(0.17	, 0.22)	0.92	0.01	(0.90	, 0.94)	0.05	0.01	(0.04	, 0.07)
2	becover	0.95	0.01	(0.94	, 0.97)	0.95	0.01	(0.94	, 0.97)	0.95	0.01	(0.93	, 0.96)	0.94	0.01	(0.93	, 0.96)
2	power	0.06	0.01	(0.04	, 0.07)	0.89	0.01	(0.87	, 0.90)	0.22	0.01	(0.19	, 0.24)	1.00	0.00	(1.00	, 1.00)

5	nsim	1000.00			1000.00			1000.00	NA	NA	NA	1000.00	NA	NA	NA
5	thetamean	-0.97			12.29			1.15	NA	NA	NA	1.94	NA	NA	NA
5	thetamedian	-1.15			12.19			1.13	NA	NA	NA	1.94	NA	NA	NA
5	se2mean	20.92			12.72			0.22	NA	NA	NA	0.09	NA	NA	NA
5	se2median	20.85			12.69			0.22	NA	NA	NA	0.09	NA	NA	NA
5	bias	-2.97	0.14	(-3.25 , -2.69)	10.29	0.11	(10.07 , 10.51)	0.46	0.01	(0.43 , 0.49)		1.25	0.01	(1.23 , 1.27)	
5	empse	4.52	0.10	(4.32 , 4.72)	3.62	0.08	(3.46 , 3.78)	0.46	0.01	(0.44 , 0.48)		0.31	0.01	(0.30 , 0.33)	
5	mse	29.20	1.22	(26.81 , 31.58)	118.95	2.48	(114.10 , 123.81)	0.42	0.02	(0.39 , 0.45)		1.66	0.02	(1.61 , 1.71)	
5	relprec	0.00	0.00	(0.00 , 0.00)	56.06	9.85	(36.75 , 75.37)	0.00	0.00	(0.00 , 0.00)		113.66	13.52	(87.17 , 140.16)	
5	modelse	4.57	0.01	(4.56 , 4.59)	3.57	0.01	(3.56 , 3.58)	0.47	0.00	(0.46 , 0.47)		0.30	0.00	(0.30 , 0.30)	
5	relerror	1.20	2.27	(-3.25 , 5.65)	-1.41	2.21	(-5.74 , 2.92)	2.20	2.29	(-2.28 , 6.68)		-2.23	2.19	(-6.52 , 2.06)	
5	cover	0.90	0.01	(0.88 , 0.92)	0.17	0.01	(0.15 , 0.20)	0.84	0.01	(0.81 , 0.86)		0.02	0.00	(0.01 , 0.03)	
5	becover	0.95	0.01	(0.94 , 0.97)	0.95	0.01	(0.94 , 0.96)	0.96	0.01	(0.95 , 0.97)		0.94	0.01	(0.93 , 0.96)	
5	power	0.05	0.01	(0.04 , 0.06)	0.94	0.01	(0.92 , 0.95)	0.72	0.01	(0.69 , 0.75)		1.00	0.00	(1.00 , 1.00)	
10	nsim	1000.00			1000.00			1000.00	NA	NA	NA	1000.00	NA	NA	NA
10	thetamean	-0.43			10.44			1.15	NA	NA	NA	1.81	NA	NA	NA
10	thetamedian	-0.34			10.47			1.16	NA	NA	NA	1.82	NA	NA	NA
10	se2mean	12.93			8.57			0.12	NA	NA	NA	0.07	NA	NA	NA
10	se2median	12.90			8.56			0.12	NA	NA	NA	0.07	NA	NA	NA
10	bias	-2.43	0.11	(-2.65 , -2.21)	8.44	0.09	(8.25 , 8.62)	0.46	0.01	(0.44 , 0.48)		1.12	0.01	(1.10 , 1.13)	
10	empse	3.57	0.08	(3.41 , 3.73)	2.99	0.07	(2.86 , 3.12)	0.35	0.01	(0.33 , 0.36)		0.25	0.01	(0.24 , 0.26)	
10	mse	18.63	0.82	(17.02 , 20.24)	80.09	1.63	(76.89 , 83.29)	0.33	0.01	(0.31 , 0.35)		1.31	0.02	(1.28 , 1.35)	
10	relprec	0.00	0.00	(0.00 , 0.00)	42.57	9.02	(24.89 , 60.24)	0.00	0.00	(0.00 , 0.00)		90.42	12.04	(66.83 , 114.01)	
10	modelse	3.60	0.00	(3.59 , 3.60)	2.93	0.00	(2.92 , 2.93)	0.35	0.00	(0.35 , 0.35)		0.26	0.00	(0.26 , 0.26)	
10	relerror	0.71	2.26	(-3.71 , 5.14)	-2.07	2.19	(-6.37 , 2.23)	0.44	2.25	(-3.96 , 4.84)		3.24	2.31	(-1.28 , 7.77)	
10	cover	0.91	0.01	(0.89 , 0.92)	0.19	0.01	(0.16 , 0.21)	0.75	0.01	(0.72 , 0.77)		0.01	0.00	(0.00 , 0.01)	
10	becover	0.95	0.01	(0.94 , 0.96)	0.95	0.01	(0.94 , 0.96)	0.94	0.01	(0.93 , 0.96)		0.96	0.01	(0.95 , 0.97)	
10	power	0.05	0.01	(0.04 , 0.06)	0.95	0.01	(0.93 , 0.96)	0.91	0.01	(0.89 , 0.93)		1.00	0.00	(1.00 , 1.00)	
25	nsim	1000.00			1000.00			1000.00	NA	NA	NA	1000.00	NA	NA	NA
25	thetamean	-0.13			9.26			1.17	NA	NA	NA	1.75	NA	NA	NA
25	thetamedian	-0.25			9.29			1.17	NA	NA	NA	1.75	NA	NA	NA
25	se2mean	9.48			6.57			0.08	NA	NA	NA	0.05	NA	NA	NA
25	se2median	9.43			6.56			0.08	NA	NA	NA	0.05	NA	NA	NA
25	bias	-2.13	0.09	(-2.32 , -1.95)	7.26	0.08	(7.11 , 7.42)	0.48	0.01	(0.46 , 0.49)		1.05	0.01	(1.04 , 1.07)	
25	empse	2.98	0.07	(2.85 , 3.11)	2.49	0.06	(2.38 , 2.60)	0.29	0.01	(0.28 , 0.30)		0.24	0.01	(0.23 , 0.25)	
25	mse	13.44	0.54	(12.39 , 14.49)	58.92	1.18	(56.60 , 61.23)	0.31	0.01	(0.29 , 0.33)		1.17	0.02	(1.14 , 1.20)	
25	relprec	0.00	0.00	(0.00 , 0.00)	43.59	9.09	(25.78 , 61.40)	0.00	0.00	(0.00 , 0.00)		48.07	9.37	(29.71 , 66.43)	
25	modelse	3.08	0.00	(3.07 , 3.09)	2.56	0.00	(2.56 , 2.57)	0.28	0.00	(0.28 , 0.28)		0.23	0.00	(0.23 , 0.23)	
25	relerror	3.21	2.31	(-1.32 , 7.74)	2.99	2.31	(-1.53 , 7.51)	-2.17	2.19	(-6.46 , 2.12)		-2.42	2.18	(-6.70 , 1.86)	
25	cover	0.90	0.01	(0.88 , 0.92)	0.19	0.01	(0.16 , 0.21)	0.61	0.02	(0.58 , 0.64)		0.00	0.00	(0.00 , 0.01)	
25	becover	0.97	0.01	(0.96 , 0.98)	0.96	0.01	(0.95 , 0.97)	0.94	0.01	(0.92 , 0.95)		0.95	0.01	(0.94 , 0.97)	
25	power	0.03	0.01	(0.02 , 0.04)	0.95	0.01	(0.94 , 0.97)	0.98	0.00	(0.97 , 0.99)		1.00	0.00	(1.00 , 1.00)	
50	nsim	1000.00			1000.00			1000.00	NA	NA	NA	1000.00	NA	NA	NA
50	thetamean	0.00			8.92			1.17	NA	NA	NA	1.73	NA	NA	NA
50	thetamedian	0.00			8.93			1.17	NA	NA	NA	1.73	NA	NA	NA
50	se2mean	8.47			6.00			0.07	NA	NA	NA	0.05	NA	NA	NA
50	se2median	8.43			5.98			0.07	NA	NA	NA	0.05	NA	NA	NA
50	bias	-2.00	0.09	(-2.17 , -1.82)	6.92	0.08	(6.77 , 7.08)	0.48	0.01	(0.46 , 0.49)		1.04	0.01	(1.03 , 1.06)	
50	empse	2.82	0.06	(2.70 , 2.95)	2.45	0.05	(2.34 , 2.55)	0.27	0.01	(0.25 , 0.28)		0.23	0.01	(0.22 , 0.24)	
50	mse	11.94	0.48	(11.00 , 12.89)	53.93	1.10	(51.78 , 56.08)	0.30	0.01	(0.28 , 0.31)		1.14	0.02	(1.11 , 1.17)	
50	relprec	0.00	0.00	(0.00 , 0.00)	33.00	8.41	(16.52 , 49.49)	0.00	0.00	(0.00 , 0.00)		35.81	8.59	(18.97 , 52.66)	
50	modelse	2.91	0.00	(2.90 , 2.92)	2.45	0.00	(2.44 , 2.45)	0.26	0.00	(0.26 , 0.26)		0.22	0.00	(0.22 , 0.22)	
50	relerror	3.11	2.31	(-1.41 , 7.64)	0.13	2.24	(-4.27 , 4.52)	-2.29	2.19	(-6.57 , 1.99)		-2.69	2.18	(-6.95 , 1.58)	
50	cover	0.90	0.01	(0.88 , 0.91)	0.19	0.01	(0.17 , 0.21)	0.55	0.02	(0.52 , 0.58)		0.01	0.00	(0.00 , 0.01)	

50	becover	0.96	0.01	(0.95 , 0.97)	0.95	0.01	(0.94 , 0.97)	0.96	0.01	(0.94 , 0.97)	0.95	0.01	(0.93 , 0.96)
50	power	0.04	0.01	(0.03 , 0.05)	0.96	0.01	(0.94 , 0.97)	1.00	0.00	(0.99 , 1.00)	1.00	0.00	(1.00 , 1.00)
100	nsim	1000.00			1000.00			1000.00	NA	NA NA	1000.00	NA	NA NA
100	thetamean	-0.01			8.79			1.16	NA	NA NA	1.72	NA	NA NA
100	thetamedian	0.23			8.82			1.16	NA	NA NA	1.72	NA	NA NA
100	se2mean	8.00			5.68			0.06	NA	NA NA	0.05	NA	NA NA
100	se2median	7.98			5.66			0.06	NA	NA NA	0.05	NA	NA NA
100	bias	-2.01	0.09	(-2.19 , -1.83)	6.79	0.07	(6.65 , 6.93)	0.47	0.01	(0.45 , 0.49)	1.02	0.01	(1.01 , 1.04)
100	empse	2.94	0.07	(2.81 , 3.07)	2.30	0.05	(2.20 , 2.40)	0.25	0.01	(0.24 , 0.26)	0.21	0.00	(0.20 , 0.22)
100	mse	12.67	0.55	(11.59 , 13.75)	51.38	1.01	(49.40 , 53.35)	0.28	0.01	(0.27 , 0.30)	1.10	0.01	(1.07 , 1.12)
100	relprec	0.00	0.00	(-0.00 , 0.00)	63.06	10.30	(42.87 , 83.25)	0.00	0.00	(0.00 , 0.00)	40.52	8.89	(23.10 , 57.95)
100	modelse	2.83	0.00	(2.82 , 2.83)	2.38	0.00	(2.38 , 2.39)	0.25	0.00	(0.25 , 0.25)	0.22	0.00	(0.22 , 0.22)
100	relerror	-3.73	2.16	(-7.96 , 0.49)	3.52	2.32	(-1.02 , 8.07)	-2.53	2.18	(-6.80 , 1.74)	1.25	2.26	(-3.19 , 5.69)
100	cover	0.86	0.01	(0.84 , 0.88)	0.17	0.01	(0.15 , 0.20)	0.53	0.02	(0.50 , 0.56)	0.00	0.00	(-0.00 , 0.01)
100	becover	0.95	0.01	(0.93 , 0.96)	0.96	0.01	(0.94 , 0.97)	0.94	0.01	(0.92 , 0.95)	0.95	0.01	(0.94 , 0.97)
100	power	0.05	0.01	(0.04 , 0.07)	0.97	0.01	(0.96 , 0.98)	1.00	0.00	(0.99 , 1.00)	1.00	0.00	(1.00 , 1.00)

Appendix Table 5.2 – Propensity Score Betas

	OKS_tot				OKS_sens			
	Logistic	**ME**	**Logistic**	**ME**	**Logistic**	**ME**	**Logistic**	**ME**
patientgender_n	1.16	1.22	1.19	1.22	1.15	1.29	1.22	1.29
ageatprimary	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.93
primarybmi	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
_Imrururb_i_2	1.13	1.06	1.08	1.06	1.38	1.22	1.27	1.22
_Imrururb_i_3	1.15	0.97	1.06	0.97	1.35	1.00	1.16	1.00
_Imrururb_i_4	0.93	0.80	0.86	0.82	1.11	0.91	0.93	0.94
_Imimd04_de_2	0.85	0.88	0.90	0.88	0.88	0.95	0.98	0.96
_Imimd04_de_3	0.87	0.93	0.96	0.94	0.99	1.16	1.19	1.17
_Imimd04_de_4	0.86	0.94	0.97	0.97	0.80	0.92	1.01	0.98
_Imimd04_de_5	1.09	1.22	1.19	1.23	1.23	1.51	1.47	1.53
_Imimd04_de_6	1.15	1.25	1.32	1.29	1.04	1.29	1.37	1.37
_Imimd04_de_7	0.80	0.90	0.92	0.92	0.80	1.02	1.07	1.08
_Imimd04_de_8	1.01	1.07	1.14	1.10	1.01	1.07	1.26	1.14
_Imimd04_de_9	0.75	0.85	0.83	0.86	0.75	0.97	0.90	1.00
_Imimd04_de_10	0.71	0.85	0.85	0.87	0.66	1.04	0.97	1.08
_Imq1_gener_1	2.52	2.71	2.76	2.83	2.20	2.11	2.89	2.65
_Imq1_gener_2	2.23	2.16	2.46	2.27	2.51	2.27	3.37	2.90
_Imq1_gener_3	2.26	2.34	2.50	2.46	2.18	2.24	2.91	2.88
_Imq1_gener_4	2.29	2.45	2.56	2.52	2.38	2.57	3.28	3.11
_Imq1_gener_5	2.85	2.74	3.05	2.74	2.60	2.41	3.36	2.69
mq1_eq5d_health_scale	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
mkr_q1_score	1.03	1.03	1.03	1.03	1.04	1.03	1.03	1.03
primaryasa_enc	0.88	0.90	0.93	0.90	0.85	0.92	0.95	0.91
_Icharlsoni_1	0.95	0.97	0.95	0.96	1.05	1.07	1.05	1.06
_Icharlsoni_2	0.71	0.67	0.73	0.67	0.79	0.68	0.84	0.68
_Icharlsoni_3	0.80	0.85	0.79	0.85	0.97	1.08	0.97	1.06
_Icharlsoni_4	0.50	0.43	0.51	0.44	0.73	0.64	0.80	0.67
gastrointestinaldiseases_3y	0.88	0.86	0.86	0.87	0.95	0.96	0.91	0.99
otherjointproblems_3y_com	1.01	1.03	0.98	1.02	1.20	1.18	1.13	1.16
mentalhealth_3y_com	0.95	1.01	0.93	0.99	0.88	0.91	0.83	0.88
respiratorydiseases_3y	1.11	1.13	1.09	1.13	1.11	1.11	1.07	1.13
cardiovasculardiseases_3y	0.84	0.84	0.85	0.85	0.89	0.88	0.92	0.89
thyroidproblems_3y	1.15	1.13	1.17	1.14	1.10	1.13	1.15	1.15
foothipspinalpain_3y	1.43	1.43	1.51	1.48	1.26	1.16	1.43	1.26
coxarthrosis_3y	0.77	0.74	0.72	0.73	0.75	0.69	0.66	0.67
neurologicaldisorders_3y	1.11	1.12	1.13	1.13	1.27	1.32	1.31	1.34
otherarthrosis_3y	0.99	0.92	0.96	0.91	0.61	0.56	0.54	0.54
polyarthrosis_3y	0.48	0.52	0.52	0.53	0.26	0.32	0.28	0.33
spondylosis_3y	0.70	0.78	0.75	0.78	0.52	0.67	0.58	0.67
lead_ukr_yearly_surgeonoutput	-		1.04	1.03			1.06	1.05
lead_tkr_yearly_surgeonoutput	-		0.99	0.99			0.99	0.99

	Rev_tot				Rev_sens			
	Logistic	ME	Logistic	ME	Logistic	ME	Logistic	ME
patientgender_n	1.16	1.24	1.22	1.25	1.13	1.23	1.23	1.26
ageatprimary	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
primarybmi	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.97
_Imrururb_i_2	1.09	1.07	1.01	1.07	1.16	1.09	1.03	1.09
_Imrururb_i_3	1.30	1.07	1.08	1.07	1.42	1.07	1.07	1.08
_Imrururb_i_4	1.29	1.15	1.15	1.15	1.36	1.14	1.13	1.15
_Imimd04_de_2	0.89	0.99	0.93	0.99	0.89	1.01	0.94	1.00
_Imimd04_de_3	0.82	0.98	0.88	0.98	0.79	0.97	0.86	0.96
_Imimd04_de_4	0.74	0.95	0.85	0.95	0.70	0.93	0.85	0.93
_Imimd04_de_5	0.76	0.95	0.84	0.95	0.75	0.96	0.85	0.96
_Imimd04_de_6	0.58	0.82	0.69	0.82	0.51	0.79	0.67	0.80
_Imimd04_de_7	0.59	0.84	0.70	0.84	0.57	0.90	0.75	0.89
_Imimd04_de_8	0.71	0.91	0.81	0.91	0.67	0.89	0.82	0.89
_Imimd04_de_9	0.71	0.94	0.80	0.94	0.69	0.96	0.82	0.95
_Imimd04_de_10	0.51	0.80	0.63	0.80	0.45	0.84	0.63	0.84
_Imq1_gener_1	3.93	4.12	4.12	4.24	4.12	4.27	4.79	4.59
_Imq1_gener_2	3.74	4.01	4.02	4.19	3.93	4.16	4.79	4.68
_Imq1_gener_3	2.26	2.62	2.57	2.80	2.30	2.68	3.06	3.20
_Imq1_gener_4	2.48	2.88	2.85	3.10	2.37	2.79	3.21	3.39
_Imq1_gener_5	7.44	8.57	7.77	8.90	7.95	8.95	9.47	10.19
mq1_eq5d_health_scale	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
mkr_q1_score	1.05	1.04	1.04	1.04	1.04	1.04	1.04	1.04
primaryasa_enc	0.78	0.80	0.81	0.80	0.78	0.84	0.83	0.83
_Icharlsoni_1	1.03	1.02	1.01	1.02	1.05	1.04	1.03	1.04
_Icharlsoni_2	1.05	1.04	0.99	1.02	1.08	1.04	0.98	0.99
_Icharlsoni_3	1.09	1.08	1.05	1.06	1.21	1.26	1.16	1.21
_Icharlsoni_4	1.26	1.25	1.26	1.24	1.32	1.32	1.33	1.29
gastrointestinaldiseases_3y	1.16	1.12	1.15	1.11	1.19	1.15	1.17	1.13
otherjointproblems_3y_com	0.99	0.97	1.00	0.97	0.97	0.96	0.98	0.95
mentalhealth_3y_com	1.40	1.38	1.39	1.35	1.41	1.41	1.39	1.33
respiratorydiseases_3y	1.08	1.11	1.10	1.10	1.06	1.10	1.09	1.10
cardiovasculardiseases_3y	0.97	0.97	0.97	0.97	1.01	1.02	1.04	1.02
thyroidproblems_3y	1.03	1.04	1.04	1.04	1.02	1.04	1.04	1.04
foothipspinalpain_3y	1.16	1.17	1.16	1.18	1.14	1.12	1.15	1.14
coxarthrosis_3y	0.71	0.72	0.69	0.71	0.73	0.74	0.68	0.71
neurologicaldisorders_3y	1.07	1.02	1.06	1.02	1.11	1.08	1.12	1.07
otherarthrosis_3y	0.96	0.96	0.91	0.95	0.97	0.98	0.89	0.97
polyarthrosis_3y	0.67	0.76	0.73	0.76	0.58	0.70	0.64	0.69
spondylosis_3y	0.84	0.97	0.90	0.96	0.83	1.03	0.95	1.03
lead_ukr_yearly_surgeonoutput			1.07	1.03			1.09	1.04
lead_tkr_yearly_surgeonoutput	-		0.99	1.00			0.99	1.00

Appendix Table 5.3 – SMDs

OKS	Full Cohort												>10 Cohort											
	Logistic			Mixed			Logistic_volume			Mixed_volume			Logistic			Mixed			Logistic_volume			Mixed_volume		
	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp	IPW	Strata	S_exp
yhat_base	0.0	0.1	0.1	0.1	0.6	0.1	0.0	0.2	0.0	0.1	0.6	0.1	0.0	0.1	0.1	0.1	0.8	0.1	0.1	0.3	0.0	0.1	0.7	0.1
patientgen~n	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0
ageatprimary	0.0	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.3	0.3	0.0	0.1	0.1	0.1	0.2	0.3	0.1
primarybmi	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0
mrururb_in~2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
mrururb_in~3	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
mrururb_in~4	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mimd04_dec~2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mimd04_dec~3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
mimd04_dec~4	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mimd04_dec~5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0
mimd04_dec~6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
mimd04_dec~7	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
mimd04_dec~8	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
mimd04_dec~9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mimd04_de~10	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
mq1_gener_1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0
mq1_gener_2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.1	0.0	0.2	0.0
mq1_gener_3	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.1	0.1	0.1
mq1_gener_4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0
mq1_gener_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
mq1_eq5d_h~e	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
mkr_q1_score	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.2	0.0	0.1	0.0	0.0	0.2	0.2	0.0
primaryasa~c	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.1	0.0	0.0	0.1	0.1	0.0
Charlson_1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Charlson_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.2	0.1	0.0
Charlson_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Charlson_4	0.0	0.0	0.0	0.8	0.0	0.0	0.1	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
gastroint~3y	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0
otherjoint~m	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.0
mentalheal~m	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
respirato~3y	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
cardiovas~3y	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.2	0.0	0.1	0.0	0.0	0.0	0.2	0.0
thyroidpr~3y	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
foothipsp~3y	0.1	0.0	0.0	0.6	0.0	0.0	0.1	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
coxarthro~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
neurologi~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
otherarth~3y	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0
polyarthr~3y	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0
spondylos~3y	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0

Revision	Full Cohort												>10 Cohort											
	Logistic			Mixed			Logistic_volume			Mixed_volume			Logistic			Mixed			Logistic_volume			Mixed_volume		
	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp	IPW	Str	Str exp
yhat_base	0.0	0.1	0.0	0.2	0.3	0.0	3.4	0.3	0.0	0.2	0.3	0.0	0.0	0.1	0.0	0.5	0.6	0.1	4.1	0.5	0.0	0.5	0.6	0.1
patientgen~n	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.5	0.0	0.0	0.1	0.0	0.0
ageatprimary	0.0	0.1	0.0	0.2	0.1	0.0	0.2	0.0	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.0	0.3	0.1	0.0	0.2	0.2	0.0
primarybmi	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mrururb_in~2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
mrururb_in~3	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0
mrururb_in~4	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0
mimd04_dec~2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0
mimd04_dec~3	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0
mimd04_dec~4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mimd04_dec~5	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mimd04_dec~6	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0
mimd04_dec~7	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0
mimd04_dec~8	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mimd04_dec~9	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mimd04_de~10	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
mq1_gener_1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
mq1_gener_2	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	1.0	0.1	0.0	0.0	0.1	0.0
mq1_gener_3	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
mq1_gener_4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
mq1_gener_5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0
mq1_eq5d_h~e	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.9	0.0	0.0	0.1	0.0	0.0
mkr_q1_score	0.1	0.0	0.0	0.1	0.0	0.0	0.8	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	1.2	0.0	0.0	0.1	0.1	0.0
primaryasa~c	0.0	0.0	0.0	0.1	0.1	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.2	0.0	0.0	0.1	0.1	0.0
Charlson_1	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.5	0.0	0.0	0.1	0.0	0.0
Charlson_2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0
Charlson_3	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Charlson_4	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
gastroint~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
otherjoint~m	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
mentalheal~m	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.5	0.0	0.0	0.1	0.0	0.0
respirato~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0
cardiovas~3y	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1	0.0
thyroidpr~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
foothipsp~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
coxarthro~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
neurologi~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
otherarth~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0
polyarthr~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
spondylos~3y	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0

Appendix Text 5.1. Code for the data generation of the Plasmode Simulation. Example with the real-world continuous variable.

```
rm(list=ls())
#install.packages("TruncatedDistributions", repos="http://R-Forge.R-project.org")
library(AER); library(parallel); library(ggplot2); library(dplyr); library(TruncatedDistributions);
library(ivtools); library(xlsx);
library(disk.frame);  library(snow); library(ParallelLogger); library(epitools); library(rsimsum);
library(speedglm); library(readr)
############ DATA GENERATION ######################
#### set seed ####
set.seed(1311)
#### function ####
PlasmodeAlbert_cont<- function(formulaOut=NULL, objectOut=NULL,formulaExp=NULL,objectExp=NULL,data,
                idVar,effect =NULL, MMOut=1,MMExp=1, nsim, size, eventRate=NULL, exposedPrev=NULL,
                truncateleft=0, truncateright=48, sd_trunc=NULL, round=FALSE, IVstrength, IV_viol=0 )

{
  outcome <- all.vars(formula(objectOut))[1] ## selects the outcome variable
  exposure <- all.vars(formula(objectOut))[2] ##selects the exposure variable
  x <- data[order(data[,exposure]),] # order according to exposure status, unexposed first
  n <- nrow(x)
  n1 <- sum(x[,exposure])     # number of exposed in real data
  n0 <- n - n1
  size1 <- round(ifelse(is.null(exposedPrev), n1*(size/n), size*exposedPrev))  # desired exposed
  size0 <- size - size1
  if(size1 > n1 | size0 > n0) stop("Number of requested exposed or unexposed exceeds observed number --
reduce size")
   # find intercept value needed to get approximate desired event rate under new parameters
  X <- model.matrix(objectOut)[order(data[,exposure]),]
  OutCoeff<-coef(objectOut)
  bnew <- c(OutCoeff[1], MMOut*OutCoeff[-1])
  bnew <- replace(bnew, names(OutCoeff) == exposure, effect)
  bnew <- replace(bnew, names(coef(objectOut)) == "iv", IV_viol)
  Xbnew <- as.vector(bnew%*%t(X))

  # Finding the exposure prevalence in base cohort
  if(is.null(exposedPrev)) exposedPrev<- n1/n
  XEXP <- model.matrix(objectExp)[order(data[,exposure]),] # there was an error here, no
[order(data[,exposure]),]

  # Calculating intercept value needed to get approximate desired exposure prevalence under new
parameters.
  ExpCoeff <- coef(objectExp)
  bnewExp <- c(ExpCoeff[1], MMExp*ExpCoeff[-1])
  bnewExp[length(bnewExp)]<-log(IVstrength)
  XbnewExp <- as.vector(bnewExp%*%t(XEXP))
  fnExp <- function(d)mean(plogis(d+XbnewExp))-exposedPrev
  deltaExp <- uniroot(fnExp, lower=-2000, upper=2000)$root
  Probexp <- plogis(deltaExp+XbnewExp)
  ResVarOut <- var(residuals(objectOut))
  #### sample and simulate
  ids <- ynew <- expnew <- data.frame(matrix(nrow = size, ncol = nsim))
  #RR <- RD <- vector('numeric', length = nsim)
    for(sim in 1:nsim) {
    #sim=1
    idxs0 <- sample(1:n0, size0, replace = TRUE) # sample unexposed (located in rows 1:n0 of x)
    ids[1:size0,sim] <- x[idxs0, idVar]
    idxs1 <- sample(n0+1:n1, size1, replace = TRUE) # sample exposed (located in rows n0 + 1:n1 of x)
    ids[size0+1:size1,sim] <- x[idxs1, idVar]
    X2 <- X[c(idxs0,idxs1),]
    expnew[,sim] <- X2[,2] <- rbinom(size,1,Probexp[c(idxs0,idxs1)])
    Xbnew2 <- X2%*%bnew
    ynew[,sim] <- Xbnew2 + rnorm(size,0,sqrt(ResVarOut)) ## Change from original code -> Sqrt ResVarOut
(sd instead of variance) of residuals

    while (length( ynew[, sim][which(ynew[, sim]<truncateleft)] )>0) {             ### truncation left
      size2=length( ynew[, sim][which(ynew[, sim]<truncateleft)] )
```

```
      ynew[, sim][which(ynew[, sim]<truncateleft)]=truncateleft + rnorm(size2, 0, sd_trunc)
    }
    while (length( ynew[, sim][which(ynew[, sim]>truncateright)] )>0) {       ### truncation right
      size2=length( ynew[, sim][which(ynew[, sim]>truncateright)] )
      ynew[, sim][which(ynew[, sim]>truncateright)]=truncateright + rnorm(size2, 0, sd_trunc)
    }
    if (round)                                ### rounding
      ynew[, sim] <- round(ynew[, sim])
  }
  ## Creating simulated data for the outcome variable
  names(ids) <- paste("ID", 1:nsim, sep = "")
  names(ynew) <- paste("OUTCOME", 1:nsim, sep = "")
  names(expnew) <- paste("EXPOSURE",1:nsim, sep = "")
  sim_out <- data.frame(ids, ynew,expnew)
  return(list(TrueOutBeta = bnew, TrueExpBeta = bnewExp, Sim_Data = sim_out))
  }

#### Import database ####
setwd( "F:/Simulation_Data/" )
data<- read_csv("F:/Simulation_Data/Database_for_simulation.csv")
data$id <- row.names(data)

#### Continous Outcome ####
data$mkr_q2_score<-as.numeric(data$mkr_q2_score)
data <- data %>%
          filter(!is.na(mkr_q2_score))
table(data$ukr)

####### select synthetic IV or real IV
#data$iv2=rtgamma(length(data$ukr),shape=0.16,scale=2, a = 0, b=1)
data$iv2=as.double(data$last30_primarylead)
data <- data %>% dplyr::filter(!is.na(iv2))
data$iv<-ifelse(data$iv2 > median(data$iv2), 1, 0)
table(data$iv)

path_databases <- paste0("F:/Simulation_Data/Simulated_Datasets/Cont_Real_IV/First_Run_data.csv")
write_csv(data,path_databases)
ids_ivs <- data %>% dplyr::select(c(id,iv2,iv))
path_databases <- paste0("F:/Simulation_Data/Simulated_Datasets/Cont_Real_IV/IDs_IVs.csv")
write_csv(ids_ivs,path_databases)

#### formulas #####
form1<- mkr_q2_score ~  ukr +  ageatprimary + primarybmi + mq1_general_health + mq1_eq5d_health_scale +
mkr_q1_score + cardiovasculardiseases_3y + coxarthrosis_3y + foothipspinalpain_3y +
gastrointestinaldiseases_3y + neurologicaldisorders_3y + otherarthrosis_3y + polyarthrosis_3y +
respiratorydiseases_3y + spondylosis_3y + thyroidproblems_3y + charlsonindex_3y + mrururb_ind_n +
  mimd04_decile_n + patientgender_n + otherjointproblems_3y_com +mentalhealth_3y_com + primaryasa_enc +
  lead_yearly_surgeonoutput + iv

form2<- ukr ~  ageatprimary + primarybmi + mq1_general_health + mq1_eq5d_health_scale + mkr_q1_score +
cardiovasculardiseases_3y + coxarthrosis_3y + foothipspinalpain_3y + gastrointestinaldiseases_3y +
neurologicaldisorders_3y + otherarthrosis_3y + polyarthrosis_3y + respiratorydiseases_3y +
spondylosis_3y + thyroidproblems_3y + charlsonindex_3y + mrururb_ind_n + mimd04_decile_n +
patientgender_n + otherjointproblems_3y_com +mentalhealth_3y_com + primaryasa_enc +
  lead_yearly_surgeonoutput + iv

#### Coeficients ####
Coeff1<-glm(form1, family="gaussian", data=data ,control=glm.control(trace=TRUE))
Coeff2<- glm(form2,family="binomial", link = "logit", data=data,control=glm.control(trace=TRUE))

write.csv(as.data.frame(Coeff1$coefficients), "E:/Users/Albert/Desktop/to present
simulations/cont_real_iv_coefs_OKS.csv", row.names = TRUE)
write.csv(as.data.frame(Coeff2$coefficients), "E:/Users/Albert/Desktop/to present
simulations/cont_real_iv_coefs_exposure.csv", row.names = TRUE)

#### Modifications/Parameters ####
Modif_out <- 1
Modif_exp <- 1
objectOut=Coeff1
```

```
objectExp=Coeff2
data=data
idVar="id"
size=nrow(data)
exposedPrev=NULL
MMOut=Modif_out
MMExp=Modif_exp
truncateleft=0
truncateright=48
sd_trunc=2
round=FALSE

effects <- c(0, 1, 2, 5)
strengths <- c(1, 1.1, 2, 5, 10, 25, 50, 100)
IV_viols <- c(0)
nsim=1000

################################################################################
############################## Several simulations ##############################
################################################################################
############## Only some exponentiated (OR) ##############################

n_loops <- length(effects)*length(strengths)*length(IV_viols)
n_loop <- 0
start.time <- Sys.time()
for (IV_viol in IV_viols) {
for (effect in effects)     {
for (IVstrength in strengths ) {
  n_loop <- n_loop+1
  print(paste0("n loop: ", n_loop, " of ", n_loops ))

  Cont_Form1<- PlasmodeAlbert_cont(formulaOut=NULL, objectOut=Coeff1,
                      formulaExp=NULL,objectExp=Coeff2,
                      data=data, idVar="id",
                      effect =effect,
                      MMOut=Modif_out, MMExp=Modif_exp,
                      nsim=nsim, size=nrow(data),
                      eventRate=NULL, exposedPrev=NULL,
                      truncateleft=truncateleft, truncateright=truncateright,
                      sd_trunc=sd_trunc, round=round,
                      IVstrength=IVstrength, IV_viol=IV_viol )

 real_str<-Cont_Form1[["TrueBeta"]][["iv"]]
  simdata<-Cont_Form1$Sim_Data
  rm(Cont_Form1)
  simdata$IVstrength <- IVstrength
  simdata$IV_viol <- IV_viol
  simdata$effect <- effect
  path_databases <- paste0("F:/Simulation_Data/Simulated_Datasets/Cont_Real_IV/First_Run_IV_N_",n_loop,
                           "_n_", nsim,
                           ".csv")
    write_csv(simdata,path_databases)


}
}
}
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken
```

Appendix B

Publications and presentations arisen from this work

Protocol

Chapters 2-4

Strauss VY, Prats-Uribe A, Kolovos S, Berencsi K, Carr A, Judge A, Silman A, Arden N, Petersen I, Douglas IJ, Wilkinson JM, Murray D, Valderas JM, Beard DJ, Lamb SE, Ali MS, Pinedo-Villanueva R, , Prieto-Alhambra D. **Risk-benefit and costs of unicompartmental (compared to total) knee replacement for patients with multiple co-morbidities: a non-randomised study, and different novel approaches to minimize confounding**. Available in: https://fundingawards.nihr.ac.uk/award/15/80/40

Publications

Chapters 2-4

Prats-Uribe A, Kolovos S, Berencsi K, Carr A, Judge A, Silman A, Arden N, Petersen I, Douglas IJ, Wilkinson JM, Murray D, Valderas JM, Beard DJ, Lamb SE, Ali MS, Pinedo-Villanueva R, Strauss VY, Prieto-Alhambra D. **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 Nov;25(66):1-126. doi: 10.3310/hta25660. PMID: 34812138.

Chapter 2

Prats-Uribe A, Tobed M, Villacampa JM, Agüero A, García-Bastida C, Tato JI, Rodrigáñez L, Holguera VD, Hernández-García E, Poletti D, Simonetti G, Villarraga V, Meler-Claramonte C, Sánchez Barrueco Á, Chiesa-Estomba C, Casasayas M, Parente-Arias P, Mata-Castro N, Rello J, Castro P, Prieto-Alhambra D, Vilaseca I, Avilés-Jurado FX; TraqueoCOVID SEORL Group. **Timing of elective tracheotomy and duration of mechanical ventilation among patients admitted to intensive care with severe COVID-19: A multicenter prospective cohort study.** Head Neck. 2021 Dec;43(12):3743-3756. doi: 10.1002/hed.26863. Epub 2021 Sep 15. PMID: 34524714; PMCID: PMC8652734.

International Conference Presentations

Chapters 2-4

2020 ICPE. Multi-level Propensity Scores In Comparative Effectiveness Research On Medical Devices: A Pragmatic Example Of Partial Vs Total Knee Replacement Using National Joint Registry Data Compared To RCT Findings

2020 ICPE. The Performance of Instrumental Variables In Medical Device Epidemiology: A Simulation Study Of Non-normally Distributed Continous Outcome Data Based On Real World Post-operative Proms

2020 ICPE. The Risk of Myocardial Infarction And Venous Thromboembolism After Partial Or Total Knee Arthroplasty: A Self Controlled Case Series

2019ICPE. Propensity scores and inverse probability weighting in device epidemiology: a comparison of different approaches to minimise confounding when emulating a surgical RCT.

2019 ICPE. The performance of preference-based instrumental variables to emulate a randomised clinical trial of comparative medical device effectiveness.

2019 EUROCIM. Different approaches to minimise confounding when emulating a surgical randomised clinical trial: an application to partial vs total knee replacement.

2019 ICPE. Propensity scores and inverse probability weighting in device epidemiology: a comparison of different approaches to minimise confounding when emulating a surgical randomised clinical trial.

2019 ISPOR. Hospital costs and outcomes of unicompartimental compared to total knee replacement for patients with multiple comorbidities: a population-based study.

2019 ACPE. Comparative effectiveness of unicompartmental and total knee replacement for patients with comorbidities.

Publications and presentations arisen from work on COVID-19 during the DPhil

During the COVID-19 pandemic, I conducted several epidemiologic research projects alongside this Thesis to support rapid response. This research falls outside the scope of the Thesis but represented an invaluable learning experience during my graduate studies. The outcomes from these projects are listed below.

Publications

2022

Xie J, Feng S, Li X, Gea-Mallorquí E, **Prats-Uribe A**, Prieto-Alhambra D. **Comparative effectiveness of the BNT162b2 and ChAdOx1 vaccines against Covid-19 in people over 50** Nat Commun. 2022 Mar 21;13(1):1519. doi: 10.1038/s41467-022-29159-x. PMID: 35314697

Williams RD, … , **Prats-Uribe A**, et al. **Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network**. BMC Med Res Methodol. 2022 Jan 30;22(1):35. doi: 10.1186/s12874-022-01505-z. PMID: 35094686

Kostka K, …, Prats-Uribe A, et al. **Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS** Clin Epi. 2022. In Press.

2021

Reyes C, …, **Prats-Uribe A**, et al. **Characteristics and outcomes of patients with COVID-19 with and without prevalent hypertension: a multinational cohort study**. BMJ Open. 2021 Dec 22;11(12):e057632. doi: 10.1136/bmjopen-2021-057632. PMID: 34937727

Sentís A, **Prats-Uribe A**, et al.; Catalan HIV and STI Surveillance Group. **The impact of the COVID-19 pandemic on Sexually Transmitted Infections surveillance data: incidence drop or artefact?** BMC Public Health. 2021 Sep 7;21(1):1637. doi: 10.1186/s12889-021-11630-x. PMID: 34493245

Roel E, …, **Prats-Uribe A**, et al. **Characteristics and Outcomes of Over 300,000 Patients with COVID-19 and History of Cancer in the United States and Spain.** Cancer Epidemiol Biomarkers Prev. 2021 Oct;30(10):1884-1894. doi: 10.1158/1055-9965.EPI-21-0266. Epub 2021 Jul 16. PMID: 34272263

Recalde M, …, Prats-Uribe A, et al. **Characteristics and outcomes of 627 044 COVID-19 patients living with and without obesity in the United States, Spain, and the**

United Kingdom. Int J Obes (Lond). 2021 Nov;45(11):2347-2357. doi: 10.1038/s41366-021-00893-4. Epub 2021 Jul 15. PMID: 34267327

Prats-Uribe A, Xie J, Prieto-Alhambra D, Petersen I. **Smoking and COVID-19 Infection and Related Mortality: A Prospective Cohort Analysis of UK Biobank Data.** Clin Epidemiol. 2021 May 25;13:357-365. doi: 10.2147/CLEP.S300597. eCollection 2021. PMID: 34079379

Duarte-Salles T, …, **Prats-Uribe A**, et al. **Thirty-Day Outcomes of Children and Adolescents With COVID-19: An International Experience.** Pediatrics. 2021 Sep;148(3):e2020042929. doi: 10.1542/peds.2020-042929. Epub 2021 May 28. PMID: 34049959

Prats-Uribe A, et al. **Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study.** BMJ. 2021 May 11;373:n1038. doi: 10.1136/bmj.n1038. PMID: 33975826

Tan EH, Sena AG, **Prats-Uribe A**, et al. **COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries.** Rheumatology (Oxford). 2021 Oct 9;60(SI):SI37-SI50. doi: 10.1093/rheumatology/keab250. PMID: 33725122

Burn E, …, **Prats-Uribe A**, et al. **The natural history of symptomatic COVID-19 during the first wave in Catalonia.** Nat Commun. 2021 Feb 3;12(1):777. doi: 10.1038/s41467-021-21100-y. PMID: 33536437

Lane JCE, …, Prats-Uribe A, et al. **Risk of depression, suicide and psychosis with hydroxychloroquine treatment for rheumatoid arthritis: a multinational network cohort study**. Rheumatology (Oxford). 2021 Jul 1;60(7):3222-3234. doi: 10.1093/rheumatology/keaa771. PMID: 33367864

Morales DR, …, **Prats-Uribe A**, et al. **Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis**. Lancet Digit Health. 2021 Feb;3(2):e98-e114. doi: 10.1016/S2589-7500(20)30289-2. Epub 2020 Dec 17. PMID: 33342754

2020

Prieto-Alhambra D, …, **Prats-Uribe A**, et al. **Filling the gaps in the characterization of the clinical management of COVID-19: 30-day hospital admission and fatality rates in a cohort of 118 150 cases diagnosed in outpatient settings in Spain.** Int J Epidemiol. 2021 Jan 23;49(6):1930-1939. doi: 10.1093/ije/dyaa190. PMID: 33118038

Burn E, …, Prats-Uribe A, et al. **Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study**. Nat Commun. 2020 Oct 6;11(1):5009. doi: 10.1038/s41467-020-18849-z. PMID: 33024122

Lane JCE, …, **Prats-Uribe A**, et al.; OHDSI-COVID-19 consortium. **Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study**. Lancet Rheumatol. 2020 Nov;2(11):e698-e711. doi: 10.1016/S2665-9913(20)30276-9. Epub 2020 Aug 21. PMID: 32864628

Coma Redon E, Mora N, **Prats-Uribe A**, Fina Avilés F, Prieto-Alhambra D, Medina M. **Excess cases of influenza and the coronavirus epidemic in Catalonia: a time-series analysis of primary-care electronic medical records covering over 6 million people.** BMJ Open. 2020 Jul 29;10(7):e039369. doi: 10.1136/bmjopen-2020-039369. PMID: 32727741