BMJ Open

The impact of longer patient travel distances and times on perioperative outcomes after complex revision knee surgery

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-085201
Article Type:	Original research
Date Submitted by the Author:	15-Feb-2024

Complete List of Authors:	Matthews, Alexander; University of Exeter Faculty of Health and Life Sciences, Department of Clinical and Biomedical Sciences; Royal Devon and Exeter NHS Foundation Trust, Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research; Royal Devon and Exeter NHS Foundation Trust, Princess Elizabeth Orthopaedics Centre Evans, Jonathan; University of Exeter, Department of Public Health and Sports Sciences Lamb, Sarah; University of Exeter, College of Medicine and Health Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Gray, William; NHS England and NHS Improvement London Briggs, Tim; NHS England and NHS Improvement London Toms, Andrew ; Princess Elizabeth Orthopaedic Centre
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, PUBLIC HEALTH, Health Services

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

TravelTimesCodeFinal.R

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

The impact of longer patient travel distances and times on perioperative outcomes after complex revision knee surgery

Names, Affiliations, and positions of all authors

Alex Matthews1,2,3, Jonathan P Evans2,3, Jonathan T Evans2,3, Sallie Lamb3 Andrew Price4,5, William Gray1, Tim Briggs1,6, Andrew Toms2,3

1. Getting It Right First Time programme, NHS England, London, UK
2. Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
3. University of Exeter, Exeter, UK
4. Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, UK
5. Nuffield Orthopaedic Centre, Oxford, UK
6. Royal National Orthopaedic Hospital, Stanmore, London, UK

Correspondence: Alexander Matthews, Getting It Right First Time programme, NHS England and NHS Improvement, Wellington House, Waterloo Road. London, UK, SE1 8UG; telephone: 07401932627 email: alexander.matthews1@nhs.net

Word count: [3316 words]

Short title: Patient Journey Distance and Time for Complex Revision Knee Surgery

Key words: social determinants of health, revision knee replacement, contextual factors in complex health interventions

Twitter handles:

Alex Matthews: @DrAlexMatthews

Tim WR Briggs: @ProfTimBriggs

WilliamKGray1: @WilliamKGray1

GIRFT programme: @NHSGIRFT

Structured Abstract (Word count suggested 250-300)

Objectives

Patients with problematic knee replacements requiring further surgery often have difficulties mobilising and increasingly rely on family support. Evolving practice in England aims to manage these patients in specialised centres with the intention of improving outcomes. This practice will result in longer travel distances and times in this frailer group of patients. We want to examine the types of distances and travel times patients can be expected to travel for complex orthopaedic surgery and to explore concerns of how these impact patient outcomes.

Design

Retrospective observational study from the Hospital Episode Statistics. Multivariable adjusted logistic regression modelling was used to compare the exposure variable with perioperative outcomes

Setting

Patients presenting to tertiary referral centres between 1st January 2016 to 31st December 2019. A tertiary referral centre was defined as a trust performing >70 revisions in the year prior.

Participants

Adult patients undergoing revision total knee replacement procedures for aseptic reasons between 1st January 2016 to 31st December 2019.

Interventions

Patient level travel distance and time was calculated using the department of health Journey Time Statistics.

Main Outcome Measures

The primary outcome is the association of travel distance and time on emergency readmission within 30 days. Secondary outcomes will focus on mortality within 90 days and length of inpatient stay.

Results

1516 patients were treated at 16 tertiary referral centres for non-infected reasons. Patients in the longest driving distance group were expected to travel a median distance of 44.55 miles (IQR 35.90 to 56.30) with an expected median journey time of 66.3 minutes (IQR 57.9 to 80.5). Overall, 30-day readmission was not statistically associated with farther travel distances or driving times.

Conclusions

Patients were expected to travel up to hour for revision knee replacement surgery. There was no association between increasing travel distance and time on perioperative outcomes.

Summary Boxes

What is already known on this topic?

- A failed primary knee replacement is a life changing event often linked to reduced mobility and depression.
- Evolving practice in revision knee replacement surgery in England aims to treat these complex frail patients in super-specialised regional hospitals.
- Subsequently patients can expect to travel longer distances and times and it is unknown what affect these will have on patient outcomes.

What does this study add?

- Patient in the longest journey time category were expected to travel over an hour at peak driving times.
- Patient's travelling farther for revision knee replacement surgery did not demonstrate any statistically worse perioperative clinical outcomes.
- This information is of utility to surgical providers and commissioners of healthcare services and can inform patient-led decision-making surrounding travelling for complex revision knee replacement surgery.

Introduction

Primary knee replacement is a successful procedure that improves quality of life for the majority of patients.(1) However, at 10 years following a primary knee replacement, about 3.5% of patients will have undergone a revision surgery.(2) The majority of these procedures are carried out due to infection or polyethylene wear of the implant.(3) A failed primary knee replacement represents a life changing transition point where individuals are likely to suffer from pain, reduced mobility as well as dependency on family members.(4) Patients often face multi-step surgery with longer hospital length of stays and higher complication rates.(5, 6)

The orthopaedic GIRFT (Getting It Right First Time) programme was launched in 2012 following the publication of the Orthopaedic National Report.(7) A key recommendation was the centralisation of complex orthopaedic surgery, including revision knee surgery, to specialist centres with the aim of improved patient outcomes. Consequently, revision total knee replacement (rTKR) surgery in the England has evolved into a regional network service model. .(8) In doing so, all hospitals performing rTKRs form a network in the respective regions. Less specialist hospitals defined by lower annual case volume thresholds are encouraged to discuss and sometimes refer their caseload to more specialist centres. Several studies based on large revision hip and knee registries has suggested this model carries a lower failure rate defined by the need for further revision surgery.(9-11) Early evidence has suggested reduced early failure rates through the adoption of revision knee networks.(12)

However, this approach to managing patients is inevitably associated with increasing travel distances between some patient's homes and their treating hospital. Expected distances are important to explore, particularly as these patients struggle with mobility, may be unable to drive and may be more reliant on family members.(4) Furthermore, greater travel distances have been associated with higher readmission rates and higher mortality rates following complex vascular surgery.(13) There is also concern that patients required to travel greater distances are more likely to be re-admitted to a different hospital resulting in clinical decisions that do not

incorporate the primary surgeon and potentially alter outcomes.(14) Subsequently the aims of this paper is to examine if the same association with longer patient travel distance and perioperative outcomes exists following complex orthopaedic surgery with a focus on revision knee replacement surgery performed in high volume tertiary referral centres.

Methods

Design

This study is a retrospective data analysis of observational data from the Hospital Episode Statistics (HES) and Office for National Statistics (ONS) databases. HES data is collected by NHS England for all patients treated at NHS hospitals in England and those treated at private hospitals where treatment was funded by the NHS. This study complies with the recommended reporting guidelines when using HES data(15) and the Strengthening of Reporting of Observational studies in Epidemiology (STROBE) guidelines.(16)

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes(17) and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.(18) The HES data were linked at a patient level to data from the ONS on deaths and date of death, which allowed the identification of patients who had died after their surgery. Linkage was achieved using a unique patient identifier using a previously validated methodology.(19)

Patient travel distances were calculated using the Journey Time Statistics reference document produced by the UK Department of Transport which modelled theoretical journey times between known Lower Layer Super Output Areas (LSOA) of residence and NHS hospital sites.(20) The Journey Time Statistics document is available in the supplementary material section.

Population

An rTKR procedure was defined as a permanent removal or exchange of knee arthroplasty components. This includes a revision of a total knee replacement and a conversion of a unicondylar knee replacement to a total knee replacement. Secondary patellar resurfacing was not included as this represents a simple revision procedure, one that can be carried out in most nonspecialised hospitals. All patients aged ≥ 18 years who underwent a rTKR in a high-volume trust between 1st January 2016 to 31st December 2019 were included in the study population. A high-volume trust was classified as a centre performing > 70 revisions per year. This revision volume threshold for classification represents those proposed by the British Association for Surgeons of the Knee (BASK) Revision Knee Working Group. (21) These are more likely to represent tertiary referral centres where the stratification of more complex work will take place. Annual case volume at each trust was defined as the number of revision cases conducted in the year prior to the index procedure. This measure was preferred over a simple calculation of average annual volume as it accounts for recent experience at the point of surgery. The Office for Population Censuses and Surveys' Classification of Interventions and Procedures version 4 (OPCS-4) codes used to identify rTKR procedures are detailed in **Supplementary material S1.** Where the procedure laterality was not specified, patients were excluded. The flow of patients, with numbers excluded at each point, is summarised in **Supplementary material S2**. To manage population heterogeneity, data were extracted for the period 1st April 2011 to 31st December 2019 and only the first revision for a specific side of the body record in this time period included. (22) Thus, any early revisions on the same side of the body in the four years and nine months preceding the start of the study period were identified and these patients excluded from the study. This aims to exclude the early revision knee replacement failures which have been shown to represent catastrophic failures potentially skewing our results. (22)We excluded revisions for infection as these represent a more variable patient group with a different complication profile (23) and this is further discussed in our study limitations.

Exposure variable

In the analysis straight line travel distance was calculated between a patient's LSOA and the postal codes for the treating hospitals. LSOAs are determined by the Office

for National Statistics and are designed for the reporting of small area statistics. Public transport and highways data for England was used to create theoretical journey distances and times from origins to destinations. The resulting travel distances and/or times for each patient were divided into quintiles following a recently reported methodology.(13) Sensitivity analyses were performed using travel distances by road and peak driving times to account for variation in travel infrastructure between rural and urban areas and to attribute more meaningful results for patients. Peak driving times were calculated by using average traffic speeds for between 7am and 10am.

Outcomes

The primary outcome was emergency readmission within 30 days of discharge from the index surgical hospital. Secondary outcomes included 90-day mortality, and hospital length of stay (LOS) above the median. The LOS outcome was dichotomised into above median or below median LOS of five days.

Statistical Analyses

Data was extracted from a secure, encrypted server controlled by NHS Digital. Data were analysed within a secure, encrypted environment using standard statistical software: R Studio version 2023.09.1+494 (Boston, Massachusetts, USA). The R code and packages used are included in **Supplementary material S3**

Crude comparisons of baseline categorical characteristics and travel distance proximity were calculated. A This data were categorical in nature and summarised as frequency and percentage. In primary analysis a logistic multivariable regression model was constructed to evaluate associations between travel distance quintiles and 30 day readmission, with adjustment for the covariates listed above. The first (shortest) travel distance quintile was used as the reference in all models.

Age, sex , comorbidities and characteristics of initial presentation were included in the logistic regression model. These variables have been shown to influence the risk of complications after R-TKA and therefore represent known confounders.(9, 10, 23) This also included data on economic deprivation measured using the Index of Multiple Deprivation (IMD).(24) The IMD gives the LSOA where the patient lives a

score based on a range of measures of deprivation. IMD was categorised into quintiles, based on all-England data, for analysis. A spearman's rank correlation was performed to investigate the relationship between IMD score and travel distances. Health co-morbidity was quantified using the Hospital Frailty Risk Score (HFRS). HFRS identifies frailty based on the occurrence of any of 109 International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during any hospital admissions in the two years prior to, and for, the index admission.

All secondary outcomes were binary and analysed using the same multivariable logistic regression. Multicollinearity was assessed with reference to variance inflation factor and Shapiro-Wilcox test of normality. Model fit was assessed with reference to the pseudo R^2 values.

Results

Demographic characteristics and co-morbidities

The 1,516 patients in the overall study population, were divided into quintiles of travel distance of 303 or 304 patients each. The median straight line travel distance for quintile one was 2.3 miles (IQR 1.3 to 3.1). For the fifth travel quintile, median distance was 33.5 miles (IQR 25.5 to 41.1). Baseline co-morbidities and demographic characteristics were broadly similar among the travel distance quintiles (**Table 1**). Travel distance was not strongly correlated with age or social deprivation (**Figure 1**)

Association between travel distance and readmission, mortality and extended hospital stay

Overall, 111 patients were readmitted within 30 days. Crude comparisons of proportions readmitted within 30 days for each travel distance quintiles revealed a higher rate of readmission for the second travel quintile. In multivariable adjusted logistic regression, there was no statistically significant association between travel distance and readmission within 30 days (**Table 2**). Odds for 30-day readmission was 1.44 (95% CI 0.71 to 2.96, P 0.32) for Q5 compared with Q1. Increased travel distance was not associated with a significant change in the odds of death within 90

days (OR for Q5 vs Q1, 1.46 (95% CI 0.49 to 4.53, P 0.682)). Travel distance quintile was not associated with prolonged length of hospital stay related to the index surgery after multivariable adjustment (OR for Q5 vs Q1, 0.96 (95% CI 0.67 to 1.39, P 0.84)).

Real world travel distance and outcomes

The above results used straight line travel distance between patient' LSOA and treating hospital. A sensitivity analysis using actual patient travel distances using the shortest possible road route was performed **(Table 3)**. The median driving distance by the shortest possible road route for the closest quintile was 3.40 miles (IQR 2.00 to 4.40). The furthest quintile median driving distance was 44.55 miles (IQR 35.90 to 56.30) This analysis showed no association between driving distance and all perioperative outcomes. The OR for readmission within 30 days in Q5 vs Q1 was 1.16(0.56-2.41, p value = 0.68).

Journey drive times and outcomes

A further sensitivity analysis using driving times was calculated (Table 4). The median drive time for quintile one was 12.6 minutes (IQR 8.7 to 15.3). For the fifth time quintile the median was 66.3 minutes (IQR 57.9 to 80.5). No statistical association was found between drive time and perioperative outcomes. The OR for readmission within 30 days in Q5 vs Q1 was 0.92 (0.45 – 1.85, p value = 0.81)

Discussion

Statement of principal findings

We present a multi-hospital site retrospective analysis of patients undergoing revision knee replacement surgery at tertiary referral centres in England. In this analysis of 1,516 patients undergoing aseptic revision knee replacement surgery, we did not observe an association between distance and time travelled for revision surgery and readmission within 30 days. Patients in the longest driving time category were expected to travel for a median time of more than one hour.

Strengths and weaknesses of the study

The findings of this study should be interpreted in view of several limitations. Firstly, this analysis used observational data from a large administrative dataset covering all NHS-funded procedures conducted in England. As with all datasets we are limited in the amount of detail provided regarding presentation. We chose to categorise a high-volume centre by trust to accurately capture surgical experience. All NHS hospitals in England are run by hospital trusts which typically involve between one and four hospitals within a catchment area standardising their practice. It is common practice for specialist orthopaedic surgeons to move between these sites delivering the same procedures. Our study involved 35 hospital sites run by 16 trusts. We acknowledge this is a weakness of our study as this may not be representative of all trusts. The indication for revision coded as mechanical complication encompasses several common indications such as aseptic loosening, instability and malalignment. Reassuringly these indications have similar length of stay, and perioperative outcomes.(23) Differences exist in their re-revision rate, however this was not an outcome of focus in our study. It is likely that the complexity of the surgery undertaken may vary within the different indications for revision. Evidence suggests that operative surgical time is related to increased length of stay in aseptic revision knee replacement.(25). There is a lack of granular data for revisions due to infection and therefore we excluded this patient group as some readmissions for this patient group may represent planned readmissions. There is also a lack of granular clinical data using HES for each readmission, therefore we cannot ascertain precise reasons for readmissions, but we assume are related to a post-surgical complication. Clinical coding practice within HES is known to vary across trusts.(26) As an example, some trusts may be more consistent in coding comorbidities, and this may have created some bias. However, this is unlikely to vary systematically with travel distances and so significantly bias our findings. We acknowledge the relatively short travel distances in this population compared to examples from the United States as such the results of this study may not be generalisable to larger geographical areas or less mature healthcare systems. However, the upper quintile in our study represents a substantial journey distance and time for our patient cohort where poor mobility is a

significant factor affecting their care. This analysis does not consider journey times of those who may not have access to a car and instead chose to take public transport.

Strengths and weaknesses in relation to other studies, discussing important differences in results

This is the first study to analyse the potential impact of patient travel distances on patients receiving complex orthopaedic surgery. The findings that longer travel distances are not associated with inferior outcomes is an important part of the evaluation of the assumptions and context behind the establishment of revision knee networks.(27) This study has shown that concerns of introducing a network in larger geographical regions, for example in Scotland where longer patient travel distances and times are common, may be less important.(28) This is particularly useful as regions explore the geography of their revision networks and during summative outcome assessment of this complex health intervention.(29)

It may be seen as surprising that no association between travel distance and prolonged length of hospital stay was identified. An expectation exists of increasing difficulties being encountered with the discharge of patients living greater distances from their treating hospital, which has been observed in patients following elective pancreatic surgery.(30) This is also an observation seen in patients being treated in specialist vascular centres in the United States which led to the recommendation of additional care coordination and follow up efforts. However, the geography of the population in these studies was much larger with significantly longer travel distances.

We did not observe a strong correlation between social deprivation status and age of the patient with longer travel distances. It is reassuring that access to treatment for older patients and those from poor socioeconomic backgrounds is unaffected by travel distance. However, there may be patients who refused to travel to a specialist centre and opted for treatment at their local centre.

Meaning of the study: possible explanations and implications for clinicians and policymakers

The organisation and delivery of revision knee services in England has recently undergone a substantial change and now such services are provided around regional networks of care. This promises substantial advantages to the increasing number of patients with problematic knee replacements in our ageing population who will benefit from regional expertise.(8) However, it is unknown the impact of patients residing farther from tertiary referral centres, particularly rural patients who may encounter additional difficulties associated with greater travel distance. A recent study following the outcomes of aortic surgery found that longer travel distances are associated with inferior perioperative outcomes(13). Similar associations have been found in postoperative colorectal surgery patients (31). As such our results are reassuring to policy makers and clinicians.

Unanswered questions and future research

There is a scarcity of evidence evaluating the patient perception of complex health interventions such as network models of care. Recent work by Kugler et al has demonstrated the willingness of patients to travel further for better outcomes in the context of total knee replacement surgery. (32) Nevertheless, patient perceptions of travelling further for their treatment should be a focus for future research in the context of revision knee patients, particularly as this is one of the top ten research priorities identified by the James Lind Alliance priority setting partnership.(33)

Conclusion

We did not observe an association in our study population between 30-day readmission rates and increasing travel distances or times between a patient's home and their treating hospital in revision knee replacement. This paper is the first to

explore the relationship between travel distance and complex orthopaedic surgery and informs some concerns regarding the creation of a centralised revision knee network. This information is of utility to surgical providers and commissioners of healthcare services. Furthermore, it can inform patient-led decision making and the exploration of perceptions surrounding travelling for complex surgery. Although this is the first assessment in complex orthopaedic surgery, a prospective analysis will be undertaken as part of the ongoing auditing of revision knee networks in England.

Supplementary material, Figures, Tables and Files

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

See separate file named supplementary material

Supplementary material S2 – Flow of patient inclusion/exclusions

-See attached file named Supplementary Material

Supplementary material S3 – R Code

See attached file named Supplementary Material

Figure 1 – Box plot showing association of social deprivation and age on travel distance quintile. Spearman's rank correlation investigating relationship between these factors with travel distance.

-See attached file names tables and figures

Table 1 – Baseline demographics and clinical characteristics and raw perioperative outcomes for patients by travel distances quintile

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of Patients	304	303	303	303	303
Deprivation Quintile					
1(Most Deprived)	54 (18%)	87 (29%)	50 (17%)	58 (19%)	57 (19%)
2	48 (16%)	61 (20%)	68 (22%)	58 (19%)	66 (22%)
3	57 (19%)	65 (21%)	63 (21%)	63 (21%)	56 (18%)
4	55 (18%)	43 (14%)	49 (16%)	80 (26%)	75 (25%)
5 (Least Deprived)	90 (30%)	47 (16%)	73 (24%)	44 (15%)	49 (16%)
Sex					
Male	121 (40%)	133 (44%)	135 (45%)	124 (41%)	141 (47%)
Age in years					
16-59	52 (17%)	42 (14%)	56 (18%)	57 (19%)	68 (22%)
60-64	46 (15%)	29 (10%)	35 (12%)	38 (13%)	47 (16%)
65-69	53 (17%)	58 (19%)	56 (18%)	52 (17%)	44 (15%)
70-74	45 (15%)	56 (18%)	53 (17%)	59 (19%)	49 (16%)
75-79	46 (15%)	43 (14%)	47 (16%)	44 (15%)	47 (16%)
>=80	62 (20%)	75 (25%)	56 (18%)	53 (17%)	48 (16%)
Diagnosis					
Mechanical Complication	172 (57%)	198 (65%)	208 (69%)	212 (70%)	192 (63%)
Fracture	26 (9%)	29 (10%)	18 (6%)	10 (3%)	31 (10%)
Progressive OA	39 (13%)	32 (11%)	24 (8%)	24 (8%)	18 (6%)
Hospital Frailty Risk Score					
None	158 (52%)	115 (38%)	149 (49%)	156 (51%)	158 (52%)
Mild	100 (33%)	123 (41%)	106 (35%)	100 (33%)	99 (33%)
Moderate	38 (13%)	56 (18%)	40 (13%)	43 (14%)	36 (12%)

Severe	8 (3%)	9 (3%)	8 (3%)	4 (1%)	10 (3%)
Annual Surgeon Volume					
Volume 0-4	30 (10%)	39 (13%)	31 (10%)	30 (10%)	24 (8%)
Volume 5-9	43 (14%)	46 (15%)	44 (15%)	41 (14%)	35 (12%)
Volume 10-14	89 (29%)	72 (24%)	55 (18%)	63 (21%)	58 (19%)
Volume 15-19	49 (16%)	56 (18%)	44 (15%)	47 (16%)	53 (17%)
Volume 20-24	48 (16%)	64 (21%)	74 (24%)	62 (20%)	47 (16%)
Volume >=25	45 (15%)	26 (9%)	55 (18%)	60 (20%)	86 (28%)
Perioperative Outcomes					
Readmission within 30 days	15(5%)	36(12%)	23(8%)	17(6%)	20(7%)
90 Day Mortality	8(3%)	15(5%)	10(3%)	7(2%)	10(3%)
Prolonged Length of Stay	135(44%)	134(44%)	135(45%)	130(43%)	132(44%)

Table 2 – Multivariable Logistic Regression showing Odds Ratios for primary and secondary outcomes by straight line travel quintile

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	2.23 (95% CI 1.19 to 4.37),p=0.01	1.55 (95% CI 0.79 to 3.13),p=0.21	1.06 (95% CI 0.51 to 2.23),p=0.87	1.44 (95% CI 0.71 to 2.96),p=0.32
90 Day Mortality	1.79 (95% CI 0.66 to 5.17),p=0.27	1.55 (95% CI 0.52 to 4.70),p=0.43	1.72 (95% CI 0.53 to 5.64),p=0.36	1.46 (95% CI 0.49 to 4.53),p=0.50
Prolonged Length of stay	0.90 (95% CI 0.62 to 1.30),p=0.57	1.02 (95% CI 0.71 to 1.46),p=0.91	0.99 (95% CI 0.69 to 1.412),p=0.95	0.96 (95% CI 0.67 to 1.39),p=0.84

Table 3 – Sensitivity analysis exploring road travel distance quintile and primary/secondary outcomes

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	2.11 (95% CI 1.14 to 4.06),p=0.02	1.41 (95% CI 0.71 to 2.84),p=0.33	1.44 (95% CI 0.73 to 2.90),p=0.29	1.16 (95% CI 0.56 to 2.4),p=0.68
90 Day Mortality	1.52 (95% CI 0.58 to 4.21),p=0.40	1.18 (95% CI 0.39 to 3.56),p=0.77	2.42 (95% CI 0.83 to 7.27),p=0.11	0.83 (95% CI 0.25 to 2.64),p=0.75
Prolonged Length of stay	0.69 (95% CI 0.48 to 0.99),p=0.04	0.94 (95% CI 0.66 to 1.35),p=0.74	0.86 (95% CI 0.60 to 1.23),p=0.41	0.95 (95% CI 0.66 to 1.36),p=0.77

Table 4- Sensitivity analysis exploring driving time quintile and primary/secondary outcomes

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	1.19 (95% CI 0.63 to 2.25),p=0.59	1.23 (95% CI 0.65 to 2.35),p=0.52	1.53 (95% CI 0.82 to 2.89),p=0.18	0.92 (95% CI 0.45 to 1.85),p=0.81
90 Day Mortality	2.29 (95% CI 0.84 to 6.76),p=0.12	1.52 (95% CI 0.50 to 4.74),p=0.46	2.63 (95% CI 0.90 to 8.16),p=0.08	0.94 (95% CI 0.26 to 3.23),p=0.91
Prolonged Length of stay	0.82 (95% CI 0.57 to 1.52),p=0.27	0.80 (95% CI 0.56 to 1.15),p=0.23	1.27 (95% CI 0.89 to 1.81),p=0.19	1.01 (95% CI 0.70 to 1.44),p=0.97

Public and Patient Involvement statement

The study's chief investigator (AT) led the James Lind Alliance 'Revision Knee Replacement' priority setting partnership. This group of patients, carers and health care professionals identified the need to investigate the best way of organising revision knee replacement surgery to improve patient outcomes as one of their top 10 research questions. Patients were therefore directly involved in the development

of the study's aims and objectives. The results of the study will be disseminated to the members of this group prior to publication.

Copyright/licence for publication Statement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing Interests

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical Approval

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data. Ethical approval was not required.

Funding

No funding was obtained to carry out this study

Data Sharing

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2,3
Introduction			

Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up (*b*) For matched studies, give matching criteria and number of exposed and unexposed	5,6 NA
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6,7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	5,7
Bias	9	Describe any efforts to address potential sources of bias	5,6
Study size	10	Explain how the study size was arrived at	6
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7,8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (*b*) Describe any methods used to examine subgroups and interactions (*c*) Explain how missing data were addressed (*d*) If applicable, explain how loss to follow-up was addressed (*e*) Describe any sensitivity analyses	7 7 n/a 7,8
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	6 6 16
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest (c) Summarise follow-up time (eg, average and total amount)	17 n/a 6
Outcome data	15*	Report numbers of outcome events or summary measures over time	8
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	8 8 n/a
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	8,9
Discussion			
Key results	18	Summarise key results with reference to study objectives	9
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	10,11
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	9,10,11
Generalisability	21	Discuss the generalisability (external validity) of the study results	10,11
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	23

References

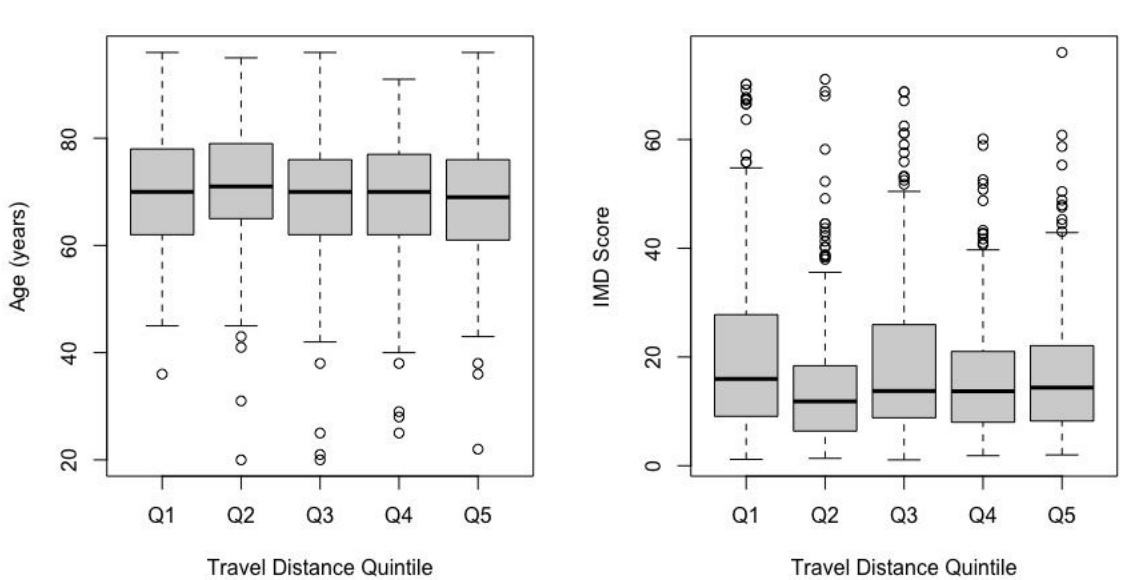
1. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet. 2018;392(10158):1672-82.
2. Sabah SA, Hedge EA, Abram SGF, Alvand A, Price AJ, Hopewell S. Patient-reported outcome measures following revision knee replacement: a review of PROM instrument utilisation and measurement properties using the COSMIN checklist. BMJ Open. 2021;11(10):e046169.
3. Sabah S, Knight R, Alvand A, Murray D, Petrou S, Beard D, Price A. No exponential rise in revision knee replacement surgery over the past 15 years: an analysis from the National Joint Registry. Osteoarthritis and Cartilage. 2022;30(12):1670-9.
4. Rowland T, Lindberg-Larsen M, Santy-Tomlinson J, Jensen CM. A qualitative study of patients' experiences before, during and after surgical treatment for periprosthetic knee joint infection; “I assumed it had to be like that … ”. International Journal of Orthopaedic and Trauma Nursing. 2023;48:100992.
5. Garbarino LJ, Gold PA, Sodhi N, Anis HK, Ehiorobo JO, Boraiah S, et al. The effect of operative time on in-hospital length of stay in revision total knee arthroplasty. Ann Transl Med. 2019;7(4):66.
6. Liodakis E, Bergeron SG, Zukor DJ, Huk OL, Epure LM, Antoniou J. Perioperative Complications and Length of Stay After Revision Total Hip and Knee Arthroplasties: An Analysis of the NSQIP Database. The Journal of Arthroplasty. 2015;30(11):1868-71.
7. Briggs TWR. A National Review of Adult Elective Orthopaedic Services in England. In: Time GIRF, editor. 2015.
8. Kalson NS, Mathews JA, Phillips JRA, Baker PN, Price AJ, Toms AD, et al. Revision knee replacement surgery in the NHS: A BASK surgical practice guideline. The Knee. 2021;29:353-64.
9. Yapp LZ, Walmsley PJ, Moran M, Clarke JV, Simpson AHRW, Scott CEH. The effect of hospital case volume on re-revision following revision total knee arthroplasty. The Bone & Joint Journal. 2021;103-B(4):602-9.
10. Halder AM, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Re-Revision Rate Following Aseptic Revision Total Knee Arthroplasty: An Analysis of 23,644 Cases. J Arthroplasty. 2020;35(4):1054-9.
11. Jeschke E, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Revision Rate and Mortality Following Revision Total Hip Arthroplasty: An Analysis of 17,773 Cases. J Arthroplasty. 2019;34(9):2045-50.
12. Bloch BV, Matar HE, Berber R, Gray WK, Briggs TWR, James PJ, Manktelow ARJ. The impact of a revision arthroplasty network on patient outcomes. Bone Joint J. 2023;105-b(6):641-8.
13. Feldman ZM, Zheng X, Mao J, Sumpio BJ, Mohebali J, Chang DC, et al. Longer patient travel distance is associated with increased non-index readmission after complex aortic surgery. Journal of Vascular Surgery. 2023;77(6):1607-17.e7.

14. Saunders RS, Fernandes-Taylor S, Kind AJ, Engelbert TL, Greenberg CC, Smith MA, et al. Rehospitalization to primary versus different facilities following abdominal aortic aneurysm repair. J Vasc Surg. 2014;59(6):1502-10, 10.e1-2.
15. Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes--systematic review and recommendations for reporting. Eur J Public Health. 2013;23(1):86-92.
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
17. NHS Digital. Hospital Episode Statistics(HES) Analysis Guide. London, UK: NHS Digital; 2018.
18. Information Standard Board for Health and Social Care. Anonymisation Standard for Publishing Health and Social Care Data Specification. London, UK: NHS Digital; 2013.
19. NHS Digital. Delivering linked datasets to support health and care delivery and research. London, UK: NHS Digital; 2018.
20. Department for Transport. Journey Time Statistics, notes and definitions. London, UK: Department for Transport; 2019.
21. Kalson NS, Mathews JA, Miles J, Bloch BV, Price AJ, Phillips JRA, et al. Provision of revision knee surgery and calculation of the effect of a network service reconfiguration: An analysis from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Knee. 2020;27(5):1593-600.
22. Deere K, Whitehouse MR, Kunutsor SK, Sayers A, Price AJ, Mason J, Blom AW. How long do revised and multiply revised knee replacements last? An analysis of the National Joint Registry. Lancet Rheumatol. 2021;3(6):e438-e46.
23. Sabah SA, Knight R, Alvand A, Palmer AJR, Middleton R, Abram SGF, et al. Patient-Relevant Outcomes Following First Revision Total Knee Arthroplasty, by Diagnosis: An Analysis of Implant Survivorship, Mortality, Serious Medical Complications, and Patient-Reported Outcome Measures Utilizing the National Joint Registry Data Set. JBJS. 2023;105(20):1611-21.
24. Kamath CC, O'Byrne TJ, Lewallen DG, Berry DJ, Maradit Kremers H. Neighborhood-Level Socioeconomic Deprivation, Rurality, and Long-Term Outcomes of Patients Undergoing Total Joint Arthroplasty: Analysis from a Large, Tertiary Care Hospital. Mayo Clin Proc Innov Qual Outcomes. 2022;6(4):337-46.
25. Lindberg-Larsen M, Jørgensen CC, Bæk Hansen T, Solgaard S, Odgaard A, Kehlet H. Re-admissions, re-operations and length of stay in hospital after aseptic revision knee replacement in Denmark: a two-year nationwide study. Bone Joint J. 2014;96-b(12):1649-56.
26. Hardy F, Heyl J, Tucker K, Hopper A, Marchã MJ, Briggs TWR, et al. Data consistency in the English Hospital Episodes Statistics database. BMJ Health Care Inform. 2022;29(1).
27. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.
28. Farrow L, Kennedy JW, Yapp L, Harding T, Walmsley P. Provision of revision knee arthroplasty services across Scotland: A national audit. Knee. 2023;42:312-9.
29. Paterson C, Baarts C, Launsø L, Verhoef MJ. Evaluating complex health interventions: a critical analysis of the 'outcomes' concept. BMC Complement Altern Med. 2009;9:18.

30. Jackson KL, Glasgow RE, Mone MC, Sheng X, Mulvihill SJ, Scaife CL. Does travel distance influence length of stay in elective pancreatic surgery? HPB (Oxford). 2014;16(6):543-9.
31. Cloyd JM, Huang L, Ma Y, Rhoads KF. Predictors of readmission to non-index hospitals after colorectal surgery. The American Journal of Surgery. 2017;213(1):18-23.
32. Kugler CM, De Santis KK, Rombey T, Goossen K, Breuing J, Könsgen N, et al. Perspective of potential patients on the hospital volume-outcome relationship and the minimum volume threshold for total knee arthroplasty: a qualitative focus group and interview study. BMC Health Serv Res. 2021;21(1):633.
33. Mathews JA, Kalson NS, Tarrant PM, Toms AD. Top ten research priorities for problematic knee arthroplasty. Bone Joint J. 2020;102-b(9):1176-82.

Tables and Figures

Figure 1 – Box plot showing association of social deprivation and age on travel distance quintile. Spearman's rank correlation investigating relationship between these factors with travel distance.



Spearman's Rank Correlation

Age and Travel Distance = rho -0.08 pvalue = 0.00126 (very weak correlation as age increases travel distance decreases)

Social Deprivation and Travel distance = rho -0.01 pvalue = 0.6

Table 1 – Baseline demographics and clinical characteristics and raw perioperative outcomes for patients by travel distances quintile

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of Patients	304	303	303	303	303
Deprivation Quintile					
1(Most Deprived)	54 (18%)	87 (29%)	50 (17%)	58 (19%)	57 (19%)
2	48 (16%)	61 (20%)	68 (22%)	58 (19%)	66 (22%)

3	57 (19%)	65 (21%)	63 (21%)	63 (21%)	56 (18%)
4	55 (18%)	43 (14%)	49 (16%)	80 (26%)	75 (25%)
5 (Least Deprived)	90 (30%)	47 (16%)	73 (24%)	44 (15%)	49 (16%)
Sex					
Male	121 (40%)	133 (44%)	135 (45%)	124 (41%)	141 (47%)
Age in years					
16-59	52 (17%)	42 (14%)	56 (18%)	57 (19%)	68 (22%)
60-64	46 (15%)	29 (10%)	35 (12%)	38 (13%)	47 (16%)
65-69	53 (17%)	58 (19%)	56 (18%)	52 (17%)	44 (15%)
70-74	45 (15%)	56 (18%)	53 (17%)	59 (19%)	49 (16%)
75-79	46 (15%)	43 (14%)	47 (16%)	44 (15%)	47 (16%)
>=80	62 (20%)	75 (25%)	56 (18%)	53 (17%)	48 (16%)
Diagnosis					
Mechanical Complication	172 (57%)	198 (65%)	208 (69%)	212 (70%)	192 (63%)
Fracture	26 (9%)	29 (10%)	18 (6%)	10 (3%)	31 (10%)
Progressive OA	39 (13%)	32 (11%)	24 (8%)	24 (8%)	18 (6%)
Hospital Frailty Risk Score					
None	158 (52%)	115 (38%)	149 (49%)	156 (51%)	158 (52%)
Mild	100 (33%)	123 (41%)	106 (35%)	100 (33%)	99 (33%)
Moderate	38 (13%)	56 (18%)	40 (13%)	43 (14%)	36 (12%)
Severe	8 (3%)	9 (3%)	8 (3%)	4 (1%)	10 (3%)
Annual Surgeon Volume					
Volume 0-4	30 (10%)	39 (13%)	31 (10%)	30 (10%)	24 (8%)
Volume 5-9	43 (14%)	46 (15%)	44 (15%)	41 (14%)	35 (12%)
Volume 10-14	89 (29%)	72 (24%)	55 (18%)	63 (21%)	58 (19%)
Volume 15-19	49 (16%)	56 (18%)	44 (15%)	47 (16%)	53 (17%)
Volume 20-24	48 (16%)	64 (21%)	74 (24%)	62 (20%)	47 (16%)
Volume >=25	45 (15%)	26 (9%)	55 (18%)	60 (20%)	86 (28%)
Perioperative Outcomes					
Readmission within 30 days	15(5%)	36(12%)	23(8%)	17(6%)	20(7%)
90 Day Mortality	8(3%)	15(5%)	10(3%)	7(2%)	10(3%)
Prolonged Length of Stay	135(44%)	134(44%)	135(45%)	130(43%)	132(44%)

Table 2 – Multivariable Logistic Regression showing Odds Ratios for primary and secondary outcomes by straight line travel quintile

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	2.23 (95% CI 1.19 to 4.37),p=0.01	1.55 (95% CI 0.79 to 3.13),p=0.21	1.06 (95% CI 0.51 to 2.23),p=0.87	1.44 (95% CI 0.71 to 2.96),p=0.32
90 Day Mortality	1.79 (95% CI 0.66 to 5.17),p=0.27	1.55 (95% CI 0.52 to 4.70),p=0.43	1.72 (95% CI 0.53 to 5.64),p=0.36	1.46 (95% CI 0.49 to 4.53),p=0.50
Prolonged Length of stay	0.90 (95% CI 0.62 to 1.30),p=0.57	1.02 (95% CI 0.71 to 1.46),p=0.91	0.99 (95% CI 0.69 to 1.412),p=0.95	0.96 (95% CI 0.67 to 1.39),p=0.84

Table 3 – Sensitivity analysis exploring road travel distance quintile and primary/secondary outcomes

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	2.11 (95% CI 1.14 to 4.06),p=0.02	1.41 (95% CI 0.71 to 2.84),p=0.33	1.44 (95% CI 0.73 to 2.90),p=0.29	1.16 (95% CI 0.56 to 2.4),p=0.68
90 Day Mortality	1.52 (95% CI 0.58 to 4.21),p=0.40	1.18 (95% CI 0.39 to 3.56),p=0.77	2.42 (95% CI 0.83 to 7.27),p=0.11	0.83 (95% CI 0.25 to 2.64),p=0.75
Prolonged Length of stay	0.69 (95% CI 0.48 to 0.99),p=0.04	0.94 (95% CI 0.66 to 1.35),p=0.74	0.86 (95% CI 0.60 to 1.23),p=0.41	0.95 (95% CI 0.66 to 1.36),p=0.77

Table 4- Sensitivity analysis exploring driving time quintile and primary/secondary outcomes

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	1.19 (95% CI 0.63 to 2.25),p=0.59	1.23 (95% CI 0.65 to 2.35),p=0.52	1.53 (95% CI 0.82 to 2.89),p=0.18	0.92 (95% CI 0.45 to 1.85),p=0.81
90 Day Mortality	2.29 (95% CI 0.84 to 6.76),p=0.12	1.52 (95% CI 0.50 to 4.74),p=0.46	2.63 (95% CI 0.90 to 8.16),p=0.08	0.94 (95% CI 0.26 to 3.23),p=0.91
Prolonged Length of stay	0.82 (95% CI 0.57 to 1.52),p=0.27	0.80 (95% CI 0.56 to 1.15),p=0.23	1.27 (95% CI 0.89 to 1.81),p=0.19	1.01 (95% CI 0.70 to 1.44),p=0.97

Supplementary Material

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

Code	Code description
OPCS-4 codes for knee revision procedures	
O180	Conversion from previous hybrid prosthetic replacement of knee joint using cement
O182	Conversion to hybrid prosthetic replacement of knee joint using cement
O183	Revision of hybrid prosthetic replacement of knee joint using cement
O184	Attention to hybrid prosthetic replacement of knee joint using cement
W400	Conversion from previous cemented total prosthetic replacement of knee joint
W402	Conversion to total prosthetic replacement of knee joint using cement
W403	Revision of total prosthetic replacement of knee joint using cement
W404	Revision of one component of total prosthetic replacement of knee joint using cement
W410	Conversion from previous uncemented total prosthetic replacement of knee joint
W412	Conversion to total prosthetic replacement of knee joint not using cement
W413	Revision of total prosthetic replacement of knee joint not using cement
W414	Revision of one component of total prosthetic replacement of knee joint not using cement
W420	Conversion from previous total prosthetic replacement of knee joint NEC

W422	Conversion to total prosthetic replacement of knee joint NEC
W423	Revision of total prosthetic replacement of knee joint NEC
W424*	Attention to total prosthetic replacement of knee joint NEC
W425	Revision of one component of total prosthetic replacement of knee joint NEC
W522†	Conversion to prosthetic replacement of articulation of bone using cement NEC
W523†	Revision of prosthetic replacement of articulation of bone using cement NEC
W532†	Conversion to prosthetic replacement of articulation of bone not using cement NEC
W533†	Revision of prosthetic replacement of articulation of bone not using cement NEC
W542†	Conversion to prosthetic replacement of articulation of bone NEC
W543†	Revision of prosthetic replacement of articulation of bone NEC
W544*†	Attention to prosthetic replacement of articulation of bone NEC
W553†	Conversion to prosthetic interposition arthroplasty of joint
W564†	Conversion to interposition arthroplasty of joint NEC
W574†	Conversion to excision arthroplasty of joint
W582†	Revision of resurfacing arthroplasty of joint
W603†	Conversion to arthrodesis and extra-articular bone graft NEC
W613†	Conversion to arthrodesis and articular bone graft NEC
W641†	Conversion to arthrodesis and internal fixation NEC
W642†	Conversion to arthrodesis and external fixation NEC

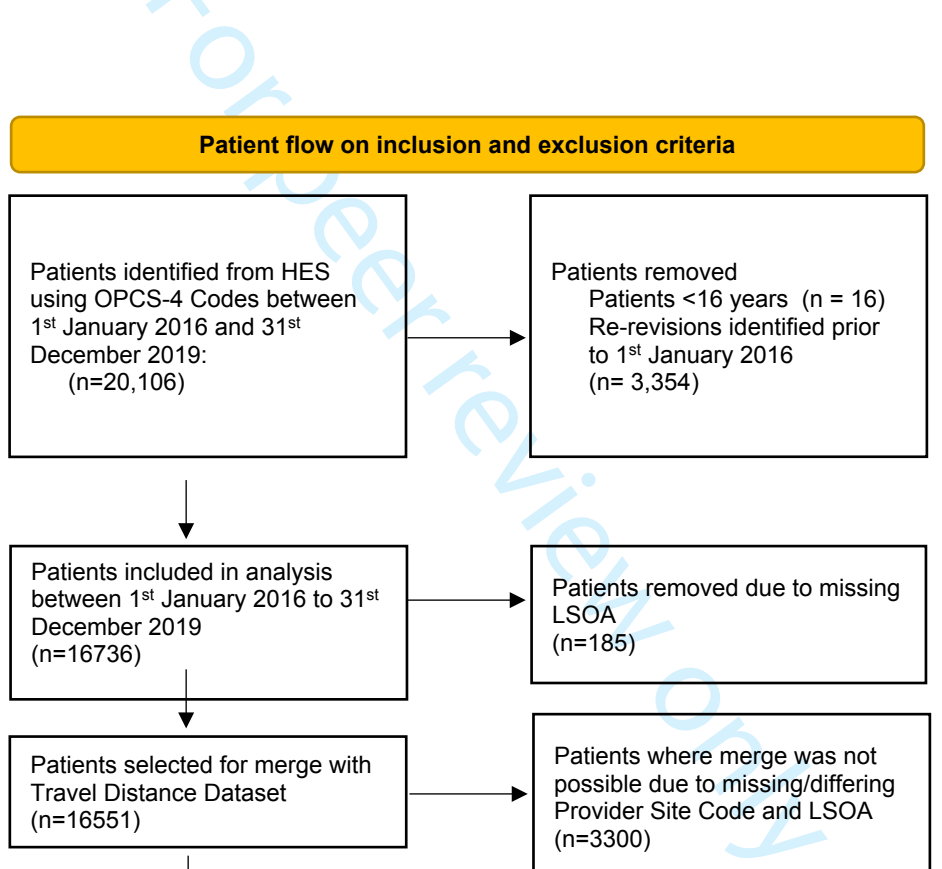
OPCS-4 codes for laterality

Z941	Bilateral
Z942	Left-sided
Z943	Right-sided
ICD-10 codes for Infection	
T845	Infection and inflammatory reaction due to internal joint prosthesis
T846	Infection and inflammatory reaction due to internal fixation device [any site]
T847	Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts
T814	Infection following a procedure, not elsewhere classified
ICD-10 codes for fracture	
M966	Fracture of bone following insertion of orthopaedic implant, joint prosthesis or bone plate
ICD-10 codes for mechanical complications	
T840	Mechanical complication of internal joint prosthesis
T841	Mechanical complication of internal fixation device of bones of limb
T842	Mechanical complication of internal fixation device of other bones
T843	Mechanical complication of other bone devices, implants and grafts
T844	Mechanical complication of other internal orthopaedic devices, imnplants and grafts
ICD-10 codes for osteoarthritis/arthrosis	
M15-	Polyarthrosis
M17-	Gonarthrosis
M19-	Other arthrosis

OPCS-4 = Office of Populations Censuses and Surveys Classification of Interventions and Procedures version 4. ICD-10 = International Statistical

Classification of Diseases and Related Health Problems, tenth revision. * Where OPCS-4 codes Y032 (renewal of prosthesis in organ NOC) or Y037 (removal of prosthesis from organ NOC) were also used. † Where OPCS-4 codes O132 (knee NEC) or Z765 (lower end of femur NEC) or Z774 (upper end of tibia NEC) or Z787 (patella) or Z844 (patellofemoral joint) or Z845 (tibiofemoral joint) or Z846 (knee joint) or Z851 (upper tibiofibular joint) were used to identify knee as the body site.

Supplementary material S2 – Flow of patient inclusion/exclusions



Supplementary material S3 – R code

See separate .R file

BMJ Open

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement for aseptic reasons: An analysis using national administrative data from Hospital Episode Statistics

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-085201.R1
Article Type:	Original research
Date Submitted by the Author:	16-Oct-2024
Complete List of Authors:	Matthews, Alexander; University of Exeter Faculty of Health and Life Sciences, Department of Clinical and Biomedical Sciences; Royal Devon and Exeter NHS Foundation Trust, Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research; Royal Devon and Exeter NHS Foundation Trust, Princess Elizabeth Orthopaedics Centre Evans, Jonathan; University of Exeter, Department of Public Health and Sports Sciences Lamb, Sarah; University of Exeter, College of Medicine and Health Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Gray, William; NHS England and NHS Improvement London Briggs, Tim; NHS England and NHS Improvement London Toms, Andrew ; Princess Elizabeth Orthopaedic Centre
Primary Subject Heading:	Public health
Secondary Subject Heading:	Surgery
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, PUBLIC HEALTH, Health Services

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement for aseptic reasons: An analysis using national administrative data from Hospital Episode Statistics

Names, Affiliations, and positions of all authors

Alex Matthews1,2,3,4, Jonathan P Evans2,3, Jonathan T Evans2,3, Sarah E Lamb3 Andrew Price4,5, William Gray1, Tim Briggs1,6, Andrew Toms2,3

1. Getting It Right First Time programme, NHS England, London, UK
2. Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
3. University of Exeter, Exeter, UK
4. Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, UK

5. Nuffield Orthopaedic Centre, Oxford, UK

6. Royal National Orthopaedic Hospital, Stanmore, London, UK

Correspondence: Alexander Matthews, Getting It Right First Time programme, NHS England and NHS Improvement, Wellington House, Waterloo Road. London, UK, SE1 8UG; telephone: 07401932627 email: alexander.matthews1@nhs.net

Word count: [3316 words]

Short title: Patient Journey Distance and Time for Complex Revision Knee Surgery

Key words: social determinants of health, revision knee replacement, contextual factors in complex health interventions

Twitter handles:

Alex Matthews: @DrAlexMatthews

Tim WR Briggs: @ProfTimBriggs

WilliamKGray1: @WilliamKGray1

GIRFT programme: @NHSGIRFT

Structured Abstract (Word count suggested 250-300)

Objectives

Patients with problematic knee replacements requiring further surgery often have difficulties mobilising and increasingly rely on family support. Evolving practice in England aims to manage these patients in specialised centres with the intention of improving outcomes. This practice will result in longer travel distances and times in this frailer group of patients. We want to examine the types of distances and travel times patients can be expected to travel for complex orthopaedic surgery and to explore concerns of how these impact patient outcomes.

Design

Retrospective observational study from the Hospital Episode Statistics. Multivariable adjusted logistic regression modelling was used to compare the exposure variable with perioperative outcomes

Setting

Patients presenting to tertiary referral centres between 1st January 2016 to 31st December 2019. A tertiary referral centre was defined as a trust performing >70 revisions in the year prior.

Participants

Adult patients undergoing revision total knee replacement procedures for aseptic reasons between 1st January 2016 to 31st December 2019.

Interventions

Patient level travel distance and time was calculated using the department of health Journey Time Statistics.

Main Outcome Measures

The primary outcome is the association of travel distance and time on emergency readmission within 30 days. Secondary outcomes will focus on mortality within 90 days and length of inpatient stay.

Results

1516 patients were treated at 16 tertiary referral centres for non-infected reasons. Patients in the longest driving distance group were expected to travel a median distance of 44.55 miles (IQR 35.90 to 56.30) with an expected median journey time of 66.3 minutes (IQR 57.9 to 80.5). Overall, 30-day readmission was not statistically associated with farther travel distances or driving times.

Conclusions

Patients were expected to travel up to hour for revision knee replacement surgery. There was no association between increasing travel distance and time on perioperative outcomes.

Strengths and limitations of this study

- Our study is the first to describe travel distance and time associations using a large revision knee replacement sample providing data across multiple years
- This data reflects revision knee replacement procedures undertaken across different geographical areas of England
- Owing to differences in the coverage of Hospital Episode Statistics, procedures in hospitals outside of England were not included in this analysis
- Clinical coding practice within HES is known to vary between trusts but this is unlikely to be vary systematically to bias our findings
- This analysis only reports travel times for patients with access to their own transport and does not consider times for those patients using public transport

Introduction

Primary knee replacement is a successful procedure that improves quality of life for the majority of patients.[1] However, at 10 years following a primary knee replacement, about 3.5% of patients will have undergone a revision surgery.[2] The majority of these procedures are carried out due to infection or polyethylene wear of the implant.[3] A failed primary knee replacement represents a life changing transition point where individuals are likely to suffer from pain, reduced mobility as well as dependency on family members.[4] Patients often face multi-step surgery with longer hospital length of stays and higher complication rates.[5, 6]

The orthopaedic GIRFT (Getting It Right First Time) programme was launched in 2012 following the publication of the Orthopaedic National Report.[7] A key recommendation was the centralisation of complex orthopaedic surgery, including revision knee surgery, to specialist centres with the aim of improved patient outcomes. Consequently, revision total knee replacement (rTKR) surgery in the England has evolved into a regional network service model.[8] In doing so, all hospitals performing rTKRs form a network in the respective regions. Less specialist hospitals defined by lower annual case volume thresholds are encouraged to discuss and sometimes refer their caseload to more specialist centres. Several studies based on large revision hip and knee registries has suggested this model carries a lower failure rate defined by the need for further revision surgery.[9-11] Early evidence has suggested reduced early failure rates through the adoption of revision knee networks.[12]

However, this approach to managing patients is inevitably associated with increasing travel distances between some patient's homes and their treating hospital. Expected distances are important to explore, particularly as these patients struggle with mobility, may be unable to drive and may be more reliant on family members.[4] Furthermore, greater travel distances have been associated with higher readmission rates and higher mortality rates following complex vascular surgery.[13] The pick-up rate of early complications, avoiding the need for readmission, may be less in areas further away from the main treatment centre. There is also concern that patients

required to travel greater distances are more likely to be re-admitted to a different hospital resulting in clinical decisions that do not incorporate the primary surgeon and potentially alter outcomes.[14] Subsequently the aims of this paper is to examine if the same association with longer patient travel distance and perioperative outcomes exists following complex orthopaedic surgery with a focus on revision knee replacement surgery performed in high volume tertiary referral centres.

Methods

Design

This study is a retrospective data analysis of observational data from the Hospital Episode Statistics (HES) and Office for National Statistics (ONS) databases. HES data is collected by NHS England for all patients treated at NHS hospitals in England and those treated at private hospitals where treatment was funded by the NHS. This study complies with the recommended reporting guidelines when using HES data[15] and the Strengthening of Reporting of Observational studies in Epidemiology (STROBE) guidelines.[16]

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes[17] and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.[18] The HES data were linked at a patient level to data from the ONS on deaths and date of death, which allowed the identification of patients who had died after their surgery. Linkage was achieved using a unique patient identifier using a previously validated methodology.[19]

Patient travel distances were calculated using the Journey Time Statistics reference document produced by the UK Department of Transport which modelled theoretical journey times between known Lower Layer Super Output Areas (LSOA) of residence and NHS hospital sites.[20] The Journey Time Statistics document is available in the supplementary material section.

Population

An rTKR procedure was defined as a permanent removal or exchange of knee arthroplasty components. This includes a revision of a total knee replacement and a conversion of a unicondylar knee replacement to a total knee replacement. Secondary patellar resurfacing was not included as this represents a simple revision procedure, one that can be carried out in most nonspecialised hospitals. All patients aged ≥ 18 years who underwent a rTKR in a high-volume trust between 1st January 2016 to 31st December 2019 were included in the study population. A high-volume trust was classified as a centre performing > 70 revisions per year. This revision volume threshold for classification represents those proposed by the British Association for Surgeons of the Knee (BASK) Revision Knee Working Group. [21] These are more likely to represent tertiary referral centres where the stratification of more complex work will take place. Annual case volume at each trust was defined as the number of revision cases conducted in the year prior to the index procedure. This measure was preferred over a simple calculation of average annual volume as it accounts for recent experience at the point of surgery. The Office for Population Censuses and Surveys' Classification of Interventions and Procedures version 4 (OPCS-4) codes used to identify rTKR procedures are detailed in **Supplementary material S1.** Where the procedure laterality was not specified, patients were excluded. The flow of patients, with numbers excluded at each point, is summarised in **Supplementary material S2**. To manage population heterogeneity, data were extracted for the period 1st April 2011 to 31st December 2019 and only the first revision for a specific side of the body record in this time period included. [22] Thus, any early revisions on the same side of the body in the four years and nine months preceding the start of the study period were identified and these patients excluded from the study. This aims to exclude the early revision knee replacement failures which have been shown to represent catastrophic failures potentially skewing our results. [22]We excluded revisions for infection as these represent a more variable patient group with a different complication profile [23] and this is further discussed in our study limitations.

Exposure variable

In the analysis straight line travel distance was calculated between a patient's LSOA and the postal codes for the treating hospitals. LSOAs are determined by the Office for National Statistics and are designed for the reporting of small area statistics. Public transport and highways data for England was used to create theoretical journey distances and times from origins to destinations. The resulting travel distances and/or times for each patient were divided into quintiles a priori, following a recently reported methodology.[13] Sensitivity analyses were performed using travel distances by road and peak driving times to account for variation in travel infrastructure between rural and urban areas and to attribute more meaningful results for patients. Peak driving times were calculated by using average traffic speeds for between 7am and 10am.

Outcomes

The primary outcome was emergency readmission within 30 days of discharge from the index surgical hospital. Secondary outcomes included 90-day mortality, and hospital length of stay (LOS) above the median. The LOS outcome was dichotomised into above median or below median LOS of five days.

Statistical Analyses

Data was extracted from a secure, encrypted server controlled by NHS Digital. Data were analysed within a secure, encrypted environment using standard statistical software: R Studio version 2023.09.1+494 (Boston, Massachusetts, USA). The R code and packages used are included in **Supplementary material S3**

Crude comparisons of baseline categorical characteristics and travel distance proximity were calculated. A This data were categorical in nature and summarised as frequency and percentage. In primary analysis a logistic multivariable regression model was constructed to evaluate associations between travel distance quintiles and 30 day readmission, with adjustment for the covariates listed above. The first (shortest) travel distance quintile was used as the reference in all models.

Age, sex , comorbidities and characteristics of initial presentation were included in the logistic regression model. These variables have been shown to influence the risk of complications after R-TKA and therefore represent known confounders.[9, 10, 23]

This also included data on economic deprivation measured using the Index of Multiple Deprivation (IMD).[24] The IMD gives the LSOA where the patient lives a score based on a range of measures of deprivation. IMD was categorised into quintiles, based on all-England data, for analysis. A spearman's rank correlation was performed to investigate the relationship between IMD score and travel distances. Health co-morbidity was quantified using the Hospital Frailty Risk Score (HFRS). HFRS identifies frailty based on the occurrence of any of 109 International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during any hospital admissions in the two years prior to, and for, the index admission.

All secondary outcomes were binary and analysed using the same multivariable logistic regression. Multicollinearity was assessed with reference to variance inflation factor and Shapiro-Wilcox test of normality. Model fit was assessed with reference to the pseudo R^2 values.

A supplementary analysis is available analysing travel times and distances as a continuous variable with the primary outcome. Please see supplementary material S4

Results

Demographic characteristics and co-morbidities

The 1,516 patients in the overall study population, were divided into quintiles of travel distance of 303 or 304 patients each. The median straight line travel distance for quintile one was 2.3 miles (IQR 1.3 to 3.1). For the fifth travel quintile, median distance was 33.5 miles (IQR 25.5 to 41.1). Baseline co-morbidities and demographic characteristics were broadly similar among the travel distance quintiles (**Table 1**). Travel distance was not strongly correlated with age or social deprivation (**Figure 1**)

Table 1 – Baseline demographics and clinical characteristics and raw perioperative outcomes for patients by travel distances quintile

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of Patients	304	303	303	303	303

Deprivation Quintile					
1(Most Deprived)	54 (18%)	87 (29%)	50 (17%)	58 (19%)	57 (19%)
2	48 (16%)	61 (20%)	68 (22%)	58 (19%)	66 (22%)
3	57 (19%)	65 (21%)	63 (21%)	63 (21%)	56 (18%)
4	55 (18%)	43 (14%)	49 (16%)	80 (26%)	75 (25%)
5 (Least Deprived)	90 (30%)	47 (16%)	73 (24%)	44 (15%)	49 (16%)
Sex					
Male	121 (40%)	133 (44%)	135 (45%)	124 (41%)	141 (47%)
Age in years					
16-59	52 (17%)	42 (14%)	56 (18%)	57 (19%)	68 (22%)
60-64	46 (15%)	29 (10%)	35 (12%)	38 (13%)	47 (16%)
65-69	53 (17%)	58 (19%)	56 (18%)	52 (17%)	44 (15%)
70-74	45 (15%)	56 (18%)	53 (17%)	59 (19%)	49 (16%)
75-79	46 (15%)	43 (14%)	47 (16%)	44 (15%)	47 (16%)
>=80	62 (20%)	75 (25%)	56 (18%)	53 (17%)	48 (16%)
Diagnosis					
Mechanical Complication	172 (57%)	198 (65%)	208 (69%)	212 (70%)	192 (63%)
Fracture	26 (9%)	29 (10%)	18 (6%)	10 (3%)	31 (10%)
Progressive OA	39 (13%)	32 (11%)	24 (8%)	24 (8%)	18 (6%)
Hospital Frailty Risk Score					
None	158 (52%)	115 (38%)	149 (49%)	156 (51%)	158 (52%)
Mild	100 (33%)	123 (41%)	106 (35%)	100 (33%)	99 (33%)
Moderate	38 (13%)	56 (18%)	40 (13%)	43 (14%)	36 (12%)
Severe	8 (3%)	9 (3%)	8 (3%)	4 (1%)	10 (3%)
Annual Surgeon Volume					
Volume 0-4	30 (10%)	39 (13%)	31 (10%)	30 (10%)	24 (8%)
Volume 5-9	43 (14%)	46 (15%)	44 (15%)	41 (14%)	35 (12%)
Volume 10-14	89 (29%)	72 (24%)	55 (18%)	63 (21%)	58 (19%)
Volume 15-19	49 (16%)	56 (18%)	44 (15%)	47 (16%)	53 (17%)
Volume 20-24	48 (16%)	64 (21%)	74 (24%)	62 (20%)	47 (16%)
Volume >=25	45 (15%)	26 (9%)	55 (18%)	60 (20%)	86 (28%)
Perioperative Outcomes					
Readmission within 30 days	15(5%)	36(12%)	23(8%)	17(6%)	20(7%)
90 Day Mortality	8(3%)	15(5%)	10(3%)	7(2%)	10(3%)
Prolonged Length of Stay	135(44%)	134(44%)	135(45%)	130(43%)	132(44%)

Association between travel distance and readmission, mortality and extended hospital stay

Overall, 111 patients were readmitted within 30 days. Crude comparisons of proportions readmitted within 30 days for each travel distance quintiles revealed a higher rate of readmission for the second travel quintile. In multivariable adjusted logistic regression, there was no statistically significant association between travel distance and readmission within 30 days (**Table 2**). Odds for 30-day readmission was 1.44 (95% CI 0.71 to 2.96, P 0.32) for Q5 compared with Q1. Increased travel distance was not associated with a significant change in the odds of death within 90 days (OR for Q5 vs Q1, 1.46 (95% CI 0.49 to 4.53, P 0.682)). Travel distance quintile was not associated with prolonged length of hospital stay related to the index surgery after multivariable adjustment (OR for Q5 vs Q1, 0.96 (95% CI 0.67 to 1.39, P 0.84)).

Table 2 – Multivariable Logistic Regression showing Odds Ratios for primary and secondary outcomes by straight line travel quintile

	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Readmission with 30 days	2.23 (95% CI 1.19 to 4.37),p=0.01	1.55 (95% CI 0.79 to 3.13),p=0.21	1.06 (95% CI 0.51 to 2.23),p=0.87	1.44 (95% CI 0.71 to 2.96),p=0.32
90 Day Mortality	1.79 (95% CI 0.66 to 5.17),p=0.27	1.55 (95% CI 0.52 to 4.70),p=0.43	1.72 (95% CI 0.53 to 5.64),p=0.36	1.46 (95% CI 0.49 to 4.53),p=0.50
Prolonged Length of stay	0.90 (95% CI 0.62 to 1.30),p=0.57	1.02 (95% CI 0.71 to 1.46),p=0.91	0.99 (95% CI 0.69 to 1.412),p=0.95	0.96 (95% CI 0.67 to 1.39),p=0.84

Real world travel distance and outcomes

The above results used straight line travel distance between patient' LSOA and treating hospital. A sensitivity analysis using actual patient travel distances using the shortest possible road route was performed **(Table 3)**. The median driving distance by the shortest possible road route for the closest quintile was 3.40 miles (IQR 2.00 to 4.40). The furthest quintile median driving distance was 44.55 miles (IQR 35.90 to 56.30) This analysis showed no association between driving distance and all perioperative outcomes. The OR for readmission within 30 days in Q5 vs Q1 was 1.16(0.56-2.41, p value = 0.68).

Table 3 – Sensitivity analysis exploring road travel distance quintile and primary/secondary outcomes

	Quintile 2	Quintile 3	Quintile 4	Quintile 5

Readmission with 30 days	2.11 (95% CI 1.14 to 4.06),p=0.02	1.41 (95% CI 0.71 to 2.84),p=0.33	1.44 (95% CI 0.73 to 2.90),p=0.29	1.16 (95% CI 0.56 to 2.4),p=0.68
90 Day Mortality	1.52 (95% CI 0.58 to 4.21),p=0.40	1.18 (95% CI 0.39 to 3.56),p=0.77	2.42 (95% CI 0.83 to 7.27),p=0.11	0.83 (95% CI 0.25 to 2.64),p=0.75
Prolonged Length of stay	0.69 (95% CI 0.48 to 0.99),p=0.04	0.94 (95% CI 0.66 to 1.35),p=0.74	0.86 (95% CI 0.60 to 1.23),p=0.41	0.95 (95% CI 0.66 to 1.36),p=0.77

Journey drive times and outcomes

A further sensitivity analysis using driving times was calculated (Table 4). The median drive time for quintile one was 12.6 minutes (IQR 8.7 to 15.3). For the fifth time quintile the median was 66.3 minutes (IQR 57.9 to 80.5). No statistical association was found between drive time and perioperative outcomes. The OR for readmission within 30 days in Q5 vs Q1 was 0.92 (0.45 – 1.85, p value = 0.81)

Table 4- Sensitivity analysis exploring driving time quintile and primary/secondary outcomes

	Quintile 2	**Quintile 3**	**Quintile 4**	**Quintile 5**
Readmission with 30 days	1.19 (95% CI 0.63 to 2.25),p=0.59	1.23 (95% CI 0.65 to 2.35),p=0.52	1.53 (95% CI 0.82 to 2.89),p=0.18	0.92 (95% CI 0.45 to 1.85),p=0.81
90 Day Mortality	2.29 (95% CI 0.84 to 6.76),p=0.12	1.52 (95% CI 0.50 to 4.74),p=0.46	2.63 (95% CI 0.90 to 8.16),p=0.08	0.94 (95% CI 0.26 to 3.23),p=0.91
Prolonged Length of stay	0.82 (95% CI 0.57 to 1.52),p=0.27	0.80 (95% CI 0.56 to 1.15),p=0.23	1.27 (95% CI 0.89 to 1.81),p=0.19	1.01 (95% CI 0.70 to 1.44),p=0.97

Discussion

Statement of principal findings

We present a multi-hospital site retrospective analysis of patients undergoing revision knee replacement surgery at tertiary referral centres in England. In this analysis of 1,516 patients undergoing aseptic revision knee replacement surgery, we did not observe an association between distance and time travelled for revision surgery and readmission within 30 days. Patients in the longest driving time category were expected to travel for a median time of more than one hour.

Strengths and weaknesses of the study

The findings of this study should be interpreted in view of several limitations. Firstly, this analysis used observational data from a large administrative dataset covering all NHS-funded procedures conducted in England. As with all datasets we are limited in the amount of detail provided regarding presentation. We chose to categorise a high-volume centre by trust to accurately capture surgical experience. All NHS hospitals in England are run by hospital trusts which typically involve between one and four hospitals within a catchment area standardising their practice. It is common practice for specialist orthopaedic surgeons to move between these sites delivering the same procedures. Our study involved 35 hospital sites run by 16 trusts. We acknowledge this is a weakness of our study as this may not be representative of all trusts. The indication for revision coded as mechanical complication encompasses several common indications such as aseptic loosening, instability and malalignment. Reassuringly these indications have similar length of stay, and perioperative outcomes.[23] Differences exist in their re-revision rate, however this was not an outcome of focus in our study. It is likely that the complexity of the surgery undertaken may vary within the different indications for revision. Evidence suggests that operative surgical time is related to increased length of stay in aseptic revision knee replacement.[25]. There is a lack of granular data for revisions due to infection and therefore we excluded this patient group as some readmissions for this patient group may represent planned readmissions. There is also a lack of granular clinical data using HES for each readmission, therefore we cannot ascertain precise reasons

for readmissions, but we assume are related to a post-surgical complication. Clinical coding practice within HES is known to vary across trusts.[26] As an example, some trusts may be more consistent in coding comorbidities, and this may have created some bias. However, this is unlikely to vary systematically with travel distances and so significantly bias our findings. We acknowledge the relatively short travel distances in this population compared to examples from the United States as such the results of this study may not be generalisable to larger geographical areas or less mature healthcare systems. However, the upper quintile in our study represents a substantial journey distance and time for our patient cohort where poor mobility is a significant factor affecting their care. This analysis does not consider journey times of those who may not have access to a car and instead chose to take public transport.

Strengths and weaknesses in relation to other studies, discussing important differences in results

This is the first study to analyse the potential impact of patient travel distances on patients receiving complex orthopaedic surgery. The findings that longer travel distances are not associated with inferior outcomes is an important part of the evaluation of the assumptions and context behind the establishment of revision knee networks.[27] This study has shown that concerns of introducing a network in larger geographical regions, for example in Scotland where longer patient travel distances and times are common, may be less important.[28] This is particularly useful as regions explore the geography of their revision networks and during summative outcome assessment of this complex health intervention.[29]

It may be seen as surprising that no association between travel distance and prolonged length of hospital stay was identified. An expectation exists of increasing difficulties being encountered with the discharge of patients living greater distances from their treating hospital, which has been observed in patients following elective pancreatic surgery.[30] This is also an observation seen in patients being treated in specialist vascular centres in the United States which led to the recommendation of additional care coordination and follow up efforts. However, the geography of the population in these studies was much larger with significantly longer travel distances.

We did not observe a strong correlation between social deprivation status and age of the patient with longer travel distances. It is reassuring that access to treatment for older patients and those from poor socioeconomic backgrounds is unaffected by travel distance. However, there may be patients who refused to travel to a specialist centre and opted for treatment at their local centre.

Meaning of the study: possible explanations and implications for clinicians and policymakers

The organisation and delivery of revision knee services in England has recently undergone a substantial change and now such services are provided around regional networks of care. This promises substantial advantages to the increasing number of patients with problematic knee replacements in our ageing population who will benefit from regional expertise.[8] However, it is unknown the impact of patients residing farther from tertiary referral centres, particularly rural patients who may encounter additional difficulties associated with greater travel distance. A recent study following the outcomes of aortic surgery found that longer travel distances are associated with inferior perioperative outcomes[13]. Similar associations have been found in postoperative colorectal surgery patients [31]. As such our results are reassuring to policy makers and clinicians.

Unanswered questions and future research

There is a scarcity of evidence evaluating the patient perception of complex health interventions such as network models of care. Recent work by Kugler et al has demonstrated the willingness of patients to travel further for better outcomes in the context of total knee replacement surgery. [32] Nevertheless, patient perceptions of travelling further for their treatment should be a focus for future research in the context of revision knee patients, particularly as this is one of the top ten research priorities identified by the James Lind Alliance priority setting partnership.[33]

Conclusion

We did not observe an association in our study population between 30-day readmission rates and increasing travel distances or times between a patient's home and their treating hospital in revision knee replacement. This paper is the first to explore the relationship between travel distance and complex orthopaedic surgery and informs some concerns regarding the creation of a centralised revision knee network. This information is of utility to surgical providers and commissioners of healthcare services. Furthermore, it can inform patient-led decision making and the exploration of perceptions surrounding travelling for complex surgery. Although this is the first assessment in complex orthopaedic surgery, a prospective analysis will be undertaken as part of the ongoing auditing of revision knee networks in England.

Supplementary material and figures

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

See separate file named supplementary material S1

Supplementary material S2 – Flow of patient inclusion/exclusions

-See attached file named Supplementary Material S2

Supplementary material S3 – R Code

See attached file named Supplementary Material S3

Supplementary material S4 – Relationship between Travel distances and times modelled as a continuous variable with primary outcome (readmission within 30 days)

See attached file named supplementary Material S4

Figure 1 – Box plot showing association of social deprivation and age on travel distance quintile. Spearman's rank correlation investigating relationship between these factors with travel distance.

See attached files called 'Figure 1 – Deprivation and Travel Distance' AND 'Figure 1 – Age and Travel Distance'

Contributorship

Alex Matthews: Conceptualisation, Methodology, Project Administration, Investigation, Data Curation, Formal Analysis, Visualisation, Writing - original draft, Writing - review and editing. This author is the guarantor and is responsible for the content

Jonathan P Evans: Conceptualisation, Supervision, Writing - review & editing

Jonathan T Evans: Supervision, Writing - review and editing

Sarah E Lamb: Conceptualisation, Supervision, Writing - review and editing

Andrew Price: Conceptualisation, Supervision, Writing - review and editing

William Gray: Conceptualisation, Supervision, Methodology, Writing - review and editing

Tim Briggs: Supervision, Writing - review and editing

Andrew Toms: Conceptualisation, Supervision, Writing - review and editing

Public and Patient Involvement statement

The study's chief investigator (AT) led the James Lind Alliance 'Revision Knee Replacement' priority setting partnership. This group of patients, carers and health care professionals identified the need to investigate the best way of organising revision knee replacement surgery to improve patient outcomes as one of their top 10 research questions. Patients were therefore directly involved in the development of the study's aims and objectives. The results of the study will be disseminated to the members of this group prior to publication.

Copyright/licence for publication Statement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing Interests

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical Approval

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data. Ethical approval was not required.

Funding

No funding was obtained to carry out this study

Data Sharing

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

References

1. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet. 2018;392(10158):1672-82.
2. Sabah SA, Hedge EA, Abram SGF, Alvand A, Price AJ, Hopewell S. Patient-reported outcome measures following revision knee replacement: a review of PROM instrument utilisation and measurement properties using the COSMIN checklist. BMJ Open. 2021;11(10):e046169.
3. Sabah S, Knight R, Alvand A, Murray D, Petrou S, Beard D, Price A. No exponential rise in revision knee replacement surgery over the past 15 years: an analysis from the National Joint Registry. Osteoarthritis and Cartilage. 2022;30(12):1670-9.
4. Rowland T, Lindberg-Larsen M, Santy-Tomlinson J, Jensen CM. A qualitative study of patients' experiences before, during and after surgical treatment for periprosthetic knee joint infection; “I assumed it had to be like that … ”. International Journal of Orthopaedic and Trauma Nursing. 2023;48:100992.
5. Garbarino LJ, Gold PA, Sodhi N, Anis HK, Ehiorobo JO, Boraiah S, et al. The effect of operative time on in-hospital length of stay in revision total knee arthroplasty. Ann Transl Med. 2019;7(4):66.
6. Liodakis E, Bergeron SG, Zukor DJ, Huk OL, Epure LM, Antoniou J. Perioperative Complications and Length of Stay After Revision Total Hip and Knee Arthroplasties: An Analysis of the NSQIP Database. The Journal of Arthroplasty. 2015;30(11):1868-71.
7. Briggs TWR. A National Review of Adult Elective Orthopaedic Services in England. In: Time GIRF, editor. 2015.
8. Kalson NS, Mathews JA, Phillips JRA, Baker PN, Price AJ, Toms AD, et al. Revision knee replacement surgery in the NHS: A BASK surgical practice guideline. The Knee. 2021;29:353-64.
9. Yapp LZ, Walmsley PJ, Moran M, Clarke JV, Simpson AHRW, Scott CEH. The effect of hospital case volume on re-revision following revision total knee arthroplasty. The Bone & Joint Journal. 2021;103-B(4):602-9.
10. Halder AM, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Re-Revision Rate Following Aseptic Revision Total Knee Arthroplasty: An Analysis of 23,644 Cases. J Arthroplasty. 2020;35(4):1054-9.
11. Jeschke E, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Revision Rate and Mortality Following Revision Total Hip Arthroplasty: An Analysis of 17,773 Cases. J Arthroplasty. 2019;34(9):2045-50.
12. Bloch BV, Matar HE, Berber R, Gray WK, Briggs TWR, James PJ, Manktelow ARJ. The impact of a revision arthroplasty network on patient outcomes. Bone Joint J. 2023;105-b(6):641-8.
13. Feldman ZM, Zheng X, Mao J, Sumpio BJ, Mohebali J, Chang DC, et al. Longer patient travel distance is associated with increased non-index readmission after complex aortic surgery. Journal of Vascular Surgery. 2023;77(6):1607-17.e7.
14. Saunders RS, Fernandes-Taylor S, Kind AJ, Engelbert TL, Greenberg CC, Smith MA, et al. Rehospitalization to primary versus different facilities following abdominal aortic aneurysm repair. J Vasc Surg. 2014;59(6):1502-10, 10.e1-2.
15. Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes--systematic review and recommendations for reporting. Eur J Public Health. 2013;23(1):86-92.

16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
17. NHS Digital. Hospital Episode Statistics(HES) Analysis Guide. London, UK: NHS Digital; 2018.
18. Information Standard Board for Health and Social Care. Anonymisation Standard for Publishing Health and Social Care Data Specification. London, UK: NHS Digital; 2013.
19. NHS Digital. Delivering linked datasets to support health and care delivery and research. London, UK: NHS Digital; 2018.
20. Department for Transport. Journey Time Statistics, notes and definitions. London, UK: Department for Transport; 2019.
21. Kalson NS, Mathews JA, Miles J, Bloch BV, Price AJ, Phillips JRA, et al. Provision of revision knee surgery and calculation of the effect of a network service reconfiguration: An analysis from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Knee. 2020;27(5):1593-600.
22. Deere K, Whitehouse MR, Kunutsor SK, Sayers A, Price AJ, Mason J, Blom AW. How long do revised and multiply revised knee replacements last? An analysis of the National Joint Registry. Lancet Rheumatol. 2021;3(6):e438-e46.
23. Sabah SA, Knight R, Alvand A, Palmer AJR, Middleton R, Abram SGF, et al. Patient-Relevant Outcomes Following First Revision Total Knee Arthroplasty, by Diagnosis: An Analysis of Implant Survivorship, Mortality, Serious Medical Complications, and Patient-Reported Outcome Measures Utilizing the National Joint Registry Data Set. JBJS. 2023;105(20):1611-21.
24. Kamath CC, O'Byrne TJ, Lewallen DG, Berry DJ, Maradit Kremers H. Neighborhood-Level Socioeconomic Deprivation, Rurality, and Long-Term Outcomes of Patients Undergoing Total Joint Arthroplasty: Analysis from a Large, Tertiary Care Hospital. Mayo Clin Proc Innov Qual Outcomes. 2022;6(4):337-46.
25. Lindberg-Larsen M, Jørgensen CC, Bæk Hansen T, Solgaard S, Odgaard A, Kehlet H. Re-admissions, re-operations and length of stay in hospital after aseptic revision knee replacement in Denmark: a two-year nationwide study. Bone Joint J. 2014;96-b(12):1649-56.
26. Hardy F, Heyl J, Tucker K, Hopper A, Marchã MJ, Briggs TWR, et al. Data consistency in the English Hospital Episodes Statistics database. BMJ Health Care Inform. 2022;29(1).
27. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.
28. Farrow L, Kennedy JW, Yapp L, Harding T, Walmsley P. Provision of revision knee arthroplasty services across Scotland: A national audit. Knee. 2023;42:312-9.
29. Paterson C, Baarts C, Launsø L, Verhoef MJ. Evaluating complex health interventions: a critical analysis of the 'outcomes' concept. BMC Complement Altern Med. 2009;9:18.
30. Jackson KL, Glasgow RE, Mone MC, Sheng X, Mulvihill SJ, Scaife CL. Does travel distance influence length of stay in elective pancreatic surgery? HPB (Oxford). 2014;16(6):543-9.
31. Cloyd JM, Huang L, Ma Y, Rhoads KF. Predictors of readmission to non-index hospitals after colorectal surgery. The American Journal of Surgery. 2017;213(1):18-23.
32. Kugler CM, De Santis KK, Rombey T, Goossen K, Breuing J, Könsgen N, et al. Perspective of potential patients on the hospital volume-outcome relationship and the

minimum volume threshold for total knee arthroplasty: a qualitative focus group and interview study. BMC Health Serv Res. 2021;21(1):633.

33. Mathews JA, Kalson NS, Tarrant PM, Toms AD. Top ten research priorities for problematic knee arthroplasty. Bone Joint J. 2020;102-b(9):1176-82.

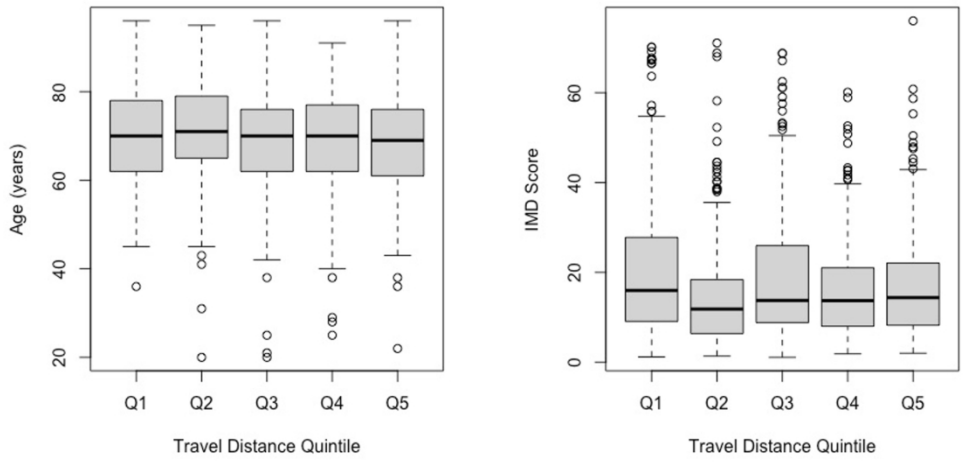


Figure 1 – Box plot showing association of social deprivation and age on travel distance quintile. Spearman's rank correlation investigating relationship between these factors with travel distance.

202x113mm (300 x 300 DPI)

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

Code	Code description
OPCS-4 codes for knee revision procedures	
O180	Conversion from previous hybrid prosthetic replacement of knee joint using cement
O182	Conversion to hybrid prosthetic replacement of knee joint using cement
O183	Revision of hybrid prosthetic replacement of knee joint using cement
O184	Attention to hybrid prosthetic replacement of knee joint using cement
W400	Conversion from previous cemented total prosthetic replacement of knee joint
W402	Conversion to total prosthetic replacement of knee joint using cement
W403	Revision of total prosthetic replacement of knee joint using cement
W404	Revision of one component of total prosthetic replacement of knee joint using cement
W410	Conversion from previous uncemented total prosthetic replacement of knee joint
W412	Conversion to total prosthetic replacement of knee joint not using cement
W413	Revision of total prosthetic replacement of knee joint not using cement
W414	Revision of one component of total prosthetic replacement of knee joint not using cement
W420	Conversion from previous total prosthetic replacement of knee joint NEC
W422	Conversion to total prosthetic replacement of knee joint NEC

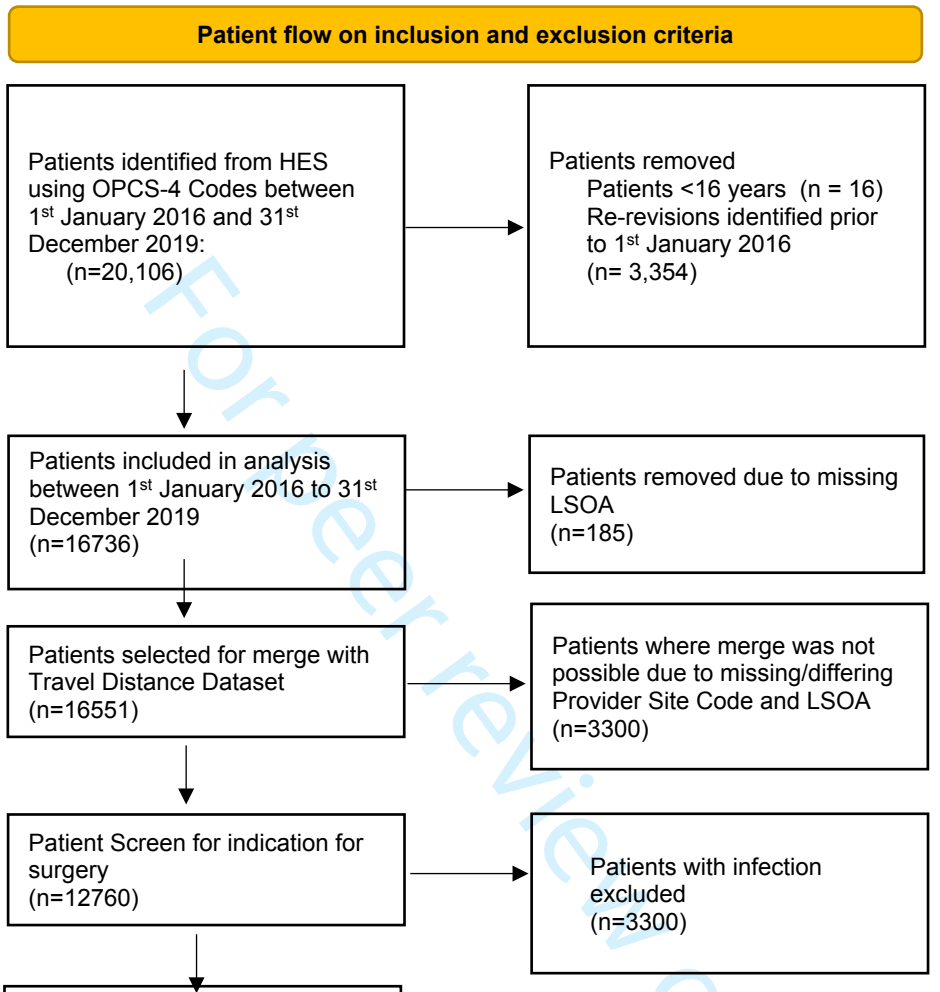
W423	Revision of total prosthetic replacement of knee joint NEC
W424*	Attention to total prosthetic replacement of knee joint NEC
W425	Revision of one component of total prosthetic replacement of knee joint NEC
W522†	Conversion to prosthetic replacement of articulation of bone using cement NEC
W523†	Revision of prosthetic replacement of articulation of bone using cement NEC
W532†	Conversion to prosthetic replacement of articulation of bone not using cement NEC
W533†	Revision of prosthetic replacement of articulation of bone not using cement NEC
W542†	Conversion to prosthetic replacement of articulation of bone NEC
W543†	Revision of prosthetic replacement of articulation of bone NEC
W544*†	Attention to prosthetic replacement of articulation of bone NEC
W553†	Conversion to prosthetic interposition arthroplasty of joint
W564†	Conversion to interposition arthroplasty of joint NEC
W574†	Conversion to excision arthroplasty of joint
W582†	Revision of resurfacing arthroplasty of joint
W603†	Conversion to arthrodesis and extra-articular bone graft NEC
W613†	Conversion to arthrodesis and articular bone graft NEC
W641†	Conversion to arthrodesis and internal fixation NEC
W642†	Conversion to arthrodesis and external fixation NEC
OPCS-4 codes for laterality	
Z941	Bilateral

Z942	Left-sided
Z943	Right-sided
ICD-10 codes for Infection	
T845	Infection and inflammatory reaction due to internal joint prosthesis
T846	Infection and inflammatory reaction due to internal fixation device [any site]
T847	Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts
T814	Infection following a procedure, not elsewhere classified
ICD-10 codes for fracture	
M966	Fracture of bone following insertion of orthopaedic implant, joint prosthesis or bone plate
ICD-10 codes for mechanical complications	
T840	Mechanical complication of internal joint prosthesis
T841	Mechanical complication of internal fixation device of bones of limb
T842	Mechanical complication of internal fixation device of other bones
T843	Mechanical complication of other bone devices, implants and grafts
T844	Mechanical complication of other internal orthopaedic devices, imnplants and grafts
ICD-10 codes for osteoarthritis/arthrosis	
M15-	Polyarthrosis
M17-	Gonarthrosis
M19-	Other arthrosis

OPCS-4 = Office of Populations Censuses and Surveys Classification of Interventions and Procedures version 4. ICD-10 = International Statistical Classification of Diseases and Related Health Problems, tenth revision. * Where

OPCS-4 codes Y032 (renewal of prosthesis in organ NOC) or Y037 (removal of prosthesis from organ NOC) were also used. † Where OPCS-4 codes O132 (knee NEC) or Z765 (lower end of femur NEC) or Z774 (upper end of tibia NEC) or Z787 (patella) or Z844 (patellofemoral joint) or Z845 (tibiofemoral joint) or Z846 (knee joint) or Z851 (upper tibiofibular joint) were used to identify knee as the body site.

Supplementary material S2 – Flow of patient inclusion/exclusions



```
####Start####

#Travel Times and Perioperative Outcomes in Revision Knee Replacement

####Preparation of Data####
#load HES data

RTKA2023 <- read.csv("~/Desktop/RTKA 06-09-23 CSV.csv")

RTKA2023 <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/RTKA 06-09-23 CSV.csv")



#table only shows first 50 columns but we know there are 51 columns. Write this generic code to change preferences


rstudioapi::writeRStudioPreference("data_viewer_max_columns", 1000L)

#Find number of incomplete cases in the data

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)

#There are 4 entried with missing data only in the age group

#check how many incomplete entries in age of patient column

sum(!complete.cases(RTKA2023$age_of_patient))

#In case of missing values there are only 4 for age of patient
#Can use imputation but given small number decision to remove
#What is the mean age of the patients

mean(RTKA2023$age_of_patient, na.rm = TRUE)

#mean age excluding missing values is 70
summary(RTKA2023$age_of_patient, na.rm = TRUE)

#Check age is normally distributed

hist(RTKA2023$age_of_patient)


#we must remove the missing data by coding it NA first
```

```
RTKA2023$age_of_patient[RTKA2023$age_of_patient ==""] <- NA



#Remove NA rows

RTKA2023 <- RTKA2023[!is.na(RTKA2023$age_of_patient),]

#Now check number of missing values

sum(!complete.cases(RTKA2023$age_of_patient))
#Now states 0 missing values

#There are other missing values for IMD decile
##In fact there are 690 IMD score missing values

sum(!complete.cases(RTKA2023$IMD_score))


hist(RTKA2023$IMD_score)
#IMD score is non normally distributed

summary(RTKA2023$IMD_score, na.rm = TURE)

#Median IMD score is 15.

#Use imputation to impute median for missing value

RTKA2023$IMD_score[is.na(RTKA2023$IMD_score)] <- 15

#Check imputation complete

sum(!complete.cases(RTKA2023$IMD_score))

#Now showing 0 missing values

#Next attach IMD decile number 6 to the missing values. As a score of 15 equates to the 6th
decile

RTKA2023$IMD_decile[is.na(RTKA2023$IMD_decile)] <- 6

#Check duplicate entry spells

duplicates <- RTKA2023[duplicated(RTKA2023),]
```

```
print(duplicates)


duplicated(RTKA2023$P_Spell_ID, fromLast = TRUE)

#No duplicates in data


#Frequencies of revisions by volume

as.numeric(RTKA2023$TV12mo)


#frequencies of revisions by trust volume
table(RTKA2023$TVcat)

#Proportions by trust volume

prop.table(table(RTKA2023$TVcat))



#Some entried are blank but are read as real values and not missing data
#The table between age and sex shows three variables here
#The dataset contains non standard missing values that are not recognised as NA
#Replace empty strings with NA

RTKA2023[RTKA2023 == ""] <- NA

#Check this has registered

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)



#Then remove IMD quintile with NA in rows as only 132 missing
#Remove this column

RTKA2023$IMD_quintile <- NULL

#Column with LSOA_2011_Code has 171 missing. To look at travel times you need to
remove these rows

RTKA2023 <- RTKA2023[!is.na(RTKA2023$LSOA_2011_Code),]

missing_data <- colSums(is.na(RTKA2023))
```

```
print(missing_data)


#Load Travel times data

TRAVELTIMES <- read.csv("~/Desktop/Drive time and Miles reference file.csv")

LSOAREF <- read.csv("~/Desktop/LSOA Matrix.csv")


#Join data but The data is too big so we need to do this using SQL

install.packages("RSQLite")
library(RSQLite)

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "times", TRAVELTIMES)
dbWriteTable(con, "lsoa", LSOAREF)

query <- "
Select *
FROM times
JOIN lsoa ON times.LSOAName = lsoa.LSOA11NM"

result <- dbGetQuery(con, query)

#Write Dataframes

write.csv(result, "~/Desktop/JOINLSOATRAVEL.csv")

result<- read.csv("~/Desktop/JOINLSOATRAVEL.csv")


#####Now join this data to your revisions spreadsheet using key identifiers LSOA and
Organisation site code

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "revisions2", RTKA2023)
dbWriteTable(con, "travel2", result)

query <- "
Select *
FROM revisions2
JOIN travel2 ON revisions2.LSOA_2011_Code = travel2.LSOA11CD AND revisions2.Sitecode =
travel2.ProviderSiteCode"
```

```
result1 <- dbGetQuery(con, query)

write.csv(result1, "~/Desktop/REVISIONSTRAVELTIMES.csv")

result2<- read.csv("~/Desktop/REVISIONSTRAVELTIMES.csv")

#Check your data for missing values

missing_data <- colSums(is.na(result1))
print(missing_data)


####Prepare Outcomes, Exposure variable and co-variates ####

#Set up outcomes

#Replace NA's in the Read columns with N

result1$Read30 <- ifelse(is.na(result1$Read30), 'N', result1$Read30)
result1$Read90 <- ifelse(is.na(result1$Read90), 'N', result1$Read90)

result1$Read30days <- ifelse(result1$Read30 == "Y", 1, 0)
#readmission for 90 days
result1$Read90days <- ifelse(result1$Read90 == "Y", 1, 0)



#Set up your co-variates

result1$HFRS_Band = as.factor(result1$HFRS_Band)
result1$HFRS_Band = relevel(result1$HFRS_Band, ref = 'None')

result1$POD = as.factor(result1$POD)
result1$POD = relevel(result1$POD, ref = 'EL')

table(result1$POD)



#Sensitivity analysis for only aseptic cases

result2 <- subset(result1, infection == 0)


#Subset the data to focus on teritary centres only determined by volume >59. Therefore
include volume categories D,E & F
```

#Trust volume was categorised as < 20, 20-39, 40-59, 60-79, 80-99 and ≥ 100 procedures in the previous year. These categories were chosen to ensure that there were more than ten trusts/surgeons represented in each category and that the categorisations were meaningful and consistent.

```
traveltimesrev <- subset(result2, TVcat == "D" | TVcat == "E" | TVcat == "F")

#≥70 a year BASK recommendations for Major Revision Centres

result2$MRC <- ifelse(result2$TV12mo > 70, 1, 0)

traveltimesrev <- subset(result2, MRC == 1)

#Create travel time quintile variable

quintiles <- quantile(traveltimesrev$DistanceMiles, probs = seq(0,1,0.2), na.rm=TRUE)

traveltimesrev$distancequintile <- cut(traveltimesrev$DistanceMiles, breaks = quintiles,
labels = c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)

#Add new outcome variable LOS>median

summary(traveltimesrev$Spell_Los)
#Spell length of stay median is 5 days

traveltimesrev$LongLOS <- ifelse(traveltimesrev$Spell_Los >5, 1,0)

#Add IMD quintiles to look at this association with the outcome

quintiles <- quantile(traveltimesrev$IMD_score, probs = seq(0,1,0.2), na.rm=TRUE)

traveltimesrev$IMD_quintile <- cut(traveltimesrev$IMD_score, breaks = quintiles, labels =
c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)

####Save final dataset
write.csv(traveltimesrev, "~/Desktop/REVISIONSTRAVELTIMESFINAL.csv")


###Load final dataset

traveltimesrev <- read.csv("/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/REVISIONSTRAVELTIMESFINAL.csv")

####Descriptive Statistics####
```

```
#Describe raw count statistics based on stratified travel quintile

table(traveltimesrev$distancequintile)

summary(traveltimesrev$DistanceMiles)

table(traveltimesrev$ageband, traveltimesrev$distancequintile)


summary(traveltimesrev$distancequintile)

table(traveltimesrev$Fractue)
table(traveltimesrev$Mechanical.complication)
table(traveltimesrev$OA)
table(traveltimesrev$RTKA_nonspecific)
table(traveltimesrev$conversion.to.TKA)
table(traveltimesrev$one_component)
table(traveltimesrev$Attention.to)

table(traveltimesrev$distancequintile, traveltimesrev$Read30days)

table(traveltimesrev$Read30days)

table(traveltimesrev$Provider_Name)

table(traveltimesrev$distancequintile, traveltimesrev$Read90days)

table(traveltimesrev$distancequintile, traveltimesrev$Mort90days)

table(traveltimesrev$distancequintile, traveltimesrev$rev1yr)

table(traveltimesrev$distancequintile, traveltimesrev$LongLOS)

#Demographics and Clinical Characteristics

table(traveltimesrev$distancequintile, traveltimesrev$IMD_quintile)

table(traveltimesrev$distancequintile, traveltimesrev$sex)

table(traveltimesrev$distancequintile, traveltimesrev$Mechanical.complication)

table(traveltimesrev$distancequintile, traveltimesrev$Fractue)

table(traveltimesrev$distancequintile, traveltimesrev$OA)

table(traveltimesrev$distancequintile, traveltimesrev$HFRS_Band)
```

```
table(traveltimesrev$distancequintile, traveltimesrev$CVcat)


####Correlations####
#Find out if IMD score or Age as continous are associated with Travel distance

#Look at median age and IMD in each of the travel distance quintiles first


new1 <- subset(traveltimesrev, distancequintile == "Q1")

new2 <- subset(traveltimesrev, distancequintile == "Q2")

new3 <- subset(traveltimesrev, distancequintile == "Q3")

new4 <- subset(traveltimesrev, distancequintile == "Q4")

new5 <- subset(traveltimesrev, distancequintile == "Q5")

#Calculate median age for each travel quintile

summary(new1$age_of_patient)
summary(new2$age_of_patient)
summary(new3$age_of_patient)
summary(new4$age_of_patient)
summary(new5$age_of_patient)

boxplot(traveltimesrev$age_of_patient ~ traveltimesrev$distancequintile, xlab = "Travel
Distance Quintile", ylab = "Age (years)")

#Calculate median IMD score for each travel quintile


summary(new1$IMD_score)
summary(new2$IMD_score)
summary(new3$IMD_score)
summary(new4$IMD_score)
summary(new5$IMD_score)

boxplot(traveltimesrev$IMD_score ~ traveltimesrev$distancequintile, xlab = "Travel
Distance Quintile", ylab = "IMD Score")

#Next do a Spearman's rank correlation between travel distance and age, and then for
travel distance and IMD score

cor.test(traveltimesrev$age_of_patient, traveltimesrev$DistanceMiles,
method="spearman")
```

```
cor.test(traveltimesrev$IMD_score, traveltimesrev$DistanceMiles, method="spearman")



#Find the median travel time for patients in Q5 travel quintile

new <- subset(traveltimesrev, distancequintile == "Q5")

summary(new$DistanceMiles)

#Find median travel distance for each travel quintile

new1 <- subset(traveltimesrev, distancequintile == "Q1")

new2 <- subset(traveltimesrev, distancequintile == "Q2")

new3 <- subset(traveltimesrev, distancequintile == "Q3")

new4 <- subset(traveltimesrev, distancequintile == "Q4")

new5 <- subset(traveltimesrev, distancequintile == "Q5")

summary(new1$DistanceMiles)
summary(new5$DistanceMiles)
#Repeat for other distance quintiles


####Modelling####

#Logistic Regression

#Primary outcome variable binary admitted within 30 days or not

model.log<-glm(Read30days ~ distancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))

install.packages("MASS")
library("MASS")

#Mass is loaded in other packages such as lmertest

OR_CI <- round(exp(cbind(coef(model.log),
                         confint(model.log))), digits = 3)
```

```
result_table <- data.frame(
  Coefficient = coef(model.log),
  P_Value = summary(model.log)$coefficients[, "Pr(>|z|)"]
)

write.csv(result_table, "~/Desktop/Sensitivty MORT.csv")


#Plot graph

#this creates a matrix, we now need to convert into a dataframe and change column names

df <- as.data.frame(OR_CI)


#Remove intercept row the first row

df = df[-1,]



#add covariate column
df$covariate <- c('Distance quintile 2 (ref: Q1)', 'Distance quintile 3 (ref: Q1)', 'Distance
quintile 4 (ref: Q1)', 'Distance quintile 5 (ref: Q1)', 'IMD_quintileQ2 (ref:Q1)',
'IMD_quintileQ3 (ref:Q1)', 'IMD_quintileQ4 (ref:Q1)', 'IMD_quintileQ5 (ref:Q1)', 'Male vs
Female', '60-64', '65-69', '70-74', '75-79', '>=80', 'Mechanical failure vs no failure', 'Fracture
vs no fracture', 'Progressive OA vs no OA', 'HFRS_Band Mild (ref: None)', 'HFRS_Band
Moderate (ref:None)', 'HFRS_Band Severe (ref:None)', 'Surgeon annual volume 5-9 (ref 0-4)',
'Surgeon annual volume 10-14 (ref 0-4)', 'Surgeon annual volume 15-19 (ref 0-4)', 'Surgeon
annual volume 20-24 (ref 0-4)', 'Surgeon annual volume >=25 (ref 0-4)')


#Save dataframe to desktop for analysis write up

write.csv(df, "~/Desktop/sensitivty MORT Log.csv")


ggplot(data=df, aes(y = df$covariate, x = df$V1, xmin=df$`2.5 %`,xmax=df$`97.5 %`))+
  geom_point()+
  geom_errorbarh(height=.1)+
  geom_vline(xintercept = 1)+
  xlab("Odds Ratio")+
  ylab("Exposure & Co-variates")+
  ggtitle("Odds for mortality within 90 days")
```

```
#Save Odds ratio's and 95% confidence intervals as new dataframe

coefficients_table <- as.data.frame(exp(cbind(OR = coef(model.log), confint(model.log, level
= 0.95))))

write.csv(coefficients_table, "~/Desktop/MultivariableLogisticRegression.csv")



#No statistical difference in 30 day readmission rates between different quintiles

#Risk of LOS>median
model.log<-glm(LongLOS ~ distancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))
#No statistical difference for LOS between quintiles adjusted


#Mortality at 90 days
model.log<-glm(Mort90days ~ distancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))

#No difference for mortality at 90 days

#Testing for model fit

null <- glm(rev1yr ~ 1, data = traveltimesrev, family = "binomial")
#or
null <- glm(Mort90days ~ 1, data = traveltimesrev, family = "binomial")
#or
null <- glm(Read30days ~ 1, data = traveltimesrev, family = "binomial")
#or
null <- glm(Read90days ~ 1, data = traveltimesrev, family = "binomial")
#or
null <- glm(Medianlos ~ 1, data = traveltimesrev, family = "binomial")

anova(model.log, null, test = "Chisq")

LRT <- model.log$null.deviance - model.log$deviance
print(LRT)
```

```
#This shows a non significant X2 statistics which shows the model has a good fit

#The best way to check for collinearity is using VIF

#variance inflation factor (or VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.


install.packages("car")
library(car)

plot(model.log)

install.packages("carData")
library(carData)

#You need to run this code for each model used in Logistic Regression

vif(model.log)

ols_vif_tol(model.log)

#For each outcome model (logistic regression) VIF is <3 therefore

#None of the VIF exceeds 5 so we can assume there is no evidence of strong multicollinearity

shapiro.test(rstandard(model.log))

#Shapiro wilcox test shows no evidence of multicollinearity, very low p value so we can reject the null hypothesis of normality

#Calculate pseudo R squared values at assess model fit

ll.full<-logLik(model.log)
ll.null<-logLik(null)
n<-length(model.log$residuals)

McFadden Test
as.numeric(1-(ll.full/ll.null))

#Evidence showing good model fit for all models


#####Sensitivty analysis Drive Distances####
#Create drive time quintile variable
```

```
quintiles <- quantile(traveltimesrev$OffPeakDriveDistanceMiles, probs = seq(0,1,0.2),
na.rm=TRUE)

traveltimesrev$drivedistancequintile <- cut(traveltimesrev$OffPeakDriveDistanceMiles,
breaks = quintiles, labels = c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)


#Find median off peak distance by quintile

new1 <- subset(traveltimesrev, drivedistancequintile == "Q1")

new2 <- subset(traveltimesrev, drivedistancequintile == "Q2")

new3 <- subset(traveltimesrev, drivedistancequintile == "Q3")

new4 <- subset(traveltimesrev, drivedistancequintile == "Q4")

new5 <- subset(traveltimesrev, drivedistancequintile == "Q5")

summary(traveltimesrev$OffPeakDriveDistanceMiles)
summary(new1$OffPeakDriveDistanceMiles)
summary(new5$OffPeakDriveDistanceMiles)

#The calculate the primary and secondary outcomes again

#Logistic Regression

#Primary outcome variable binary admitted within 30 days or not

model.log<-glm(Read30days ~ drivedistancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))

#Risk of LOS>median
model.log<-glm(LongLOS ~ drivedistancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))
#No statistical difference for LOS between quintiles adjusted

#Mortality at 90 days
model.log<-glm(Mort90days ~ drivedistancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
```

```
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))



####Sensitivty analysis drive times####

#Create drive time quintiles

quintiles <- quantile(traveltimesrev$PeakDriveTime, probs = seq(0,1,0.2), na.rm=TRUE)

traveltimesrev$timequintile <- cut(traveltimesrev$PeakDriveTime, breaks = quintiles, labels
= c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)

#Find median off peak times by quintile

new1 <- subset(traveltimesrev, timequintile == "Q1")

new2 <- subset(traveltimesrev, timequintile == "Q2")

new3 <- subset(traveltimesrev, timequintile == "Q3")

new4 <- subset(traveltimesrev, timequintile == "Q4")

new5 <- subset(traveltimesrev, timequintile == "Q5")

summary(traveltimesrev$timequintile)
summary(new1$PeakDriveTime)
summary(new5$PeakDriveTime)

#The calculate the primary and secondary outcomes again

#Logistic Regression

#Primary outcome variable binary admitted within 30 days or not

model.log<-glm(Read30days ~ timequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))

#Risk of LOS>median
model.log<-glm(LongLOS ~ timequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
```

```
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))
#No statistical difference for LOS between quintiles adjusted

#Mortality at 90 days
model.log<-glm(Mort90days ~ timequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model.log)
exp(cbind(OR = coef(model.log), confint(model.log, level = 0.95)))


####Sup Material S4 Crude Rates ####
#Supplementary Tables travel as continous variable

#Plot crude rates of 30 day readmission and road distances with off peak journeys in mind

# Calculate failure rates by surgical unit
hospital_failure_rates <- traveltimesrev %>%
  group_by(OffPeakDriveDistanceMiles) %>%
  summarise(
    total_surgeries = n(),
    total_failures = sum(Read30days, na.rm = TRUE),
    failure_rate = total_failures / total_surgeries
  )


# Remove any rows with NA values in relevant columns before fitting
hospital_failure_rates_clean <- hospital_failure_rates %>%
  filter(!is.na(OffPeakDriveDistanceMiles), !is.na(failure_rate))

# Fit the LOESS model to the cleaned data
loess_fit <- loess(failure_rate ~ OffPeakDriveDistanceMiles, data =
hospital_failure_rates_clean)

# Make predictions on the cleaned data
predictions <- predict(loess_fit, newdata = hospital_failure_rates_clean, se = TRUE)

# Add the predictions back to the cleaned dataset
hospital_failure_rates_clean$fit <- predictions$fit
hospital_failure_rates_clean$se <- predictions$se.fit


ggplot(hospital_failure_rates_clean, aes(x = OffPeakDriveDistanceMiles, y = failure_rate)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = fit), color = "blue") +  # Add the fitted line
  geom_ribbon(aes(ymin = pmax(0, fit - 1.96 * se), ymax = fit + 1.96 * se), alpha = 0.2) +  #
95% CI with lower bound constrained to 0
```

```
  labs(
    x = "Off Peak Drive Distance (Miles)",
    y = "Readmission within 30 days",
    title = "LOESS Fit: Re-admission within 30 days by travel distance"
  ) +
  scale_y_continuous(labels = scales::percent_format(), limits = c(0, NA)) +
  scale_x_continuous(limits = c(0, max(hospital_failure_rates$OffPeakDriveDistanceMiles)))


#Crude rates and travel distance as crow flies

# Calculate failure rates by surgical unit
hospital_failure_rates <- traveltimesrev %>%
  group_by(DistanceMiles) %>%
  summarise(
    total_surgeries = n(),
    total_failures = sum(Read30days, na.rm = TRUE),
    failure_rate = total_failures / total_surgeries
  )


# Remove any rows with NA values in relevant columns before fitting
hospital_failure_rates_clean <- hospital_failure_rates %>%
  filter(!is.na(DistanceMiles), !is.na(failure_rate))

# Fit the LOESS model to the cleaned data
loess_fit <- loess(failure_rate ~ DistanceMiles, data = hospital_failure_rates_clean)

# Make predictions on the cleaned data
predictions <- predict(loess_fit, newdata = hospital_failure_rates_clean, se = TRUE)

# Add the predictions back to the cleaned dataset
hospital_failure_rates_clean$fit <- predictions$fit
hospital_failure_rates_clean$se <- predictions$se.fit


ggplot(hospital_failure_rates_clean, aes(x = DistanceMiles, y = failure_rate)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = fit), color = "blue") +  # Add the fitted line
  geom_ribbon(aes(ymin = pmax(0, fit - 1.96 * se), ymax = fit + 1.96 * se), alpha = 0.2) +  #
95% CI with lower bound constrained to 0
  labs(
    x = "As Crow Flies Travel Distance (Miles)",
    y = "Readmission within 30 days",
    title = "LOESS Fit: Re-admission within 30 days by travel distance"
  ) +
  scale_y_continuous(labels = scales::percent_format(), limits = c(0, NA)) +
```

```
  scale_x_continuous(limits = c(0, max(hospital_failure_rates$DistanceMiles)))



#Crude rates peak drive time and 30 day re-admission

# Calculate failure rates by surgical unit
hospital_failure_rates <- traveltimesrev %>%
  group_by(PeakDriveTime) %>%
  summarise(
    total_surgeries = n(),
    total_failures = sum(Read30days, na.rm = TRUE),
    failure_rate = total_failures / total_surgeries
  )


# Remove any rows with NA values in relevant columns before fitting
hospital_failure_rates_clean <- hospital_failure_rates %>%
  filter(!is.na(PeakDriveTime), !is.na(failure_rate))

# Fit the LOESS model to the cleaned data
loess_fit <- loess(failure_rate ~ PeakDriveTime, data = hospital_failure_rates_clean)

# Make predictions on the cleaned data
predictions <- predict(loess_fit, newdata = hospital_failure_rates_clean, se = TRUE)

# Add the predictions back to the cleaned dataset
hospital_failure_rates_clean$fit <- predictions$fit
hospital_failure_rates_clean$se <- predictions$se.fit


ggplot(hospital_failure_rates_clean, aes(x = PeakDriveTime, y = failure_rate)) +
  geom_point(alpha = 0.5) +
  geom_line(aes(y = fit), color = "blue") +  # Add the fitted line
  geom_ribbon(aes(ymin = pmax(0, fit - 1.96 * se), ymax = fit + 1.96 * se), alpha = 0.2) +  #
95% CI with lower bound constrained to 0
  labs(
    x = "Peak Drive Time (Minutes)",
    y = "Readmission within 30 days",
    title = "LOESS Fit: Re-admission within 30 days by travel times"
  ) +
  scale_y_continuous(labels = scales::percent_format(), limits = c(0, NA)) +
  scale_x_continuous(limits = c(0, max(hospital_failure_rates$PeakDriveTime)))
```

```
####Supp Material S4 Logistic Regression####
#Logisic Regression Model Distance Miles Primary Outcome

#####Model 1 unadjusted


model <- glm(Read30days ~ DistanceMiles, family = binomial(link = "logit"), data =
traveltimesrev)

summary(model)

#p value 0.763, -0.002439 coef, AIC 797.95

#null models
#re-revision at 2 yrs
null <- glm(Read30days ~ 1, data = traveltimesrev, family = "binomial")


ll.full <- logLik(model)
ll.null <- logLik(null)
n<- length(model$residuals)
as.numeric(1-(ll.full/ll.null))

#r squared 0.000117

#Patient Factors

model<-glm(Read30days ~ DistanceMiles + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band, data=traveltimesrev, family =
"binomial")
summary(model)

#AIC 796.67, p value 0.976. r squared 0.04

#Surgeon Factors

model<-glm(Read30days ~ DistanceMiles + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)
exp(cbind(OR = coef(model), confint(model, level = 0.95)))

#AIC 797.31, p value 0.912, coef 0.009223, r squared 0.0538



#No statistical relationship between as crow flies travel distance and primary outcome
```

```
#Is travel distance linear or non linear

#Box Tidwell

model <- glm(Read30days ~ DistanceMiles, family = binomial(link = "logit"), data =
traveltimesrev)

coef_summary <- summary(model)

box_tidwell_Mean_Unit <- coef_summary$coefficient[2, "Pr(>|z|)"]

print(box_tidwell_Mean_Unit)

#p value 0.762, it is not non-linear, no indication to model with splines

#Model as categorical quintiles and assess model fit for comparison

model<-glm(Read30days ~ distancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)


#AIC 794.68, r squared 0.064 (improved model fit)

#Logistic regression travel distance by road

#####Model 1 unadjusted


model <- glm(Read30days ~ OffPeakDriveDistanceMiles, family = binomial(link = "logit"),
data = traveltimesrev)

summary(model)

#p value 0.544, -0.003686 coef, AIC 797.66

#null models
#re-revision at 2 yrs
null <- glm(Read30days ~ 1, data = traveltimesrev, family = "binomial")


ll.full <- logLik(model)
ll.null <- logLik(null)
n<- length(model$residuals)
as.numeric(1-(ll.full/ll.null))
```

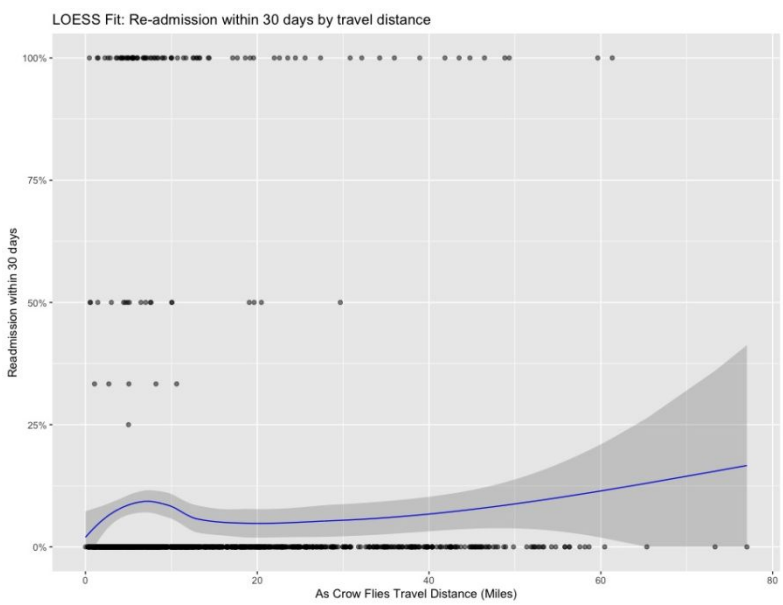
```
#r squared 0.000475

#Patient Factors

model<-glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band, data=traveltimesrev, family =
"binomial")
summary(model)

#AIC 796.6, p value 0.787. r squared 0.0421, coef -0.00168

#Surgeon Factors

model<-glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)

#AIC 797.3, p value 0.882, coef -0.00093, r squared 0.0538

exp(cbind(OR = coef(model), confint(model, level = 0.95)))


#No statistical relationship between as crow road travel distance and primary outcome

#Is travel distance linear or non linear

#Box Tidwell

model <- glm(Read30days ~ OffPeakDriveDistanceMiles, family = binomial(link = "logit"),
data = traveltimesrev)

coef_summary <- summary(model)

box_tidwell_Mean_Unit <- coef_summary$coefficient[2, "Pr(>|z|)"]

print(box_tidwell_Mean_Unit)

#p value 0.544, it is not non-linear, no indication to model with splines

#Model as categorical quintiles and assess model fit for comparison

model<-glm(Read30days ~ distancequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)
```

```
#AIC 794.68, r squared 0.064 (improved model fit)




#Logistic regression Road Travel Times

#####Model 1 unadjusted


model <- glm(Read30days ~ PeakDriveTime, family = binomial(link = "logit"), data =
traveltimesrev)

summary(model)

# AIC 797.77, p value 0.608, coef -0.002451

#null models
#re-revision at 2 yrs
null <- glm(Read30days ~ 1, data = traveltimesrev, family = "binomial")


ll.full <- logLik(model)
ll.null <- logLik(null)
n<- length(model$residuals)
as.numeric(1-(ll.full/ll.null))

#r squared 0.000336

#Patient Factors

model<-glm(Read30days ~ PeakDriveTime + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band, data=traveltimesrev, family =
"binomial")
summary(model)

#AIC 796.65, p value 0.863. r squared 0.042

#Surgeon Factors

model<-glm(Read30days ~ PeakDriveTime + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)
```

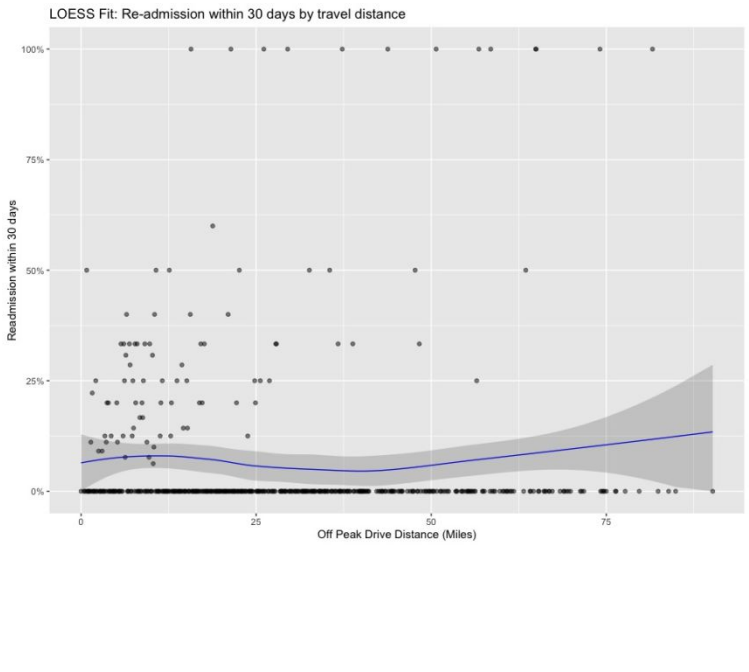
```
exp(cbind(OR = coef(model), confint(model, level = 0.95)))


#AIC 797.32, p value 0.968, coef 0.0002027, r squared 0.0538


#No statistical relationship between drive time and primary outcome

#Is travel distance linear or non linear

#Box Tidwell

model <- glm(Read30days ~ PeakDriveTime, family = binomial(link = "logit"), data =
traveltimesrev)

coef_summary <- summary(model)

box_tidwell_Mean_Unit <- coef_summary$coefficient[2, "Pr(>|z|)"]

print(box_tidwell_Mean_Unit)

#p value 0.608, it is not non-linear, no indication to model with splines

#Model as categorical quintiles and assess model fit for comparison

model<-glm(Read30days ~ timequintile + IMD_quintile + sex + ageband +
Mechanical.complication + Fractue + OA + HFRS_Band + CVcat, data=traveltimesrev, family
= "binomial")
summary(model)

#AIC 800, r squared 0.058

####END####
```

Supplementary material S4 – Relationship between Travel distances and times modelled as a continuous variable with primary outcome (readmission within 30 days)

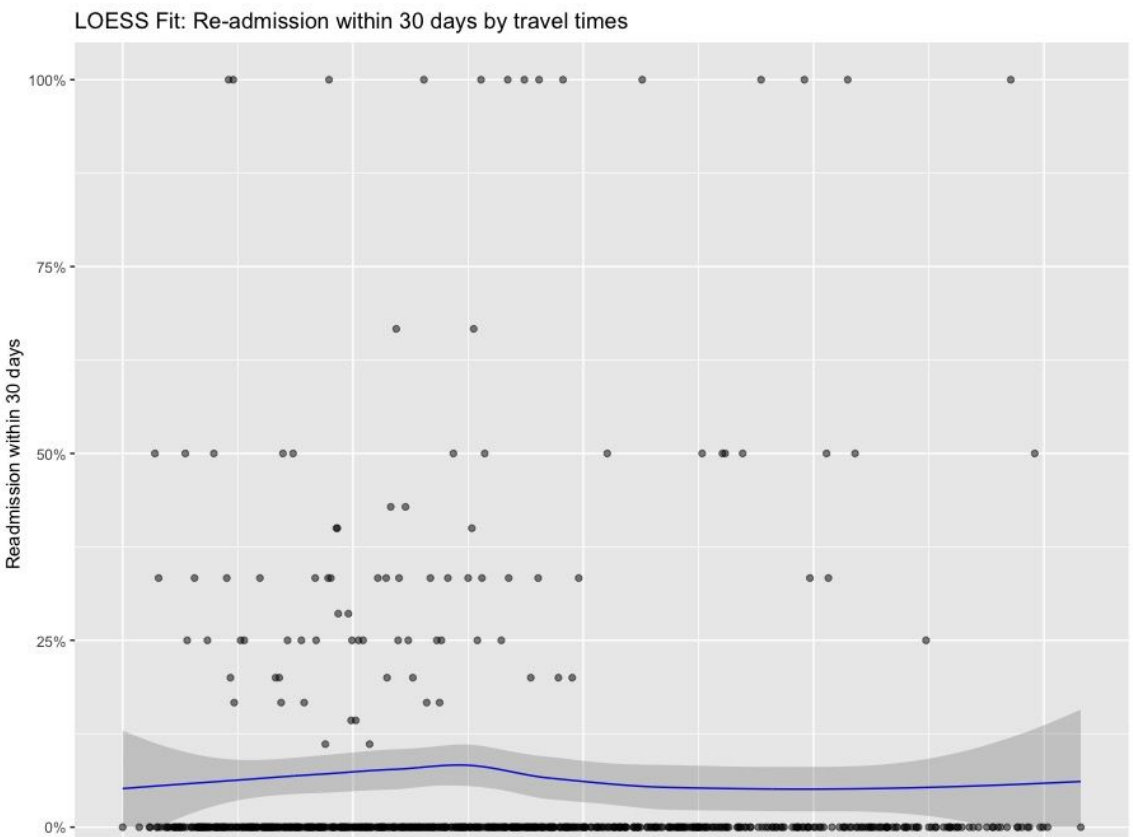
Crude Rates of 30-day readmission and Straight-Line Travel Distance (Locally Estimated Scatterplot Smoothing Fit used to estimate trends with standard errors)



Crude Rates of 30-day readmission and Travel Distance by Road (Locally Estimated Scatterplot Smoothing Fit used to estimate trends with standard errors)



Crude Rates of 30-day readmission and Travel Times by Road (Locally Estimated Scatterplot Smoothing Fit used to estimate trends with standard errors)



Multiple Variable Logistic Regression for Travel Distances and Times (continous) and readmission within 30 days.

	Association between Travel Distances and Times (continuous) and Readmission within 30 days		
	Odds Ratio/Coefficient estimate (95% confidence intervals)	**p value**	**R^2**
Straight Line Travel Distance	1.00 (0.98 to 1.02)	0.91	5.38%

Travel Distance by Road	1.00 (0.99 to 1.01)	0.88	5.38%
Peak Road Travel Times	1.00 (0.99 to 1.01)	0.97	5.38%

BMJ Open

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement: An analysis using national administrative data from Hospital Episode Statistics

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-085201.R2
Article Type:	Original research
Date Submitted by the Author:	16-Jan-2025
Complete List of Authors:	Matthews, Alexander; University of Exeter Faculty of Health and Life Sciences, Department of Clinical and Biomedical Sciences; Royal Devon and Exeter NHS Foundation Trust, Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research; Royal Devon and Exeter NHS Foundation Trust, Princess Elizabeth Orthopaedics Centre Evans, Jonathan; University of Exeter, Department of Public Health and Sports Sciences Lamb, Sarah; University of Exeter, College of Medicine and Health Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Gray, William; NHS England and NHS Improvement London Briggs, Tim; NHS England and NHS Improvement London Toms, Andrew ; Princess Elizabeth Orthopaedic Centre
Primary Subject Heading:	Public health
Secondary Subject Heading:	Surgery
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, PUBLIC HEALTH, Health Services

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement: An analysis using national administrative data from Hospital Episode Statistics

Names, Affiliations, and positions of all authors

Alex Matthews1,2,3,4, Jonathan P Evans2,3, Jonathan T Evans2,3, Sarah E Lamb3
Andrew Price4,5, William Gray1, Tim Briggs1,6, Andrew Toms2,3

1. Getting It Right First Time programme, NHS England, London, UK
2. Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
3. University of Exeter, Exeter, UK
4. Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, UK
5. Nuffield Orthopaedic Centre, Oxford, UK
6. Royal National Orthopaedic Hospital, Stanmore, London, UK

Correspondence: Alexander Matthews, Getting It Right First Time programme, NHS England and NHS Improvement, Wellington House, Waterloo Road. London, UK, SE1 8UG; telephone: 07401932627 email: alexander.matthews1@nhs.net

Word count: [3316 words]

Short title: Patient Journey Distance and Time for Complex Revision Knee Surgery

Key words: social determinants of health, revision knee replacement, contextual factors in complex health interventions

Twitter handles:

Alex Matthews: @DrAlexMatthews

Tim WR Briggs: @ProfTimBriggs

WilliamKGray1: @WilliamKGray1

GIRFT programme: @NHSGIRFT

Structured Abstract

Objectives

Patients undergoing revision total knee replacement (RevKR) surgery often have difficulties mobilising and increasingly rely on family support. Evolving practice in England aims to manage these patients in specialised centres with the intention of improving outcomes. This practice will result in longer travel distances and times in this frailer group of patients. We want to examine the types of distances and travel times patients can be expected to travel for this complex orthopaedic surgery and to explore concerns of how these impact patient outcomes.

Design

Retrospective observational study from the Hospital Episode Statistics. Pooled multivariable adjusted logistic regression models were used to investigate the relationship between patient travel distances and times with perioperative outcomes.

Setting

Patients presenting to tertiary referral centres between 1st January 2016 to 31st December 2019. A tertiary referral centre was defined as a trust performing >49 revisions in the year prior.

Participants

Adult patients undergoing RevKR procedures for any reason between 1st January 2016 to 31st December 2019.

Exposure

The shortest patient level travel distance and time was calculated using the department of health Journey Time Statistics using TRACC software and Dijkstra's algorithm.

Main Outcome Measures

The primary outcome is emergency readmission within 30 days. Secondary outcomes are mortality within 90 days and length of inpatient stay.

Results

6,880 patients underwent RevKR at 36 tertiary referral centres. There was a weak correlation between social deprivation and travel distance, with patients from the most deprived areas travelling longer distances. Overall, 30-day readmission was not statistically associated longer driving distance (OR 1.00 95% CI 0.99 to 1.02) or peak driving times (OR 1.00 95% CI 0.99 to 1.01).

Conclusions

There was no association between increasing travel distance and time on perioperative outcomes for RevKR patients.

Strengths and limitations of this study

- Our study is the first to describe patient travel distance and time associations using a large longitudinal dataset.
- This data reflects revision knee replacement procedures undertaken across different geographical areas of England
- Owing to differences in the coverage of Hospital Episode Statistics, procedures in hospitals outside of England were not included in this analysis
- Clinical coding practice within HES is known to vary between trusts but this is unlikely to be vary systematically to bias our findings
- This analysis only reports travel times for patients with access to their own transport and does not consider times for those patients using public transport

Introduction

Primary knee replacement is a successful procedure that improves quality of life for the majority of patients.[1] However, at 10 years following a primary knee replacement, about 3.5% of patients will have undergone a revision surgery.[2] The majority of these procedures are carried out due to infection or polyethylene wear of the implant.[3] A failed primary knee replacement represents a life changing transition point where individuals are likely to suffer from pain, reduced mobility as well as dependency on family members.[4] Patients often face multi-step surgery with longer hospital length of stays and higher complication rates.[5, 6]

The Getting It Right First Time (GIRFT) programme orthopaedic National Report was published in 2015.[7] A key recommendation was the centralisation of complex orthopaedic surgery, including revision knee surgery, to specialist centres with the aim of improved patient outcomes. Consequently, revision total knee replacement (RevKR) surgery in the England has evolved into a regional network service model.[8] All hospitals performing RevKR form a network in the respective regions. Less specialist hospitals, defined by lower annual case volume thresholds, are encouraged to discuss and sometimes refer their caseload to more specialist centres. Several studies based on large revision hip and knee registries have suggested this model carries a lower failure rate defined by the need for further revision surgery.[9-11] Early evidence has suggested reduced early failure rates through the adoption of revision knee networks.[12]

However, for some patients, this approach to managing patients is inevitably associated with increasing travel distances between patient's homes and their treating hospital. Travel distance has been shown to be an important factor in patient choice when selecting a surgeon for joint replacement surgery. It may be even more important for those awaiting revision joint replacement surgery as these patients struggle with mobility, may be unable to drive and may be more reliant on family members.[4] Evidence suggests that patients considering joint replacement are prepared to travel longer distances to obtain the best possible outcomes. A requisite in making such a decision requires data on outcomes of patients travelling greater

distances. Patients travel longer distances have been found to have higher readmission rates and higher mortality rates when undergoing other types of specialised surgery.[13] The pick-up rate of early complications, avoiding the need for readmission, may be less in areas further away from the main treatment centre. There is also concern that patients required to travel greater distances are more likely to be re-admitted to a different hospital than that where surgery was undertaken, resulting in clinical decisions that do not incorporate the primary surgeon and so potentially leading to poorer outcomes.[14] There is an absence of evidence in the literature to support or refute this argument in the context of patients undergoing RevKR. Therefore the aim of this paper is to investigate the relationship between longer patient travel distances and perioperative outcomes following RevKR performed in high volume tertiary referral centres.

Methods

Design

This study is a retrospective data analysis of observational data from the Hospital Episode Statistics (HES) and Office for National Statistics (ONS) databases. HES data is collected by NHS England for all patients treated at NHS hospitals in England and those treated at private hospitals where treatment was funded by the NHS. This study complies with the recommended reporting guidelines when using HES data[15] and the Strengthening of Reporting of Observational studies in Epidemiology (STROBE) guidelines.[16]

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes[17] and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.[18] The HES data were linked at a patient level to data from the ONS on deaths and date of death, which allowed the identification of patients who had died after their surgery. Linkage was achieved using a unique pseudonymised patient identifier using a previously validated methodology.[19]

Patient travel distances were calculated using the Journey Time Statistics reference document produced by the UK Department of Transport which modelled theoretical journey times between known centroids of Lower Layer Super Output Areas (LSOA)

of residence and NHS hospital sites.[20] Please refer to **Supplementary material S1** for Journey Times Statistics reference document.

Population

An RevKR procedure was defined as a permanent removal or exchange of knee arthroplasty components. This includes a revision of a total knee replacement and a conversion of a unicondylar knee replacement to a total knee replacement. Secondary patellar resurfacing was not included as this represents a simple revision procedure, one that can be carried out in most nonspecialised hospitals. All patients aged ≥ 18 years who underwent a RevKR in a high-volume trust between 1st January 2016 to 31st December 2019 were included in the study population. A high-volume trust was classified as a centre performing > 49 revisions per year. This revision volume threshold for classification represents that proposed by the British Association for Surgeons of the Knee (BASK) Revision Knee Working Group and is a mandatory requirement for all highly specialist centres co-ordinating regional networks. [21] As such centres attaining this threshold are more likely to represent tertiary referral centres where the stratification of more complex work will take place. Annual case volume at each trust was defined as the number of revision cases conducted in the year prior to the index procedure. This measure was preferred over a simple calculation of average annual volume as it accounts for recent experience at the point of surgery. The Office for Population Censuses and Surveys' Classification of Interventions and Procedures version 4 (OPCS-4) codes used to identify RevKR procedures are detailed in **Supplementary material S2.** Since laterality was needed to identify re-revisions, patients were excluded where the procedure laterality was not specified. The flow of patients, with numbers excluded at each point, is summarised in **Supplementary material S3**. To manage population heterogeneity, data were extracted for the period 1st April 2011 to 31st December 2019 and only the first revision for a specific side of the body record in this time period included. [22] Thus, any early revisions on the same side of the body in the four years and nine months preceding the start of the study period were identified and these patients excluded from the study. This aims to exclude the early revision knee replacement failures which have been shown to represent catastrophic failures potentially skewing our results.[22] We included revisions for infection as, despite

these representing a more variable patient group, presence of infection was thought to be unrelated to how far a patient lives from a specialised referral centre.

Exposure variable

Travel distances and times were calculated between a patient's LSOA and the postal codes for the treating hospitals. LSOAs are determined by the Office for National Statistics and are designed for the reporting of small area statistics. Public transport and highways data for England were used to create theoretical journey distances and times from origins to destinations. A network of journey distances and times from origins to destinations was produced using a software package called Transport Accessibility and Connectivity Calculator (TRACC). The Dijkstra's algorithm calculated the shortest route between these points. Data linkage was achieved with our clinical dataset following a reproducible workflow. The resulting travel distances and/or times for each patient were analysed as continuous variables. Three exposure variables were used. Straight line travel distance represented the distance "as the crow flies" between a patient's LSOA and treating hospital. Off peak driving distance represented the shortest driving distance between a patients LSOA and treating hospital. Finally peak driving times were calculated using average traffic speeds between 7am and 10am for the shortest possible road route between a patients LSOA and treating hospital. These three variables were used to account for variation in travel infrastructure between rural and urban areas and to attribute more meaningful results for patients.

Co-variates and cluster variable

The following groups of known or potential confounding variables were chosen a priori for inclusion in our multivariable logistic regression modelling:

Patient factors: Age in years (continuous), sex (male/female). Health co-morbidity was quantified using the Hospital Frailty Risk Score (HFRS). HFRS identifies frailty based on the occurrence of any of 109 International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during any hospital admissions in the two years prior to, and for, the index admission. Deprivation was measured using the Index of Multiple Deprivation (IMD).[23] The

IMD gives the LSOA where the patient lives a score based on a range of measures of deprivation. IMD was analysed as a continuous variable.

Clinical factors: Defined by the presence or absence of infection as the primary indication for RevKR. This was identified from the International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during the admission.

Surgical factors: Surgeon and hospital volume (both continuous) was defined as the number of RevKRs performed by a consultant or hospital in the 365 days prior to each index procedure across the entire cohort. This was calculated before any exclusion criteria was applied.

Temporal factors: Financial year of procedure (2015/16, 2016/17, 2017/18, 2018/19, 2019/20).

Hospital Provider: Clustering of patients by hospital provider was initially modelled using random effects. However, despite variability between hospital providers with primary and secondary outcomes, instability in the model estimates were observed. To address the possibility of clustering at this level, a fixed effects model was adopted with hospital provider as a covariate.

Outcomes

The primary outcome was emergency readmission within 30 days of discharge from the index surgical hospital. Readmission in this early period is very likely related to a complication of the surgical procedure. It has been used as a marker of perioperative outcomes in similar studies investigating the relationship between patient travel distance and outcomes following surgery. [13]

Secondary outcomes were:

90-day all-cause mortality, identified using linked data from Civil Registrations (Mortality) dataset;

Inpatient length of hospital stay was attributed from continuous inpatient spells (CIPS), which is the preferred estimate of length of stay. This refers to the length of first stay after the operation regardless of any transfers across providers. The

median length of stay was calculated after visually inspecting the distribution and this was dichotomized into prolonged length of stay if longer than the median stay.

Statistical Analyses

Data was extracted from a secure, encrypted server controlled by NHS England. Data were analysed within a secure, encrypted environment using standard statistical software: R Studio version 2023.09.1+494 (Boston, Massachusetts, USA). The R code and packages used are included in **Supplementary material S4**.

Missing data were managed according to its extent and relevance to the aims of this study. Age and IMD score were imputed for the small number of missing cases using the mean of the entire study cohort. Given the central role of LSOA in estimating travel distances and times and fewer than 5% of cases with missing data, these cases were excluded to avoid the introduction of bias. Following data linkage, approximately 36% (n = 5,838) of cases did not match with travel data. Multiple imputation was performed using predictive mean matching based on the entire cohort of patients with the following predictors: age, sex, HFRS score, IMD score, hospital provider code, hospital volume and surgeon volume. Dependent variables including readmission at 30 days, mortality at 90 days and length of stay were also used in the imputation following a recommended approach using preditive mean matching[24]. A total of five imputations were randomly chosen and subsequent regression analyses were pooled.[25] Imputed data is shown in **Supplementary material S5**.

Patient travel distances were categorised into quintiles for interpretation of baseline demographics and clinical characteristics. Subsequent analysis of travel distances and times were performed as continuous variables. Spearman's rank correlation was performed to investigate the relationship between IMD score and patient age with travel distances.

Straight line travel distance was modelled with restricted cubic splines to allow for the non-linear effects when testing the association with the primary outcome. All exposures were modelled with restricted cubic splines to allow for the non-linear effects when testing the association with prolonged length of stay. The Akaike

Information Criterion was used to select the most parsimonious specification of restricted cubic splines using the final adjusted model.

Fixed effects logistic regression models were used for the outcomes of readmission at 30 days, mortality at 90 days and prolonged length of stay. Adjustment for confounding was undertaken incrementally, adjusting for each of the five groups of confounding variables to explore their influence on the effect at each stage with reference to model fit statistics. This was done following an apriori methodology with addition and or removal of factors in the following order: patient factors, clinical factors, surgical factors, temporal factors and the hospital provider. The ultimate decision on the preferred statistical model was assessed using the Akaike Information Criterion (AIC) accepting the model with the lowest AIC. Co-variates were modelled as either linear or categorical terms to simplify the model and aid interpretability. Multicollinearity was assessed using eigenvalues, variance inflation factors and by examination of model parameter estimates with stepwise addition and removal of covariates. Odds ratios with 95% CIs and associated p-values were reported. A p-value of < 0.05 was taken to indicate statistical significance.

Results

Overview of results

A total of 16,736 patients met the inclusion criteria. Excluding missing LSOA data (n=171), 16,565 patients were included in the analysis. Following data linkage with department of transport journey times statistics, 10,727 patients had complete data linkage and data were imputed for the remaining 5,838 (35.2%). Of the 16,565 patients, 41.5% (n=6,880) presented to a tertiary referral centre and these data formed our analysis cohort. Patients were operated on across 181 hospital sites and 38 hospital trust providers. The baseline demographic and clinical characteristics of the patients were broadly similar between quintiles of straight-line travel distance. (Table 1). Higher hospital volumes were seen in patients travelling longer distances. Straight line travel distance was weakly correlated with age and social deprivation (**Figure 1**). Older patients were less likely to travel farther distances. Patients from the least deprived areas travelled shorter distances.

Table 1 – Baseline patient demographics and clinical characteristics stratified by travel distance quintiles from first imputed dataset

	Travel Distance Quintile				
	1	2	3	4	5
Distance (Miles)	2.09 (1.35 to 2.75)	4.42 (3.91 to 5.00)	7.08 (6.34 to 7.99)	11.39 (10.11 to 12.74)	22.42 (18.09 to 32.19)
Driving Time (Minutes)	13 (9.3 to 17)	20.45 (17 to 25)	26.30 (21.98 to 31.13)	34.10 (29.68 to 40.20)	52.05 (42.68 to 66.83)
Number of patients	1376	1376	1376	1376	1376
Tertiary Providers	37 (97.37%)	38 (100%)	36 (94.74%)	35 (92.11%)	37 (97.37%)
Age Mean (SD)	69.71 (10.81)	69.96 (10.71)	69.66 (10.92)	68.84 (11.01)	68.58 (10.75)
Female Sex	762 (55.38%)	768 (55.81%)	729 (52.98%)	722 (52.47%)	734 (53.34%)
HFRS None	647 (47.02%)	620 (45.06%)	614 (44.62%)	666 (48.40%)	676 (49.13%)
HFRS Mild	438 (31.83%)	474 (34.45%)	485 (35.25%)	465 (33.79%)	433 (31.47%)
HFRS Moderate	241 (17.51%)	236 (17.15%)	243 (17.66%)	198 (14.39%)	230 (16.72%)
HFRS Severe	50 (3.63%)	46 (3.34%)	34 (2.47%)	47 (3.42%)	37 (2.69%)

Infection Present	314 (22.82%)	331 (24.06%)	310 (22.53%)	334 (24.27%)	355 (25.80%)
Surgeon Volume	7 (3 to 13)	7 (3 to 13)	8 (3 to 15)	8 (3 to 16)	9 (4 to 17)
Hospital Volume	73 (60 to 87)	74 (60 to 89)	79 (63 to 97)	79 (63 to 99)	85 (68.75 to 112)
IMD Score	16.44 (8.73 to 28.67)	14.30 (7.96 to 24.57)	14.50 (8.47 to 21.36)	14.83 (9.23 to 21.74)	14.752 (8.78 to 21.45)
Year 2015/16	104 (7.56%)	94 (6.83%)	94 (6.83%)	89 (6.47%)	92 (6.69%)
Year 2016/17	383 (27.83%)	354 (25.73%)	348 (25.29%)	338 (24.56%)	353 (25.65%)
Year 2017/18	384 (27.91%)	365 (26.53%)	339 (24.64%)	360 (26.16%)	336 (24.42%)
Year 2018/19	269 (19.55%)	325 (23.62%)	347 (25.22%)	354 (25.73%)	339 (24.64%)
Year 2019/20	236 (17.15%)	238 (17.30%)	248 (18.02%)	235 (17.08%)	256 (18.60%)

Outcomes

The primary and secondary outcomes are summarised in table 2.

The observed rate of readmission at 30 days was 8.3% (568/6880). There was a negative association between higher straight line travel distances and emergency readmission at 30 days (Figure 2). However wide confidence intervals precluded statical inferences. In addition, higher travel distance by road and longer drive times

were not associated with statistically worse readmission rates at 30 days. The rate of mortality at 90 days was only 3.2% (217/6880). No statistically significant relationship was observed between the distance a patient travels by road or the time a patient spends travelling at peak driving times with rates of mortality at 90 days. 49.7% (3421/6880) of patients reported hospital stays more than 5 days. Following adjustment of confounding factors, we observed no associations between prolonged length of stay and patient travel distance (Figures 3-5)

Table 2 – Adjusted pooled Multivariable Logistic Regression showing Odds Ratios for primary and secondary outcomes by exposure variables

	Straight line travel distance (OR, 95% CI)	**Travel distance by shortest road route (OR, 95% CI)**	**Peak Travel times by shortest road route (OR, 95% CI)**
Readmission with 30 days	Figure 2	1.00 (0.99 to 1.02), p value = 0.81	1.00 (0.99 to 1.01), p value = 0.69
90 Day Mortality	1.00 (0.98 to 1.02), p value = 0.87	1.00 (0.99 to 1.01), p value = 0.86	1.00 (0.99 to 1.01), p value = 0.89
Prolonged Length of stay	Figure 3	Figure 4	Figure 5

•Odds ratios have been adjusted for patient age, sex, HFRS score,

Discussion

Statement of principal findings

We present a multi-hospital site retrospective analysis of patients undergoing revision knee replacement surgery at tertiary referral centres in England. In this analysis of 6,880 patients undergoing RevKR, we did not observe a statistical

association between distance and time travelled for revision surgery and readmission within 30 days.

Strengths and weaknesses of the study

The findings of this study should be interpreted in view of several limitations. Firstly, this analysis used observational data from a large administrative dataset covering all NHS-funded procedures conducted in England. As with all administrative datasets we are limited in the amount of detail provided regarding presentation. We chose to categorise a high-volume centre by trust to accurately capture surgical experience. All NHS hospitals in England are run by hospital trusts which typically involve between one and four hospitals within a catchment area standardising their practice. It is common practice for specialist orthopaedic surgeons to move between these sites delivering the same procedures. Our study involved 187 hospital sites run by 38 trusts. We acknowledge this is a weakness of our study as this may not be representative of all trusts. We included all indications for RevKR in our patient cohort because indication was not thought to be related to how far a patient lives from a hospital. However, we acknowledge the rate of complications is higher in patients with infection and we subsequently adjusted for indication for revision in our analyses. [26] It is likely that because we did not exclude previous revision knee arthroplasty patients, the complexity of the surgery undertaken in our cohort varied. We recognise this is a limitation of the study however we assume case mix was unrelated patient travel distance.

There were many missing patients (approximately 36%) following the linkage of HES data with Journey Time Statistics. To account for this, assumed that the data was missing at random and used multiple imputation to estimate missing travel distances. It is likely the imputed values may introduce bias, however we modelled these based on predictors and dependent variables to improve our estimates. We do not present a sample size calculation, rather we have used all available data and our sample size was set by our inclusion criteria. We controlled for the clustered nature of our data between hospital providers through inclusion as a covariate in our modelling. To ensure consistency in our definition of tertiary referral hospitals, only hospitals performing >49 revisions/year were included. These are likely to treat a similar case

mix of patients and potentially have similar access to resources within a national healthcare system. This approach allowed us to control for variation across providers. However, we acknowledge it does not fully account for the hierarchical nature of the data with differences in treatment protocols and hospital specialisation among factors which may influence patient outcomes.

There is a lack of granular data for revisions due to infection and therefore we excluded this patient group as some readmissions for this patient group may represent planned readmissions. There is also a lack of granular clinical data using HES for each readmission, therefore we cannot ascertain precise reasons for readmissions, but we assume are related to a post-surgical complication. Clinical coding practice within HES is known to vary across trusts.[27] As an example, some trusts may be more consistent in coding comorbidities, and this may have created some bias. However, this is unlikely to vary systematically with travel distances and so significantly bias our findings. We acknowledge the relatively short travel distances in this population compared to examples from the United States as such the results of this study may not be generalisable to larger geographical areas or less mature healthcare systems. However, the upper quintile in our study represents a substantial journey distance and time for our patient cohort where poor mobility is a significant factor affecting their care. This analysis does not consider journey times of those who may not have access to a car and instead chose to take public transport.

Strengths and weaknesses in relation to other studies, discussing important differences in results

This is the first study to analyse the potential impact of patient travel distances on patients receiving RevKR. The findings that longer travel distances are not associated with inferior outcomes is an important part of the evaluation of the assumptions and context behind the establishment of revision knee networks.[28] This study has shown that concerns of introducing a network in larger geographical regions, for example in Scotland where longer patient travel distances and times are common, may be less important.[29] This is particularly useful as regions explore the

geography of their revision networks and during summative outcome assessment of this complex health intervention.[30] Despite there being a potential negative association between straight line travel distance and emergency readmission at 30 days, there was a lack of association involving driving distances and times which present real world challenges for patients.

It may be seen as surprising that no association between travel distance and prolonged length of hospital stay was identified. An expectation exists of increasing difficulties being encountered with the discharge of patients living greater distances from their treating hospital, which has been observed in patients following elective pancreatic surgery.[31] This is also an observation seen in patients being treated in specialist vascular centres in the United States which led to the recommendation of additional care coordination and follow up efforts. However, the geography of the population in these studies was much larger with significantly longer travel distances.

We did observe a weak but statistically significant correlation between social deprivation status and age of the patient with longer travel distances. Patients from poorer sociodemographic background may be expected to travel further for RevKR. This highlights the additional care coordination and follow-up efforts that should accompany the widening reach of regional revision knee networks. It is reassuring that access to treatment for older patients is unaffected by travel distance. However, there may be patients who refused to travel to a specialist centre and opted for treatment at their local centre.

Meaning of the study: possible explanations and implications for clinicians and policymakers

The organisation and delivery of revision knee services in England has recently undergone a substantial change and now such services are provided around regional networks of care. This promises substantial advantages to the increasing number of patients with problematic knee replacements in our ageing population who will benefit from regional expertise.[8] However, it is unknown the impact of patients

residing farther from tertiary referral centres, particularly rural patients who may encounter additional difficulties associated with greater travel distance. A recent study following the outcomes of aortic surgery found that longer travel distances are associated with inferior perioperative outcomes[13]. Similar associations have been found in postoperative colorectal surgery patients [32]. As such our results are reassuring to policy makers and clinicians.

Unanswered questions and future research

There is a scarcity of evidence evaluating the patient perception of complex health interventions such as network models of care. Recent work by Kugler et al has demonstrated the willingness of patients to travel further for better outcomes in the context of total knee replacement surgery. [33] Nevertheless, patient perceptions of travelling further for their treatment should be a focus for future research in the context of revision knee patients, particularly as this is one of the top ten research priorities identified by the James Lind Alliance priority setting partnership.[34]

Conclusion

We did not observe an association in our study population between 30-day readmission rates and increasing travel distances or times between a patient's home and their treating hospital in revision knee replacement. This paper is the first to explore the relationship between travel distance and complex orthopaedic surgery and informs some concerns regarding the creation of a centralised revision knee network. This information is of utility to surgical providers and commissioners of healthcare services. Furthermore, it can inform patient-led decision making and the exploration of perceptions surrounding travelling for complex surgery. Although this is the first assessment in complex orthopaedic surgery, a prospective analysis will be undertaken as part of the ongoing auditing of revision knee networks in England.

Supplementary material and figures

Supplementary material S1 – Journey Time Statistics Reference Document

Supplementary material S2 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

See separate file named supplementary material S2

Supplementary material S3 – Flow of patient inclusion/exclusions

-See attached file named Supplementary Material S3

Supplementary material S4 – R Code

See attached file named Supplementary Material S4

Supplementary material S5 –Scatterplot for imputed data: A comparison between imputed values and observed values following multiple random imputation. Imputed values in “blue”, observed values in “grey”. Imputation 0 on X axis refers to original dataset. Subsequent random imputations labelled 1 to 5 on x axis.

Figure 1 -

(Left) Scatterplot showing correlation between patient age and travel distance. Red line represents linear regression trend. Spearman's rank correlation is presented in chart.

(Right) Scatterplot showing correlation between social deprivation and patient travel distance. Red line represents linear regression trend. Spearman's rank correlation is presented in chart.

Figure 2 - **Predicted probability of emergency readmission at 30 days by straight line patient travel distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 3 knots at 5%, 50% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 3 - **Predicted probability of prolonged length of inpatient stay at by patient straight line travel distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 4 - **Predicted probability of prolonged length of inpatient stay at by patient driving distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 5 - **Predicted probability of prolonged length of inpatient stay at by patient driving time from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Contributorship

Alex Matthews: Conceptualisation, Methodology, Project Administration, Investigation, Data Curation, Formal Analysis, Visualisation, Writing - original draft, Writing - review and editing. This author is the guarantor and is responsible for the content

Jonathan P Evans: Conceptualisation, Supervision, Writing - review & editing

Jonathan T Evans: Supervision, Writing - review and editing

Sarah E Lamb: Conceptualisation, Supervision, Writing - review and editing

Andrew Price: Conceptualisation, Supervision, Writing - review and editing

William Gray: Conceptualisation, Supervision, Methodology, Writing - review and editing

Tim Briggs: Supervision, Writing - review and editing

Andrew Toms: Conceptualisation, Supervision, Writing - review and editing

Public and Patient Involvement statement

The study's chief investigator (AT) led the James Lind Alliance 'Revision Knee Replacement' priority setting partnership. This group of patients, carers and health care professionals identified the need to investigate the best way of organising revision knee replacement surgery to improve patient outcomes as one of their top 10 research questions. Patients were therefore directly involved in the development of the study's aims and objectives. The results of the study will be disseminated to the members of this group prior to publication.

Copyright/licence for publication Statement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing Interests

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical Approval

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data. Ethical approval was not required.

Funding

No funding was obtained to carry out this study

Data Sharing

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

References

1. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet. 2018;392(10158):1672-82.
2. Sabah SA, Hedge EA, Abram SGF, Alvand A, Price AJ, Hopewell S. Patient-reported outcome measures following revision knee replacement: a review of PROM instrument utilisation and measurement properties using the COSMIN checklist. BMJ Open. 2021;11(10):e046169.
3. Sabah S, Knight R, Alvand A, Murray D, Petrou S, Beard D, Price A. No exponential rise in revision knee replacement surgery over the past 15 years: an analysis from the National Joint Registry. Osteoarthritis and Cartilage. 2022;30(12):1670-9.
4. Rowland T, Lindberg-Larsen M, Santy-Tomlinson J, Jensen CM. A qualitative study of patients' experiences before, during and after surgical treatment for periprosthetic knee joint infection; "I assumed it had to be like that … ". International Journal of Orthopaedic and Trauma Nursing. 2023;48:100992.
5. Garbarino LJ, Gold PA, Sodhi N, Anis HK, Ehiorobo JO, Boraiah S, et al. The effect of operative time on in-hospital length of stay in revision total knee arthroplasty. Ann Transl Med. 2019;7(4):66.
6. Liodakis E, Bergeron SG, Zukor DJ, Huk OL, Epure LM, Antoniou J. Perioperative Complications and Length of Stay After Revision Total Hip and Knee Arthroplasties: An Analysis of the NSQIP Database. The Journal of Arthroplasty. 2015;30(11):1868-71.
7. Briggs TWR. A National Review of Adult Elective Orthopaedic Services in England. In: Time GIRF, editor. 2015.
8. Kalson NS, Mathews JA, Phillips JRA, Baker PN, Price AJ, Toms AD, et al. Revision knee replacement surgery in the NHS: A BASK surgical practice guideline. The Knee. 2021;29:353-64.
9. Yapp LZ, Walmsley PJ, Moran M, Clarke JV, Simpson AHRW, Scott CEH. The effect of hospital case volume on re-revision following revision total knee arthroplasty. J Bone Joint Surg Br. 2021;103-B(4):602-9.
10. Halder AM, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Re-Revision Rate Following Aseptic Revision Total Knee Arthroplasty: An Analysis of 23,644 Cases. J Arthroplasty. 2020;35(4):1054-9.
11. Jeschke E, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Revision Rate and Mortality Following Revision Total Hip Arthroplasty: An Analysis of 17,773 Cases. J Arthroplasty. 2019;34(9):2045-50.
12. Bloch BV, Matar HE, Berber R, Gray WK, Briggs TWR, James PJ, Manktelow ARJ. The impact of a revision arthroplasty network on patient outcomes. J Bone Joint Surg Br. 2023;105-b(6):641-8.
13. Feldman ZM, Zheng X, Mao J, Sumpio BJ, Mohebali J, Chang DC, et al. Longer patient travel distance is associated with increased non-index readmission after complex aortic surgery. Journal of Vascular Surgery. 2023;77(6):1607-17.e7.
14. Saunders RS, Fernandes-Taylor S, Kind AJ, Engelbert TL, Greenberg CC, Smith MA, et al. Rehospitalization to primary versus different facilities following abdominal aortic aneurysm repair. J Vasc Surg. 2014;59(6):1502-10, 10.e1-2.
15. Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes--systematic review and recommendations for reporting. Eur J Public Health. 2013;23(1):86-92.

16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
17. NHS Digital. Hospital Episode Statistics(HES) Analysis Guide. London, UK: NHS Digital; 2018.
18. Information Standard Board for Health and Social Care. Anonymisation Standard for Publishing Health and Social Care Data Specification. London, UK: NHS Digital; 2013.
19. NHS Digital. Delivering linked datasets to support health and care delivery and research. London, UK: NHS Digital; 2018.
20. Department for Transport. Journey Time Statistics, notes and definitions. London, UK: Department for Transport; 2019.
21. Kalson NS, Mathews JA, Miles J, Bloch BV, Price AJ, Phillips JRA, et al. Provision of revision knee surgery and calculation of the effect of a network service reconfiguration: An analysis from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Knee. 2020;27(5):1593-600.
22. Deere K, Whitehouse MR, Kunutsor SK, Sayers A, Price AJ, Mason J, Blom AW. How long do revised and multiply revised knee replacements last? An analysis of the National Joint Registry. Lancet Rheumatol. 2021;3(6):e438-e46.
23. Kamath CC, O'Byrne TJ, Lewallen DG, Berry DJ, Maradit Kremers H. Neighborhood-Level Socioeconomic Deprivation, Rurality, and Long-Term Outcomes of Patients Undergoing Total Joint Arthroplasty: Analysis from a Large, Tertiary Care Hospital. Mayo Clin Proc Innov Qual Outcomes. 2022;6(4):337-46.
24. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17(1):162.
25. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016;4(2):30.
26. Sabah SA, Knight R, Alvand A, Palmer AJR, Middleton R, Abram SGF, et al. Patient-Relevant Outcomes Following First Revision Total Knee Arthroplasty, by Diagnosis: An Analysis of Implant Survivorship, Mortality, Serious Medical Complications, and Patient-Reported Outcome Measures Utilizing the National Joint Registry Data Set. J Bone Joint Surg Am. 2023;105(20):1611-21.
27. Hardy F, Heyl J, Tucker K, Hopper A, Marchã MJ, Briggs TWR, et al. Data consistency in the English Hospital Episodes Statistics database. BMJ Health Care Inform. 2022;29(1).
28. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.
29. Farrow L, Kennedy JW, Yapp L, Harding T, Walmsley P. Provision of revision knee arthroplasty services across Scotland: A national audit. Knee. 2023;42:312-9.
30. Paterson C, Baarts C, Launsø L, Verhoef MJ. Evaluating complex health interventions: a critical analysis of the 'outcomes' concept. BMC Complement Altern Med. 2009;9:18.
31. Jackson KL, Glasgow RE, Mone MC, Sheng X, Mulvihill SJ, Scaife CL. Does travel distance influence length of stay in elective pancreatic surgery? HPB (Oxford). 2014;16(6):543-9.
32. Cloyd JM, Huang L, Ma Y, Rhoads KF. Predictors of readmission to non-index hospitals after colorectal surgery. The American Journal of Surgery. 2017;213(1):18-23.

33. Kugler CM, De Santis KK, Rombey T, Goossen K, Breuing J, Könsgen N, et al. Perspective of potential patients on the hospital volume-outcome relationship and the minimum volume threshold for total knee arthroplasty: a qualitative focus group and interview study. BMC Health Serv Res. 2021;21(1):633.
34. Mathews JA, Kalson NS, Tarrant PM, Toms AD. Top ten research priorities for problematic knee arthroplasty. Bone Joint J. 2020;102-b(9):1176-82.

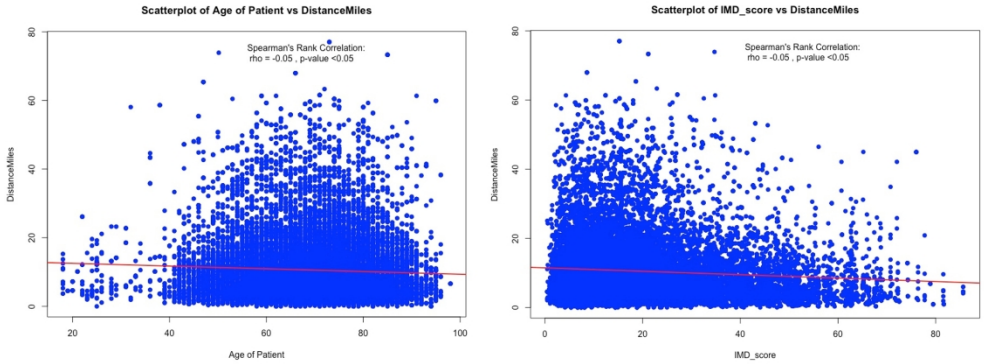


Figure 1

354x136mm (300 x 300 DPI)

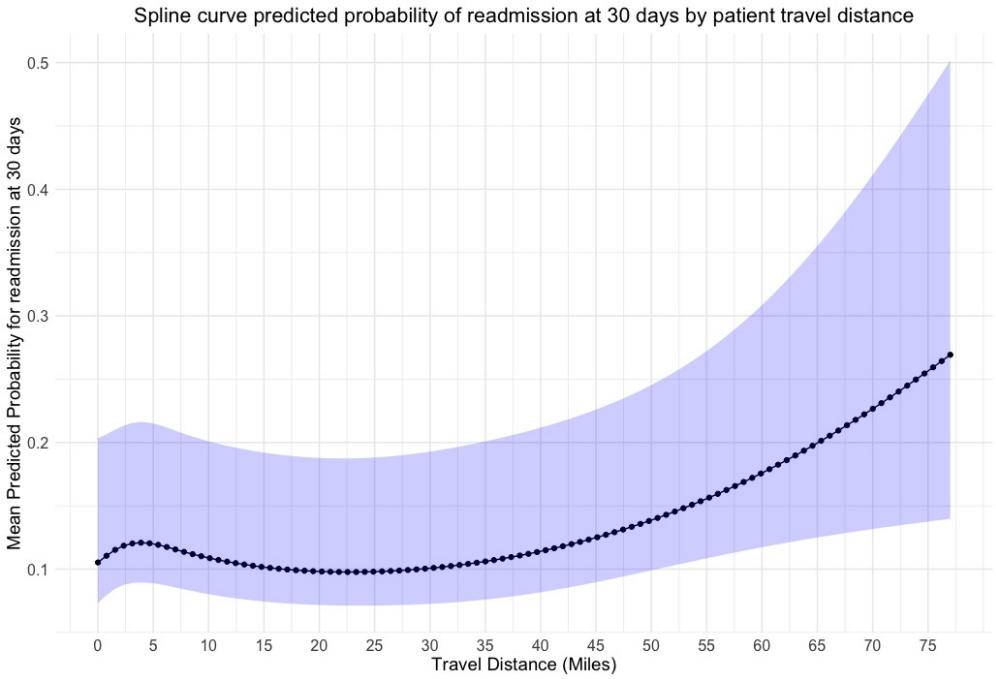


Figure 2

87x59mm (300 x 300 DPI)

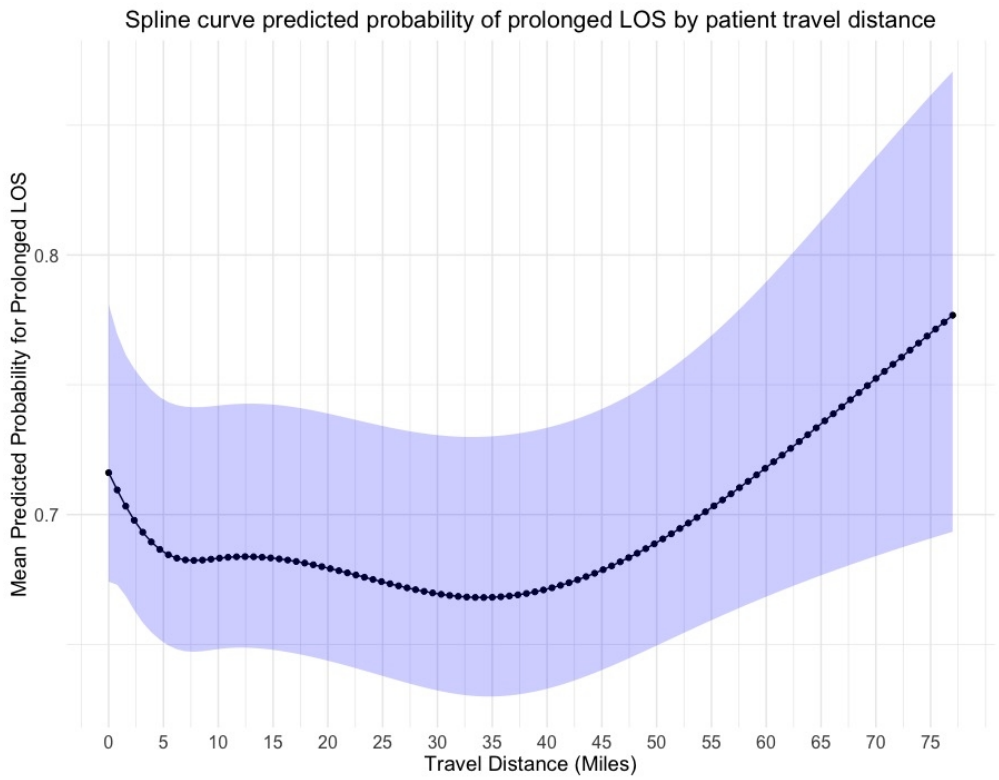


Figure 3

76x59mm (300 x 300 DPI)

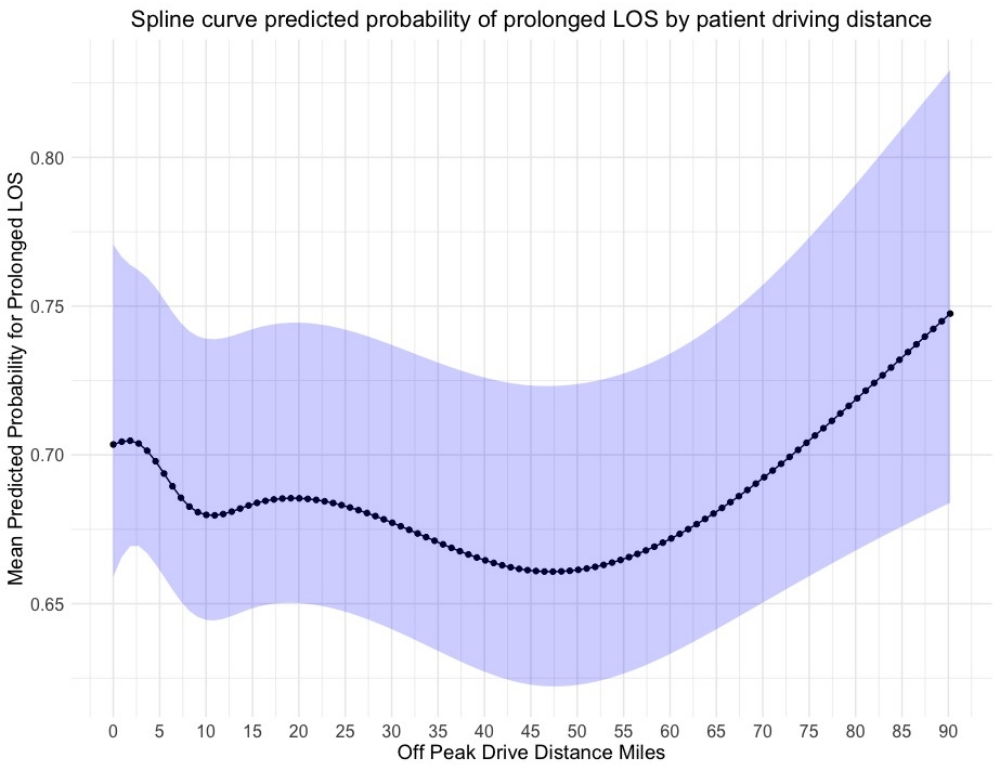


Figure 4

78x60mm (300 x 300 DPI)

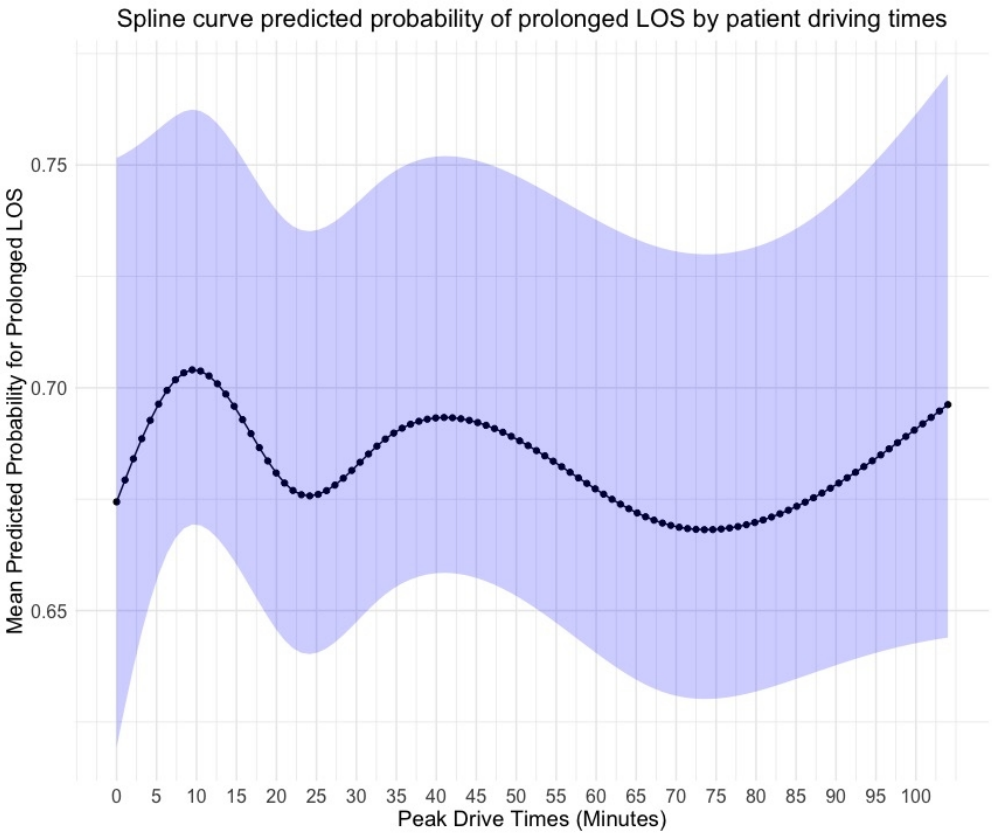


Figure 5

72x60mm (300 x 300 DPI)

Journey Time Statistics: Notes and Definitions

About this release

This publication supports the latest statistics on journey times.

In this publication

Further information

Public enquiries
020 7944 3077
vehicles.stats@dft.gov.uk
Media enquiries
020 7944 3066

Overview

This note provides information on the methodology used, the source data and definitions of key terms for calculating Journey Time Statistics.

These annual statistics were first published in December 2015 for the year 2014 and have been developed from the earlier Accessibility Statistics published for 2007 to 2013.

The Journey Time Statistics produced by DfT consists of theoretical journey times calculated by modelling journeys between known sets of origins and destinations. It uses information on the road network, traffic speeds and public transport timetables in England.

The relevant Journey Time Statistics calculation is varied for origins and destination to meet a variety of needs. Two sets of analysis are published:

- Access to key services; and
- Connectivity

Origin indicators

These indicators measure the number of different services in a particular area that users can reach within a given time.

Destination indicators

These indicators measure the proportion of users that can access a service within a certain time.

The 'user' populations for each service in the destination indicators are:

Employment	16-74 year olds
Primary schools	5-10 year olds
Secondary schools	11-15 year olds
Further education	16-19 year olds
All other services	All households

Key services

- Employment centres: Data used are the number of jobs in a Lower Super Output Area (LSOA). The data tables include results for employment centres of 3 different sizes (100-499 jobs, 500-4,999 jobs and at least 5,000 jobs). For the key services average, the 500-4,999 jobs definition is used for employment.
- Education: Locations of all open Primary schools, Secondary schools, Further Education and Sixth Form Colleges.
- General Practice (GP) surgeries: For 2017 based on the Patients Registered at a GP Practice dataset released by NHS Digital – previously this was based on a filtered dataset of NHS prescribers released by NHS Digital.
- Hospitals: Based on hospitals that are registered with the Care Quality Commission (CQC) and are managed by Acute Trusts.
- Food stores: Locations of grocery, supermarkets or convenience stores.
- Town centres: Locations of Town centres using a central focal point for the town mapped to the nearest road.

Geography

- Local authorities

In some parts of England there are two tiers of local authorities, and in others a single unitary authority. Statistics have been calculated for both types of authority - around 360 in all. These vary considerably in size, from a population of a few tens of thousands to over a million.

- Lower Layer Super Output Areas (LSOA)

LSOAs are small areas designed to be of a similar population size, with an average of approximately 1,500 residents or 650 households. There are 32,844 Lower-layer Super Output Areas (LSOAs) in England. They were determined by the Office for National Statistics for the reporting of small area statistics and are derived from the 2011 Census.

- Urban and rural definitions

This report uses the Defra Rural-Urban Classification, based on 2011 Census Output Areas. The Rural-Urban Classification defines areas as rural if they fall outside of settlements with more than 10,000 resident population. See Defra's Definitions and Local Authority Classification for more details.

Journey time calculations

The journey time calculations are carried out using a commercially available software package called TRACC, owned by Basemap. TRACC is a desktop application that uses public transport and highways data to create journey times from origins to destinations. It uses timetable information showing both arrival and departure times at stops from public transport services against a specific time/day period. Highways information from road networks are used to fill the gaps between public transport services by creating a linear network that connects the origins, destinations and stops together. This provides a fully routable network of nodes and lines which is saved on file as a graph network. The graph network has various constraints which can be altered to suit the user need such as distance travelled, interchange delays on public transport and stopping limitations on road networks. The TRACC software then queries the graph network with origin and destination co-ordinates and uses the Dijkstra shortest path algorithm to route between these points. This is an algorithm for finding the shortest distance for travel between the graph networks.

For a public transport journey, the journey time produced includes all walking elements of the journey, i.e. the walk from the origin of the journey to the road, from the road to public transport stops, any interchange of public transport using the road and then from the final stop to the destination via the road, and finally from the nearest point on the road network to the destination. The journey assumes arrival at the first stop one minute before the initial departure, with any subsequent interchange waiting times included as part of the final journey time.

Car, cycle or walk only journeys are similar except that once the road network is reached the journey proceeds link by link along the road network at speeds governed by data held in the model. These are specific to the mode, the road type, and in some cases the individual road link.

The 10 shortest journey times from each origin (i.e. Output Area) are calculated for each destination type. For the public transport / walking mode these consist of the 10 shortest journey times by either walking or public transport, after applying a 5 minute penalty for any journeys using public transport (to represent travellers arriving slightly early at the first stop).

The journey times are representative of the ‘morning peak’. This is made explicit for public transport / walking by requiring the journey to be completed between 7 and 10am, and for car journeys by using average traffic speeds for between 7 and 10am. For the cycle mode no actual speed data are available. The cycle speeds used are default assumptions, and are not based on a particular time of day.

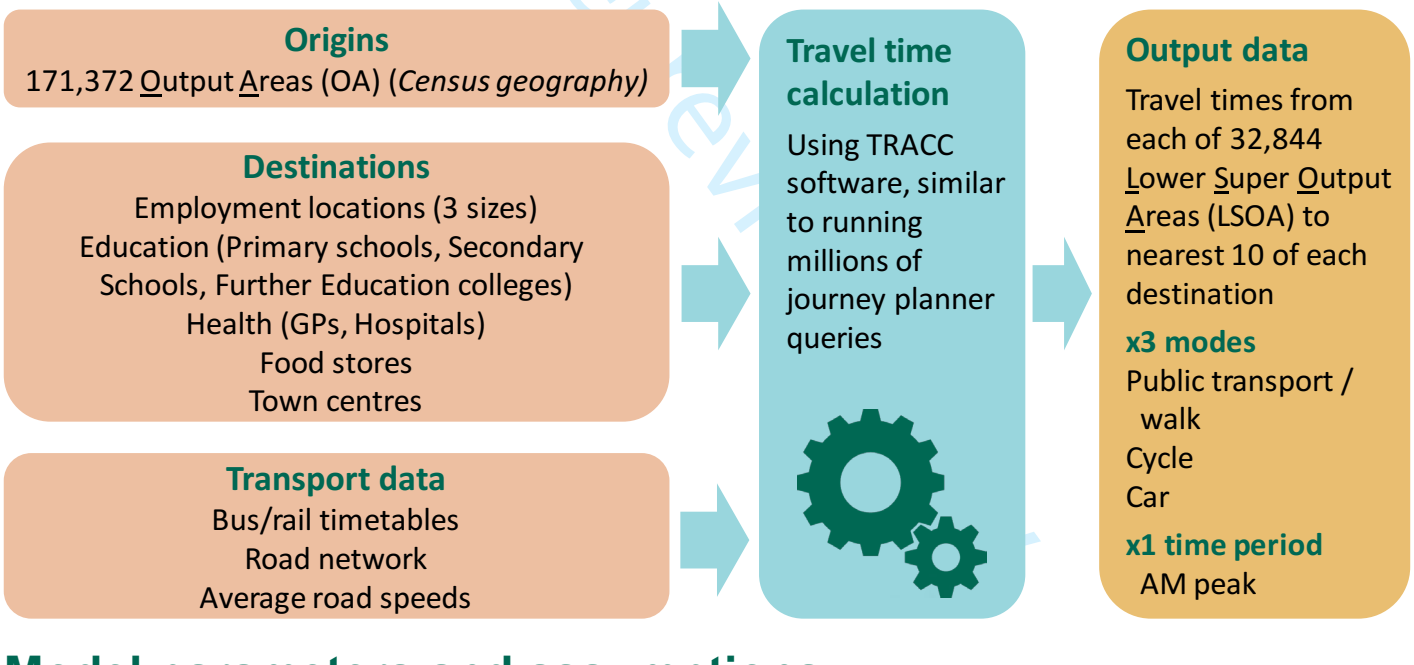
Access to key services

The Access to Services analysis applies the Journey Times methodology to origins consisting of residential neighbourhoods and destinations consisting of centres of employment and a range of key local services. Journey times are calculated for three modes of transport: public transport; driving; and cycling. These journey times are then used to generate further indicators, as described in the **Outputs Section**.

The Access to Services calculation process and the coverage of the data set are very similar to those of the Accessibility Statistics from which they were developed. However, the calculation algorithm and a number of other features of the design are different, so the results are not directly comparable.

The statistics are designed to represent as much as possible the situation on a **Tuesday in October** of **the year to which they relate**. Data for the second week of October are used in the analysis, since this provides a fairly typical week, unaffected by major national holidays, school holidays or other seasonal effects. The origins, destinations and public transport timetables used are as far as possible for this date. The traffic data are averages for the preceding 12 months up to and including August. The road networks are those current at the start of the traffic data year.

Outline of access to services calculation process



Model parameters and assumptions

General parameters

Maximum journey time of **2 hours.**

Maximum journey distance of **100km.**

Walking

These apply to both:

- walking between origin / destination and the transport networks at both ends of a journey by

any mode;

► walk only journeys as part of the public transport / walk mode.

Maximum straight line distance between origin / destination and road network of **2km.** The algorithm will always use nearest point on network. For cycle or car modes, travel by cycle or car begins from this point. For public transport/walk, traveller walks along road network to the most suitable public transport stop, or direct to the destination if this is quicker.

Walking speed on road/path network of **4.8km/h.**

Walking speed off road/path network of **4.0km/h.**

Public transport

Interval within which door-to-door journey must be completed (required for timetable selection) is **7am to 10am on a Tuesday.**

Maximum walk distance of **3km** - this applies to walks from origin to first public transport stop, from last stop to destination, and also walking directly from origin to destination without using public transport at all.

Maximum number of potential first public transport stops considered in routing algorithm is **100** (starting with the closest to origin).

Allowance for catching first public transport service is **5 minutes** - added to any journey that involves boarding one or more public transport services.

Public transport speed – this is provided implicitly by the timetable information.

Interchange time of **5 minutes** (minimum interval allowed between arriving at a stop and catching another service).

Maximum straight line distance between public transport interchanges of **500m**.

Stop clustering at **150m** – groups together public transport stops within this distance of one another to speed up processing. The individual timetables for each service are retained.

Cycling speeds

Road Type	Speed
Motorway	0.0 km/h
Urban Motorway	0.0 km/h
A road	16.0 km/h
B road	16.0 km/h
Minor road	16.0 km/h
Local street	16.0 km/h
Private road – restricted access	4.8 km/h
Private road – public access	16.0 km/h
Pedestrian street	4.8 km/h
Alley	4.8 km/h

Parking time of **5 minutes** - added to all cycle journeys.

Car speeds

Type of road	2014	2015	2016	2017
	Default speeds (km/h)			
Motorway	79.5	77.0	77.5	77.6
Urban Motorway	79.5	77.0	77.5	77.6
A road	42.7	43.7	43.3	43.2
B road	41.6	43.0	42.2	41.9
Minor road	36.8	37.5	36.8	36.3
Local street	19.2	17.8	18.8	18.3
Private road – restricted access	17.0	16.7	16.2	15.3
Private road – public access	14.8	15.2	15.1	13.6
Pedestrian street	0.0	0.0	0.0	0.0
Alley	0.0	0.0	0.0	0.0

Car speeds are calculated for specific links where more than 200 records exist otherwise the default speeds are used. Minimum journey time for a journey that uses a car is **5 minutes**.

Time at junctions

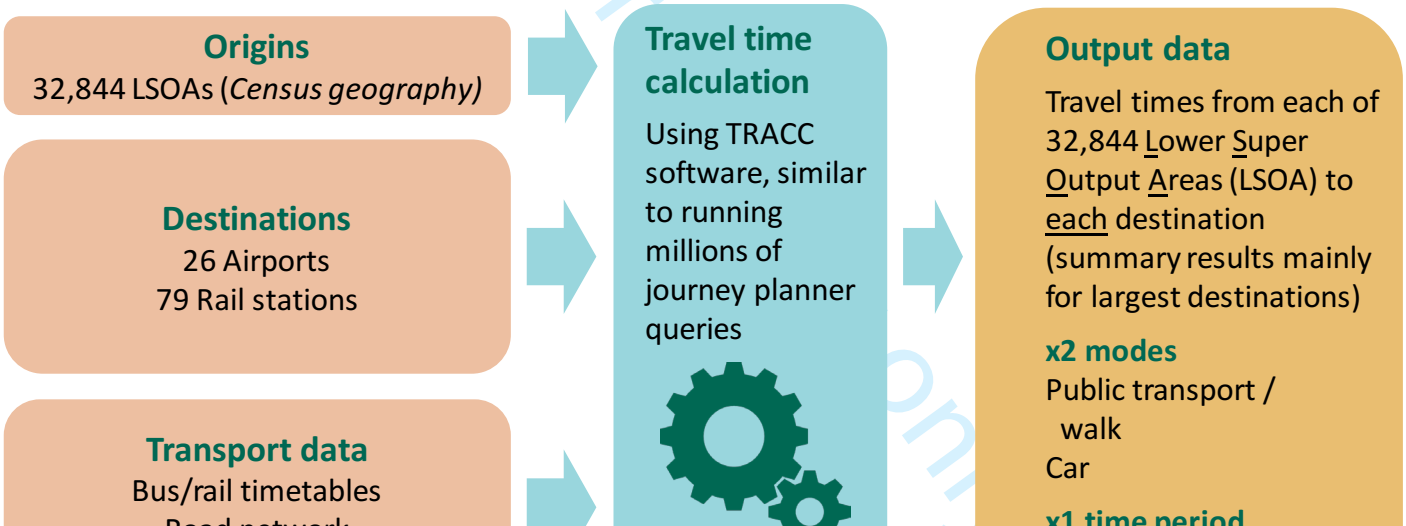
Road normalisation is used for all modes of transport which converts each road link to a straight line to speed up processing. The true link length is retained for accurate speed/time calculations, but there could be a small effect on the calculation of shortest distance from the road network to destination points. Effect for origins is minimal due to origins being constrained to road nodes.

Connectivity

These experimental analyses are intended to apply the Journey Times methodology to a range of more strategic or economically significant destinations than the primarily local services covered by the Access to Services analyses; including airports and railway stations. The principle difference in the Connectivity approach from that of the Access to Services analyses is that journey times are calculated, as far as possible, to all accessible locations, rather than to just the nearest 10 examples. This tends to result in a much larger data set being generated. In some cases a longer maximum journey time may be allowed although this may depend on what is considered reasonable for the type of destination. Given these factors, a less detailed origin data set may be used than for Access to Services. This is both necessary, to limit the size of the data set, and acceptable where the typical journey lengths are longer.

The first connectivity analyses published using the new Journey Time methods were released in Journey Time Statistics 2015, published in April 2017, for two destination sets – airports and rail stations. These analyses using the Journey Times methods superseded two earlier Connectivity Statistics reports published in 2014 and 2015 based on the old accessibility statistics methods, in the same way that the new Access to Services analyses have replaced the earlier Accessibility Statistics. Again, the connectivity results produced using the old and new methods are not directly comparable.

Outline of Connectivity calculation



Model parameters and assumptions

Origins	Population weighted centroids (the central point) of 32,844 English LSOAs as specified in the 2011 Census geography. These points were then constrained to the nearest road node, as for Access to Services method.

Journey Time Calculation	As for Access to Services, for public transport / walking and car modes only, except that a maximum journey time of 240 minutes and maximum straight line distance of 400km is allowed.
Outputs	Generally similar to Access to Services, with different journey time classifications as appropriate. Journey time results to specific destinations are included – this is the key difference in the Connectivity analyses. 'Average journey times' and 'nearest' destinations should be used with caution. The average journey times exclude results for areas with no available connection under 240 minutes, which may become significant in remote areas and for destinations are a great distance from the origin. The 'nearest' destination is the destination with the shortest average journey time across the whole area considered – which will be relatively large in the case of local authority level results.

Data sources

Origins

The origins used for all Access to Services calculations are the 171,372 English Output Areas (OA) as specified in the 2011 Census geography.

To provide the actual journey start point in each OA, the population weighted centroid of the OA was shifted to the nearest node (i.e. junction) on the road network. This was to avoid biasing the journey time results where the centroid of the OA was a long way from a road. In fact it is rare for an OA centroid to be more than about 100 metres from a road – only a tiny handful of OA in remote areas have centroids as much as 1km from a road. The OA centroids have been shifted onto the nearest road node rather than the nearest point on a road in order to reduce issues arising from normalising the road network.

Origin	Data source for the origin points
All	Data: Population centroid of each Output Area in 2011. Source: ONS 2011 Census Boundaries. Further information: http://geoportal.statistics.gov.uk

Destinations

The destinations used consist of three different sizes of employment centre and the locations of seven other types of key local service. For each of these key services a nationally consistent data set has been identified or derived – further information on these is provided in this section.

Each destination is located by a 6-figure National Grid reference. For the employment destinations this is taken to be the population weighted centroid of the LSOA.

Destination	Number of locations			
	2014	2015	2016	2017
Employment centres (small)	16,465	16,625	16,930	17,194
Employment centres (medium)	9,235	9,460	9,707	10,241
Employment centres (large)	645	676	719	785
Primary schools	16,463	16,484	16,655	16,927
Secondary schools	3,365	3,376	3,381	3,174
Further education colleges	2,624	2,606	2,418	2,304
GPs	9,257	11,167	9,128	7,353
Hospitals	296	278	278	277
Food stores	19,549	19,746	21,665	20,987
Town centres	1,211	1,211	1,211	1,211

The data source for GP surgeries was reviewed and replaced for 2017.

Access to key services

Destinations 2017	Data source for the locations of the service	Data source for users of the service
Employment	Data: Number of jobs available in a LSOA in the year before the calculation year.	Data: Number of 16-74 year olds in each output area.
	Source: ONS Business Register Employment Survey.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://www.nomisweb.co.uk/default.asp	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Primary schools	Data: Location of all open primary schools in September of calculation year.	Data: Number of 5-10 year olds in each output area.
	Source: The Department for Education (DfE) Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Secondary schools	Data: Location of all open secondary schools in September of calculation year.	Data: Number of 11-15 year olds in schools in each output area.
	Source: DfE Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Further education colleges	Data: Location of all open further education and sixth form colleges/school sixth form in September of calculation year.	Data: Number of 16-19 year olds in each output area.
	Source: DfE Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.

Destinations 2017	Data source for the locations of the service	Data source for users of the service
GPs	Data: Locations of GP surgeries with registered patients in October of calculation year.	Data: Number of households in each output area.
	Source: NHS Digital table of Registered patients at GP practices	Source: 2011 Census + Local Authority (LA) updates from the Ministry of Housing, Communities & Local Government (MHCLG) mid-year household projections of calculation year.
	Further information: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections
Hospitals	Data: Location of hospitals.	Data: Number of households in each output area.
	Source: Care Quality Commission - Directory of places that provide care.	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: http://www.cqc.org.uk/content/how-get-and-re-use-cqc-information-and-data	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Destinations 2017	Data source for the locations of the service	Data source for users of the service
Food stores	Data: Location of grocery/ supermarkets or convenience stores in October of calculation year.	Data: Number of households in each output area.
	Source: The Local Data Company	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: https://www.localdatacompany.com/	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections
Town centres	Data: Location of town centres in 2004.	Data: Number of households in each output area.
	Source: MHCLG Town Centre and retail planning statistics for England and Wales.	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: https://data.gov.uk/dataset/ed07b21f-0a33-49e2-9578-83ccbc6a20db/english-town-centres-2004	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

GP destination data

The GP surgery destinations used from 2014 to 2016 are based on the list of practices maintained by the Organisational Data Service of the Health & Social Care Information Centre, and published at https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data. This was supplemented with information on branch surgeries from the same source. Grid references were derived from the postcode using the Office for National Statistics (ONS) Postcode Address File. Practices with identical postcodes were taken to be duplicates or co-located, and all additional records after the first were removed.

From 2017, the list of GP locations is taken from the NHS Digital publication of Registered patients at GP practices for October of the calculation year. This had the effect of reducing the number of locations in the dataset, but removed the need for manual adjustments and produces a more stable list defined as GP practices with registered patients. Grid references were derived from the postcode using the Office for National Statistics (ONS) Postcode Address File.

Hospital destination data

The starting point for hospital sites is the Care Quality Commission's (CQC) list of 'active locations' dataset, which is thought to be the most-up-to date and freely available source of data on individual National Health Service (NHS) and social care 'sites' or hospitals. A criteria was developed in consultation with the Department of Health to reduce the list down to capture only the key hospitals. The following have been removed and individual records have been inspected to remove further examples of these cases and for any duplicates:

- care home records;
- non-NHS providers;
- sites not associated with acute providers;
- any remaining sites that are associated with Specialist Trusts (usually single speciality Trusts or Sites);
- records where it is evident from the name that the record is not a hospital (e.g. headquarters, specialist units.)

This gave a final list of 278 hospitals in 2017 run by Acute (non-specialist) Trusts. As well as covering all general hospitals this will still include some with a largely or entirely community or rehabilitation role, where these happen to be managed by an Acute Trust. It was considered on balance better to leave these in the list, rather than risk adding further subjectivity to the selection. Whilst not perfect, it is considered that the resulting list is a significant improvement on that used previously.

Steps taken to produce hospital data set
Remove records where **Care Home** = Y
Remove records where **Provider ID** begins 1-
Keep records where **Benchmark Group** is Care Home or **Cluster Group** is Acute
Filter the trust site locations by name to remove obvious non-hospital sites. Key words used for this process are: birth, dental, house, clinic, grange, lodge, infirmary, health, community, unit, surgery, centre
Manual review of remaining locations

Employment destination data

The employment centres are defined by the number of jobs existing in each English LSOA, taken from the Business Register Employment Survey. Large Employment Centres are defined as those with 5,000 or more jobs, Medium Employment Centres as those with 500 or more jobs, up to 4,999 and Small Employment Centres as those with 100 or more jobs, up to 499.

Data are downloaded from the Nomis website; although LSOA level BRES data has safeguarded access, access can be requested through the site. The chosen data download options are LSOA2011 geography, date as calculation year, variable as employment status where the value is employed, and the measure chosen is a count.

For the 2016 destination set, the BRES changed from 2001 census geography to 2011 census geography. The majority of LSOA boundaries are unchanged between these datasets, but some have been merged or split. Therefore the employment destination indicators are not strictly comparable between 2015 and 2016 Journey Time statistics. See https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography for further information.

Education destination data

The education destination datasets are taken from the Department for Education database of educational establishments. The database was filtered to remove those establishments that were not open during the school year starting in September of the calculation year. Further filters were applied to remove special educational establishments, boarding schools and selective schools, and then to select schools at each phase of education for primary and secondary schools and further educational establishments. The following table lists the filters used.

Phase of Education	Code Variable	Variable	Selected codes and values	
All Schools	OpenDate			30/08/17 or earlier; NULL
	CloseDate			30/08/18 or later; NULL
	TypeOfEstablishment_Code_	TypeOfEstablishment	1	Community school
			2	Voluntary aided school
			3	Voluntary controlled school
			5	Foundation school
			6	City technology college
			12	Foundation special school
			18	Further education
			28	Academy sponsor led
			29	Higher education institutions
			31	Sixth form centres
			32	Special post 16 institution
			34	Academy converter
			35	Free schools
			36	Free schools special
			39	Free schools 16 to 19
			40	University technical college
			41	Studio schools
			45	Academy 16-19 converter
			46	Academy 16 to 19 sponsor led
	Boarders_Code_	Boarders	0	Not applicable
			1	No boarders
			9	NULL
	AdmissionsPolicy_Code_	AdmissionsPolicy	0	Not applicable
			4	Non-selective
			9	NULL
Primary schools	PhaseOfEducation_Code_	PhaseOfEducation	2	Primary
			3	Middle deemed primary
			7	All through

Phase of Education	Code Variable	Variable	Selected codes and values	
Secondary schools	PhaseOfEducation_Code_	PhaseOfEducation	0	Not applicable
			4	Secondary
			5	Middle deemed secondary
			7	All through
	Statutory High age		>=16	
	Statutory Low age		< 16	
FE	PhaseOfEducation_Code_	PhaseOfEducation	4	Secondary
			5	Middle deemed secondary
			6	16 plus
			7	All through
	Statutory High age		>16	
	OfficialSixthForm_Code_	OfficialSixthForm	0	Not applicable
			1	Has a sixth form
			9	NULL
OR				
FE	EstablishmentTypeGroup__code_	EstablishmentTypeGroup	1	Colleges

Food Stores destination data

The food stores destination dataset is purchased from The Local Data Company and includes all branches of multiple food store chains. Although some data are available for independent food stores, this only exists within town centres and so has not been included.

Connectivity

Destinations	Data source for the locations of the service	Data source for users of the service
Airports	Data: Location of GB airports excluding highlands and islands of Scotland Source: National Public Transport Access Nodes Further information: https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan	Data: Number of households in each output area. Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year. Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Destinations	Data source for the locations of the service	Data source for users of the service
Railway stations	Data: Location of larger (category A, B and C1) rail stations in GB Source: Network rail classification Further information: http://webarchive.nationalarchives.gov.uk/20101007153226/http://www.dft.gov.uk/pgr/rail/passenger/stations/betterrailstations/ http://archive.nr.co.uk/browse%20documents/rus%20documents/route%20utilisation%20strategies/network/working%20group%202%20-%20stations/networkrusstations.pdf	Data: Number of households in each output area. Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year. Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Transport network data

Travellers moved between their original and their destination via one or more of the following transport networks, depending on the mode of transport being modelled. For all modes, travellers will probably also need to walk between their origin / destination and the transport network. For some short journeys, it may be quicker for travellers to walk directly to their destination, rather than using public transport at all – this is why public transport / walking results are modelled as a combined mode.

Public transport

National public transport timetable data are publically available. Data for bus, local coach and other local transport services (e.g. light rail, metro, and ferry) are captured in the Traveline National Data Set (TNDS), rail timetable data are published by the Association of Train Operating Companies (ATOC), and national coach services in the National Coach Data Set (NCDS).

Walk

The walking network is represented by the road and urban path elements of the Integrated Transport Network produced by the Ordnance Survey.

Cycle

The cycling network is represented by the road network including cycle paths and bridleways from the Integrated Transport Network. Cycle journeys are also allowed to use footpaths at walking pace.

Car

The car network is represented by the road component of the Integrated Transport Network.

Data on actual vehicle speeds on each road network link (generally the stretch of road between 2 nodes, or junctions) is obtained from Trafficmaster Satnav devices and are used to estimate car speeds. These data are used to calculate annual average traffic speeds on each link of the road network (by direction if the link is bi-directional). These are used as the link speeds for cars in the modelling. Where the Trafficmaster sample for an individual link is too small, national averages of the same data for the particular road type are used instead. This is an innovation from 2014. Previously the sample was too small and the model reverted to default assumptions for car speed based on road type which were much higher than the Trafficmaster averages, resulting in some inconsistency in the model.

Outputs

The journey time results are used to create the following indicators for publication:

Indicator	Description
Minimum journey time	The shortest of the ten journey time results.
Origin indicators	Four measures, the number of destinations (up to the maximum of 10) that can be reached from a given origin within 15, 30, 45 and 60 minutes.
Destination indicators	Four measures, the percentages of service users within the given geographical area who can access at least one service location within 15, 30, 45 and 60 minutes.

Each of these indicators is calculated for each mode and each destination type, and at a number of geographical scales as follows:

- England
- Region
- Local Authorities, including London Boroughs, Metropolitan districts, Unitary authorities, Counties and non-Metropolitan districts, also Inner and Outer London and former Metropolitan counties
- 2011 Lower layer Super Output Area
- 2011 Defra Rural/Urban Classification

The indicators for each geography are calculated as population weighted averages. In other words, the average minimum journey time for an area, B, is:

$$mjt(B)= \sum(i=1)^n(mjt(OAi)\times pop(OA\_i))/pop(B)$$

where mjt(B) is the minimum journey time in area B, mjt(OAi) is the minimum journey time of the ith of n output areas making up area B, and pop(B) and pop(OAi) are the user populations resident in area B and output area i respectively.

The service user populations used in the above weighting, and in the destination indicators, depend on the destination type, as follows:

Destination type	Service user population basis
Employment centres	Resident population of working age (16-74 years)
Primary schools	Population aged 5-10
Secondary schools	Population aged 11-15
Further education colleges	Population aged 16-19
GPs, hospitals, food stores, town centres	Number of households
Average key services	Resident population of working age (16-74 years)

Strengths and Weaknesses

In using the data, the following points should be kept in mind:

► All journey times are compiled on a consistent basis across the country.

► The statistics are based on the calculation of theoretical journey times, they are not based on real journeys. They are however based on actual public transport times, and average traffic speeds on the road network.

► Although the statistics are calculated to a high level of geographical detail, some assumptions and simplifications are necessary in the modelling (for example assigning the start point of journeys to a single point in each Output Area, road speeds, interchange times for public transport).

► For 2016 we have used the 2015 BRES data to designate Lower Super Output Areas as employment centres. The 2015 BRES is the first year to use LSOAs based on the 2011 census, and although the majority of these are an exact match to the 2001 LSOAs, there are some that were merged, split or had other boundary changes. For these areas journey times from earlier years are not comparable to the 2016 journey times. This effect is more pronounced for large employment centres, as there are fewer destinations to route to.

► For particular areas, local authorities and other experts may have more detailed information allowing them to produce more accurate or detailed models of the local situation.

► Demand responsive services (e.g. bus services which have to be booked) are only included to the extent that they can be plausibly modelled, in the Traveline National Data Set.

► Since new journey calculation software was adopted for 2014, along with a significant number of other changes to the methodology, from 2014 results are not directly comparable with those for earlier years.

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

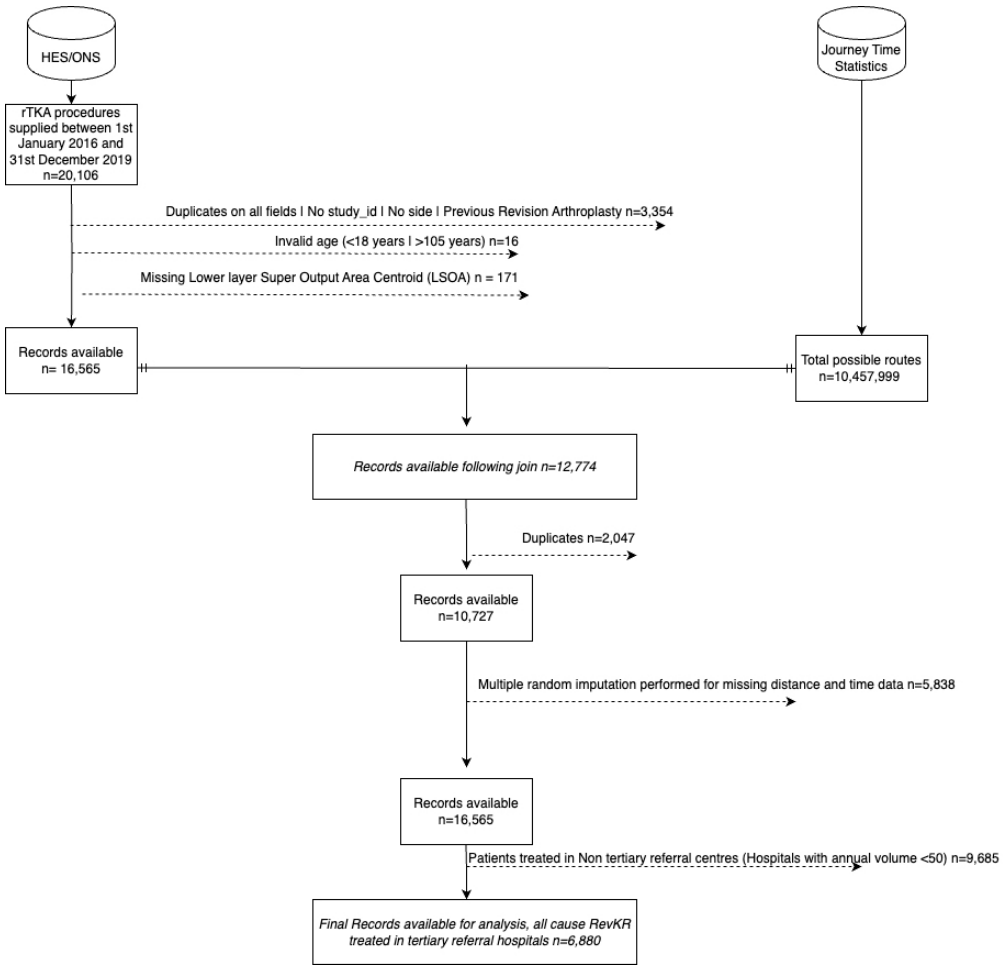
Code	Code description
OPCS-4 codes for knee revision procedures	
O180	Conversion from previous hybrid prosthetic replacement of knee joint using cement
O182	Conversion to hybrid prosthetic replacement of knee joint using cement
O183	Revision of hybrid prosthetic replacement of knee joint using cement
O184	Attention to hybrid prosthetic replacement of knee joint using cement
W400	Conversion from previous cemented total prosthetic replacement of knee joint
W402	Conversion to total prosthetic replacement of knee joint using cement
W403	Revision of total prosthetic replacement of knee joint using cement
W404	Revision of one component of total prosthetic replacement of knee joint using cement
W410	Conversion from previous uncemented total prosthetic replacement of knee joint
W412	Conversion to total prosthetic replacement of knee joint not using cement
W413	Revision of total prosthetic replacement of knee joint not using cement
W414	Revision of one component of total prosthetic replacement of knee joint not using cement
W420	Conversion from previous total prosthetic replacement of knee joint NEC
W422	Conversion to total prosthetic replacement of knee joint NEC

W423	Revision of total prosthetic replacement of knee joint NEC
W424*	Attention to total prosthetic replacement of knee joint NEC
W425	Revision of one component of total prosthetic replacement of knee joint NEC
W522†	Conversion to prosthetic replacement of articulation of bone using cement NEC
W523†	Revision of prosthetic replacement of articulation of bone using cement NEC
W532†	Conversion to prosthetic replacement of articulation of bone not using cement NEC
W533†	Revision of prosthetic replacement of articulation of bone not using cement NEC
W542†	Conversion to prosthetic replacement of articulation of bone NEC
W543†	Revision of prosthetic replacement of articulation of bone NEC
W544*†	Attention to prosthetic replacement of articulation of bone NEC
W553†	Conversion to prosthetic interposition arthroplasty of joint
W564†	Conversion to interposition arthroplasty of joint NEC
W574†	Conversion to excision arthroplasty of joint
W582†	Revision of resurfacing arthroplasty of joint
W603†	Conversion to arthrodesis and extra-articular bone graft NEC
W613†	Conversion to arthrodesis and articular bone graft NEC
W641†	Conversion to arthrodesis and internal fixation NEC
W642†	Conversion to arthrodesis and external fixation NEC
OPCS-4 codes for laterality	
Z941	Bilateral

Z942	Left-sided
Z943	Right-sided
ICD-10 codes for Infection	
T845	Infection and inflammatory reaction due to internal joint prosthesis
T846	Infection and inflammatory reaction due to internal fixation device [any site]
T847	Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts
T814	Infection following a procedure, not elsewhere classified
ICD-10 codes for fracture	
M966	Fracture of bone following insertion of orthopaedic implant, joint prosthesis or bone plate
ICD-10 codes for mechanical complications	
T840	Mechanical complication of internal joint prosthesis
T841	Mechanical complication of internal fixation device of bones of limb
T842	Mechanical complication of internal fixation device of other bones
T843	Mechanical complication of other bone devices, implants and grafts
T844	Mechanical complication of other internal orthopaedic devices, imnplants and grafts
ICD-10 codes for osteoarthritis/arthrosis	
M15-	Polyarthrosis
M17-	Gonarthrosis
M19-	Other arthrosis

OPCS-4 = Office of Populations Censuses and Surveys Classification of Interventions and Procedures version 4. ICD-10 = International Statistical Classification of Diseases and Related Health Problems, tenth revision. * Where

OPCS-4 codes Y032 (renewal of prosthesis in organ NOC) or Y037 (removal of prosthesis from organ NOC) were also used. † Where OPCS-4 codes O132 (knee NEC) or Z765 (lower end of femur NEC) or Z774 (upper end of tibia NEC) or Z787 (patella) or Z844 (patellofemoral joint) or Z845 (tibiofemoral joint) or Z846 (knee joint) or Z851 (upper tibiofibular joint) were used to identify knee as the body site.



77x73mm (300 x 300 DPI)

Supplementary material S4 – R Code

```
#Travel Times and Perioperative Outcomes in Revision Knee Replacement

setwd("/Users/alexandermatthews//OneDrive - University of Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/")


####Preparation of Data####
#load HES data

RTKA2023 <- read.csv("~/Desktop/RTKA 06-09-23 CSV.csv")

RTKA2023 <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/RTKA 06-09-23 CSV.csv")



#table only shows first 50 columns but we know there are 51 columns. Write this generic code to change preferences


rstudioapi::writeRStudioPreference("data_viewer_max_columns", 1000L)

#Some entried are blank but are read as real values and not missing data
#The table between age and sex shows three variables here
#The dataset contains non standard missing values that are not recognised as NA
#Replace empty strings with NA

RTKA2023[RTKA2023 == ""] <- NA

#Find number of incomplete cases in the data

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)

#There are 14 entries with missing data only in the age group

#check how many incomplete entries in age of patient column

sum(!complete.cases(RTKA2023$age_of_patient))

#In case of missing values there are only 14 for age of patient
#Can use imputation based on mean age
#What is the mean age of the patients
```

```
mean(RTKA2023$age_of_patient, na.rm = TRUE)

#mean age excluding missing values is 70
summary(RTKA2023$age_of_patient, na.rm = TRUE)

#Check age is normally distributed

hist(RTKA2023$age_of_patient)

#Input mean for missing values for age

RTKA2023$age_of_patient[is.na(RTKA2023$age_of_patient)] <- 69.82


#Now check number of missing values

sum(!complete.cases(RTKA2023$age_of_patient))
#Now states 0 missing values

#There are other missing values for IMD decile
##In fact there are 439 IMD score missing values

sum(!complete.cases(RTKA2023$IMD_score))


hist(RTKA2023$IMD_score)
#IMD score is non normally distributed

summary(RTKA2023$IMD_score, na.rm = TURE)

#Median IMD score is 15.543

#Use imputation to impute median for missing value

RTKA2023$IMD_score[is.na(RTKA2023$IMD_score)] <- 15.543

#Check imputation complete

sum(!complete.cases(RTKA2023$IMD_score))

#Now showing 0 missing values

#Next attach IMD decile number 6 to the missing values. As a score of 15 equates to the 6th
decile

RTKA2023$IMD_decile[is.na(RTKA2023$IMD_decile)] <- 6
```

```
#Check duplicate entry spells

duplicates <- RTKA2023[duplicated(RTKA2023),]


#No duplicates in data


#Frequencies of revisions by volume

as.numeric(RTKA2023$TV12mo)


#frequencies of revisions by trust volume
table(RTKA2023$TVcat)

#Proportions by trust volume

prop.table(table(RTKA2023$TVcat))

#Some entried are blank but are read as real values and not missing data
#The table between age and sex shows three variables here
#The dataset contains non standard missing values that are not recognised as NA
#Replace empty strings with NA

RTKA2023[RTKA2023 == ""] <- NA

#Check this has registered

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)


#Column with LSOA_2011_Code has 171 missing.

#LSOA is part of primary exposure variable, small number of missing cases. Decision to
remove rows rather than estimate from imputation because factor variable and dependent
on provider code. Multiple imputation was used later to estimate missing travel data for
these multiple rows where LSOA and site code was availble

#Remove missing data in dataframe combined_data for column LSOA_2011_Code with
missing fields = 171

RTKA2023<- RTKA2023[!is.na(RTKA2023$LSOA_2011_Code), ]
```

```
#16,565 patients before link with TRACC travel data

#Load Travel times data

TRAVELTIMES <- read.csv("~/Desktop/Drive time and Miles reference file.csv")

LSOAREF <- read.csv("~/Desktop/LSOA Matrix.csv")

LSOAREF <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/LSOA Matrix.csv")


#Join data but The data is too big so we need to do this using SQL

install.packages("RSQLite")
library(RSQLite)

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "times", TRAVELTIMES)
dbWriteTable(con, "lsoa", LSOAREF)

query <- "
Select *
FROM times
JOIN lsoa ON times.LSOAName = lsoa.LSOA11NM"

result <- dbGetQuery(con, query)

#10million 457 thousand and 999 possible combinations

#Write Dataframes

write.csv(result, "~/Desktop/JOINLSOATRAVEL.csv")

result<- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/JOINLSOATRAVEL.csv")


#####Now join this data to your revisions spreadsheet using key identifiers LSOA and
Organisation site code

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "revisions3", RTKA2023)
dbWriteTable(con, "travel3", result)
```

```
query <- "
Select *
FROM revisions3
JOIN travel3 ON revisions3.LSOA_2011_Code = travel3.LSOA11CD AND revisions3.Sitecode =
travel3.ProviderSiteCode"

result_join <- dbGetQuery(con, query)

#Number of patients following join 12,774


result1 <- result_join
#Check your data for missing values

missing_data <- colSums(is.na(result1))
print(missing_data)

#Check data for duplicates

duplicates <- RTKA2023[duplicated(RTKA2023$Epikey), ]


# Check for duplicates in the 'epikey' column
duplicates <- result1[duplicated(result1$Epikey), ]

#There are 2,047 duplicates

#Remove duplicates in result 1


# Remove duplicates: Keep only the first occurrence of each 'Epikey'
result1 <- result1[!duplicated(result1$Epikey), ]

#final dataframe is 10,727


write.csv(result1, "/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/FinalJOIN.csv")


####Prepare Outcomes, Exposure variable and co-variates ####

#Set up outcomes

#Replace NA's in the Read columns with N
```

```
result1$Read30 <- ifelse(is.na(result1$Read30), 'N', result1$Read30)
result1$Read90 <- ifelse(is.na(result1$Read90), 'N', result1$Read90)

result1$Read30days <- ifelse(result1$Read30 == "Y", 1, 0)
#readmission for 90 days
result1$Read90days <- ifelse(result1$Read90 == "Y", 1, 0)



#Set up your co-variates

result1$HFRS_Band = as.factor(result1$HFRS_Band)
result1$HFRS_Band = relevel(result1$HFRS_Band, ref = 'None')

result1$POD = as.factor(result1$POD)
result1$POD = relevel(result1$POD, ref = 'EL')

table(result1$POD)


#I've joined two dataframes based on a shared field. But some rows have not jointed

#Journey times statistics - 10,457,999 rows

#12,774 following join with revisions and travel data called "result1" but had duplicates
2,047 so remove these (duplicates due to slightly different latitude and longitude for same
Site codes in journey times statistics )

#Final results 1 following removal of duplicates is 10,727

#Original dataframe is 16,736 called RTKA2023 following removal of early revisions,
excluding missing LSOA was 16565

#Missing data for travel seen in 5,838 patients or 35% of patients

#Use multiple imputation to impute missing distance values for cases without join

#How many unmatched rows?

unmatched_rows <- RTKA2023[!(RTKA2023$Epikey %in% result1$Epikey), ]

#There are 5,838 unmatched rows

#I want to create a dataframe showing both matched and unmatched fields based on this.

# Identify columns that are in result1 but not in RTKA2023
missing_cols <- setdiff(names(result1), names(RTKA2023))
```

```
# Add missing columns to RTKA2023 with NA values
for (col in missing_cols) {
  RTKA2023[[col]] <- NA
}

# Ensure column order is the same as result1
RTKA2023 <- RTKA2023[, names(result1)]

# Identify unmatched rows
unmatched_rows <- RTKA2023[!(RTKA2023$Epikey %in% result1$Epikey), ]

# Combine matched rows (result1) with unmatched rows
combined_data <- rbind(result1, unmatched_rows)

duplicates <- combined_data[duplicated(combined_data$Epikey), ]

#0 duplicates

write.csv(combined_data, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/FinalJOINCombined.csv")


combined_data <- read.csv("/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/FinalJOINCombined.csv")

#Replace NA's in the Read columns with N

combined_data$Read30 <- ifelse(is.na(combined_data$Read30), 'N',
combined_data$Read30)


combined_data$Read30days <- ifelse(combined_data$Read30 == "Y", 1, 0)



#Now have dataframe displaying both matched and unmatched rows

missing_data <- colSums(is.na(combined_data))
print(missing_data)

#How many patients in high volume centres >49

combined_data$MRC <- ifelse(combined_data$TV12mo > 49, 1, 0)
```

```
nopatients <- subset(combined_data, MRC == 1)

#6880 patients

missing_data <- colSums(is.na(nopatients))
print(missing_data)

# Count unique levels of ProvCode
n_levels <- length(unique(nopatients$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")
#38 providers

#How many sites
# Count unique levels of ProvCode
n_levels <- length(unique(nopatients$Sitecode))
cat("Number of unique sites (Sitecode):", n_levels, "\n")

#187 sites

#rates of readmission 30 days

table(nopatients$Read30days)

#568/6880 8.3%

#rates of mortality at 90 days

table(nopatients$Mort90days)

#217/6880 3.2%

#Rates of length of stay above median. Remember median calculated across entire cohort

summary(combined_data$Spell_Los) #Median of 5

nopatients$Long_Los <- ifelse(nopatients$Spell_Los > 5, 1, 0)

table(nopatients$Long_Los)

#3421/6880 49.7%


#3157 travel data not available

#16,565 observations in entire dataframe not limited to teriatry referral centres

#CV12mo missing 71 cases. Imputation using median due to positive skew
```

```
hist(combined_data$CV12mo)

#mean age excluding missing values is 70
summary(combined_data$CV12mo, na.rm = TRUE)


#Input median of 6 for missing data

combined_data$CV12mo[is.na(combined_data$CV12mo)] <- 6

#Now need to use multiple imputation method to estimate travel data for columns
"DistanceMiles", "OffPeakDriveDistanceMiles", "PeakDriveTimes' based on associated
predictors:

#Refer to this resource "https://bookdown.org/mwheymans/bookmi/multiple-
imputation.html#setting-the-imputation-methods"

#And this resource for context
https://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html

# https://www.ebpi.uzh.ch/dam/jcr:dc0cef17-29c7-4e61-8d33-
e690561ab7ae/mi_intro20191001.pdf (Advice on multi level modelling and imputation)

# Install packages if they are not already installed
install.packages(c("mice", "ggplot2", "naniar"))

# Load the packages
library(mice)
library(ggplot2)
library(naniar)

#assuming missing data is due to random chance, LSOA and SiteCode are related to the
exposure but also include all other variables linked to your analysis
#Subset dataframe called combined_date with only with relevant columns: age_of_patient,
sex, HFRS_Band IMD_Score, IMD_Decile, infection,  TVcat, CVcat, SiteCode, ProvCode, FinY,
DistanceMiles, OffPeakDriveDistanceMiles, PeakDriveTime, Mort90days, Read30, Spell_Los
#decision not to include site code and LSOA as likely not present in missing data
"LSOA_2011_Code", "Sitecode"



# Specify the relevant columns I've included TV12mo as may be related to outcome,
ProvCode for clustering,
relevant_columns <- c(
  "age_of_patient", "sex", "HFRS_Band", "IMD_score",
  "infection", "TV12mo", "CV12mo", "ProvCode", "FinY",
```

```
  "DistanceMiles", "OffPeakDriveDistanceMiles", "PeakDriveTime",
  "Mort90days", "Read30days", "Spell_Los"
)

# Subset the dataframe with only the relevant columns
subset_combined_data <- combined_data[, relevant_columns]

#Currently sex, HFRS_Band, TVCat, Sitecode, ProvCode, FinY are not incorporated in model
as character variables

#convert these to factors


# Convert variables to factors
subset_combined_data$sex <- as.factor(subset_combined_data$sex)
subset_combined_data$ProvCode <- as.factor(subset_combined_data$ProvCode)
subset_combined_data$FinY <- as.factor(subset_combined_data$FinY)
subset_combined_data$HFRS_Band <- as.factor(subset_combined_data$HFRS_Band)

subset_combined_data$Sitecode <- as.factor(subset_combined_data$Sitecode)
subset_combined_data$LSOA_2011_Code <-
as.factor(subset_combined_data$LSOA_2011_Code)



# Check the structure of the dataframe to confirm
str(subset_combined_data[, c("sex", "Sitecode", "ProvCode", "FinY", "HFRS_Band",
"LSOA_2011_Code")])


#visualise missing data

vis_miss(subset_combined_data)

#35% missing travel data

# Set the seed for reproducibility
set.seed(123)


# Perform Multiple Imputation

imp <- mice(subset_combined_data, m=5, method='pmm')

#Check for imputation values

imp$imp$OffPeakDriveDistanceMiles
```

```
#visualise imputed values

imp$imp

#Means of the imputed values

imp$chainMean

#What are the predictors

imp$predictorMatrix

#Plot imputation values against observed values.

my_plot <- stripplot(imp, col=c("grey", "blue"), pch = c(1, 20))

my_plot

#Guidelines for imputation model suggest all variables in the analysis should be included,
inclusive of dependent or outcome variables

#Ensure TVCat is not a predictor variable

pred <-imp$predictorMatrix
pred["TVcat"] <- 0
pred


#Plot the convergence (how equal is the variance to the mean)

plot(imp)

#Stack the imputed values into a single dataset and include original data

imp2 <- complete(imp, "long", inc = TRUE)

#Save imp2

write.csv(imp2, "/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/imp2.csv")

#Read it back in here:

imp2 <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/imp2.csv")
```

```
#Save as Supplemenatry figure

#Filter data by tertiary hospitals only


#But current guidelines suggest >49 is a high volume centre called a major revision centre
and probably represents a unit with tertiary specialisation

imp2$MRC <- ifelse(imp2$TV12mo > 49, 1, 0)

tertiary_revisions <- subset(imp2, MRC == 1)

tertiary_revisions$Long_Los <- ifelse(tertiary_revisions$Spell_Los > 5, 1, 0)


#declare the imputed data to be mids again, the format MICE is expecting for regression
analyses
tertiary_revisions <- as.mids(tertiary_revisions)


#Now run your regression model using a multivariable model

#A priori co-variates chosen based on evidence of predictors for readmission

####Primary Outcome 30 day readmission ####

#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ DistanceMiles + IMD_score +
HFRS_Band +
                sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
              family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the long data including all imputations for VIF

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)


# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = tertiary_revisions, family = "binomial")




# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?

#Box Tidwell

#Recode back into correct format

tertiary_revisions <- as.mids(tertiary_revisions)

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}
```

```
# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ DistanceMiles + Interaction,
data = tert
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.03 evidence of non linearity

#Are spline terms significant for DistanceMiles if using 3 knots, 4 knots and 5 knots

#Use data of all imputations in long format

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)



# Load the required library
library(splines)

#AIC of non spline model

model <- glm(Read30days ~ DistanceMiles, data = tertiary_revisions, family = binomial)
summary(model)
```

```
#AIC 21862

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$DistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Read30days ~ ns(DistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)

  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
```

```
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#AIC better fit 21806
#Model with 3 knots, significant terms but greater knots do not improve the model fit. Non linear relationship is evident and should be modelled with splines




#Prepare predictors for model prediction


#you need to ensure that the predicted probabilities align with the corresponding observations
#Explore the data for missing values
sum(!complete.cases(tertiary_revisions$DistanceMiles))
#Unimputed dataset is missing, so exclude these

tertiary_revisions <- tertiary_revisions[!is.na(tertiary_revisions$DistanceMiles),]


sum(!complete.cases(tertiary_revisions$sex))

sum(!complete.cases(tertiary_revisions$Read30days))

sum(!complete.cases(tertiary_revisions$HFRS_Band))

sum(!complete.cases(tertiary_revisions$IMD_score))

sum(!complete.cases(tertiary_revisions$infection))

#Currently infection as numeric - ensure is factor

tertiary_revisions$infection <- as.factor(tertiary_revisions$infection)
tertiary_revisions$HFRS_Band <- as.factor(tertiary_revisions$HFRS_Band)
tertiary_revisions$sex <- as.factor(tertiary_revisions$sex)
tertiary_revisions$FinY <- as.factor(tertiary_revisions$FinY)
tertiary_revisions$ProvCode <- as.factor(tertiary_revisions$ProvCode)
tertiary_revisions$DistanceMiles <- as.numeric(tertiary_revisions$DistanceMiles)
tertiary_revisions$age_of_patient <- as.numeric(tertiary_revisions$age_of_patient)
tertiary_revisions$IMD_score <- as.numeric(tertiary_revisions$IMD_score)
tertiary_revisions$TV12mo <- as.numeric(tertiary_revisions$TV12mo)
```

```
tertiary_revisions$CV12mo <- as.numeric(tertiary_revisions$CV12mo)

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$DistanceMiles, probs = c(0.05, 0.50, 0.95), na.rm =
TRUE)
print(knots)
#Knots at 53, 69 and 84
spline_terms <- ns(tertiary_revisions$DistanceMiles, knots = knots)



model_with_custom_splines <- glm(Read30days ~ ns(DistanceMiles, knots = knots) +
HFRS_Band + IMD_score +
                    sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                  family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$DistanceMiles),
max(tertiary_revisions$DistanceMiles), length.out = 100)

new_data <- expand.grid(
  DistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
```

```
    TV12mo = mean(tertiary_revisions$TV12mo),
    infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))
```

```
# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(DistanceMiles = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$DistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




# Combine mean_unit_range, predicted_probs, ci_lower, and ci_upper into plot_data
plot_data <- data.frame(DistanceMiles = DistanceMiles_range,
                        predicted_prob = predicted_probs,
                        ci_lower = boot_results$ci_lower,
                        ci_upper = boot_results$ci_upper)

library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
```

```
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )


# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Travel Distance (Miles)", y = "Mean Predicted Probability for readmission at 30
days", title = "Spline curve predicted probability of readmission at 30 days by patient travel
distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )


#Spline curve does appear to show the predicted probability of emergency readmission at
30 days increases with travel distance but wide confidence intervals

#Model Distance Miles and 30 day readmission with 3 knot splines
```

```
####First Imputation and descriptive stats####

#Use first imputed data for clinical and demographic characteristic summary

#complete_data is the first imputation

# Count unique levels of ProvCode
n_levels <- length(unique(complete_data$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")


# Count unique levels of sites
n_levels <- length(unique(complete_data))
cat("Number of unique providers (ProvCode):", n_levels, "\n")

# Count unique levels of ProvCode
n_levels <- length(unique(tertiary_revisions$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")

#38 unique providers



#Number of sites

# Count unique levels of Sites but need to use original dataframe as sites not included in
imputation analysis

#Find all those attending tertirary referral centre from original data
tertiary_all <- subset(combined_data, MRC == 1)

#Find number of sites
n_levels <- length(unique(tertiary_all$Sitecode))
cat("Number of unique providers (Sites):", n_levels, "\n")

#187 sites

#Back to first imputation dataset. Calculate median number of miles straight line distance

summary(complete_data$DistanceMiles)

#Median is 7.1 IQR is 3.9 to 12.7. Range 0 to 77.1 miles.
```

```
#Driving distances

summary(complete_data$OffPeakDriveDistanceMiles)

#Median 10.4 miles, IQR is 5.8 to 18.3 miles

#Calculate median driving times

summary(complete_data$PeakDriveTime)

#Median is 27 minutes IQR is 18.4 to 38.4. Maximum 104 minutes


#Create travel time quintile variable

quintiles <- quantile(complete_data$DistanceMiles, probs = seq(0,1,0.2), na.rm=TRUE)

complete_data$distancequintile <- cut(complete_data$DistanceMiles, breaks = quintiles,
labels = c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)

#Tabulate descriptive stats

hist(tertiary_all$Spell_Los)
summary(tertiary_all$Spell_Los)

# Total number of revisions
total_revisions <- nrow(complete_data)

# Create a summary table
summary_stats <- complete_data %>%
 group_by(distancequintile) %>%
 summarise(
  # Count of observations
  Count = n(),

  # Distinct Providers
  Distinct_Units = n_distinct(ProvCode),
  Total_Distinct_Units = n_distinct(complete_data$ProvCode),
  Distinct_Units_Percent = (Distinct_Units / Total_Distinct_Units) * 100,

  #Median distance

  Distance_LowerQuartile = quantile(DistanceMiles, 0.25, na.rm = TRUE),
  Distance_Median = median(DistanceMiles, na.rm = TRUE),
  Distance_UpperQuartile = quantile(DistanceMiles, 0.75, na.rm = TRUE),
```

```
    #Mean driving time
    DrivingTime_LowerQuartile = quantile(PeakDriveTime, 0.25, na.rm = TRUE),
    DrivingTime_Median = median(PeakDriveTime, na.rm = TRUE),
    DdrivingTime_UpperQuartile = quantile(PeakDriveTime, 0.75, na.rm = TRUE),



    # Age: Mean and standard deviation
    Age_Mean = mean(age_of_patient, na.rm = TRUE),
    Age_SD = sd(age_of_patient, na.rm = TRUE),

    # Age: Mean ± SD (concatenated)
    Age_Mean_SD = paste(round(mean(age_of_patient, na.rm = TRUE), 2), "±",
round(sd(age_of_patient, na.rm = TRUE), 2)),



    # Gender: frequency and percentage
    Female_Freq = sum(sex == "Female", na.rm = TRUE),
    Female_Percent = sum(sex == "Female", na.rm = TRUE) / n() * 100,
    Male_Freq = sum(sex == "Male", na.rm = TRUE),
    Male_Percent = sum(sex == "Male", na.rm = TRUE) / n() * 100,

    # ASA: frequency and percentage for each level
    HFRS_None_Freq = sum(HFRS_Band == "None", na.rm = TRUE),
    HFRS_None_Percent = sum(HFRS_Band == "None", na.rm = TRUE) / n() * 100,
    HFRS_Mild_Freq = sum(HFRS_Band == "Mild", na.rm = TRUE),
    HFRS_Mild_Percent = sum(HFRS_Band == "Mild", na.rm = TRUE) / n() * 100,
    HFRS_Moderate_Freq = sum(HFRS_Band == "Moderate", na.rm = TRUE),
    HFRS_Moderate_Percent = sum(HFRS_Band == "Moderate", na.rm = TRUE) / n() * 100,
    HFRS_Severe_Freq = sum(HFRS_Band == "Severe", na.rm = TRUE),
    HFRS_Severe_Percent = sum(HFRS_Band == "Severe", na.rm = TRUE) / n() * 100,



    #Infection

    Infection_Freq = sum(infection == "1", na.rm = TRUE),
    Infection_Percent = sum(infection == "1", na.rm = TRUE) / n() * 100,



    # Year: frequency and percentage for each year from 2009 to 2019
    Year_2015_2016_Freq = sum(FinY == "2015/16", na.rm = TRUE),
    Year_2015_2016_Percent = sum(FinY == "2015/16", na.rm = TRUE) / n() * 100,
    Year_2016_2017_Freq = sum(FinY == "2016/17", na.rm = TRUE),
    Year_2016_2017_Percent = sum(FinY == "2016/17", na.rm = TRUE) / n() * 100,
    Year_2017_2018_Freq = sum(FinY == "2017/18", na.rm = TRUE),
    Year_2017_2018_Percent = sum(FinY == "2017/18", na.rm = TRUE) / n() * 100,
    Year_2018_2019_Freq = sum(FinY == "2018/19", na.rm = TRUE),
```

```
    Year_2018_2019_Percent = sum(FinY == "2018/19", na.rm = TRUE) / n() * 100,
    Year_2019_2020_Freq = sum(FinY== "2019/20", na.rm = TRUE),
    Year_2019_2020_Percent = sum(FinY == "2019/20", na.rm = TRUE) / n() * 100,


    # Median Surgeon Volume: lower quartile, median, and upper quartile
    Surgeon_LowerQuartile = quantile(CV12mo, 0.25, na.rm = TRUE),
    Surgeon_Median = median(CV12mo, na.rm = TRUE),
    Surgeon_UpperQuartile = quantile(CV12mo, 0.75, na.rm = TRUE),


    #Median hospital volume

    Hospital_LowerQuartile = quantile(TV12mo, 0.25, na.rm = TRUE),
    Hospital_Median = median(TV12mo, na.rm = TRUE),
    Hospital_UpperQuartile = quantile(TV12mo, 0.75, na.rm = TRUE),


    #Median IMD Score

    IMD_LowerQuartile = quantile(IMD_score, 0.25, na.rm = TRUE),
    IMD_Median = median(IMD_score, na.rm = TRUE),
    IMD_UpperQuartile = quantile(IMD_score, 0.75, na.rm = TRUE),

  )



# Print the summary table
print(summary_stats)

write.csv(summary_stats, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/Summary_stats.csv")




####Cluster Variable ####

# Compute the mean outcome for each cluster
library(dplyr)
prov_means <- tertiary_revisions %>%
  group_by(ProvCode) %>%
  summarize(mean_outcome = mean(Read30days, na.rm = TRUE))

# Plot variability
```

```
boxplot(mean_outcome ~ ProvCode, data = prov_means, xlab = "ProvCode", ylab = "Mean
Outcome")

# Summary statistics of variability
summary(prov_means$mean_outcome)

#There is evidence of variability between providers


# Fit logistic regression on imputed datasets
m3.mi <- with(tertiary_revisions, glmer(Read30days ~ DistanceMiles + IMD_score +
HFRS_Band +
                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY + (1 |
ProvCode),
                      family = "binomial"))


print(m3.mi)

#Including ProvCode as a random effect was tested but led to convergence issues likely due
to numerical instability between providers so a decision was made to accept the fixed
effects model which may account for clustering at the provider level but is a limitation of
the study


#Was travel distance strongly correlated with IMD_score or age?



#Next do a Spearman's rank correlation between travel distance and age, and then for
travel distance and IMD score

imp2$MRC <- ifelse(imp2$TV12mo > 49, 1, 0)

tertiary_revisions <- subset(imp2, MRC == 1)


write.csv(tertiary_revisions, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/tertiary_revisions.csv")


tertiary_revisions <- as.mids(tertiary_revisions)

tertiary_revisions$age_of_patient <-
as.numeric(as.character(tertiary_revisions$age_of_patient))
```

```
tertiary_revisions$DistanceMiles <-
as.numeric(as.character(tertiary_revisions$DistanceMiles))



#Age and travel distance, Cannot pool the results based on the multiple imputations as cor
test not compatible.Therefore stack all imputations together and calculate correlation


# Scatterplot with linear regression line
plot(tertiary_revisions$age_of_patient, tertiary_revisions$DistanceMiles,
    main = "Scatterplot of Age of Patient vs DistanceMiles",
    xlab = "Age of Patient", ylab = "DistanceMiles",
    pch = 19, col = "blue")

# Add a linear trendline
abline(lm(DistanceMiles ~ age_of_patient, data = tertiary_revisions), col = "red", lwd = 2)

# Calculate Spearman's rank correlation
spearman_test <- cor.test(tertiary_revisions$age_of_patient,
tertiary_revisions$DistanceMiles, method = "spearman")

# Extract rho and p-value
rho <- round(spearman_test$estimate, 2)
p_value <- spearman_test$p.value
p_value_text <- ifelse(p_value < 0.05, "<0.05", paste0("=", round(p_value, 3)))

# Add a legend with Spearman's rank correlation information
legend("topright", legend = paste("Spearman's Rank Correlation:\n",
                          "rho =", rho, ", p-value", p_value_text),
     col = c("blue", "red"), lty = c(NA, 1), pch = c(19, NA), lwd = c(NA, 2), bty = "n")



#IMD score and travel distance

# Scatterplot with trendline
plot(tertiary_revisions$IMD_score, tertiary_revisions$DistanceMiles,
    main = "Scatterplot of IMD_score vs DistanceMiles",
    xlab = "IMD_score", ylab = "DistanceMiles",
    pch = 19, col = "blue")

# Add a linear trendline (for visualizing the general trend)
abline(lm(DistanceMiles ~ IMD_score, data = tertiary_revisions), col = "red", lwd = 2)

# Calculate Spearman's rank correlation
spearman_test <- cor.test(tertiary_revisions$IMD_score, tertiary_revisions$DistanceMiles,
method = "spearman")
```

```
# Extract rho and p-value
rho <- round(spearman_test$estimate, 2)
p_value <- spearman_test$p.value
p_value_text <- ifelse(p_value < 0.05, "<0.05", paste0("=", round(p_value, 3)))

# Add a legend with Spearman's rank correlation information
legend("topright", legend = paste("Spearman's Rank Correlation:\n",
                                  "rho =", rho, ", p-value", p_value_text),
       col = c("blue", "red"), lty = c(NA, 1), pch = c(19, NA), lwd = c(NA, 2), bty = "n")

#Exposure 2 - OffPeakDriveDistanceMiles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ OffPeakDriveDistanceMiles +
IMD_score + HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)
```

```
# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?


# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ OffPeakDriveDistanceMiles +
Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.05. There is no evidence of non linearity

#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ PeakDriveTime + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)
```

```
# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ PeakDriveTime + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?



# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ PeakDriveTime + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.13 not evidence of non linearity




####Secondary Outcome mortality 90 days ####


#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ DistanceMiles + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?
```

```
library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Mort90days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity


#Is there evidence of non linearity?

library(mice)

tertiary_revisions <- as.mids(tertiary_revisions)


# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Mort90days ~ DistanceMiles + Interaction,
                                               family = binomial(link = "logit")))
```

```
# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


# P value 0.95

#Exposure 2 - OffPeakDriveDistanceMiles



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ OffPeakDriveDistanceMiles +
IMD_score + HFRS_Band +
                                      sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                    family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)
```

```
#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

tertiary_revisions <- as.mids(tertiary_revisions)

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
```

```
model <- with(tertiary_revisions_modified, glm(Mort90days ~ OffPeakDriveDistanceMiles +
Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#0.989



#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ PeakDriveTime + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
```

```
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Mort90days ~ PeakDriveTime + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
```

```
model <- with(tertiary_revisions_modified, glm(Mort90days ~ PeakDriveTime + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


# P avlue 0.78

####Secondary outcome prolonged LOS ####

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)

tertiary_revisions$Long_Los <- ifelse(tertiary_revisions$Spell_Los > 5, 1, 0)

tertiary_revisions <- as.mids(tertiary_revisions)

#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ DistanceMiles + IMD_score + HFRS_Band +
                                      sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                    family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Long_Los ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)
```

```
# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ DistanceMiles + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


#P value 0.002 Non linear

# Load the required library
library(splines)

#AIC of non spline model

model <- glm(Long_Los ~ DistanceMiles, data = tertiary_revisions, family = binomial)
summary(model)

#AIC 52853

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$DistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(DistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)

  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
```

```
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52769, model with four knots best fit and improved fit from original linear model

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$DistanceMiles, probs = c(0.05, 0.35, 0.65, 0.95), na.rm =
TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$DistanceMiles, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(DistanceMiles, knots = knots) +
HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$DistanceMiles),
max(tertiary_revisions$DistanceMiles), length.out = 100)

new_data <- expand.grid(
  DistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
                        TV12mo = mean(tertiary_revisions$TV12mo),
                        infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
```

```
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(DistanceMiles = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals
```

```
# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$DistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )
```

```
# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Travel Distance (Miles)", y = "Mean Predicted Probability for Prolonged LOS", title
= "Spline curve predicted probability of prolonged LOS by patient travel distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),  # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )



#Exposure 2 - OffPeakDriveDistanceMiles



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ OffPeakDriveDistanceMiles + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)
```

```
# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
```

```
                    lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                           add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ OffPeakDriveDistanceMiles +
Interaction,
                           family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#0.003

#AIC of non spline model

model <- glm(Long_Los ~ OffPeakDriveDistanceMiles, data = tertiary_revisions, family =
binomial)
summary(model)

#AIC 52853

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$OffPeakDriveDistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(OffPeakDriveDistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)
```

```
  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52718, model with four knots best fit and significant spline terms

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$OffPeakDriveDistanceMiles, probs = c(0.05, 0.35, 0.65,
0.95), na.rm = TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$OffPeakDriveDistanceMiles, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(OffPeakDriveDistanceMiles, knots =
knots) + HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$OffPeakDriveDistanceMiles),
max(tertiary_revisions$OffPeakDriveDistanceMiles), length.out = 100)

new_data <- expand.grid(
  OffPeakDriveDistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
                        TV12mo = mean(tertiary_revisions$TV12mo),
                        infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
```

```
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
```

```
plot_data <- data.frame(OffPeakDriveDistanceMiles = DistanceMiles_range, predicted_prob
= predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$OffPeakDriveDistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
```

```
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )


# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Off Peak Drive Distance Miles", y = "Mean Predicted Probability for Prolonged
LOS", title = "Spline curve predicted probability of prolonged LOS by patient driving
distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )




#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ PeakDriveTime + IMD_score + HFRS_Band
+
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))
```

```
print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)
```

```
# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                               lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                      add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ PeakDriveTime + Interaction,
                                  family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#P value 0.000916

#AIC of non spline model

model <- glm(Long_Los ~ PeakDriveTime, data = tertiary_revisions, family = binomial)
summary(model)

#AIC 52843

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$PeakDriveTime, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(PeakDriveTime, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)
```

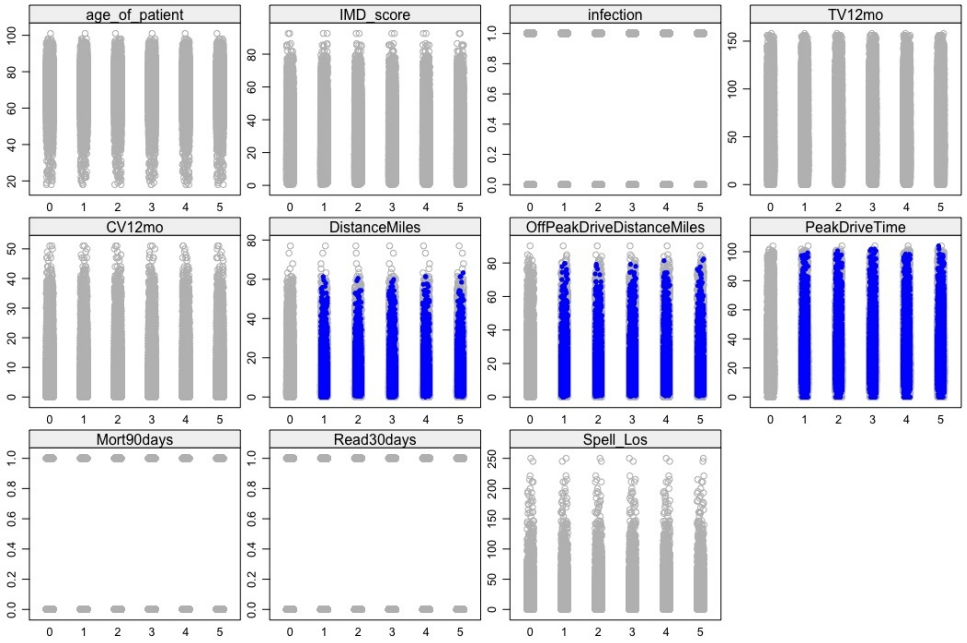
```
  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52715, model with four knots best fit and significant spline terms and most parsimonious

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$PeakDriveTime, probs = c(0.05, 0.35, 0.65, 0.95), na.rm
= TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$PeakDriveTime, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(PeakDriveTime, knots = knots) +
HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$PeakDriveTime),
max(tertiary_revisions$PeakDriveTime), length.out = 100)

new_data <- expand.grid(
  PeakDriveTime = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
```

```
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(PeakDriveTime = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit
```

```
# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DriveTime = new_data$PeakDriveTime,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DriveTime)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DriveTime) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )



# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DriveTime, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DriveTime, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
```

```
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Peak Drive Times (Minutes)", y = "Mean Predicted Probability for Prolonged LOS",
title = "Spline curve predicted probability of prolonged LOS by patient driving times") +
  scale_x_continuous(limits = c(0, max(mean_data$DriveTime, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )









####END####
```



91x63mm (300 x 300 DPI)

BMJ Open

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement: A retrospective observational study using data for England from Hospital Episode Statistics

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-085201.R3
Article Type:	Original research
Date Submitted by the Author:	10-Mar-2025
Complete List of Authors:	Matthews, Alexander; University of Exeter Faculty of Health and Life Sciences, Department of Clinical and Biomedical Sciences; Royal Devon and Exeter NHS Foundation Trust, Evans, Jonathan; University of Exeter Medical School, Health Services and Policy Research; Royal Devon and Exeter NHS Foundation Trust, Princess Elizabeth Orthopaedics Centre Evans, Jonathan; University of Exeter, Department of Public Health and Sports Sciences Lamb, Sarah; University of Exeter, College of Medicine and Health Price, Andrew; University of Oxford, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences Gray, William; NHS England and NHS Improvement London Briggs, Tim; NHS England and NHS Improvement London Toms, Andrew ; Princess Elizabeth Orthopaedic Centre
Primary Subject Heading:	Public health
Secondary Subject Heading:	Surgery
Keywords:	Knee < ORTHOPAEDIC & TRAUMA SURGERY, PUBLIC HEALTH, Health Services

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Title

What is the impact of longer patient travel distances and times on perioperative outcomes following revision knee replacement: A retrospective observational study using data for England from Hospital Episode Statistics

Names, Affiliations, and positions of all authors

Alex Matthews1,2,3,4, Jonathan P Evans2,3, Jonathan T Evans2,3, Sarah E Lamb3
Andrew Price4,5, William Gray1, Tim Briggs1,6, Andrew Toms2,3

1. Getting It Right First Time programme, NHS England, London, UK
2. Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK
3. University of Exeter, Exeter, UK
4. Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, UK

5. Nuffield Orthopaedic Centre, Oxford, UK

6. Royal National Orthopaedic Hospital, Stanmore, London, UK

Correspondence: Alexander Matthews, Getting It Right First Time programme, NHS England and NHS Improvement, Wellington House, Waterloo Road. London, UK, SE1 8UG; telephone: 07401932627 email: alexander.matthews1@nhs.net

Word count: [3316 words]

Short title: Patient Journey Distance and Time for Complex Revision Knee Surgery

Key words: social determinants of health, revision knee replacement, contextual factors in complex health interventions

Twitter handles:

Alex Matthews: @DrAlexMatthews

Tim WR Briggs: @ProfTimBriggs

WilliamKGray1: @WilliamKGray1

GIRFT programme: @NHSGIRFT

Structured Abstract

Objectives

Patients undergoing revision total knee replacement (RevKR) surgery often have difficulties mobilising and increasingly rely on family support. Evolving practice in England aims to manage these patients in specialised centres with the intention of improving outcomes. This practice will result in longer travel distances and times in this frailer group of patients. We want to examine the types of distances and travel times patients can be expected to travel for this complex orthopaedic surgery and to explore concerns of how these impact patient outcomes.

Design

Retrospective observational study from the Hospital Episode Statistics. Multivariable adjusted logistic regression models were used to investigate the relationship between patient travel distances and times with perioperative outcomes.

Setting

Patients presenting to tertiary referral centres between 1st January 2016 to 31st December 2019. A tertiary referral centre was defined as a trust performing >49 revisions in the year prior.

Participants

Adult patients undergoing RevKR procedures for any reason between 1st January 2016 to 31st December 2019.

Exposure

The shortest patient level travel distance and time was calculated using the department of health Journey Time Statistics using TRACC software and Dijkstra's algorithm.

Main Outcome Measures

The primary outcome is emergency readmission within 30 days. Secondary outcomes are mortality within 90 days and length of inpatient stay.

Results

6,880 patients underwent RevKR at 36 tertiary referral centres. There was a weak correlation between social deprivation and travel distance, with patients from the most deprived areas travelling longer distances. Overall, 30-day readmission was not statistically associated longer driving distance (OR 1.00 95% CI 0.99 to 1.02) or peak driving times (OR 1.00 95% CI 0.99 to 1.01).

Conclusions

There was no association between increasing travel distance and time on perioperative outcomes for RevKR patients.

Strengths and limitations of this study

- This study is one of the largest studies in the literature investigating outcomes following revision knee replacement.
- This data reflects revision knee replacement procedures undertaken across different geographical areas of England
- Owing to differences in the coverage of Hospital Episode Statistics, procedures in hospitals outside of England were not included in this analysis
- Clinical coding practice is known to vary across trusts, with some trusts more consistent in coding than others which may have created some bias in the model estimates
- This analysis only reports travel times for patients with access to their own transport and does not consider times for those patients using public transport

Introduction

Primary knee replacement is a successful procedure that improves quality of life for the majority of patients.[1] However, at 10 years following a primary knee replacement, about 3.5% of patients will have undergone a revision surgery.[2] The majority of these procedures are carried out due to infection or polyethylene wear of the implant.[3] A failed primary knee replacement represents a life changing transition point where individuals are likely to suffer from pain, reduced mobility as well as dependency on family members.[4] Patients often face multi-step surgery with longer hospital length of stays and higher complication rates.[5, 6]

The Getting It Right First Time (GIRFT) programme orthopaedic National Report was published in 2015.[7] A key recommendation was the centralisation of complex orthopaedic surgery, including revision knee surgery, to specialist centres with the aim of improved patient outcomes. Consequently, revision total knee replacement (RevKR) surgery in the England has evolved into a regional network service model.[8] All hospitals performing RevKR form a network in the respective regions. Less specialist hospitals, defined by lower annual case volume thresholds, are encouraged to discuss and sometimes refer their caseload to more specialist centres. Several studies based on large revision hip and knee registries have suggested this model carries a lower failure rate defined by the need for further revision surgery.[9-11] Early evidence has suggested reduced early failure rates through the adoption of revision knee networks.[12]

However, for some patients, this approach to managing patients is inevitably associated with increasing travel distances between patient's homes and their treating hospital. Travel distance has been shown to be an important factor in patient choice when selecting a surgeon for joint replacement surgery. It may be even more important for those awaiting revision joint replacement surgery as these patients struggle with mobility, may be unable to drive and may be more reliant on family members.[4] Evidence suggests that patients considering joint replacement are prepared to travel longer distances to obtain the best possible outcomes. A requisite in making such a decision requires data on outcomes of patients travelling greater

distances. Patients travel longer distances have been found to have higher readmission rates and higher mortality rates when undergoing other types of specialised surgery.[13] The pick-up rate of early complications, avoiding the need for readmission, may be less in areas further away from the main treatment centre. There is also concern that patients required to travel greater distances are more likely to be re-admitted to a different hospital than that where surgery was undertaken, resulting in clinical decisions that do not incorporate the primary surgeon and so potentially leading to poorer outcomes.[14] There is an absence of evidence in the literature to support or refute this argument in the context of patients undergoing RevKR. Therefore the aim of this paper is to investigate the relationship between longer patient travel distances and perioperative outcomes following RevKR performed in high volume tertiary referral centres.

Methods

Design

This study is a retrospective data analysis of observational data from the Hospital Episode Statistics (HES) and Office for National Statistics (ONS) databases. HES data is collected by NHS England for all patients treated at NHS hospitals in England and those treated at private hospitals where treatment was funded by the NHS. This study complies with the recommended reporting guidelines when using HES data[15] and the Strengthening of Reporting of Observational studies in Epidemiology (STROBE) guidelines.[16]

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes[17] and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data.[18] The HES data were linked at a patient level to data from the ONS on deaths and date of death, which allowed the identification of patients who had died after their surgery. Linkage was achieved using a unique pseudonymised patient identifier using a previously validated methodology.[19]

Patient travel distances were calculated using the Journey Time Statistics reference document produced by the UK Department of Transport which modelled theoretical journey times between known centroids of Lower Layer Super Output Areas (LSOA)

of residence and NHS hospital sites.[20] Please refer to **Supplementary material S1** for Journey Times Statistics reference document.

Population

An RevKR procedure was defined as a permanent removal or exchange of knee arthroplasty components. This includes a revision of a total knee replacement and a conversion of a unicondylar knee replacement to a total knee replacement. Secondary patellar resurfacing was not included as this represents a simple revision procedure, one that can be carried out in most nonspecialised hospitals. All patients aged ≥ 18 years who underwent a RevKR in a high-volume trust between 1st January 2016 to 31st December 2019 were included in the study population. A high-volume trust was classified as a centre performing > 49 revisions per year. This revision volume threshold for classification represents that proposed by the British Association for Surgeons of the Knee (BASK) Revision Knee Working Group and is a mandatory requirement for all highly specialist centres co-ordinating regional networks. [21] As such centres attaining this threshold are more likely to represent tertiary referral centres where the stratification of more complex work will take place. Annual case volume at each trust was defined as the number of revision cases conducted in the year prior to the index procedure. This measure was preferred over a simple calculation of average annual volume as it accounts for recent experience at the point of surgery. The Office for Population Censuses and Surveys' Classification of Interventions and Procedures version 4 (OPCS-4) codes used to identify RevKR procedures are detailed in **Supplementary material S2.** Since laterality was needed to identify re-revisions, patients were excluded where the procedure laterality was not specified. The flow of patients, with numbers excluded at each point, is summarised in **Supplementary material S3**. To manage population heterogeneity, data were extracted for the period 1st April 2011 to 31st December 2019 and only the first revision for a specific side of the body record in this time period included. [22] Thus, any early revisions on the same side of the body in the four years and nine months preceding the start of the study period were identified and these patients excluded from the study. This aims to exclude the early revision knee replacement failures which have been shown to represent catastrophic failures potentially skewing our results.[22] We included revisions for infection as, despite

these representing a more variable patient group, presence of infection was thought to be unrelated to how far a patient lives from a specialised referral centre.

Exposure variable

Travel distances and times were calculated between a patient's LSOA and the postal codes for the treating hospitals. LSOAs are determined by the Office for National Statistics and are designed for the reporting of small area statistics. Public transport and highways data for England were used to create theoretical journey distances and times from origins to destinations. A network of journey distances and times from origins to destinations was produced using a software package called Transport Accessibility and Connectivity Calculator (TRACC). The Dijkstra's algorithm calculated the shortest route between these points. Data linkage between the HES/ONS dataset and the travel times dataset was achieved using two shared data fields; LSOA and hospital site. The resulting travel distances and/or times for each patient were analysed as continuous variables. Three exposure variables were used. Straight line travel distance represented the distance "as the crow flies" between a patient's LSOA and treating hospital. Off peak driving distance represented the shortest driving distance between a patients LSOA and treating hospital. Finally peak driving times were calculated using average traffic speeds between 7am and 10am for the shortest possible road route between a patients LSOA and treating hospital. These three variables were used to account for variation in travel infrastructure between rural and urban areas and to attribute more meaningful results for patients.

Co-variates and cluster variable

The following groups of known or potential confounding variables were chosen a priori for inclusion in our multivariable logistic regression modelling:

Patient factors: Age in years (continuous), sex (male/female). Health co-morbidity was quantified using the Hospital Frailty Risk Score (HFRS). HFRS identifies frailty based on the occurrence of any of 109 International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during any hospital admissions in the two years prior to, and for, the index admission. Deprivation was measured using the Index of Multiple Deprivation (IMD).[23] The

IMD gives the LSOA where the patient lives a score based on a range of measures of deprivation. IMD was analysed as a continuous variable.

Clinical factors: Defined by the presence or absence of infection as the primary indication for RevKR. This was identified from the International Statistical Classification of Diseases and Related Health Problems, tenth revision (ICD-10) codes used during the admission.

Surgical factors: Surgeon and hospital volume (both continuous) was defined as the number of RevKRs performed by a consultant or hospital in the 365 days prior to each index procedure across the entire cohort. This was calculated before any exclusion criteria was applied.

Temporal factors: Financial year of procedure (2015/16, 2016/17, 2017/18, 2018/19, 2019/20).

Hospital Provider: Clustering of patients by hospital provider was initially modelled using random effects. However, despite variability between hospital providers with primary and secondary outcomes, instability in the model estimates were observed. To address the possibility of clustering at this level, a fixed effects model was adopted with hospital provider as a covariate.

Outcomes

The primary outcome was emergency readmission within 30 days of discharge from the index surgical hospital. Readmission in this early period is very likely related to a complication of the surgical procedure. It has been used as a marker of perioperative outcomes in similar studies investigating the relationship between patient travel distance and outcomes following surgery. [13]

Secondary outcomes were:

90-day all-cause mortality, identified using linked data from Civil Registrations (Mortality) dataset;

Inpatient length of hospital stay was attributed from continuous inpatient spells (CIPS), which is the preferred estimate of length of stay. This refers to the length of first stay after the operation regardless of any transfers across providers. The

median length of stay was calculated after visually inspecting the distribution and this was dichotomized into prolonged length of stay if longer than the median stay.

Statistical Analyses

Data was extracted from a secure, encrypted server controlled by NHS England. Data were analysed within a secure, encrypted environment using standard statistical software: R Studio version 2023.09.1+494 (Boston, Massachusetts, USA). The R code and packages used are included in **Supplementary material S4**.

Missing data were managed according to its extent and relevance to the aims of this study. Age and IMD score were imputed for the small number of missing cases using the mean of the entire study cohort. Given the central role of LSOA in estimating travel distances and times and fewer than 5% of cases with missing data, these cases were excluded to avoid the introduction of bias. Following data linkage between the HES/ONS dataset and the travel times dataset, approximately 36% (n = 5,838) of cases did not match. Multiple imputation was performed using predictive mean matching based on the entire cohort of patients with the following predictors: age, sex, HFRS score, IMD score, hospital provider code, hospital volume and surgeon volume. Dependent variables including readmission at 30 days, mortality at 90 days and length of stay were also used in the imputation following a recommended approach using preditive mean matching[24]. A total of five imputations were randomly chosen and subsequent regression analyses were performed.[25] Imputed data is shown in **Supplementary material S5**.

Patient travel distances were categorised into quintiles for interpretation of baseline demographics and clinical characteristics. Subsequent analysis of travel distances and times were performed as continuous variables. Spearman's rank correlation was performed to investigate the relationship between IMD score and patient age with travel distances.

Straight line travel distance was modelled with restricted cubic splines to allow for the non-linear effects when testing the association with the primary outcome. All exposures were modelled with restricted cubic splines to allow for the non-linear effects when testing the association with prolonged length of stay. The Akaike Information Criterion was used to select the most parsimonious specification of

restricted cubic splines using the final adjusted model. Fixed effects logistic regression models were used for the outcomes of readmission at 30 days, mortality at 90 days and prolonged length of stay. Where implemented, the use of splines was used to create figures depicting the association between travel distance or times and probability of the outcomes. Only adjusted spline models were used to depict these associations. All co-variates were included in the adjusted models. Multicollinearity was assessed using eigenvalues, variance inflation factors and by examination of model parameter estimates with the unadjusted model. Odds ratios with 95% CIs and associated p-values were reported. A p-value of < 0.05 was taken to indicate statistical significance.

Results

Overview of results

A total of 16,736 patients met the inclusion criteria. Excluding missing LSOA data (n=171), 16,565 patients were included in the analysis. Following data linkage with department of transport journey times statistics, 10,727 patients had complete data linkage and data were imputed for the remaining 5,838 (35.2%). Of the 16,565 patients, 41.5% (n=6,880) presented to a tertiary referral centre and these data formed our analysis cohort. Patients were operated on across 181 hospital sites and 38 hospital trust providers. The baseline demographic and clinical characteristics of the patients were broadly similar between quintiles of straight-line travel distance. (Table 1). Higher hospital volumes were seen in patients travelling longer distances. Figure 1 shows that straight line travel distance was weakly correlated with age (r= -0.05, p value <0.05) and social deprivation(r= -0.05, p value <0.05). Older patients were less likely to travel farther distances. Patients from the least deprived areas travelled shorter distances.

Table 1 – Baseline patient demographics and clinical characteristics stratified by travel distance quintiles from first imputed dataset

348

	Travel Distance Quintile				
	1	2	3	4	5
Distance (Miles)	2.09 (1.35 to 2.75)	4.42 (3.91 to 5.00)	7.08 (6.34 to 7.99)	11.39 (10.11 to 12.74)	22.42 (18.09 to 32.19)
Driving Time (Minutes)	13 (9.3 to 17)	20.45 (17 to 25)	26.30 (21.98 to 31.13)	34.10 (29.68 to 40.20)	52.05 (42.68 to 66.83)
Number of patients	1376	1376	1376	1376	1376
Tertiary Providers	37 (97.37%)	38 (100%)	36 (94.74%)	35 (92.11%)	37 (97.37%)
Age Mean (SD)	69.71 (10.81)	69.96 (10.71)	69.66 (10.92)	68.84 (11.01)	68.58 (10.75)
Female Sex	762 (55.38%)	768 (55.81%)	729 (52.98%)	722 (52.47%)	734 (53.34%)
HFRS None	647 (47.02%)	620 (45.06%)	614 (44.62%)	666 (48.40%)	676 (49.13%)
HFRS Mild	438 (31.83%)	474 (34.45%)	485 (35.25%)	465 (33.79%)	433 (31.47%)
HFRS Moderate	241 (17.51%)	236 (17.15%)	243 (17.66%)	198 (14.39%)	230 (16.72%)
HFRS Severe	50 (3.63%)	46 (3.34%)	34 (2.47%)	47 (3.42%)	37 (2.69%)
Infection Present	314 (22.82%)	331 (24.06%)	310 (22.53%)	334 (24.27%)	355 (25.80%)
Surgeon Volume	7 (3 to 13)	7 (3 to 13)	8 (3 to 15)	8 (3 to 16)	9 (4 to 17)

Hospital Volume	73 (60 to 87)	74 (60 to 89)	79 (63 to 97)	79 (63 to 99)	85 (68.75 to 112)
IMD Score	16.44 (8.73 to 28.67)	14.30 (7.96 to 24.57)	14.50 (8.47 to 21.36)	14.83 (9.23 to 21.74)	14.752 (8.78 to 21.45)
Year 2015/16	104 (7.56%)	94 (6.83%)	94 (6.83%)	89 (6.47%)	92 (6.69%)
Year 2016/17	383 (27.83%)	354 (25.73%)	348 (25.29%)	338 (24.56%)	353 (25.65%)
Year 2017/18	384 (27.91%)	365 (26.53%)	339 (24.64%)	360 (26.16%)	336 (24.42%)
Year 2018/19	269 (19.55%)	325 (23.62%)	347 (25.22%)	354 (25.73%)	339 (24.64%)
Year 2019/20	236 (17.15%)	238 (17.30%)	248 (18.02%)	235 (17.08%)	256 (18.60%)

Outcomes

The primary and secondary outcomes are summarised in table 2.

The observed rate of readmission at 30 days was 8.3% (568/6880). There was a negative association between higher straight line travel distances and emergency readmission at 30 days (Figure 2). However wide confidence intervals precluded statical inferences. In addition, higher travel distance by road and longer drive times were not associated with statistically worse readmission rates at 30 days. The rate of mortality at 90 days was only 3.2% (217/6880). No statistically significant relationship was observed between the distance a patient travels by road or the time a patient spends travelling at peak driving times with rates of mortality at 90 days. 49.7% (3421/6880) of patients reported hospital stays more than 5 days. Following

adjustment of confounding factors, we observed no associations between prolonged length of stay and patient travel distance (Figures 3-5)

Table 2 – Adjusted pooled Multivariable Logistic Regression showing Odds Ratios for primary and secondary outcomes by exposure variables

	Straight line travel distance (OR, 95% CI)	Travel distance by shortest road route (OR, 95% CI)	Peak Travel times by shortest road route (OR, 95% CI)
Readmission with 30 days	Figure 2	1.00 (0.99 to 1.02), p value = 0.81	1.00 (0.99 to 1.01), p value = 0.69
90 Day Mortality	1.00 (0.98 to 1.02), p value = 0.87	1.00 (0.99 to 1.01), p value = 0.86	1.00 (0.99 to 1.01), p value = 0.89
Prolonged Length of stay	Figure 3	Figure 4	Figure 5

•Odds ratios have been adjusted for patient age, sex, HFRS score,

Discussion

Statement of principal findings

We present a multi-hospital site retrospective analysis of patients undergoing revision knee replacement surgery at tertiary referral centres in England. In this analysis of 6,880 patients undergoing RevKR, we did not observe a statistical association between distance and time travelled for revision surgery and readmission within 30 days.

Strengths and weaknesses of the study

The findings of this study should be interpreted in view of several limitations. Firstly, this analysis used observational data from a large administrative dataset covering all NHS-funded procedures conducted in England. As with all administrative datasets we are limited in the amount of detail provided regarding presentation. We chose to categorise a high-volume centre by trust to accurately capture surgical experience. All NHS hospitals in England are run by hospital trusts which typically involve between one and four hospitals within a catchment area standardising their practice. It is common practice for specialist orthopaedic surgeons to move between these sites delivering the same procedures. Our study involved 187 hospital sites run by 38 trusts. We acknowledge this is a weakness of our study as this may not be representative of all trusts. We included all indications for RevKR in our patient cohort because indication was not thought to be related to how far a patient lives from a hospital. However, we acknowledge the rate of complications is higher in patients with infection and we subsequently adjusted for indication for revision in our analyses. [26] It is likely that because we did not exclude previous revision knee arthroplasty patients, the complexity of the surgery undertaken in our cohort varied. We recognise this is a limitation of the study however we assume case mix was unrelated patient travel distance.

There were many missing patients (approximately 36%) following the linkage of HES data with Journey Time Statistics. To account for this, assumed that the data was missing at random and used multiple imputation to estimate missing travel distances. It is likely the imputed values may introduce bias, however we modelled these based on predictors and dependent variables to improve our estimates. We do not present a sample size calculation, rather we have used all available data and our sample size was set by our inclusion criteria. We controlled for the clustered nature of our data between hospital providers through inclusion as a covariate in our modelling. To ensure consistency in our definition of tertiary referral hospitals, only hospitals performing >49 revisions/year were included. These are likely to treat a similar case mix of patients and potentially have similar access to resources within a national healthcare system. This approach allowed us to control for variation across providers. However, we acknowledge it does not fully account for the hierarchical

nature of the data with differences in treatment protocols and hospital specialisation among factors which may influence patient outcomes.

There is a lack of granular clinical data using HES for each readmission. Therefore, we cannot ascertain precise reasons for readmissions, but we assume are related to a post-surgical complication. Information on the exact date of readmission and death was also not available. Therefore, a time-to-event approach in outcome analysis was not possible. Clinical coding practice within HES is known to vary across trusts.[27] As an example, some trusts may be more consistent in coding comorbidities, and this may have created some bias. However, this is unlikely to vary systematically with travel distances and so significantly bias our findings. We acknowledge the relatively short travel distances in this population compared to examples from the United States as such the results of this study may not be generalisable to larger geographical areas or less mature healthcare systems. However, the upper quintile in our study represents a substantial journey distance and time for our patient cohort where poor mobility is a significant factor affecting their care. This analysis does not consider journey times of those who may not have access to a car and instead chose to take public transport.

Strengths and weaknesses in relation to other studies, discussing important differences in results

This is the first study to analyse the potential impact of patient travel distances on patients receiving RevKR. The findings that longer travel distances are not associated with inferior outcomes is an important part of the evaluation of the assumptions and context behind the establishment of revision knee networks.[28] This study has shown that concerns of introducing a network in larger geographical regions, for example in Scotland where longer patient travel distances and times are common, may be less important.[29] This is particularly useful as regions explore the geography of their revision networks and during summative outcome assessment of this complex health intervention.[30] Despite there being a potential negative association between straight line travel distance and emergency readmission at 30

days, there was a lack of association involving driving distances and times which present real world challenges for patients.

It may be seen as surprising that no association between travel distance and prolonged length of hospital stay was identified. An expectation exists of increasing difficulties being encountered with the discharge of patients living greater distances from their treating hospital, which has been observed in patients following elective pancreatic surgery.[31] This is also an observation seen in patients being treated in specialist vascular centres in the United States which led to the recommendation of additional care coordination and follow up efforts. However, the geography of the population in these studies was much larger with significantly longer travel distances.

We did observe a weak but statistically significant correlation between social deprivation status and age of the patient with longer travel distances. Patients from poorer sociodemographic background may be expected to travel further for RevKR. This highlights the additional care coordination and follow-up efforts that should accompany the widening reach of regional revision knee networks. It is reassuring that access to treatment for older patients is unaffected by travel distance. However, there may be patients who refused to travel to a specialist centre and opted for treatment at their local centre.

Meaning of the study: possible explanations and implications for clinicians and policymakers

The organisation and delivery of revision knee services in England has recently undergone a substantial change and now such services are provided around regional networks of care. This promises substantial advantages to the increasing number of patients with problematic knee replacements in our ageing population who will benefit from regional expertise.[8] However, it is unknown the impact of patients residing farther from tertiary referral centres, particularly rural patients who may encounter additional difficulties associated with greater travel distance. A recent study following the outcomes of aortic surgery found that longer travel distances are

associated with inferior perioperative outcomes[13]. Similar associations have been found in postoperative colorectal surgery patients [32]. As such our results are reassuring to policy makers and clinicians.

Unanswered questions and future research

There is a scarcity of evidence evaluating the patient perception of complex health interventions such as network models of care. Recent work by Kugler et al has demonstrated the willingness of patients to travel further for better outcomes in the context of total knee replacement surgery. [33] Nevertheless, patient perceptions of travelling further for their treatment should be a focus for future research in the context of revision knee patients, particularly as this is one of the top ten research priorities identified by the James Lind Alliance priority setting partnership.[34]

Conclusion

We did not observe an association in our study population between 30-day readmission rates and increasing travel distances or times between a patient's home and their treating hospital in revision knee replacement. This paper is the first to explore the relationship between travel distance and complex orthopaedic surgery and informs some concerns regarding the creation of a centralised revision knee network. This information is of utility to surgical providers and commissioners of healthcare services. Furthermore, it can inform patient-led decision making and the exploration of perceptions surrounding travelling for complex surgery. Although this is the first assessment in complex orthopaedic surgery, a prospective analysis will be undertaken as part of the ongoing auditing of revision knee networks in England.

Supplementary material and figures

Supplementary material S1 – Journey Time Statistics Reference Document

Supplementary material S2 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

See separate file named supplementary material S2

Supplementary material S3 – Flow of patient inclusion/exclusions

-See attached file named Supplementary Material S3

Supplementary material S4 – R Code

See attached file named Supplementary Material S4

Supplementary material S5 –Scatterplot for imputed data: A comparison between imputed values and observed values following multiple random imputation. Imputed values in “blue”, observed values in “grey”. Imputation 0 on X axis refers to original dataset. Subsequent random imputations labelled 1 to 5 on x axis.

Figure 1 -

(Left) Scatterplot showing correlation between patient age and travel distance. Red line represents linear regression trend. Spearman's rank correlation is presented in chart.

(Right) Scatterplot showing correlation between social deprivation and patient travel distance. Red line represents linear regression trend. Spearman's rank correlation is presented in chart.

Figure 2 - **Predicted probability of emergency readmission at 30 days by straight line patient travel distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 3 knots at 5%, 50% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 3 - **Predicted probability of prolonged length of inpatient stay at by patient straight line travel distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 4 - **Predicted probability of prolonged length of inpatient stay at by patient driving distance from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Figure 5 - **Predicted probability of prolonged length of inpatient stay at by patient driving time from hospital after RevKR**

A Fixed effects multivariable logistic regression model using 4 knots at 5%, 35%, 65% and 95% centiles of mean unit volume. 95% confidence intervals represented by blue shaded line

Contributorship

Alex Matthews: Conceptualisation, Methodology, Project Administration, Investigation, Data Curation, Formal Analysis, Visualisation, Writing - original draft, Writing - review and editing. This author is the guarantor and is responsible for the content

Jonathan P Evans: Conceptualisation, Supervision, Writing - review & editing

Jonathan T Evans: Supervision, Writing - review and editing

Sarah E Lamb: Conceptualisation, Supervision, Writing - review and editing

Andrew Price: Conceptualisation, Supervision, Writing - review and editing

William Gray: Conceptualisation, Supervision, Methodology, Writing - review and editing

Tim Briggs: Supervision, Writing - review and editing

Andrew Toms: Conceptualisation, Supervision, Writing - review and editing

Public and Patient Involvement statement

The study's chief investigator (AT) led the James Lind Alliance 'Revision Knee Replacement' priority setting partnership. This group of patients, carers and health care professionals identified the need to investigate the best way of organising revision knee replacement surgery to improve patient outcomes as one of their top 10 research questions. Patients were therefore directly involved in the development of the study's aims and objectives. The results of the study will be disseminated to the members of this group prior to publication.

Copyright/licence for publication Statement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing Interests

All authors have completed the Unified Competing Interest form (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical Approval

The analysis and presentation of data follows current NHS England guidance for the use of HES data for research purposes and is anonymised to the level required by ISB1523 Anonymisation Standard for Publishing Health and Social Care Data. Ethical approval was not required.

Funding

No funding was obtained to carry out this study

Data Sharing

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

References

1. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet. 2018;392(10158):1672-82.
2. Sabah SA, Hedge EA, Abram SGF, Alvand A, Price AJ, Hopewell S. Patient-reported outcome measures following revision knee replacement: a review of PROM instrument utilisation and measurement properties using the COSMIN checklist. BMJ Open. 2021;11(10):e046169.
3. Sabah S, Knight R, Alvand A, Murray D, Petrou S, Beard D, Price A. No exponential rise in revision knee replacement surgery over the past 15 years: an analysis from the National Joint Registry. Osteoarthritis and Cartilage. 2022;30(12):1670-9.
4. Rowland T, Lindberg-Larsen M, Santy-Tomlinson J, Jensen CM. A qualitative study of patients' experiences before, during and after surgical treatment for periprosthetic knee joint infection; "I assumed it had to be like that … ". International Journal of Orthopaedic and Trauma Nursing. 2023;48:100992.
5. Garbarino LJ, Gold PA, Sodhi N, Anis HK, Ehiorobo JO, Boraiah S, et al. The effect of operative time on in-hospital length of stay in revision total knee arthroplasty. Ann Transl Med. 2019;7(4):66.
6. Liodakis E, Bergeron SG, Zukor DJ, Huk OL, Epure LM, Antoniou J. Perioperative Complications and Length of Stay After Revision Total Hip and Knee Arthroplasties: An Analysis of the NSQIP Database. The Journal of Arthroplasty. 2015;30(11):1868-71.
7. Briggs TWR. A National Review of Adult Elective Orthopaedic Services in England. In: Time GIRF, editor. 2015.
8. Kalson NS, Mathews JA, Phillips JRA, Baker PN, Price AJ, Toms AD, et al. Revision knee replacement surgery in the NHS: A BASK surgical practice guideline. The Knee. 2021;29:353-64.
9. Yapp LZ, Walmsley PJ, Moran M, Clarke JV, Simpson AHRW, Scott CEH. The effect of hospital case volume on re-revision following revision total knee arthroplasty. J Bone Joint Surg Br. 2021;103-B(4):602-9.
10. Halder AM, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Re-Revision Rate Following Aseptic Revision Total Knee Arthroplasty: An Analysis of 23,644 Cases. J Arthroplasty. 2020;35(4):1054-9.
11. Jeschke E, Gehrke T, Günster C, Heller KD, Leicht H, Malzahn J, et al. Low Hospital Volume Increases Revision Rate and Mortality Following Revision Total Hip Arthroplasty: An Analysis of 17,773 Cases. J Arthroplasty. 2019;34(9):2045-50.
12. Bloch BV, Matar HE, Berber R, Gray WK, Briggs TWR, James PJ, Manktelow ARJ. The impact of a revision arthroplasty network on patient outcomes. J Bone Joint Surg Br. 2023;105-b(6):641-8.
13. Feldman ZM, Zheng X, Mao J, Sumpio BJ, Mohebali J, Chang DC, et al. Longer patient travel distance is associated with increased non-index readmission after complex aortic surgery. Journal of Vascular Surgery. 2023;77(6):1607-17.e7.
14. Saunders RS, Fernandes-Taylor S, Kind AJ, Engelbert TL, Greenberg CC, Smith MA, et al. Rehospitalization to primary versus different facilities following abdominal aortic aneurysm repair. J Vasc Surg. 2014;59(6):1502-10, 10.e1-2.
15. Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (Hospital Episode Statistics) to assess health-care outcomes--systematic review and recommendations for reporting. Eur J Public Health. 2013;23(1):86-92.

16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
17. NHS Digital. Hospital Episode Statistics(HES) Analysis Guide. London, UK: NHS Digital; 2018.
18. Information Standard Board for Health and Social Care. Anonymisation Standard for Publishing Health and Social Care Data Specification. London, UK: NHS Digital; 2013.
19. NHS Digital. Delivering linked datasets to support health and care delivery and research. London, UK: NHS Digital; 2018.
20. Department for Transport. Journey Time Statistics, notes and definitions. London, UK: Department for Transport; 2019.
21. Kalson NS, Mathews JA, Miles J, Bloch BV, Price AJ, Phillips JRA, et al. Provision of revision knee surgery and calculation of the effect of a network service reconfiguration: An analysis from the National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. Knee. 2020;27(5):1593-600.
22. Deere K, Whitehouse MR, Kunutsor SK, Sayers A, Price AJ, Mason J, Blom AW. How long do revised and multiply revised knee replacements last? An analysis of the National Joint Registry. Lancet Rheumatol. 2021;3(6):e438-e46.
23. Kamath CC, O'Byrne TJ, Lewallen DG, Berry DJ, Maradit Kremers H. Neighborhood-Level Socioeconomic Deprivation, Rurality, and Long-Term Outcomes of Patients Undergoing Total Joint Arthroplasty: Analysis from a Large, Tertiary Care Hospital. Mayo Clin Proc Innov Qual Outcomes. 2022;6(4):337-46.
24. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17(1):162.
25. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016;4(2):30.
26. Sabah SA, Knight R, Alvand A, Palmer AJR, Middleton R, Abram SGF, et al. Patient-Relevant Outcomes Following First Revision Total Knee Arthroplasty, by Diagnosis: An Analysis of Implant Survivorship, Mortality, Serious Medical Complications, and Patient-Reported Outcome Measures Utilizing the National Joint Registry Data Set. J Bone Joint Surg Am. 2023;105(20):1611-21.
27. Hardy F, Heyl J, Tucker K, Hopper A, Marchã MJ, Briggs TWR, et al. Data consistency in the English Hospital Episodes Statistics database. BMJ Health Care Inform. 2022;29(1).
28. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.
29. Farrow L, Kennedy JW, Yapp L, Harding T, Walmsley P. Provision of revision knee arthroplasty services across Scotland: A national audit. Knee. 2023;42:312-9.
30. Paterson C, Baarts C, Launsø L, Verhoef MJ. Evaluating complex health interventions: a critical analysis of the 'outcomes' concept. BMC Complement Altern Med. 2009;9:18.
31. Jackson KL, Glasgow RE, Mone MC, Sheng X, Mulvihill SJ, Scaife CL. Does travel distance influence length of stay in elective pancreatic surgery? HPB (Oxford). 2014;16(6):543-9.
32. Cloyd JM, Huang L, Ma Y, Rhoads KF. Predictors of readmission to non-index hospitals after colorectal surgery. The American Journal of Surgery. 2017;213(1):18-23.

33. Kugler CM, De Santis KK, Rombey T, Goossen K, Breuing J, Könsgen N, et al. Perspective of potential patients on the hospital volume-outcome relationship and the minimum volume threshold for total knee arthroplasty: a qualitative focus group and interview study. BMC Health Serv Res. 2021;21(1):633.
34. Mathews JA, Kalson NS, Tarrant PM, Toms AD. Top ten research priorities for problematic knee arthroplasty. Bone Joint J. 2020;102-b(9):1176-82.

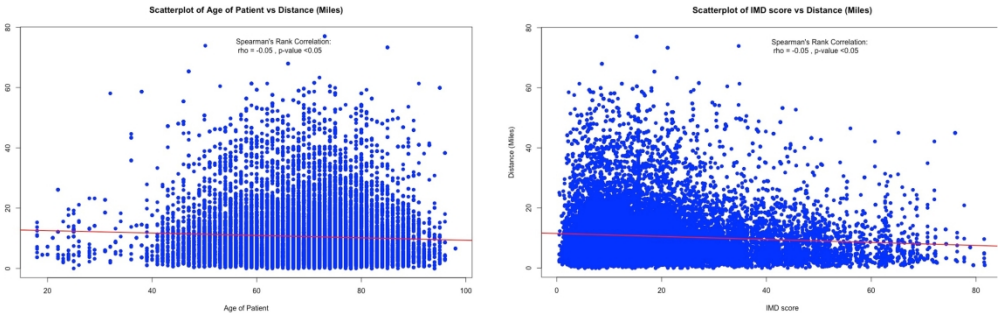


Figure 1

360x119mm (300 x 300 DPI)

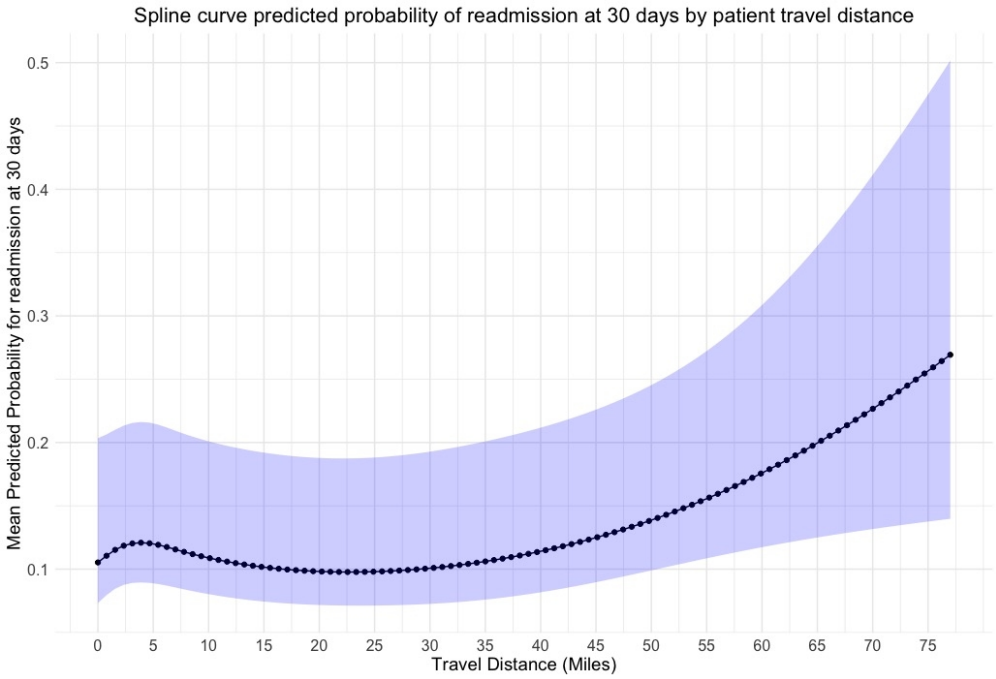


Figure 2

87x59mm (300 x 300 DPI)

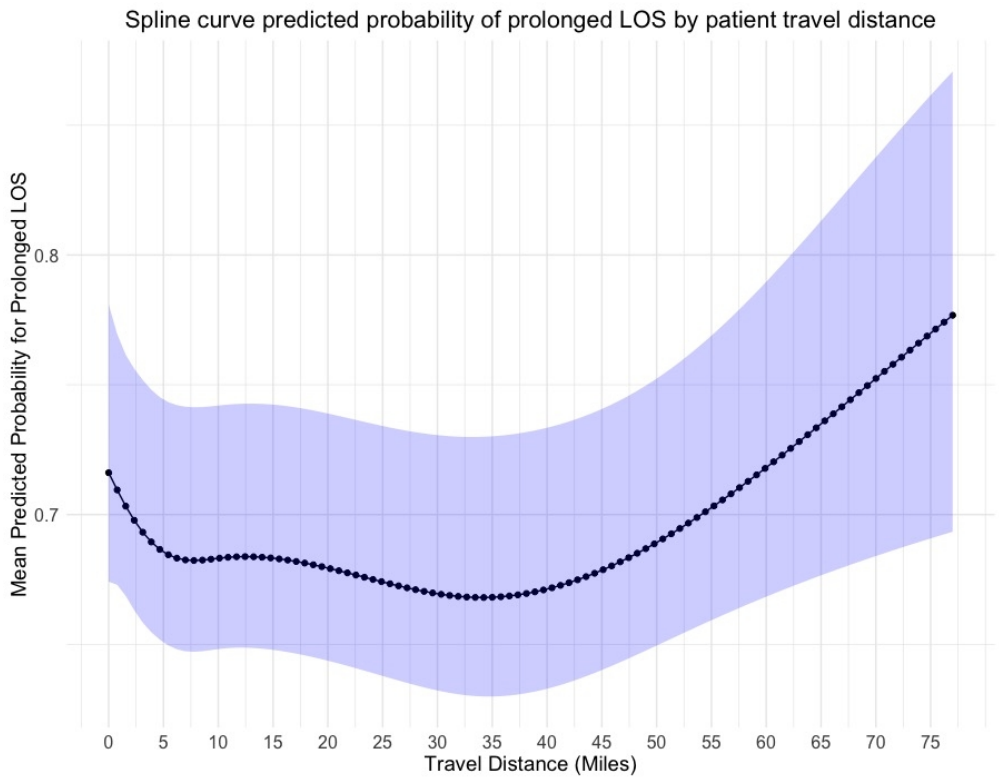


Figure 3

76x59mm (300 x 300 DPI)

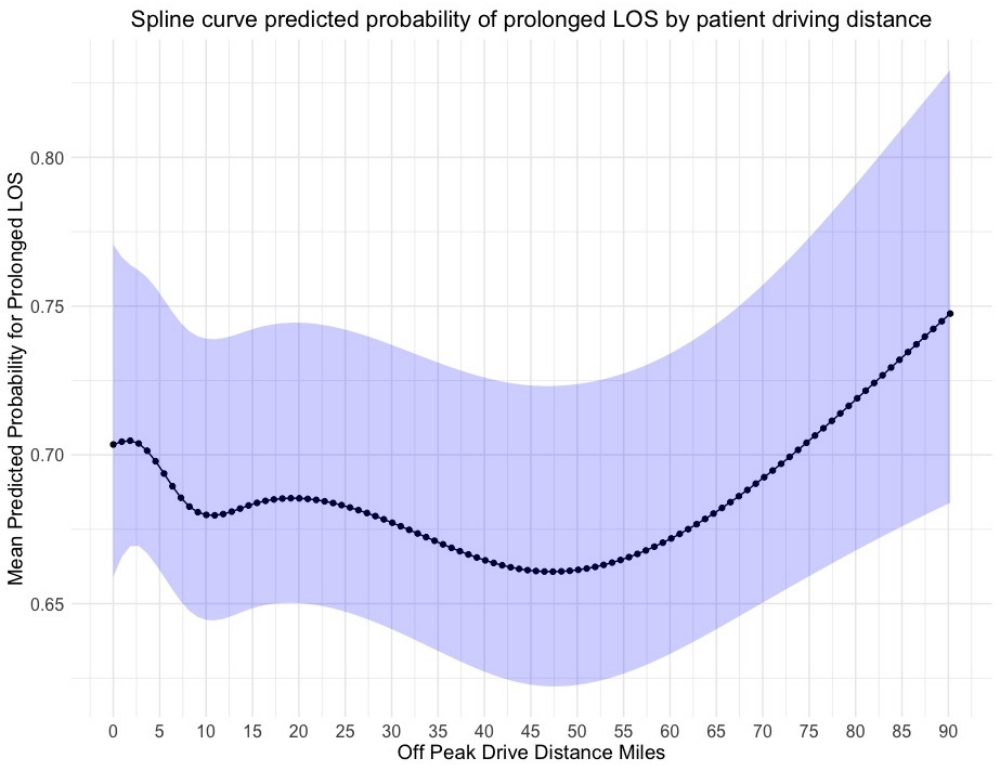


Figure 4

78x60mm (300 x 300 DPI)

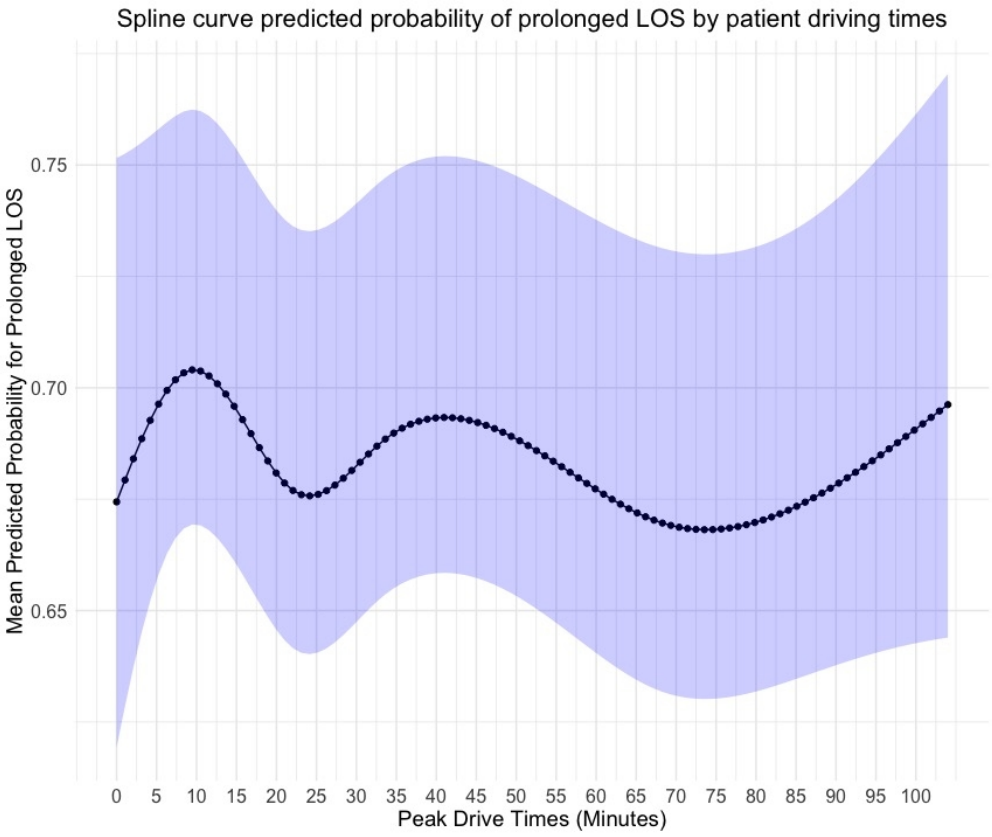


Figure 5

72x60mm (300 x 300 DPI)

Journey Time Statistics: Notes and Definitions

About this release

This publication supports the latest statistics on journey times.

In this publication

Further information

Public enquiries
020 7944 3077
vehicles.stats@dft.gov.uk
Media enquiries
020 7944 3066

Overview

This note provides information on the methodology used, the source data and definitions of key terms for calculating Journey Time Statistics.

These annual statistics were first published in December 2015 for the year 2014 and have been developed from the earlier Accessibility Statistics published for 2007 to 2013.

The Journey Time Statistics produced by DfT consists of theoretical journey times calculated by modelling journeys between known sets of origins and destinations. It uses information on the road network, traffic speeds and public transport timetables in England.

The relevant Journey Time Statistics calculation is varied for origins and destination to meet a variety of needs. Two sets of analysis are published:

- Access to key services; and
- Connectivity

Origin indicators

These indicators measure the number of different services in a particular area that users can reach within a given time.

Destination indicators

These indicators measure the proportion of users that can access a service within a certain time.

The 'user' populations for each service in the destination indicators are:

Employment	16-74 year olds
Primary schools	5-10 year olds
Secondary schools	11-15 year olds
Further education	16-19 year olds
All other services	All households

Key services

- Employment centres: Data used are the number of jobs in a Lower Super Output Area (LSOA). The data tables include results for employment centres of 3 different sizes (100-499 jobs, 500-4,999 jobs and at least 5,000 jobs). For the key services average, the 500-4,999 jobs definition is used for employment.
- Education: Locations of all open Primary schools, Secondary schools, Further Education and Sixth Form Colleges.
- General Practice (GP) surgeries: For 2017 based on the Patients Registered at a GP Practice dataset released by NHS Digital – previously this was based on a filtered dataset of NHS prescribers released by NHS Digital.
- Hospitals: Based on hospitals that are registered with the Care Quality Commission (CQC) and are managed by Acute Trusts.
- Food stores: Locations of grocery, supermarkets or convenience stores.
- Town centres: Locations of Town centres using a central focal point for the town mapped to the nearest road.

Geography

- Local authorities

In some parts of England there are two tiers of local authorities, and in others a single unitary authority. Statistics have been calculated for both types of authority - around 360 in all. These vary considerably in size, from a population of a few tens of thousands to over a million.

- Lower Layer Super Output Areas (LSOA)

LSOAs are small areas designed to be of a similar population size, with an average of approximately 1,500 residents or 650 households. There are 32,844 Lower-layer Super Output Areas (LSOAs) in England. They were determined by the Office for National Statistics for the reporting of small area statistics and are derived from the 2011 Census.

- Urban and rural definitions

This report uses the Defra Rural-Urban Classification, based on 2011 Census Output Areas. The Rural-Urban Classification defines areas as rural if they fall outside of settlements with more than 10,000 resident population. See Defra's Definitions and Local Authority Classification for more details.

Journey time calculations

The journey time calculations are carried out using a commercially available software package called TRACC, owned by Basemap. TRACC is a desktop application that uses public transport and highways data to create journey times from origins to destinations. It uses timetable information showing both arrival and departure times at stops from public transport services against a specific time/day period. Highways information from road networks are used to fill the gaps between public transport services by creating a linear network that connects the origins, destinations and stops together. This provides a fully routable network of nodes and lines which is saved on file as a graph network. The graph network has various constraints which can be altered to suit the user need such as distance travelled, interchange delays on public transport and stopping limitations on road networks. The TRACC software then queries the graph network with origin and destination co-ordinates and uses the Dijkstra shortest path algorithm to route between these points. This is an algorithm for finding the shortest distance for travel between the graph networks.

For a public transport journey, the journey time produced includes all walking elements of the journey, i.e. the walk from the origin of the journey to the road, from the road to public transport stops, any interchange of public transport using the road and then from the final stop to the destination via the road, and finally from the nearest point on the road network to the destination. The journey assumes arrival at the first stop one minute before the initial departure, with any subsequent interchange waiting times included as part of the final journey time.

Car, cycle or walk only journeys are similar except that once the road network is reached the journey proceeds link by link along the road network at speeds governed by data held in the model. These are specific to the mode, the road type, and in some cases the individual road link.

The 10 shortest journey times from each origin (i.e. Output Area) are calculated for each destination type. For the public transport / walking mode these consist of the 10 shortest journey times by either walking or public transport, after applying a 5 minute penalty for any journeys using public transport (to represent travellers arriving slightly early at the first stop).

The journey times are representative of the ‘morning peak’. This is made explicit for public transport / walking by requiring the journey to be completed between 7 and 10am, and for car journeys by using average traffic speeds for between 7 and 10am. For the cycle mode no actual speed data are available. The cycle speeds used are default assumptions, and are not based on a particular time of day.

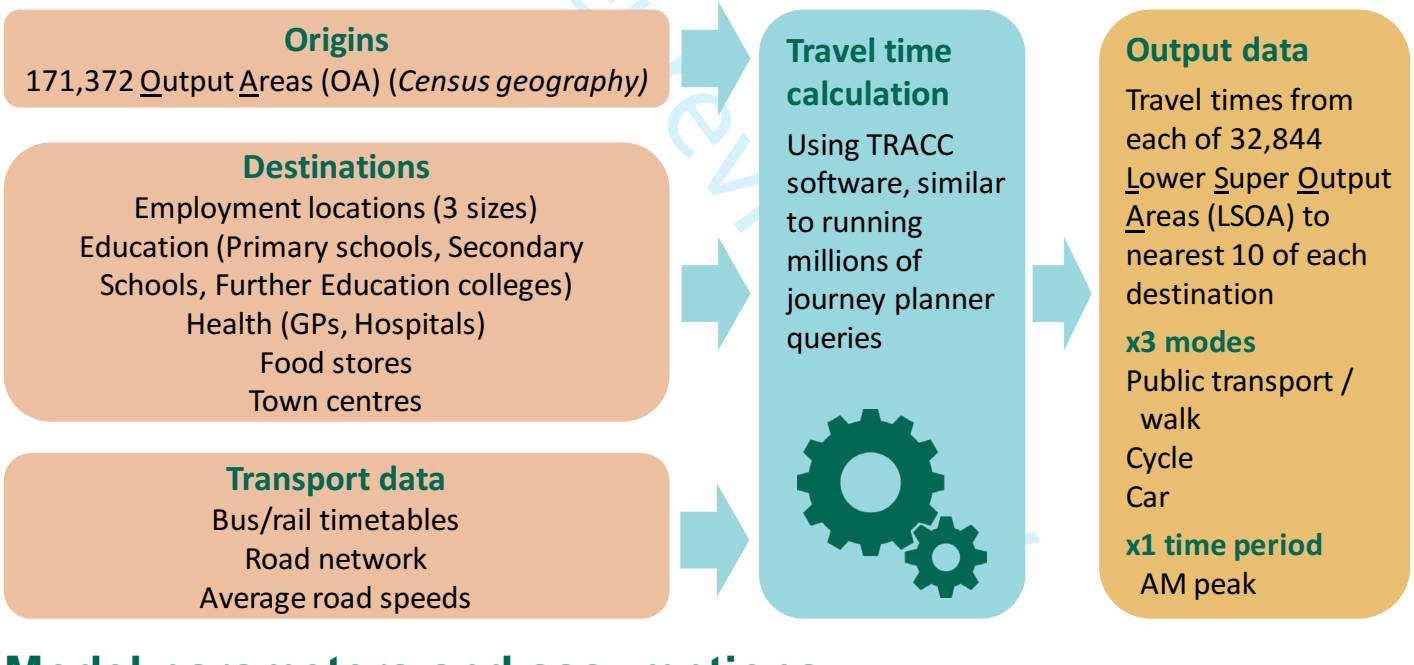
Access to key services

The Access to Services analysis applies the Journey Times methodology to origins consisting of residential neighbourhoods and destinations consisting of centres of employment and a range of key local services. Journey times are calculated for three modes of transport: public transport; driving; and cycling. These journey times are then used to generate further indicators, as described in the **Outputs Section**.

The Access to Services calculation process and the coverage of the data set are very similar to those of the Accessibility Statistics from which they were developed. However, the calculation algorithm and a number of other features of the design are different, so the results are not directly comparable.

The statistics are designed to represent as much as possible the situation on a **Tuesday in October** of **the year to which they relate**. Data for the second week of October are used in the analysis, since this provides a fairly typical week, unaffected by major national holidays, school holidays or other seasonal effects. The origins, destinations and public transport timetables used are as far as possible for this date. The traffic data are averages for the preceding 12 months up to and including August. The road networks are those current at the start of the traffic data year.

Outline of access to services calculation process



Model parameters and assumptions

General parameters

Maximum journey time of **2 hours.**

Maximum journey distance of **100km.**

Walking

These apply to both:

- walking between origin / destination and the transport networks at both ends of a journey by

any mode;

► walk only journeys as part of the public transport / walk mode.

Maximum straight line distance between origin / destination and road network of **2km.** The algorithm will always use nearest point on network. For cycle or car modes, travel by cycle or car begins from this point. For public transport/walk, traveller walks along road network to the most suitable public transport stop, or direct to the destination if this is quicker.

Walking speed on road/path network of **4.8km/h.**

Walking speed off road/path network of **4.0km/h.**

Public transport

Interval within which door-to-door journey must be completed (required for timetable selection) is **7am to 10am on a Tuesday.**

Maximum walk distance of **3km** - this applies to walks from origin to first public transport stop, from last stop to destination, and also walking directly from origin to destination without using public transport at all.

Maximum number of potential first public transport stops considered in routing algorithm is **100** (starting with the closest to origin).

Allowance for catching first public transport service is **5 minutes** - added to any journey that involves boarding one or more public transport services.

Public transport speed – this is provided implicitly by the timetable information.

Interchange time of **5 minutes** (minimum interval allowed between arriving at a stop and catching another service).

Maximum straight line distance between public transport interchanges of **500m**.

Stop clustering at **150m** – groups together public transport stops within this distance of one another to speed up processing. The individual timetables for each service are retained.

Cycling speeds

Road Type	Speed
Motorway	0.0 km/h
Urban Motorway	0.0 km/h
A road	16.0 km/h
B road	16.0 km/h
Minor road	16.0 km/h
Local street	16.0 km/h
Private road – restricted access	4.8 km/h
Private road – public access	16.0 km/h
Pedestrian street	4.8 km/h
Alley	4.8 km/h

Parking time of **5 minutes** - added to all cycle journeys.

Car speeds

Type of road	2014	2015	2016	2017
	Default speeds (km/h)			
Motorway	79.5	77.0	77.5	77.6
Urban Motorway	79.5	77.0	77.5	77.6
A road	42.7	43.7	43.3	43.2
B road	41.6	43.0	42.2	41.9
Minor road	36.8	37.5	36.8	36.3
Local street	19.2	17.8	18.8	18.3
Private road – restricted access	17.0	16.7	16.2	15.3
Private road – public access	14.8	15.2	15.1	13.6
Pedestrian street	0.0	0.0	0.0	0.0
Alley	0.0	0.0	0.0	0.0

Car speeds are calculated for specific links where more than 200 records exist otherwise the default speeds are used. Minimum journey time for a journey that uses a car is **5 minutes**.

Time at junctions

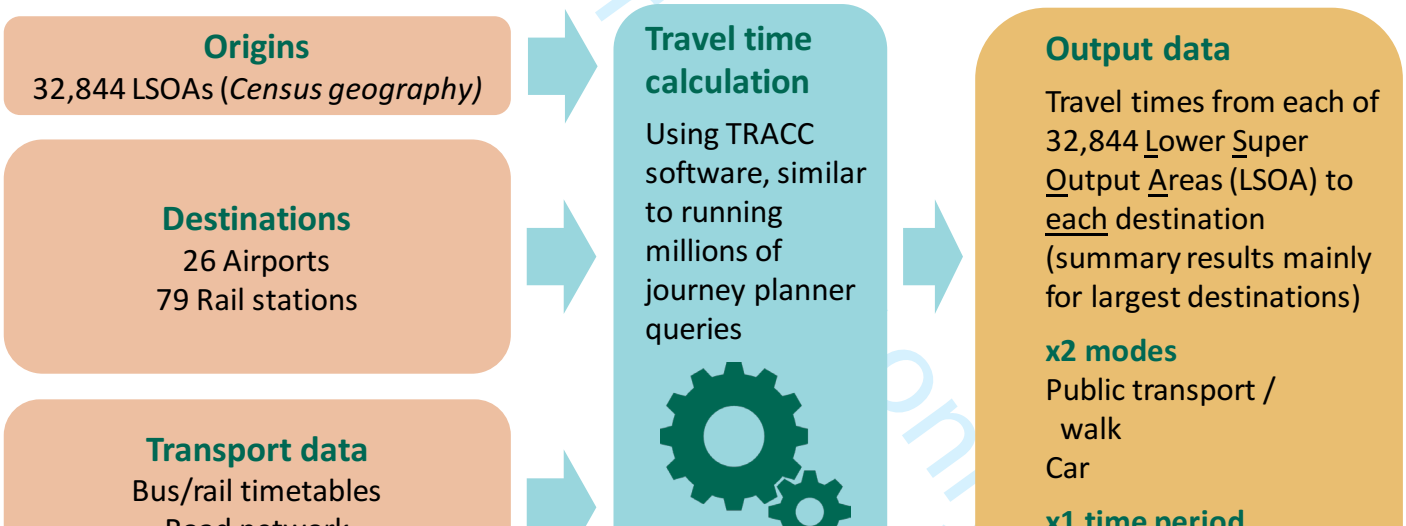
Road normalisation is used for all modes of transport which converts each road link to a straight line to speed up processing. The true link length is retained for accurate speed/time calculations, but there could be a small effect on the calculation of shortest distance from the road network to destination points. Effect for origins is minimal due to origins being constrained to road nodes.

Connectivity

These experimental analyses are intended to apply the Journey Times methodology to a range of more strategic or economically significant destinations than the primarily local services covered by the Access to Services analyses; including airports and railway stations. The principle difference in the Connectivity approach from that of the Access to Services analyses is that journey times are calculated, as far as possible, to all accessible locations, rather than to just the nearest 10 examples. This tends to result in a much larger data set being generated. In some cases a longer maximum journey time may be allowed although this may depend on what is considered reasonable for the type of destination. Given these factors, a less detailed origin data set may be used than for Access to Services. This is both necessary, to limit the size of the data set, and acceptable where the typical journey lengths are longer.

The first connectivity analyses published using the new Journey Time methods were released in Journey Time Statistics 2015, published in April 2017, for two destination sets – airports and rail stations. These analyses using the Journey Times methods superseded two earlier Connectivity Statistics reports published in 2014 and 2015 based on the old accessibility statistics methods, in the same way that the new Access to Services analyses have replaced the earlier Accessibility Statistics. Again, the connectivity results produced using the old and new methods are not directly comparable.

Outline of Connectivity calculation



Model parameters and assumptions

Origins	Population weighted centroids (the central point) of 32,844 English LSOAs as specified in the 2011 Census geography. These points were then constrained to the nearest road node, as for Access to Services method.

Journey Time Calculation	As for Access to Services, for public transport / walking and car modes only, except that a maximum journey time of 240 minutes and maximum straight line distance of 400km is allowed.
Outputs	Generally similar to Access to Services, with different journey time classifications as appropriate. Journey time results to specific destinations are included – this is the key difference in the Connectivity analyses. 'Average journey times' and 'nearest' destinations should be used with caution. The average journey times exclude results for areas with no available connection under 240 minutes, which may become significant in remote areas and for destinations are a great distance from the origin. The 'nearest' destination is the destination with the shortest average journey time across the whole area considered – which will be relatively large in the case of local authority level results.

Data sources

Origins

The origins used for all Access to Services calculations are the 171,372 English Output Areas (OA) as specified in the 2011 Census geography.

To provide the actual journey start point in each OA, the population weighted centroid of the OA was shifted to the nearest node (i.e. junction) on the road network. This was to avoid biasing the journey time results where the centroid of the OA was a long way from a road. In fact it is rare for an OA centroid to be more than about 100 metres from a road – only a tiny handful of OA in remote areas have centroids as much as 1km from a road. The OA centroids have been shifted onto the nearest road node rather than the nearest point on a road in order to reduce issues arising from normalising the road network.

Origin	Data source for the origin points
All	Data: Population centroid of each Output Area in 2011. Source: ONS 2011 Census Boundaries. Further information: http://geoportal.statistics.gov.uk

Destinations

The destinations used consist of three different sizes of employment centre and the locations of seven other types of key local service. For each of these key services a nationally consistent data set has been identified or derived – further information on these is provided in this section.

Each destination is located by a 6-figure National Grid reference. For the employment destinations this is taken to be the population weighted centroid of the LSOA.

Destination	Number of locations			
	2014	2015	2016	2017
Employment centres (small)	16,465	16,625	16,930	17,194
Employment centres (medium)	9,235	9,460	9,707	10,241
Employment centres (large)	645	676	719	785
Primary schools	16,463	16,484	16,655	16,927
Secondary schools	3,365	3,376	3,381	3,174
Further education colleges	2,624	2,606	2,418	2,304
GPs	9,257	11,167	9,128	7,353
Hospitals	296	278	278	277
Food stores	19,549	19,746	21,665	20,987
Town centres	1,211	1,211	1,211	1,211

The data source for GP surgeries was reviewed and replaced for 2017.

Access to key services

Destinations 2017	Data source for the locations of the service	Data source for users of the service
Employment	Data: Number of jobs available in a LSOA in the year before the calculation year.	Data: Number of 16-74 year olds in each output area.
	Source: ONS Business Register Employment Survey.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://www.nomisweb.co.uk/default.asp	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Primary schools	Data: Location of all open primary schools in September of calculation year.	Data: Number of 5-10 year olds in each output area.
	Source: The Department for Education (DfE) Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Secondary schools	Data: Location of all open secondary schools in September of calculation year.	Data: Number of 11-15 year olds in schools in each output area.
	Source: DfE Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.
Further education colleges	Data: Location of all open further education and sixth form colleges/school sixth form in September of calculation year.	Data: Number of 16-19 year olds in each output area.
	Source: DfE Edubase.	Source: ONS mid-year population estimates for calculation year.
	Further information: https://get-information-schools.service.gov.uk/	Further information: ONS mid-year population estimates: http://www.ons.gov.uk/ons/taxonomy/index.

Destinations 2017	Data source for the locations of the service	Data source for users of the service
GPs	Data: Locations of GP surgeries with registered patients in October of calculation year.	Data: Number of households in each output area.
	Source: NHS Digital table of Registered patients at GP practices	Source: 2011 Census + Local Authority (LA) updates from the Ministry of Housing, Communities & Local Government (MHCLG) mid-year household projections of calculation year.
	Further information: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections
Hospitals	Data: Location of hospitals.	Data: Number of households in each output area.
	Source: Care Quality Commission - Directory of places that provide care.	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: http://www.cqc.org.uk/content/how-get-and-re-use-cqc-information-and-data	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Destinations 2017	Data source for the locations of the service	Data source for users of the service
Food stores	Data: Location of grocery/ supermarkets or convenience stores in October of calculation year.	Data: Number of households in each output area.
	Source: The Local Data Company	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: https://www.localdatacompany.com/	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections
Town centres	Data: Location of town centres in 2004.	Data: Number of households in each output area.
	Source: MHCLG Town Centre and retail planning statistics for England and Wales.	Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year.
	Further information: https://data.gov.uk/dataset/ed07b21f-0a33-49e2-9578-83ccbc6a20db/english-town-centres-2004	Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

GP destination data

The GP surgery destinations used from 2014 to 2016 are based on the list of practices maintained by the Organisational Data Service of the Health & Social Care Information Centre, and published at https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data. This was supplemented with information on branch surgeries from the same source. Grid references were derived from the postcode using the Office for National Statistics (ONS) Postcode Address File. Practices with identical postcodes were taken to be duplicates or co-located, and all additional records after the first were removed.

From 2017, the list of GP locations is taken from the NHS Digital publication of Registered patients at GP practices for October of the calculation year. This had the effect of reducing the number of locations in the dataset, but removed the need for manual adjustments and produces a more stable list defined as GP practices with registered patients. Grid references were derived from the postcode using the Office for National Statistics (ONS) Postcode Address File.

Hospital destination data

The starting point for hospital sites is the Care Quality Commission's (CQC) list of 'active locations' dataset, which is thought to be the most-up-to date and freely available source of data on individual National Health Service (NHS) and social care 'sites' or hospitals. A criteria was developed in consultation with the Department of Health to reduce the list down to capture only the key hospitals. The following have been removed and individual records have been inspected to remove further examples of these cases and for any duplicates:

- care home records;
- non-NHS providers;
- sites not associated with acute providers;
- any remaining sites that are associated with Specialist Trusts (usually single speciality Trusts or Sites);
- records where it is evident from the name that the record is not a hospital (e.g. headquarters, specialist units.)

This gave a final list of 278 hospitals in 2017 run by Acute (non-specialist) Trusts. As well as covering all general hospitals this will still include some with a largely or entirely community or rehabilitation role, where these happen to be managed by an Acute Trust. It was considered on balance better to leave these in the list, rather than risk adding further subjectivity to the selection. Whilst not perfect, it is considered that the resulting list is a significant improvement on that used previously.

Steps taken to produce hospital data set
Remove records where **Care Home** = Y
Remove records where **Provider ID** begins 1-
Keep records where **Benchmark Group** is Care Home or **Cluster Group** is Acute
Filter the trust site locations by name to remove obvious non-hospital sites. Key words used for this process are: birth, dental, house, clinic, grange, lodge, infirmary, health, community, unit, surgery, centre
Manual review of remaining locations

Employment destination data

The employment centres are defined by the number of jobs existing in each English LSOA, taken from the Business Register Employment Survey. Large Employment Centres are defined as those with 5,000 or more jobs, Medium Employment Centres as those with 500 or more jobs, up to 4,999 and Small Employment Centres as those with 100 or more jobs, up to 499.

Data are downloaded from the Nomis website; although LSOA level BRES data has safeguarded access, access can be requested through the site. The chosen data download options are LSOA2011 geography, date as calculation year, variable as employment status where the value is employed, and the measure chosen is a count.

For the 2016 destination set, the BRES changed from 2001 census geography to 2011 census geography. The majority of LSOA boundaries are unchanged between these datasets, but some have been merged or split. Therefore the employment destination indicators are not strictly comparable between 2015 and 2016 Journey Time statistics. See https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography for further information.

Education destination data

The education destination datasets are taken from the Department for Education database of educational establishments. The database was filtered to remove those establishments that were not open during the school year starting in September of the calculation year. Further filters were applied to remove special educational establishments, boarding schools and selective schools, and then to select schools at each phase of education for primary and secondary schools and further educational establishments. The following table lists the filters used.

Phase of Education	Code Variable	Variable	Selected codes and values	
All Schools	OpenDate			30/08/17 or earlier; NULL
	CloseDate			30/08/18 or later; NULL
	TypeOfEstablishment_Code_	TypeOfEstablishment	1	Community school
			2	Voluntary aided school
			3	Voluntary controlled school
			5	Foundation school
			6	City technology college
			12	Foundation special school
			18	Further education
			28	Academy sponsor led
			29	Higher education institutions
			31	Sixth form centres
			32	Special post 16 institution
			34	Academy converter
			35	Free schools
			36	Free schools special
			39	Free schools 16 to 19
			40	University technical college
			41	Studio schools
			45	Academy 16-19 converter
			46	Academy 16 to 19 sponsor led
	Boarders_Code_	Boarders	0	Not applicable
			1	No boarders
			9	NULL
	AdmissionsPolicy_Code_	AdmissionsPolicy	0	Not applicable
			4	Non-selective
			9	NULL
Primary schools	PhaseOfEducation_Code_	PhaseOfEducation	2	Primary
			3	Middle deemed primary
			7	All through

Phase of Education	Code Variable	Variable	Selected codes and values	
Secondary schools	PhaseOfEducation_Code_	PhaseOfEducation	0	Not applicable
			4	Secondary
			5	Middle deemed secondary
			7	All through
	Statutory High age		>=16	
	Statutory Low age		< 16	
FE	PhaseOfEducation_Code_	PhaseOfEducation	4	Secondary
			5	Middle deemed secondary
			6	16 plus
			7	All through
	Statutory High age		>16	
	OfficialSixthForm_Code_	OfficialSixthForm	0	Not applicable
			1	Has a sixth form
			9	NULL
OR				
FE	EstablishmentTypeGroup__code_	EstablishmentTypeGroup	1	Colleges

Food Stores destination data

The food stores destination dataset is purchased from The Local Data Company and includes all branches of multiple food store chains. Although some data are available for independent food stores, this only exists within town centres and so has not been included.

Connectivity

Destinations	Data source for the locations of the service	Data source for users of the service
Airports	Data: Location of GB airports excluding highlands and islands of Scotland Source: National Public Transport Access Nodes Further information: https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan	Data: Number of households in each output area. Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year. Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Destinations	Data source for the locations of the service	Data source for users of the service
Railway stations	Data: Location of larger (category A, B and C1) rail stations in GB Source: Network rail classification Further information: http://webarchive.nationalarchives.gov.uk/20101007153226/http://www.dft.gov.uk/pgr/rail/passenger/stations/betterrailstations/ http://archive.nr.co.uk/browse%20documents/rus%20documents/route%20utilisation%20strategies/network/working%20group%202%20-%20stations/networkrusstations.pdf	Data: Number of households in each output area. Source: 2011 Census + LA updates from MHCLG mid-year household projections of calculation year. Further information: 2011 Census: http://www.nomisweb.co.uk/census/2011 MHCLG mid-year household projections: https://www.gov.uk/government/statistical-data-sets/live-tables-on-household-projections

Transport network data

Travellers moved between their original and their destination via one or more of the following transport networks, depending on the mode of transport being modelled. For all modes, travellers will probably also need to walk between their origin / destination and the transport network. For some short journeys, it may be quicker for travellers to walk directly to their destination, rather than using public transport at all – this is why public transport / walking results are modelled as a combined mode.

Public transport

National public transport timetable data are publically available. Data for bus, local coach and other local transport services (e.g. light rail, metro, and ferry) are captured in the Traveline National Data Set (TNDS), rail timetable data are published by the Association of Train Operating Companies (ATOC), and national coach services in the National Coach Data Set (NCDS).

Walk

The walking network is represented by the road and urban path elements of the Integrated Transport Network produced by the Ordnance Survey.

Cycle

The cycling network is represented by the road network including cycle paths and bridleways from the Integrated Transport Network. Cycle journeys are also allowed to use footpaths at walking pace.

Car

The car network is represented by the road component of the Integrated Transport Network.

Data on actual vehicle speeds on each road network link (generally the stretch of road between 2 nodes, or junctions) is obtained from Trafficmaster Satnav devices and are used to estimate car speeds. These data are used to calculate annual average traffic speeds on each link of the road network (by direction if the link is bi-directional). These are used as the link speeds for cars in the modelling. Where the Trafficmaster sample for an individual link is too small, national averages of the same data for the particular road type are used instead. This is an innovation from 2014. Previously the sample was too small and the model reverted to default assumptions for car speed based on road type which were much higher than the Trafficmaster averages, resulting in some inconsistency in the model.

Outputs

The journey time results are used to create the following indicators for publication:

Indicator	Description
Minimum journey time	The shortest of the ten journey time results.
Origin indicators	Four measures, the number of destinations (up to the maximum of 10) that can be reached from a given origin within 15, 30, 45 and 60 minutes.
Destination indicators	Four measures, the percentages of service users within the given geographical area who can access at least one service location within 15, 30, 45 and 60 minutes.

Each of these indicators is calculated for each mode and each destination type, and at a number of geographical scales as follows:

- England
- Region
- Local Authorities, including London Boroughs, Metropolitan districts, Unitary authorities, Counties and non-Metropolitan districts, also Inner and Outer London and former Metropolitan counties
- 2011 Lower layer Super Output Area
- 2011 Defra Rural/Urban Classification

The indicators for each geography are calculated as population weighted averages. In other words, the average minimum journey time for an area, B, is:

$$mjt(B)= \sum(i=1)^n(mjt(OAi)\times pop(OA\_i))/pop(B)$$

where mjt(B) is the minimum journey time in area B, mjt(OAi) is the minimum journey time of the ith of n output areas making up area B, and pop(B) and pop(OAi) are the user populations resident in area B and output area i respectively.

The service user populations used in the above weighting, and in the destination indicators, depend on the destination type, as follows:

Destination type	Service user population basis
Employment centres	Resident population of working age (16-74 years)
Primary schools	Population aged 5-10
Secondary schools	Population aged 11-15
Further education colleges	Population aged 16-19
GPs, hospitals, food stores, town centres	Number of households
Average key services	Resident population of working age (16-74 years)

Strengths and Weaknesses

In using the data, the following points should be kept in mind:

► All journey times are compiled on a consistent basis across the country.

► The statistics are based on the calculation of theoretical journey times, they are not based on real journeys. They are however based on actual public transport times, and average traffic speeds on the road network.

► Although the statistics are calculated to a high level of geographical detail, some assumptions and simplifications are necessary in the modelling (for example assigning the start point of journeys to a single point in each Output Area, road speeds, interchange times for public transport).

► For 2016 we have used the 2015 BRES data to designate Lower Super Output Areas as employment centres. The 2015 BRES is the first year to use LSOAs based on the 2011 census, and although the majority of these are an exact match to the 2001 LSOAs, there are some that were merged, split or had other boundary changes. For these areas journey times from earlier years are not comparable to the 2016 journey times. This effect is more pronounced for large employment centres, as there are fewer destinations to route to.

► For particular areas, local authorities and other experts may have more detailed information allowing them to produce more accurate or detailed models of the local situation.

► Demand responsive services (e.g. bus services which have to be booked) are only included to the extent that they can be plausibly modelled, in the Traveline National Data Set.

► Since new journey calculation software was adopted for 2014, along with a significant number of other changes to the methodology, from 2014 results are not directly comparable with those for earlier years.

Supplementary material S1 – OPCS-4 code criteria used for Hospital Episode Statistics data extraction

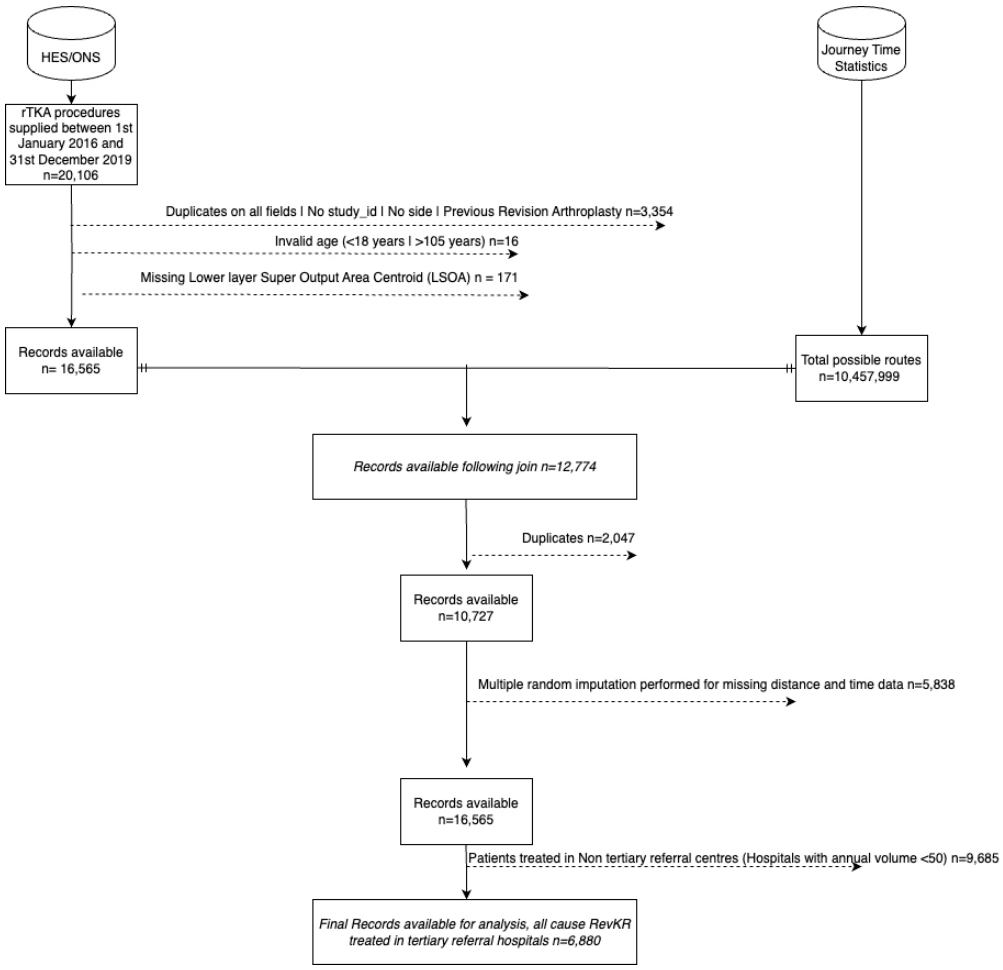
Code	Code description
OPCS-4 codes for knee revision procedures	
O180	Conversion from previous hybrid prosthetic replacement of knee joint using cement
O182	Conversion to hybrid prosthetic replacement of knee joint using cement
O183	Revision of hybrid prosthetic replacement of knee joint using cement
O184	Attention to hybrid prosthetic replacement of knee joint using cement
W400	Conversion from previous cemented total prosthetic replacement of knee joint
W402	Conversion to total prosthetic replacement of knee joint using cement
W403	Revision of total prosthetic replacement of knee joint using cement
W404	Revision of one component of total prosthetic replacement of knee joint using cement
W410	Conversion from previous uncemented total prosthetic replacement of knee joint
W412	Conversion to total prosthetic replacement of knee joint not using cement
W413	Revision of total prosthetic replacement of knee joint not using cement
W414	Revision of one component of total prosthetic replacement of knee joint not using cement
W420	Conversion from previous total prosthetic replacement of knee joint NEC
W422	Conversion to total prosthetic replacement of knee joint NEC

W423	Revision of total prosthetic replacement of knee joint NEC
W424*	Attention to total prosthetic replacement of knee joint NEC
W425	Revision of one component of total prosthetic replacement of knee joint NEC
W522†	Conversion to prosthetic replacement of articulation of bone using cement NEC
W523†	Revision of prosthetic replacement of articulation of bone using cement NEC
W532†	Conversion to prosthetic replacement of articulation of bone not using cement NEC
W533†	Revision of prosthetic replacement of articulation of bone not using cement NEC
W542†	Conversion to prosthetic replacement of articulation of bone NEC
W543†	Revision of prosthetic replacement of articulation of bone NEC
W544*†	Attention to prosthetic replacement of articulation of bone NEC
W553†	Conversion to prosthetic interposition arthroplasty of joint
W564†	Conversion to interposition arthroplasty of joint NEC
W574†	Conversion to excision arthroplasty of joint
W582†	Revision of resurfacing arthroplasty of joint
W603†	Conversion to arthrodesis and extra-articular bone graft NEC
W613†	Conversion to arthrodesis and articular bone graft NEC
W641†	Conversion to arthrodesis and internal fixation NEC
W642†	Conversion to arthrodesis and external fixation NEC
OPCS-4 codes for laterality	
Z941	Bilateral

Z942	Left-sided
Z943	Right-sided
ICD-10 codes for Infection	
T845	Infection and inflammatory reaction due to internal joint prosthesis
T846	Infection and inflammatory reaction due to internal fixation device [any site]
T847	Infection and inflammatory reaction due to other internal orthopaedic prosthetic devices, implants and grafts
T814	Infection following a procedure, not elsewhere classified
ICD-10 codes for fracture	
M966	Fracture of bone following insertion of orthopaedic implant, joint prosthesis or bone plate
ICD-10 codes for mechanical complications	
T840	Mechanical complication of internal joint prosthesis
T841	Mechanical complication of internal fixation device of bones of limb
T842	Mechanical complication of internal fixation device of other bones
T843	Mechanical complication of other bone devices, implants and grafts
T844	Mechanical complication of other internal orthopaedic devices, imnplants and grafts
ICD-10 codes for osteoarthritis/arthrosis	
M15-	Polyarthrosis
M17-	Gonarthrosis
M19-	Other arthrosis

OPCS-4 = Office of Populations Censuses and Surveys Classification of Interventions and Procedures version 4. ICD-10 = International Statistical Classification of Diseases and Related Health Problems, tenth revision. * Where

OPCS-4 codes Y032 (renewal of prosthesis in organ NOC) or Y037 (removal of prosthesis from organ NOC) were also used. † Where OPCS-4 codes O132 (knee NEC) or Z765 (lower end of femur NEC) or Z774 (upper end of tibia NEC) or Z787 (patella) or Z844 (patellofemoral joint) or Z845 (tibiofemoral joint) or Z846 (knee joint) or Z851 (upper tibiofibular joint) were used to identify knee as the body site.



77x73mm (300 x 300 DPI)

Supplementary material S4 – R Code

```
#Travel Times and Perioperative Outcomes in Revision Knee Replacement

setwd("/Users/alexandermatthews//OneDrive - University of Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/")


####Preparation of Data####
#load HES data

RTKA2023 <- read.csv("~/Desktop/RTKA 06-09-23 CSV.csv")

RTKA2023 <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/RTKA 06-09-23 CSV.csv")



#table only shows first 50 columns but we know there are 51 columns. Write this generic code to change preferences


rstudioapi::writeRStudioPreference("data_viewer_max_columns", 1000L)

#Some entried are blank but are read as real values and not missing data
#The table between age and sex shows three variables here
#The dataset contains non standard missing values that are not recognised as NA
#Replace empty strings with NA

RTKA2023[RTKA2023 == ""] <- NA

#Find number of incomplete cases in the data

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)

#There are 14 entries with missing data only in the age group

#check how many incomplete entries in age of patient column

sum(!complete.cases(RTKA2023$age_of_patient))

#In case of missing values there are only 14 for age of patient
#Can use imputation based on mean age
#What is the mean age of the patients
```

```
mean(RTKA2023$age_of_patient, na.rm = TRUE)

#mean age excluding missing values is 70
summary(RTKA2023$age_of_patient, na.rm = TRUE)

#Check age is normally distributed

hist(RTKA2023$age_of_patient)

#Input mean for missing values for age

RTKA2023$age_of_patient[is.na(RTKA2023$age_of_patient)] <- 69.82


#Now check number of missing values

sum(!complete.cases(RTKA2023$age_of_patient))
#Now states 0 missing values

#There are other missing values for IMD decile
##In fact there are 439 IMD score missing values

sum(!complete.cases(RTKA2023$IMD_score))


hist(RTKA2023$IMD_score)
#IMD score is non normally distributed

summary(RTKA2023$IMD_score, na.rm = TURE)

#Median IMD score is 15.543

#Use imputation to impute median for missing value

RTKA2023$IMD_score[is.na(RTKA2023$IMD_score)] <- 15.543

#Check imputation complete

sum(!complete.cases(RTKA2023$IMD_score))

#Now showing 0 missing values

#Next attach IMD decile number 6 to the missing values. As a score of 15 equates to the 6th
decile

RTKA2023$IMD_decile[is.na(RTKA2023$IMD_decile)] <- 6
```

```
#Check duplicate entry spells

duplicates <- RTKA2023[duplicated(RTKA2023),]


#No duplicates in data


#Frequencies of revisions by volume

as.numeric(RTKA2023$TV12mo)


#frequencies of revisions by trust volume
table(RTKA2023$TVcat)

#Proportions by trust volume

prop.table(table(RTKA2023$TVcat))

#Some entried are blank but are read as real values and not missing data
#The table between age and sex shows three variables here
#The dataset contains non standard missing values that are not recognised as NA
#Replace empty strings with NA

RTKA2023[RTKA2023 == ""] <- NA

#Check this has registered

missing_data <- colSums(is.na(RTKA2023))
print(missing_data)


#Column with LSOA_2011_Code has 171 missing.

#LSOA is part of primary exposure variable, small number of missing cases. Decision to
remove rows rather than estimate from imputation because factor variable and dependent
on provider code. Multiple imputation was used later to estimate missing travel data for
these multiple rows where LSOA and site code was availble

#Remove missing data in dataframe combined_data for column LSOA_2011_Code with
missing fields = 171

RTKA2023<- RTKA2023[!is.na(RTKA2023$LSOA_2011_Code), ]
```

```
#16,565 patients before link with TRACC travel data

#Load Travel times data

TRAVELTIMES <- read.csv("~/Desktop/Drive time and Miles reference file.csv")

LSOAREF <- read.csv("~/Desktop/LSOA Matrix.csv")

LSOAREF <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/LSOA Matrix.csv")


#Join data but The data is too big so we need to do this using SQL

install.packages("RSQLite")
library(RSQLite)

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "times", TRAVELTIMES)
dbWriteTable(con, "lsoa", LSOAREF)

query <- "
Select *
FROM times
JOIN lsoa ON times.LSOAName = lsoa.LSOA11NM"

result <- dbGetQuery(con, query)

#10million 457 thousand and 999 possible combinations

#Write Dataframes

write.csv(result, "~/Desktop/JOINLSOATRAVEL.csv")

result<- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/JOINLSOATRAVEL.csv")


#####Now join this data to your revisions spreadsheet using key identifiers LSOA and
Organisation site code

con <- dbConnect(RSQLite::SQLite(),
                 dbname = "mydatabase1.db")
dbWriteTable(con, "revisions3", RTKA2023)
dbWriteTable(con, "travel3", result)
```

```
query <- "
Select *
FROM revisions3
JOIN travel3 ON revisions3.LSOA_2011_Code = travel3.LSOA11CD AND revisions3.Sitecode =
travel3.ProviderSiteCode"

result_join <- dbGetQuery(con, query)

#Number of patients following join 12,774


result1 <- result_join
#Check your data for missing values

missing_data <- colSums(is.na(result1))
print(missing_data)

#Check data for duplicates

duplicates <- RTKA2023[duplicated(RTKA2023$Epikey), ]


# Check for duplicates in the 'epikey' column
duplicates <- result1[duplicated(result1$Epikey), ]

#There are 2,047 duplicates

#Remove duplicates in result 1


# Remove duplicates: Keep only the first occurrence of each 'Epikey'
result1 <- result1[!duplicated(result1$Epikey), ]

#final dataframe is 10,727


write.csv(result1, "/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/FinalJOIN.csv")


####Prepare Outcomes, Exposure variable and co-variates ####

#Set up outcomes

#Replace NA's in the Read columns with N
```

```
result1$Read30 <- ifelse(is.na(result1$Read30), 'N', result1$Read30)
result1$Read90 <- ifelse(is.na(result1$Read90), 'N', result1$Read90)

result1$Read30days <- ifelse(result1$Read30 == "Y", 1, 0)
#readmission for 90 days
result1$Read90days <- ifelse(result1$Read90 == "Y", 1, 0)



#Set up your co-variates

result1$HFRS_Band = as.factor(result1$HFRS_Band)
result1$HFRS_Band = relevel(result1$HFRS_Band, ref = 'None')

result1$POD = as.factor(result1$POD)
result1$POD = relevel(result1$POD, ref = 'EL')

table(result1$POD)


#I've joined two dataframes based on a shared field. But some rows have not jointed

#Journey times statistics - 10,457,999 rows

#12,774 following join with revisions and travel data called "result1" but had duplicates
2,047 so remove these (duplicates due to slightly different latitude and longitude for same
Site codes in journey times statistics )

#Final results 1 following removal of duplicates is 10,727

#Original dataframe is 16,736 called RTKA2023 following removal of early revisions,
excluding missing LSOA was 16565

#Missing data for travel seen in 5,838 patients or 35% of patients

#Use multiple imputation to impute missing distance values for cases without join

#How many unmatched rows?

unmatched_rows <- RTKA2023[!(RTKA2023$Epikey %in% result1$Epikey), ]

#There are 5,838 unmatched rows

#I want to create a dataframe showing both matched and unmatched fields based on this.

# Identify columns that are in result1 but not in RTKA2023
missing_cols <- setdiff(names(result1), names(RTKA2023))
```

```
# Add missing columns to RTKA2023 with NA values
for (col in missing_cols) {
  RTKA2023[[col]] <- NA
}

# Ensure column order is the same as result1
RTKA2023 <- RTKA2023[, names(result1)]

# Identify unmatched rows
unmatched_rows <- RTKA2023[!(RTKA2023$Epikey %in% result1$Epikey), ]

# Combine matched rows (result1) with unmatched rows
combined_data <- rbind(result1, unmatched_rows)

duplicates <- combined_data[duplicated(combined_data$Epikey), ]

#0 duplicates

write.csv(combined_data, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/FinalJOINCombined.csv")


combined_data <- read.csv("/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/FinalJOINCombined.csv")

#Replace NA's in the Read columns with N

combined_data$Read30 <- ifelse(is.na(combined_data$Read30), 'N',
combined_data$Read30)


combined_data$Read30days <- ifelse(combined_data$Read30 == "Y", 1, 0)



#Now have dataframe displaying both matched and unmatched rows

missing_data <- colSums(is.na(combined_data))
print(missing_data)

#How many patients in high volume centres >49

combined_data$MRC <- ifelse(combined_data$TV12mo > 49, 1, 0)
```

```
nopatients <- subset(combined_data, MRC == 1)

#6880 patients

missing_data <- colSums(is.na(nopatients))
print(missing_data)

# Count unique levels of ProvCode
n_levels <- length(unique(nopatients$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")
#38 providers

#How many sites
# Count unique levels of ProvCode
n_levels <- length(unique(nopatients$Sitecode))
cat("Number of unique sites (Sitecode):", n_levels, "\n")

#187 sites

#rates of readmission 30 days

table(nopatients$Read30days)

#568/6880 8.3%

#rates of mortality at 90 days

table(nopatients$Mort90days)

#217/6880 3.2%

#Rates of length of stay above median. Remember median calculated across entire cohort

summary(combined_data$Spell_Los) #Median of 5

nopatients$Long_Los <- ifelse(nopatients$Spell_Los > 5, 1, 0)

table(nopatients$Long_Los)

#3421/6880 49.7%


#3157 travel data not available

#16,565 observations in entire dataframe not limited to teriatry referral centres

#CV12mo missing 71 cases. Imputation using median due to positive skew
```

```
hist(combined_data$CV12mo)

#mean age excluding missing values is 70
summary(combined_data$CV12mo, na.rm = TRUE)


#Input median of 6 for missing data

combined_data$CV12mo[is.na(combined_data$CV12mo)] <- 6

#Now need to use multiple imputation method to estimate travel data for columns
"DistanceMiles", "OffPeakDriveDistanceMiles", "PeakDriveTimes' based on associated
predictors:

#Refer to this resource "https://bookdown.org/mwheymans/bookmi/multiple-
imputation.html#setting-the-imputation-methods"

#And this resource for context
https://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html

# https://www.ebpi.uzh.ch/dam/jcr:dc0cef17-29c7-4e61-8d33-
e690561ab7ae/mi_intro20191001.pdf (Advice on multi level modelling and imputation)

# Install packages if they are not already installed
install.packages(c("mice", "ggplot2", "naniar"))

# Load the packages
library(mice)
library(ggplot2)
library(naniar)

#assuming missing data is due to random chance, LSOA and SiteCode are related to the
exposure but also include all other variables linked to your analysis
#Subset dataframe called combined_date with only with relevant columns: age_of_patient,
sex, HFRS_Band IMD_Score, IMD_Decile, infection,  TVcat, CVcat, SiteCode, ProvCode, FinY,
DistanceMiles, OffPeakDriveDistanceMiles, PeakDriveTime, Mort90days, Read30, Spell_Los
#decision not to include site code and LSOA as likely not present in missing data
"LSOA_2011_Code", "Sitecode"



# Specify the relevant columns I've included TV12mo as may be related to outcome,
ProvCode for clustering,
relevant_columns <- c(
  "age_of_patient", "sex", "HFRS_Band", "IMD_score",
  "infection", "TV12mo", "CV12mo", "ProvCode", "FinY",
```

```
  "DistanceMiles", "OffPeakDriveDistanceMiles", "PeakDriveTime",
  "Mort90days", "Read30days", "Spell_Los"
)

# Subset the dataframe with only the relevant columns
subset_combined_data <- combined_data[, relevant_columns]

#Currently sex, HFRS_Band, TVCat, Sitecode, ProvCode, FinY are not incorporated in model
as character variables

#convert these to factors


# Convert variables to factors
subset_combined_data$sex <- as.factor(subset_combined_data$sex)
subset_combined_data$ProvCode <- as.factor(subset_combined_data$ProvCode)
subset_combined_data$FinY <- as.factor(subset_combined_data$FinY)
subset_combined_data$HFRS_Band <- as.factor(subset_combined_data$HFRS_Band)

subset_combined_data$Sitecode <- as.factor(subset_combined_data$Sitecode)
subset_combined_data$LSOA_2011_Code <-
as.factor(subset_combined_data$LSOA_2011_Code)



# Check the structure of the dataframe to confirm
str(subset_combined_data[, c("sex", "Sitecode", "ProvCode", "FinY", "HFRS_Band",
"LSOA_2011_Code")])


#visualise missing data

vis_miss(subset_combined_data)

#35% missing travel data

# Set the seed for reproducibility
set.seed(123)


# Perform Multiple Imputation

imp <- mice(subset_combined_data, m=5, method='pmm')

#Check for imputation values

imp$imp$OffPeakDriveDistanceMiles
```

```
#visualise imputed values

imp$imp

#Means of the imputed values

imp$chainMean

#What are the predictors

imp$predictorMatrix

#Plot imputation values against observed values.

my_plot <- stripplot(imp, col=c("grey", "blue"), pch = c(1, 20))

my_plot

#Guidelines for imputation model suggest all variables in the analysis should be included,
inclusive of dependent or outcome variables

#Ensure TVCat is not a predictor variable

pred <-imp$predictorMatrix
pred["TVcat"] <- 0
pred


#Plot the convergence (how equal is the variance to the mean)

plot(imp)

#Stack the imputed values into a single dataset and include original data

imp2 <- complete(imp, "long", inc = TRUE)

#Save imp2

write.csv(imp2, "/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/imp2.csv")

#Read it back in here:

imp2 <- read.csv("/Users/alexandermatthews//OneDrive - University of Exeter/Alex
Matthews MD/Revision Knee Networks MD/Travel Times Analysis_/imp2.csv")
```

```
#Save as Supplemenatry figure

#Filter data by tertiary hospitals only


#But current guidelines suggest >49 is a high volume centre called a major revision centre
and probably represents a unit with tertiary specialisation

imp2$MRC <- ifelse(imp2$TV12mo > 49, 1, 0)

tertiary_revisions <- subset(imp2, MRC == 1)

tertiary_revisions$Long_Los <- ifelse(tertiary_revisions$Spell_Los > 5, 1, 0)


#declare the imputed data to be mids again, the format MICE is expecting for regression
analyses
tertiary_revisions <- as.mids(tertiary_revisions)


#Now run your regression model using a multivariable model

#A priori co-variates chosen based on evidence of predictors for readmission

####Primary Outcome 30 day readmission ####

#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ DistanceMiles + IMD_score +
HFRS_Band +
                sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
              family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the long data including all imputations for VIF

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)


# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = tertiary_revisions, family = "binomial")




# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?

#Box Tidwell

#Recode back into correct format

tertiary_revisions <- as.mids(tertiary_revisions)

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}
```

```
# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ DistanceMiles + Interaction,
data = tert
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.03 evidence of non linearity

#Are spline terms significant for DistanceMiles if using 3 knots, 4 knots and 5 knots

#Use data of all imputations in long format

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)



# Load the required library
library(splines)

#AIC of non spline model

model <- glm(Read30days ~ DistanceMiles, data = tertiary_revisions, family = binomial)
summary(model)
```

```
#AIC 21862

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$DistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Read30days ~ ns(DistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)

  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
```

```
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#AIC better fit 21806
#Model with 3 knots, significant terms but greater knots do not improve the model fit. Non linear relationship is evident and should be modelled with splines




#Prepare predictors for model prediction


#you need to ensure that the predicted probabilities align with the corresponding observations
#Explore the data for missing values
sum(!complete.cases(tertiary_revisions$DistanceMiles))
#Unimputed dataset is missing, so exclude these

tertiary_revisions <- tertiary_revisions[!is.na(tertiary_revisions$DistanceMiles),]


sum(!complete.cases(tertiary_revisions$sex))

sum(!complete.cases(tertiary_revisions$Read30days))

sum(!complete.cases(tertiary_revisions$HFRS_Band))

sum(!complete.cases(tertiary_revisions$IMD_score))

sum(!complete.cases(tertiary_revisions$infection))

#Currently infection as numeric - ensure is factor

tertiary_revisions$infection <- as.factor(tertiary_revisions$infection)
tertiary_revisions$HFRS_Band <- as.factor(tertiary_revisions$HFRS_Band)
tertiary_revisions$sex <- as.factor(tertiary_revisions$sex)
tertiary_revisions$FinY <- as.factor(tertiary_revisions$FinY)
tertiary_revisions$ProvCode <- as.factor(tertiary_revisions$ProvCode)
tertiary_revisions$DistanceMiles <- as.numeric(tertiary_revisions$DistanceMiles)
tertiary_revisions$age_of_patient <- as.numeric(tertiary_revisions$age_of_patient)
tertiary_revisions$IMD_score <- as.numeric(tertiary_revisions$IMD_score)
tertiary_revisions$TV12mo <- as.numeric(tertiary_revisions$TV12mo)
```

```
tertiary_revisions$CV12mo <- as.numeric(tertiary_revisions$CV12mo)

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$DistanceMiles, probs = c(0.05, 0.50, 0.95), na.rm =
TRUE)
print(knots)
#Knots at 53, 69 and 84
spline_terms <- ns(tertiary_revisions$DistanceMiles, knots = knots)



model_with_custom_splines <- glm(Read30days ~ ns(DistanceMiles, knots = knots) +
HFRS_Band + IMD_score +
                    sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                  family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$DistanceMiles),
max(tertiary_revisions$DistanceMiles), length.out = 100)

new_data <- expand.grid(
  DistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
```

```
    TV12mo = mean(tertiary_revisions$TV12mo),
    infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))
```

```
# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(DistanceMiles = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$DistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




# Combine mean_unit_range, predicted_probs, ci_lower, and ci_upper into plot_data
plot_data <- data.frame(DistanceMiles = DistanceMiles_range,
                        predicted_prob = predicted_probs,
                        ci_lower = boot_results$ci_lower,
                        ci_upper = boot_results$ci_upper)

library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
```

```
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )


# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Travel Distance (Miles)", y = "Mean Predicted Probability for readmission at 30
days", title = "Spline curve predicted probability of readmission at 30 days by patient travel
distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )


#Spline curve does appear to show the predicted probability of emergency readmission at
30 days increases with travel distance but wide confidence intervals

#Model Distance Miles and 30 day readmission with 3 knot splines
```

```
####First Imputation and descriptive stats####

#Use first imputed data for clinical and demographic characteristic summary

#complete_data is the first imputation

# Count unique levels of ProvCode
n_levels <- length(unique(complete_data$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")


# Count unique levels of sites
n_levels <- length(unique(complete_data))
cat("Number of unique providers (ProvCode):", n_levels, "\n")

# Count unique levels of ProvCode
n_levels <- length(unique(tertiary_revisions$ProvCode))
cat("Number of unique providers (ProvCode):", n_levels, "\n")

#38 unique providers



#Number of sites

# Count unique levels of Sites but need to use original dataframe as sites not included in
imputation analysis

#Find all those attending tertirary referral centre from original data
tertiary_all <- subset(combined_data, MRC == 1)

#Find number of sites
n_levels <- length(unique(tertiary_all$Sitecode))
cat("Number of unique providers (Sites):", n_levels, "\n")

#187 sites

#Back to first imputation dataset. Calculate median number of miles straight line distance

summary(complete_data$DistanceMiles)

#Median is 7.1 IQR is 3.9 to 12.7. Range 0 to 77.1 miles.
```

```
#Driving distances

summary(complete_data$OffPeakDriveDistanceMiles)

#Median 10.4 miles, IQR is 5.8 to 18.3 miles

#Calculate median driving times

summary(complete_data$PeakDriveTime)

#Median is 27 minutes IQR is 18.4 to 38.4. Maximum 104 minutes


#Create travel time quintile variable

quintiles <- quantile(complete_data$DistanceMiles, probs = seq(0,1,0.2), na.rm=TRUE)

complete_data$distancequintile <- cut(complete_data$DistanceMiles, breaks = quintiles,
labels = c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)

#Tabulate descriptive stats

hist(tertiary_all$Spell_Los)
summary(tertiary_all$Spell_Los)

# Total number of revisions
total_revisions <- nrow(complete_data)

# Create a summary table
summary_stats <- complete_data %>%
 group_by(distancequintile) %>%
 summarise(
  # Count of observations
  Count = n(),

  # Distinct Providers
  Distinct_Units = n_distinct(ProvCode),
  Total_Distinct_Units = n_distinct(complete_data$ProvCode),
  Distinct_Units_Percent = (Distinct_Units / Total_Distinct_Units) * 100,

  #Median distance

  Distance_LowerQuartile = quantile(DistanceMiles, 0.25, na.rm = TRUE),
  Distance_Median = median(DistanceMiles, na.rm = TRUE),
  Distance_UpperQuartile = quantile(DistanceMiles, 0.75, na.rm = TRUE),
```

```
    #Mean driving time
    DrivingTime_LowerQuartile = quantile(PeakDriveTime, 0.25, na.rm = TRUE),
    DrivingTime_Median = median(PeakDriveTime, na.rm = TRUE),
    DdrivingTime_UpperQuartile = quantile(PeakDriveTime, 0.75, na.rm = TRUE),



    # Age: Mean and standard deviation
    Age_Mean = mean(age_of_patient, na.rm = TRUE),
    Age_SD = sd(age_of_patient, na.rm = TRUE),

    # Age: Mean ± SD (concatenated)
    Age_Mean_SD = paste(round(mean(age_of_patient, na.rm = TRUE), 2), "±",
round(sd(age_of_patient, na.rm = TRUE), 2)),



    # Gender: frequency and percentage
    Female_Freq = sum(sex == "Female", na.rm = TRUE),
    Female_Percent = sum(sex == "Female", na.rm = TRUE) / n() * 100,
    Male_Freq = sum(sex == "Male", na.rm = TRUE),
    Male_Percent = sum(sex == "Male", na.rm = TRUE) / n() * 100,

    # ASA: frequency and percentage for each level
    HFRS_None_Freq = sum(HFRS_Band == "None", na.rm = TRUE),
    HFRS_None_Percent = sum(HFRS_Band == "None", na.rm = TRUE) / n() * 100,
    HFRS_Mild_Freq = sum(HFRS_Band == "Mild", na.rm = TRUE),
    HFRS_Mild_Percent = sum(HFRS_Band == "Mild", na.rm = TRUE) / n() * 100,
    HFRS_Moderate_Freq = sum(HFRS_Band == "Moderate", na.rm = TRUE),
    HFRS_Moderate_Percent = sum(HFRS_Band == "Moderate", na.rm = TRUE) / n() * 100,
    HFRS_Severe_Freq = sum(HFRS_Band == "Severe", na.rm = TRUE),
    HFRS_Severe_Percent = sum(HFRS_Band == "Severe", na.rm = TRUE) / n() * 100,



    #Infection

    Infection_Freq = sum(infection == "1", na.rm = TRUE),
    Infection_Percent = sum(infection == "1", na.rm = TRUE) / n() * 100,



    # Year: frequency and percentage for each year from 2009 to 2019
    Year_2015_2016_Freq = sum(FinY == "2015/16", na.rm = TRUE),
    Year_2015_2016_Percent = sum(FinY == "2015/16", na.rm = TRUE) / n() * 100,
    Year_2016_2017_Freq = sum(FinY == "2016/17", na.rm = TRUE),
    Year_2016_2017_Percent = sum(FinY == "2016/17", na.rm = TRUE) / n() * 100,
    Year_2017_2018_Freq = sum(FinY == "2017/18", na.rm = TRUE),
    Year_2017_2018_Percent = sum(FinY == "2017/18", na.rm = TRUE) / n() * 100,
    Year_2018_2019_Freq = sum(FinY == "2018/19", na.rm = TRUE),
```

```
    Year_2018_2019_Percent = sum(FinY == "2018/19", na.rm = TRUE) / n() * 100,
    Year_2019_2020_Freq = sum(FinY== "2019/20", na.rm = TRUE),
    Year_2019_2020_Percent = sum(FinY == "2019/20", na.rm = TRUE) / n() * 100,


    # Median Surgeon Volume: lower quartile, median, and upper quartile
    Surgeon_LowerQuartile = quantile(CV12mo, 0.25, na.rm = TRUE),
    Surgeon_Median = median(CV12mo, na.rm = TRUE),
    Surgeon_UpperQuartile = quantile(CV12mo, 0.75, na.rm = TRUE),


    #Median hospital volume

    Hospital_LowerQuartile = quantile(TV12mo, 0.25, na.rm = TRUE),
    Hospital_Median = median(TV12mo, na.rm = TRUE),
    Hospital_UpperQuartile = quantile(TV12mo, 0.75, na.rm = TRUE),


    #Median IMD Score

    IMD_LowerQuartile = quantile(IMD_score, 0.25, na.rm = TRUE),
    IMD_Median = median(IMD_score, na.rm = TRUE),
    IMD_UpperQuartile = quantile(IMD_score, 0.75, na.rm = TRUE),

  )



# Print the summary table
print(summary_stats)

write.csv(summary_stats, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/Summary_stats.csv")




####Cluster Variable ####

# Compute the mean outcome for each cluster
library(dplyr)
prov_means <- tertiary_revisions %>%
  group_by(ProvCode) %>%
  summarize(mean_outcome = mean(Read30days, na.rm = TRUE))

# Plot variability
```

```
boxplot(mean_outcome ~ ProvCode, data = prov_means, xlab = "ProvCode", ylab = "Mean
Outcome")

# Summary statistics of variability
summary(prov_means$mean_outcome)

#There is evidence of variability between providers


# Fit logistic regression on imputed datasets
m3.mi <- with(tertiary_revisions, glmer(Read30days ~ DistanceMiles + IMD_score +
HFRS_Band +
                sex + age_of_patient + infection + TV12mo + CV12mo + FinY + (1 |
ProvCode),
                family = "binomial"))


print(m3.mi)

#Including ProvCode as a random effect was tested but led to convergence issues likely due
to numerical instability between providers so a decision was made to accept the fixed
effects model which may account for clustering at the provider level but is a limitation of
the study


#Was travel distance strongly correlated with IMD_score or age?



#Next do a Spearman's rank correlation between travel distance and age, and then for
travel distance and IMD score

imp2$MRC <- ifelse(imp2$TV12mo > 49, 1, 0)

tertiary_revisions <- subset(imp2, MRC == 1)


write.csv(tertiary_revisions, "/Users/alexandermatthews//OneDrive - University of
Exeter/Alex Matthews MD/Revision Knee Networks MD/Travel Times
Analysis_/tertiary_revisions.csv")


tertiary_revisions <- as.mids(tertiary_revisions)

tertiary_revisions$age_of_patient <-
as.numeric(as.character(tertiary_revisions$age_of_patient))
```

```
tertiary_revisions$DistanceMiles <-
as.numeric(as.character(tertiary_revisions$DistanceMiles))



#Age and travel distance, Cannot pool the results based on the multiple imputations as cor
test not compatible.Therefore stack all imputations together and calculate correlation


# Scatterplot with linear regression line
plot(tertiary_revisions$age_of_patient, tertiary_revisions$DistanceMiles,
    main = "Scatterplot of Age of Patient vs DistanceMiles",
    xlab = "Age of Patient", ylab = "DistanceMiles",
    pch = 19, col = "blue")

# Add a linear trendline
abline(lm(DistanceMiles ~ age_of_patient, data = tertiary_revisions), col = "red", lwd = 2)

# Calculate Spearman's rank correlation
spearman_test <- cor.test(tertiary_revisions$age_of_patient,
tertiary_revisions$DistanceMiles, method = "spearman")

# Extract rho and p-value
rho <- round(spearman_test$estimate, 2)
p_value <- spearman_test$p.value
p_value_text <- ifelse(p_value < 0.05, "<0.05", paste0("=", round(p_value, 3)))

# Add a legend with Spearman's rank correlation information
legend("topright", legend = paste("Spearman's Rank Correlation:\n",
                          "rho =", rho, ", p-value", p_value_text),
      col = c("blue", "red"), lty = c(NA, 1), pch = c(19, NA), lwd = c(NA, 2), bty = "n")



#IMD score and travel distance

# Scatterplot with trendline
plot(tertiary_revisions$IMD_score, tertiary_revisions$DistanceMiles,
    main = "Scatterplot of IMD_score vs DistanceMiles",
    xlab = "IMD_score", ylab = "DistanceMiles",
    pch = 19, col = "blue")

# Add a linear trendline (for visualizing the general trend)
abline(lm(DistanceMiles ~ IMD_score, data = tertiary_revisions), col = "red", lwd = 2)

# Calculate Spearman's rank correlation
spearman_test <- cor.test(tertiary_revisions$IMD_score, tertiary_revisions$DistanceMiles,
method = "spearman")
```

```
# Extract rho and p-value
rho <- round(spearman_test$estimate, 2)
p_value <- spearman_test$p.value
p_value_text <- ifelse(p_value < 0.05, "<0.05", paste0("=", round(p_value, 3)))

# Add a legend with Spearman's rank correlation information
legend("topright", legend = paste("Spearman's Rank Correlation:\n",
                                  "rho =", rho, ", p-value", p_value_text),
       col = c("blue", "red"), lty = c(NA, 1), pch = c(19, NA), lwd = c(NA, 2), bty = "n")

#Exposure 2 - OffPeakDriveDistanceMiles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ OffPeakDriveDistanceMiles +
IMD_score + HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)
```

```
# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?


# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ OffPeakDriveDistanceMiles +
Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.05. There is no evidence of non linearity

#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Read30days ~ PeakDriveTime + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)
```

```
# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ PeakDriveTime + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there a non linear relationship?



# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Read30days ~ PeakDriveTime + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

# p value = 0.13 not evidence of non linearity




####Secondary Outcome mortality 90 days ####


#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ DistanceMiles + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?
```

```
library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Mort90days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity


#Is there evidence of non linearity?

library(mice)

tertiary_revisions <- as.mids(tertiary_revisions)


# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Mort90days ~ DistanceMiles + Interaction,
                                               family = binomial(link = "logit")))
```

```
# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


# P value 0.95

#Exposure 2 - OffPeakDriveDistanceMiles



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ OffPeakDriveDistanceMiles +
IMD_score + HFRS_Band +
                                      sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                    family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)
```

```
#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ OffPeakDriveDistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

tertiary_revisions <- as.mids(tertiary_revisions)

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
```

```
model <- with(tertiary_revisions_modified, glm(Mort90days ~ OffPeakDriveDistanceMiles +
Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#0.989



#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Mort90days ~ PeakDriveTime + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
```

```
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Mort90days ~ PeakDriveTime + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
```

```
model <- with(tertiary_revisions_modified, glm(Mort90days ~ PeakDriveTime + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


# P avlue 0.78

####Secondary outcome prolonged LOS ####

tertiary_revisions <- complete(tertiary_revisions, "long", inc = TRUE)

tertiary_revisions$Long_Los <- ifelse(tertiary_revisions$Spell_Los > 5, 1, 0)

tertiary_revisions <- as.mids(tertiary_revisions)

#Exposure 1 - Distance Miles

library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ DistanceMiles + IMD_score + HFRS_Band +
                                      sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                    family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)
```

```
# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Long_Los ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_DistanceMiles <- log(data$DistanceMiles)  # Add log-transformed variable
  data$Interaction <- data$DistanceMiles * data$Log_DistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                                       lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                              add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)
```

```
# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ DistanceMiles + Interaction,
                                               family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)


#P value 0.002 Non linear

# Load the required library
library(splines)

#AIC of non spline model

model <- glm(Long_Los ~ DistanceMiles, data = tertiary_revisions, family = binomial)
summary(model)

#AIC 52853

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$DistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(DistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)

  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
```

```
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52769, model with four knots best fit and improved fit from original linear model

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$DistanceMiles, probs = c(0.05, 0.35, 0.65, 0.95), na.rm =
TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$DistanceMiles, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(DistanceMiles, knots = knots) +
HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$DistanceMiles),
max(tertiary_revisions$DistanceMiles), length.out = 100)

new_data <- expand.grid(
  DistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
                        TV12mo = mean(tertiary_revisions$TV12mo),
                        infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
```

```
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(DistanceMiles = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals
```

```
# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$DistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )
```

```
# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Travel Distance (Miles)", y = "Mean Predicted Probability for Prolonged LOS", title
= "Spline curve predicted probability of prolonged LOS by patient travel distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),  # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )



#Exposure 2 - OffPeakDriveDistanceMiles



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ OffPeakDriveDistanceMiles + IMD_score +
HFRS_Band +
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))


print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)
```

```
# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#No evidence of multi-collinearity

#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_OffPeakDriveDistanceMiles <- log(data$OffPeakDriveDistanceMiles)  # Add log-
transformed variable
  data$Interaction <- data$OffPeakDriveDistanceMiles *
data$Log_OffPeakDriveDistanceMiles  # Add interaction term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)

# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
```

```
                    lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                           add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ OffPeakDriveDistanceMiles +
Interaction,
                           family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#0.003

#AIC of non spline model

model <- glm(Long_Los ~ OffPeakDriveDistanceMiles, data = tertiary_revisions, family =
binomial)
summary(model)

#AIC 52853

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$OffPeakDriveDistanceMiles, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(OffPeakDriveDistanceMiles, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)
```

```
  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52718, model with four knots best fit and significant spline terms

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$OffPeakDriveDistanceMiles, probs = c(0.05, 0.35, 0.65,
0.95), na.rm = TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$OffPeakDriveDistanceMiles, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(OffPeakDriveDistanceMiles, knots =
knots) + HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$OffPeakDriveDistanceMiles),
max(tertiary_revisions$OffPeakDriveDistanceMiles), length.out = 100)

new_data <- expand.grid(
  OffPeakDriveDistanceMiles = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)

# Create a new dataset with a range of distances and miles and all other predictor variables
new_data <- expand.grid(DistanceMiles = DistanceMiles_range,
                        sex = unique(tertiary_revisions$sex),
                        age_of_patient = mean(tertiary_revisions$age_of_patient),
                        HFRS_Band = unique(tertiary_revisions$HFRS_Band),
                        IMD_score = mean(tertiary_revisions$IMD_score),
                        FinY = unique(tertiary_revisions$FinY),
                        CV12mo = mean(tertiary_revisions$CV12mo),
                        TV12mo = mean(tertiary_revisions$TV12mo),
                        infection = unique(tertiary_revisions$infection))


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
```

```
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
```

```
plot_data <- data.frame(OffPeakDriveDistanceMiles = DistanceMiles_range, predicted_prob
= predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit

# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DistanceMiles = new_data$OffPeakDriveDistanceMiles,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DistanceMiles)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DistanceMiles) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
```

```
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )


# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DistanceMiles, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DistanceMiles, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Off Peak Drive Distance Miles", y = "Mean Predicted Probability for Prolonged
LOS", title = "Spline curve predicted probability of prolonged LOS by patient driving
distance") +
  scale_x_continuous(limits = c(0, max(mean_data$DistanceMiles, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )




#Exposure 3 - PeakDriveTime



library("lme4")

# Fit logistic regression on imputed datasets include ProvCode in fixed effects to account for
clustering
m3.mi <- with(tertiary_revisions, glm(Long_Los ~ PeakDriveTime + IMD_score + HFRS_Band
+
                                        sex + age_of_patient + infection + TV12mo + CV12mo + FinY +
ProvCode,
                                      family = "binomial"))
```

```
print(m3.mi)


# Pool results across imputed datasets
pooled_results <- pool(m3.mi)

# Summarize pooled results with confidence intervals
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Add Odds Ratios to the summary
summary_pooled$OR <- exp(summary_pooled$estimate)
summary_pooled$Lower_CI <- exp(summary_pooled$`2.5 %`)
summary_pooled$Upper_CI <- exp(summary_pooled$`97.5 %`)

# Display the final table with Odds Ratios and Confidence Intervals
print(summary_pooled)

#check for evidence of multicollinearity?

library(car)

# Use the first imputed dataset for the VIF calculation
complete_data <- complete(tertiary_revisions, 1)

# Fit a logistic regression model on the complete dataset
vif_model <- glm(Read30days ~ DistanceMiles + IMD_score + HFRS_Band +
                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                 data = complete_data, family = "binomial")

# Calculate VIF
vif_values <- vif(vif_model)
print(vif_values)


#Is there evidence of non linearity?

# Custom function to add log-transformed variable and interaction term
add_interaction <- function(data) {
  data$Log_PeakDriveTime <- log(data$PeakDriveTime)  # Add log-transformed variable
  data$Interaction <- data$PeakDriveTime * data$Log_PeakDriveTime  # Add interaction
term
  return(data)
}

# Extract the long-format data including the original data
tertiary_revisions_modified <- complete(tertiary_revisions, action = "long", include = TRUE)
```

```
# Apply the transformation to each imputed dataset
tertiary_revisions_modified <- do.call("rbind",
                               lapply(split(tertiary_revisions_modified,
tertiary_revisions_modified$.imp),
                                      add_interaction))

# Convert back to mids object
tertiary_revisions_modified <- as.mids(tertiary_revisions_modified)

# Fit the logistic regression model with the interaction term
model <- with(tertiary_revisions_modified, glm(Long_Los ~ PeakDriveTime + Interaction,
                                  family = binomial(link = "logit")))

# Pool the results
pooled_results <- pool(model)

# Summarize pooled results
summary_pooled <- summary(pooled_results, conf.int = TRUE)

# Extract the p-value for the interaction term
box_tidwell_p <- summary_pooled[summary_pooled$term == "Interaction", "p.value"]

# Print the p-value
print(box_tidwell_p)

#P value 0.000916

#AIC of non spline model

model <- glm(Long_Los ~ PeakDriveTime, data = tertiary_revisions, family = binomial)
summary(model)

#AIC 52843

# Define a function to fit and evaluate spline models with knots based on centiles
evaluate_centile_splines <- function(centiles, data) {
  # Calculate knots based on the specified centiles
  knots <- quantile(data$PeakDriveTime, probs = centiles, na.rm = TRUE)

  # Fit a logistic regression model with natural splines using the calculated knots
  model_spline <- glm(Long_Los ~ ns(PeakDriveTime, knots = knots),
                      family = binomial(link = "logit"),
                      data = data)

  # Summarize the model
  summary_model <- summary(model_spline)
```

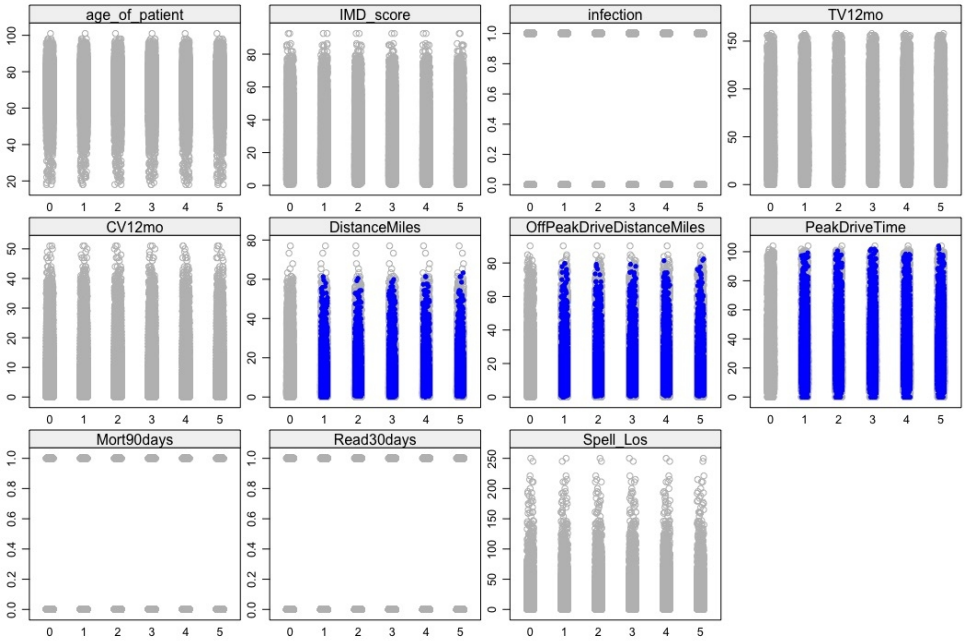
```
  # Extract p-values for the spline terms
  p_values <- summary_model$coefficients[-1, "Pr(>|z|)"]  # Exclude the intercept

  # Print the results
  cat("\nResults for centiles", centiles, ":\n")
  print(p_values)

  # Return the model and calculated knots for further inspection if needed
  return(list(model = model_spline, p_values = p_values, knots = knots))
}

# Example centile configurations for 3, 4, and 5 knots
centiles_3_knots <- c(0.05, 0.50, 0.95)  # 5th, 50th, and 95th percentiles
centiles_4_knots <- c(0.05, 0.35, 0.65, 0.95)  # Custom centiles for 4 knots
centiles_5_knots <- c(0.05, 0.25, 0.50, 0.75, 0.95)  # 5 knots centiles

# Evaluate models with centile-based knots using your dataset
results_3_knots <- evaluate_centile_splines(centiles = centiles_3_knots, data =
tertiary_revisions)
results_4_knots <- evaluate_centile_splines(centiles = centiles_4_knots, data =
tertiary_revisions)
results_5_knots <- evaluate_centile_splines(centiles = centiles_5_knots, data =
tertiary_revisions)

# Compare models with centile-based knots
cat("\nComparing models with different centile-based knots:\n")
anova(results_3_knots$model, results_4_knots$model, results_5_knots$model, test =
"Chisq")

# Print the calculated knot locations for each model
cat("\nKnot locations for 3 knots:\n")
print(results_3_knots$knots)
cat("\nKnot locations for 4 knots:\n")
print(results_4_knots$knots)
cat("\nKnot locations for 5 knots:\n")
print(results_5_knots$knots)

#52715, model with four knots best fit and significant spline terms and most parsimonious

#Run spline model with adjusted data excluding missing data
library(splines)
# For example, let's say you want 3 knots at specific percentiles
knots <- quantile(tertiary_revisions$PeakDriveTime, probs = c(0.05, 0.35, 0.65, 0.95), na.rm
= TRUE)
print(knots)

spline_terms <- ns(tertiary_revisions$PeakDriveTime, knots = knots)
```

```
model_with_custom_splines <- glm(Long_Los ~ ns(PeakDriveTime, knots = knots) +
HFRS_Band + IMD_score +
                                   sex + age_of_patient + infection + TV12mo + CV12mo + FinY + ProvCode,
                                 family = "binomial", data = tertiary_revisions)


summary(model_with_custom_splines)

#Generate a sequence of mean unit values for predicting

DistanceMiles_range <- seq(min(tertiary_revisions$PeakDriveTime),
max(tertiary_revisions$PeakDriveTime), length.out = 100)

new_data <- expand.grid(
  PeakDriveTime = DistanceMiles_range,
  sex = levels(tertiary_revisions$sex),  # Ensure it takes all factor levels
  age_of_patient = mean(tertiary_revisions$age_of_patient, na.rm = TRUE),
  HFRS_Band = levels(tertiary_revisions$HFRS_Band),  # Ensuring correct factor levels
  IMD_score = mean(tertiary_revisions$IMD_score, na.rm = TRUE),
  FinY = levels(tertiary_revisions$FinY),  # Ensuring correct factor levels
  CV12mo = mean(tertiary_revisions$CV12mo, na.rm = TRUE),
  TV12mo = mean(tertiary_revisions$TV12mo, na.rm = TRUE),
  ProvCode = levels(tertiary_revisions$ProvCode),  # Ensuring correct factor levels
  infection = levels(tertiary_revisions$infection)  # Ensuring correct factor levels
)


# Align the levels of ProvCode in new_data to match the training data
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Align the levels of all relevant categorical variables
new_data$HFRS_Band <- factor(new_data$HFRS_Band, levels =
levels(tertiary_revisions$HFRS_Band))
new_data$sex <- factor(new_data$sex, levels = levels(tertiary_revisions$sex))
new_data$FinY <- factor(new_data$FinY, levels = levels(tertiary_revisions$FinY))
new_data$infection <- factor(new_data$infection, levels =
levels(tertiary_revisions$infection))

#Factors are consistent with model

levels(new_data$HFRS_Band)
levels(tertiary_revisions$HFRS_Band)

levels(new_data$sex)
```

```
levels(tertiary_revisions$sex)

levels(new_data$FinY)
levels(tertiary_revisions$FinY)

levels(new_data$ProvCode)
levels(tertiary_revisions$ProvCode)

levels(new_data$infection)
levels(tertiary_revisions$infection)

# Check levels of ProvCode in both datasets
setdiff(levels(new_data$ProvCode), levels(tertiary_revisions$ProvCode))  # Levels in
new_data but not in tertiary_revisions
setdiff(levels(tertiary_revisions$ProvCode), levels(new_data$ProvCode))  # Levels in
tertiary_revisions but not in new_data

new_data$ProvCode <- droplevels(new_data$ProvCode)
# Check for missing values in factor variables
sum(is.na(new_data$ProvCode))  # Number of missing values in ProvCode

# Ensure that ProvCode is a factor
new_data$ProvCode <- factor(new_data$ProvCode, levels =
levels(tertiary_revisions$ProvCode))

# Now try the prediction again
predicted_probs <- predict(model_with_custom_splines, newdata = new_data, type =
"response")




# Combine mean_unit_range and predicted_probs into a data frame
plot_data <- data.frame(PeakDriveTime = DistanceMiles_range, predicted_prob =
predicted_probs)

#Calculate 95% confidence intervals

# Obtain predicted values and standard errors for the new data
predictions <- predict(model_with_custom_splines, newdata = new_data, type = "link",
se.fit = TRUE)

# Calculate the confidence intervals for the log-odds scale (link scale)
# Use a 95% confidence level (z-value = 1.96 for a 95% CI)
z_value <- 1.96
log_odds_lower <- predictions$fit - z_value * predictions$se.fit
log_odds_upper <- predictions$fit + z_value * predictions$se.fit
```

```
# Convert the log-odds confidence intervals to probabilities
# First, apply the inverse link function (logistic function) to the log-odds
lower_prob <- plogis(log_odds_lower)
upper_prob <- plogis(log_odds_upper)

# Combine the predicted probabilities and their confidence intervals into a data frame
plot_data <- data.frame(
  DriveTime = new_data$PeakDriveTime,
  predicted_prob = plogis(predictions$fit),  # Logistic transformation of the link
  ci_lower = lower_prob,
  ci_upper = upper_prob
)




library(ggplot2)
# Plot the spline curve with confidence intervals
ggplot(plot_data, aes(x = DriveTime)) +
  geom_line(aes(y = predicted_prob), color = "blue", size = 1) +
  geom_ribbon(aes(ymin = ci_lower, ymax = ci_upper), fill = "blue", alpha = 0.2) +
  labs(x = "Distance (Miles)", y = "Predicted Probability Readmission at 30 days") +
  theme_minimal()

library(dplyr)

# Group by mean_unit and calculate mean predicted_prob and corresponding confidence
intervals
mean_data <- plot_data %>%
  group_by(DriveTime) %>%
  summarise(
    mean_predicted_prob = mean(predicted_prob, na.rm = TRUE),
    mean_ci_lower = mean(ci_lower, na.rm = TRUE),
    mean_ci_upper = mean(ci_upper, na.rm = TRUE)
  )



# Define specific breaks (e.g., 25, 50, 75, ..., up to the maximum)
breaks_seq <- seq(0, max(mean_data$DriveTime, na.rm = TRUE), by = 5)

library(ggplot2)
# Plot with specified increments on x-axis
ggplot(mean_data, aes(x = DriveTime, y = mean_predicted_prob)) +
  geom_point() +  # Add points for mean_predicted_prob
  geom_line() +   # Connect points with a line
```

```
  geom_ribbon(aes(ymin = mean_ci_lower, ymax = mean_ci_upper), fill = "blue", alpha =
0.2) +  # Add ribbon for confidence intervals
  labs(x = "Peak Drive Times (Minutes)", y = "Mean Predicted Probability for Prolonged LOS",
title = "Spline curve predicted probability of prolonged LOS by patient driving times") +
  scale_x_continuous(limits = c(0, max(mean_data$DriveTime, na.rm = TRUE)), breaks =
breaks_seq) +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 14),  # Increase x-axis title font size
    axis.title.y = element_text(size = 14),  # Increase y-axis title font size
    axis.text.x = element_text(size = 12),   # Increase x-axis tick label font size
    axis.text.y = element_text(size = 12),   # Increase y-axis tick label font size
    plot.title = element_text(size = 16, hjust = 0.5)  # Increase plot title font size and center it
  )









####END####
```



91x63mm (300 x 300 DPI)