

Improving Hip Fracture Outcomes Using Routinely Collected Health Data



David Metcalfe
Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2019

Improving Hip Fracture Outcomes Using Routinely Collected Health Data

David Metcalfe

DPhil in Musculoskeletal Science
Balliol College

Hip fracture is associated with high morbidity and mortality but outcomes can be improved through dedicated clinical pathways that deliver evidence-based multi-disciplinary care. The National Hip Fracture Database (NHFD) is a national clinical audit that exists to improve hip fracture care by providing hospital-level performance data through online dashboards and an annual report. The aim of this thesis was to explore how national clinical audit can improve hip fracture care and outcomes.

In Chapter 2, a systematic review found little existing evidence to show that releasing performance data into the public domain can change healthcare decisions, clinical performance, or patient outcomes. Importantly, the evidence could not *exclude* an effect and it seems likely that data will continue to be released to the public in efforts to improve performance. Chapter 3 shows that specific identification of performance outliers is relatively insensitive to the model used for risk-adjustment and may not be enhanced by routine linkage of the NHFD to other datasets.

In Chapter 4, an interrupted time series and difference-in-differences study did not find any evidence that introduction of the NHFD alone improved performance measures or patient outcomes. However, integration of the NHFD with the Hip Fracture Best Practice Tariff (BPT) - which pays hospitals a supplement for meeting evidence-based care standards - *was* associated with marked and sustained improvements across both performance measures and patient outcomes. This suggests that an integrated system for rewarding best practice can improve outcomes beyond that of a voluntary audit of national clinical standards.

The BPT rewards hospitals for meeting seven care standards. In Chapter 5, an observational study using NHFD data found substantial deviation from national guidance around provision of total hip arthroplasty (THA), which is not yet a BPT standard. Under-provision of THA appears to disproportionately affect patients living in deprived areas and those presenting to hospital at weekends.

One possible reason for under-provision of THA is that the evidence is uncertain. There is concern that randomised trials may overstate the advantages of THA when used in routine clinical practice. However, Chapter 6 combined meta-analysis of trial data with a propensity score matched cohort study using the NHFD and did not find evidence that outcomes were worse outside clinical trials. Importantly, both studies reported an association between THA and reduced 12-month mortality, which clearly requires urgent further investigation.

This thesis shows that the NHFD is a valuable resource both for auditing standards of care and embedding studies aimed at learning from the outcomes of patients with hip fractures. If the findings in Chapter 6 are supported by further work, the provision of THA to eligible patients should be considered as a future quality standard that must be met for payment of the BPT.

Dedication

To my wife, Mina, who created two awesome little boys in the time it took to produce this thesis.

Declaration

The work presented in this thesis was undertaken by David Metcalfe and has not previously been submitted for the award of a degree by any university. All contributions by others have been acknowledged. This work was performed under the supervision of Professors Matthew Costa (Oxford, UK), Daniel Perry (Oxford, UK), Andrew Judge (Bristol, UK) and Belinda Gabbe (Melbourne, Australia), and completed within the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS) at the University of Oxford¹.

¹49,100 words excluding bibliography and appendices

Acknowledgements

This DPhil was funded by an Oxford-UCB Prize Fellowship in Biomedical Research. I am grateful to UCB (Brussels, Belgium) for supporting this work and their commitment to improving outcomes for patients with fragility fractures.

My doctoral work has been supported by an academic “dream team” of supervisors: Matthew Costa (University of Oxford), Daniel Perry (University of Oxford), Andrew Judge (University of Bristol), and Belinda Gabbe (Monash University, Melbourne, Australia). They each supervised different aspects of this DPhil and helped develop my research towards independence. I am also grateful to Daniel Prieto-Alhambra and Dominic Furniss (both University of Oxford) whose independent perspectives on my early DPhil plans shifted the emphasis from case mix adjustment to the (arguably much more interesting) work in Chapters 4-6.

The work in Chapter 2 was undertaken under the auspices of the Cochrane Collaboration. I am therefore grateful to my co-authors (Arturo Rios Diaz at Thomas Jefferson University Hospital, Philadelphia, USA; Olubode Olufajo at Howard University College of Medicine, Washington DC, USA; Sofia Massa at Oxford; Nicole Ketelaar at the Saxion University of Applied Sciences, The Netherlands; Signe Flottorp at the Norwegian Institute of Public Health, Oslo, Norway; and Daniel Perry at Oxford) as well as to Sasha Shepperd (UK Co-ordinating Editor), Gillian Leng and Luciana Ballini (Editors), Julia Worswick (Managing Editor), and Paul Miller (Information Specialist), all of whom work at the Cochrane Effective Practice and Organisation of Care (EPOC) group. Cheryl Zogg (Yale University, Connecticut, USA) guided me through the analysis of interrupted time series data in support

of Chapter 4 and provided statistical advice for analyses undertaken in Chapter 6. Mr Rik Smith (NHS National Services Scotland) aggregated patient-level data from Scottish Morbidity Record (SMR01), which greatly expedited the release of data necessary to complete Chapter 4. Dr May Ee Png (University of Oxford) helped screen randomized trials (RTs) written in Chinese for possible inclusion in Chapter 6.

The work in Chapters 2, 4, 5, and 6 has already been published. I am therefore grateful to John Wiley & Sons Inc, BMJ Publishing Group Ltd, BioMed Central, and the British Editorial Society of Bone & Joint Surgery Ltd for permission to reproduce text and figures in this thesis.

Access to the data used in Chapter 6 was funded by a Royal College of Surgeons of Edinburgh Small Grant Award in addition to the Oxford-UCB Prize Fellowship.

Finally, I am grateful to the many people (particularly those at the Falls and Fragility Fracture Audit Programme (FFFAP), Healthcare Quality Improvement Partnership (HQIP), NHS Digital, Information Services Division (ISD) Scotland, and the Office for National Statistics (ONS)) who helped facilitate access to the data that made this thesis possible.

Abbreviations

AAOS American Academy of Orthopaedic Surgeons.

ACEi angiotensin-converting enzyme inhibitor.

ACS American College of Surgeons.

AHRQ Agency for Healthcare Research and Quality.

AMI acute myocardial infarction.

AMTS Abbreviated Mental Test Score.

aOR adjusted Odds Ratio.

ARB angiotensin receptor blocker.

ARIMA AutoRegressive Integrated Moving Average.

ASA American Society of Anesthesiologists.

AUROC area under the receiver operating characteristic.

BGS British Geriatric Society.

BMD bone mineral density.

BOA British Orthopaedic Association.

BPT Best Practice Tariff.

CAG Confidentiality Advisory Group.

CBA controlled before-after.

CCG Clinical Commissioning Group.

CCI Charlson co-morbidity index.

CCU coronary care unit.

CDC Centers for Disease Control and Prevention.

CEU Clinical Effectiveness Unit.

CG124 Clinical Guideline 124.

CHF congestive heart failure.

CI confidence interval.

CLABSI central line-associated blood stream infection.

CMS Centers for Medicare & Medicaid Services.

cNRT cluster-non-randomized trial.

COPD chronic obstructive pulmonary disease.

CORNET Collaborative Orthopaedic Research Network.

CPRD Clinical Practice Research Datalink.

cRT cluster-randomized trial.

DARE Database of Abstracts of Reviews of Effects.

DARG Data Access Request Group.

DARS Data Access Request Service.

DID difference-in-differences.

DXA dual-energy x-ray absorptiometry.

ECI Elixhauser co-morbidity index.

ED emergency department.

eFI Electronic Frailty Index.

EPOC Effective Practice and Organisation of Care.

EQ-5D EuroQol Group 5D.

EU European Union.

FFAP Falls and Fragility Fracture Audit Programme.

FLS Fracture Liaison Service.

FRAX Fracture Risk Assessment Tool.

GAfREC Governance Arrangements for Research Ethics Committees.

GDPR General Data Protection Regulation (EU 2016/6709).

GLM generalized linear model.

GP General Practitioner.

GRADE Grading of Recommendations Assessment, Development and Evaluation.

HA hemiarthroplasty.

HCUP Healthcare Cost and Utilization Project.

HEALTH Hip fracture Evaluation with ALternatives of Total hip arthroplasty
versus Hemiarthroplasty.

HES Hospital Episode Statistics.

HES APC Hospital Episode Statistics Admitted Patient Care.

HFRS Hospital Frailty Risk Score.

HIRA Health Insurance Review Assessment Service.

HIV human immunodeficiency virus.

HQIP Healthcare Quality Improvement Partnership.

HR hazard ratio.

HRG Healthcare Resource Group.

HSCIC Health and Social Care Information Centre.

IAT implicit association test.

ICD-10 International Classification of Diseases, 10th Revision.

ICD-9-CM International Statistical Classification of Diseases 9th Revision Clinical Modification.

ICTRP International Clinical Trials Registry Platform.

ICU intensive care unit.

IGARD Independent Group Advising on the Release of Data.

IMD index of multiple deprivation.

IQR interquartile range.

ISAC Independent Scientific Advisory Committee.

ISD Information Services Division.

ITS interrupted time series.

ITSA interrupted time series analysis.

LOS length of stay.

LOWESS locally weighted scatterplot smoothing.

LSOA lower layer super output area.

LVD left ventricular dysfunction.

LVF left ventricular dysfunction.

MAR missing at random.

MCAR missing completely at random.

MRP Micro Panel Release Panel.

NCEPOD National Confidential Enquiry into Patient Outcome and Death.

NELA National Emergency Laparotomy Audit.

NHFD National Hip Fracture Database.

NHFS Nottingham Hip Fracture Score.

NHS National Health Service.

NHSN National Healthcare Safety Network.

NICE National Institute for Health and Care Excellence.

NIH National Institutes for Health.

NJR National Joint Registry.

NRT non-randomized trial.

NSQIP National Surgical Quality Improvement Program.

NSTEMI non-ST-elevation myocardial infarction.

ONS Office for National Statistics.

OPCS Office of Population Censuses and Surveys.

OR odds ratio.

PbR Payment-by-Results.

PDI12 paediatric safety indicator.

PEDW Patient Episode Database for Wales.

PHIS Paediatric Health Information System.

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

PROGRESS PROGnosis REsearch Strategy.

ROC receiver operating characteristic.

RP recursive partitioning.

RR risk ratio.

RT randomized trial.

SD standard deviation.

SF36 Short-Form 36.

SHR standardized hazard ratio.

SIGN Scottish Intercollegiate Guidelines Network.

SMD standardized mean difference.

SMR01 Scottish Morbidity Record.

STEMI ST-elevation myocardial infarction.

TARN Trauma Audit & Research Network.

THA total hip arthroplasty.

THIN The Health Improvement Network.

TUG timed up and go.

UK United Kingdom.

US United States.

USA United States of America.

VA Veterans Affairs.

VBP value-based purchasing.

VTE venous thromboembolism.

WHiTE World Hip Trauma Evaluation.

WHO World Health Organization.

WOMAC Western Ontario and McMaster Universities Osteoarthritis Index.

ZonMw The Netherlands Organisation for Health Research and Development.

Contents

1	Introduction	19
1.1	Epidemiology of fragility fractures	19
1.2	Risk factors for hip fracture	20
1.2.1	Osteoporosis	20
1.2.2	Falls	21
1.3	Anatomy of the hip	21
1.4	Hip fracture classification and management	22
1.5	Components of high-quality hip fracture care	27
1.5.1	Acute and multi-disciplinary rehabilitation care	28
1.5.2	Secondary prevention of fragility fractures	32
1.5.3	Data collection and clinical audit	32
1.6	Improving healthcare quality through audit	32
1.6.1	Why is audit necessary?	34
1.6.2	Can audit improve performance?	34
1.6.3	Components of a successful audit programme	35
1.7	National clinical audit in England	37
1.7.1	The National Hip Fracture Database	38
1.8	Thesis aim and structure	38
1.8.1	Core chapters	39
2	Impact of public release of performance data on the quality of care and patient outcomes	42

2.1	Introduction	42
2.2	Methods	44
2.2.1	Study inclusion criteria	44
2.3	Results	51
2.3.1	Results of the search	54
2.3.2	Included studies	54
2.3.3	Excluded studies	57
2.3.4	Risk of bias in included studies	57
2.3.5	Primary outcomes	61
2.3.6	Secondary outcomes	65
2.4	Discussion	66
2.4.1	Summary of main results	66
2.4.2	Overall completeness and applicability of evidence	67
2.4.3	Certainty of the evidence	67
2.4.4	Potential biases in the review process	68
2.4.5	Comparison with other studies or reviews	69
2.4.6	Conclusions	69
3	Optimising case mix adjustment in the NHFD	72
3.1	Introduction	72
3.1.1	Risk adjustment in the NHFD	73
3.1.2	ASA Physical Status Classification System	73
3.1.3	Alternative co-morbidity measures	74
3.1.4	Aims	76
3.2	Methods	76
3.2.1	Data source	76
3.2.2	Inclusion criteria	77
3.2.3	Variables and outcomes	77
3.2.4	Statistical analysis	78
3.3	Results	80

3.3.1	Distribution and missingness across key variables	80
3.3.2	Current NHFD risk adjustment model	81
3.3.3	Alternative co-morbidity summary measures	86
3.3.4	Incorporating AMTS	87
3.3.5	Identification of high mortality outliers	88
3.4	Discussion	93
3.4.1	ASA substituted for alternative co-morbidity measures	93
3.4.2	ASA supplemented with additional co-morbidity measures . .	96
3.4.3	Incorporating AMTS	97
3.4.4	Identification of high mortality outliers	98
3.4.5	Limitations	98
3.4.6	Conclusions	100
4	Pay-for-performance and hip fracture outcomes	101
4.1	Introduction	101
4.1.1	Pay-for-performance in the NHS	102
4.1.2	The Hip Fracture Best Practice Tariff	103
4.1.3	Objectives	104
4.2	Methods	104
4.2.1	Study design	104
4.2.2	Data sources	104
4.2.3	Setting and population	105
4.2.4	Intervention	106
4.2.5	Outcomes	107
4.2.6	Statistical analysis	107
4.3	Results	109
4.3.1	30-day mortality	113
4.3.2	60-day mortality	113
4.3.3	90-day mortality	113
4.3.4	365-day mortality	114

4.3.5	Re-admissions	115
4.3.6	Time to operation	116
4.3.7	Length of stay	116
4.4	Discussion	117
4.4.1	Strengths and limitations	121
4.4.2	Conclusions	123
5	Inequalities in the use of total hip arthroplasty for hip fracture	124
5.1	Introduction	124
5.1.1	Arthroplasty for displaced intracapsular hip fractures	124
5.2	Methods	126
5.2.1	Data source	126
5.2.2	Inclusion criteria	126
5.2.3	Variables and outcomes	127
5.2.4	Statistical analysis	128
5.3	Results	129
5.3.1	Explaining the variation	129
5.3.2	Predictors of receiving THA if NICE-eligible	134
5.3.3	Predictors of receiving THA if NICE-ineligible	134
5.4	Discussion	136
5.4.1	Barriers to increased total hip arthroplasty (THA) provision	136
5.4.2	Strengths and limitations of the study	138
5.4.3	Implications of the study findings	139
5.4.4	Impact of the study	146
5.4.5	Conclusion	147
6	Total hip arthroplasty versus hemiarthroplasty for intracapsular hip fractures	148
6.1	Introduction	148
6.1.1	HA versus THA	149

6.2	Methods	150
6.2.1	Systematic review and meta-analysis	150
6.2.2	Observational cohort study	151
6.3	Results	156
6.3.1	Meta-analysis of randomized trials	156
6.3.2	Observational cohort study	157
6.3.3	Primary outcomes	165
6.3.4	Secondary outcomes	167
6.4	Discussion	170
6.4.1	Primary outcomes	170
6.4.2	Secondary outcomes	173
6.4.3	Use of propensity score matching	174
6.4.4	Conclusion	175
7	Conclusion	177
7.1	Review of core chapters	177
7.1.1	Impact of public release of performance data on quality of care and patient outcomes	177
7.1.2	Risk adjustment in the National Hip Fracture Database	178
7.1.3	Pay-for-performance and hip fracture outcomes	178
7.1.4	Inequalities in the use of THA for hip fracture	179
7.1.5	THA versus HA for intracapsular hip fractures	180
7.2	Themes and recommendations	181
7.2.1	Public release of performance data	181
7.2.2	Risk adjustment in the NHFD	181
7.2.3	Pay-for-performance and hip fracture outcomes	182
7.2.4	Total hip arthroplasty for hip fracture	182
7.3	Access to data	184
7.4	Future work	186

A Chapter 2 - Characteristics of included studies	230
B Chapter 2 - Risk of bias assessments	240
C Chapter 2 - Certainty of the evidence	250
D Chapter 3 - Supplementary plots	252
E Chapter 4 - Best Practice Tariff criteria	260
F Chapter 5 - Mobility score reconciliation	262
G Chapter 6 - Diagnostic codes	266
H Chapter 6 - Systematic review	270
I Chapter 6 - Propensity score matching	274
J Conclusion - Data access obstacles	279
J.1 Inability to link primary care data	279
J.2 Delays linking English secondary care data	280
J.3 Access to data from Scotland	282
K Published works	283

Chapter 1

Introduction

1.1 Epidemiology of fragility fractures

Osteoporosis is a generalized skeletal disorder characterized by low bone mass and micro-architectural deterioration, which leads to bone fragility and increased risk of fracture^{1,2}. There are approximately 1.5 million fragility fractures in the United States³ and 3.5 million in the European Union⁴ each year. In the United Kingdom, over 50% of women and 20% of men older than 50 years are expected to sustain a fracture during their remaining lifetime⁵.

The global fragility fracture burden is increasing, largely due to the rising median age of the world population⁶. There are currently around 962 million people aged over 60 years worldwide, although this is projected to increase to 1.4 billion by 2030 and 2.1 billion by 2050⁷. Although age-adjusted fragility fracture incidence has stabilized over the last 20 years in many developed countries, this is not universal and the absolute numbers continue to rise⁸.

Fragility fractures have a long-term detrimental impact on health-related quality of life^{9–11}. They are associated with pain and fear of falling¹², as well as loss of independence, prolonged hospital admissions, disability, and mortality^{13–15}. In addition to their humanitarian effects, the economic costs of fragility fractures are high. The direct annual cost of fragility fracture treatment in the United States is estimated at \$17.9 billion per year³. In the European Union, the combined health and social

care cost of fragility fractures is around €37 billion per year⁴.

1.2 Risk factors for hip fracture

Hip fractures typically occur in older adults with bone fragility (e.g. caused by osteoporosis) following a low-energy fall.

1.2.1 Osteoporosis

Before the 1980s, osteoporosis could only be diagnosed following a low-energy fracture. However, with the development of dual-energy x-ray absorptiometry (DXA), it became possible to measure bone mineral density (BMD) and potentially diagnose osteoporosis without waiting for a fracture to occur¹⁶. In the mid-1990s, a working group convened by the World Health Organization (WHO) defined “osteoporosis” as a femoral neck BMD measured using DXA as 2.5 standard deviations below the reference standard of a young adult female¹⁷. However, it soon became clear that it was not sufficient to predict fracture risk based on BMD alone¹⁶. Although low BMD is associated with increased fracture risk, many age-related hip fractures occur in patients with BMD levels that do not fall below the WHO threshold for osteoporosis. For example, in one United States (US) cohort study, less than half of the women that sustained a hip fracture would have been diagnosed with osteoporosis using DXA¹⁸. The accepted explanation for this paradox is that individuals with normal BMD are simply more numerous in the population than those with low bone mass. However, this observation has driven investigators to search for other bone abnormalities (e.g. microarchitectural and geometric properties using high-resolution peripheral computed tomography)¹⁹ and to design fragility fracture risk prediction tools that incorporate clinical factors, such as the Fracture Risk Assessment Tool (FRAX)²⁰, QFracture²¹, and Garvan²² algorithms. All three tools attempt to combine estimated bone fragility with factors that are associated with falling.

1.2.2 Falls

Although osteoporosis may predispose a person to an increased risk of fracture, the precipitating event is usually a fall. It has even been argued that falling is the underlying cause of fragility fracture with fewer than a third of hip fractures attributable to bone fragility²³. The single question “do you have impaired balance?” is a better predictor of subsequent hip fracture than osteoporosis based on BMD criteria^{16,24,25}. However, as discussed in Subsection 1.2.1 on the preceding page, bone fragility is not solely a property of BMD and it is notable that hip fractures rarely occur in young people following low-energy falls²⁶.

The causes of falling amongst older adults are multifactorial but have been conceptualized as an interaction between intrinsic factors (such as adverse drug effects, reduced mobility, and poor eyesight), extrinsic factors (e.g. loose rugs and low level furniture), and situational factors (e.g. rushing to answer the telephone)^{27,28}.

1.3 Anatomy of the hip

The hip is a multi-axial ball-and-socket synovial joint in which the rounded head of the femur articulates with the concave acetabulum of the pelvis (Figure 1.1 on the next page). The hip joint is second only to the shoulder in terms of range of movement but some flexibility has been sacrificed in favour of additional stability. This strength is necessary as the hips must sustain the full weight of the body, both when standing and in motion².

Figure 1.2 on page 23 shows that the femoral neck projects inferolaterally from the femoral head and attaches to the shaft of the femur at an angle of approximately 125 degrees³⁰. A region of dense cancellous bone composes the *calcar femorale* and acts as a strut that helps transfer stress across the joint³¹. A dense fibrous capsule (composed of the strong iliofemoral, ischiofemoral, and pubofemoral ligaments) stretches from the acetabular margins to the intertrochanteric line anteriorly and part-way along the femoral neck posteriorly³². The intertrochanteric line is named

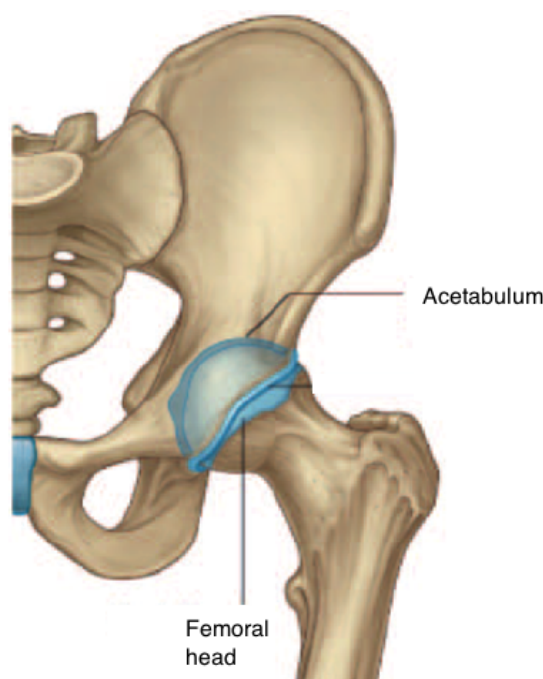


Figure 1.1: Head of the femur articulating with the concave acetabulum. *Reproduced from Drake 2009²⁹ with permission from Elsevier Inc.*

for two bony projections that lie outside the hip joint proper – the quadrangular-shaped “greater trochanter” laterally and the triangular lesser trochanter medially³³.

The blood supply to the femoral head is fundamental to surgical decision making after a hip fracture³⁴. The principal blood supply arises from the deep branch of the medial femoral circumflex artery, itself a branch of the *profunda femoris* or – less commonly – the common femoral artery³⁵. In addition, there is an anastomosis around the hip joint itself, which variably receives blood from the obturator, inferior gluteal, superior gluteal, and lateral femoral circumflex arteries as well as the medial femoral circumflex artery³⁴. There is also occasionally a small contribution from the artery of the *ligamentum teres*, although this is often small or obliterated in adults³⁶.

1.4 Hip fracture classification and management

Fractures of the proximal femur are conventionally known as “hip fractures”, even those that are outside and some distance from the hip joint, e.g. the sub-trochanteric fractures.

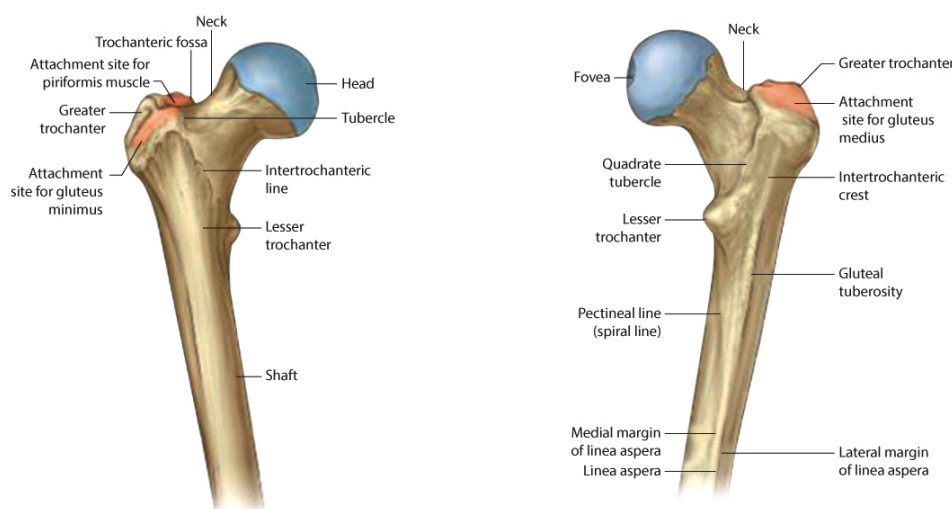


Figure 1.2: Bony anatomy of the proximal femur: anteroposterior view (left) and posteroanterior view (right). *Reproduced from Drake 2009²⁹ with permission from Elsevier Inc.*

Almost all hip fractures in the developed world undergo surgical treatment. The natural history of an untreated hip fracture is poor with many patients developing complications, such as pneumonia, decubitus ulcers, and venous thromboembolism (VTE)³⁷. There have been few RTs³⁸ but data from large cohort studies suggest that non-operative treatment is associated with twice the mortality of surgery^{39,40}. Surgical treatment is also associated with less deformity, shorter hospital length of stay, and lower total healthcare costs⁴¹. The Scottish Intercollegiate Guidelines Network (SIGN) hip fracture guideline starkly says that “conservative treatment with prolonged bed rest is not practised in this country”⁴². However, the precise *type* of operative treatment selected depends on the anatomical position and pattern of the fracture as well as the physiological status of the patient.

In contemporary clinical practice, the most useful classification of a hip fracture is in relation to the joint capsule. A fracture either occurs along the femoral neck, which is within the capsule (henceforth “intracapsular”), or outside the hip capsule (“extracapsular”)⁴³.

Undisplaced intracapsular fractures may be fixed *in situ* with devices that permit the fracture to compress during weight-bearing while the bone heals⁴³. These devices include cannulated hip screws and small (e.g. 2-hole) compression hip screws, which

achieved similar results at two years follow-up in one large RT⁴⁴. Surgical fixation is a minimally invasive operation that has the advantage of retaining the patient's own femoral head. There is however an accepted risk of fixation failure and need for re-operation, which may be as high as 20%⁴⁴.

Displaced intracapsular fractures (Figure 1.3) are encountered more frequently with approximately half of all cases in the National Hip Fracture Database (NHFD) falling into this category (see Table 3.1 on page 82). The principal concern about these fractures is that displacement may interrupt the blood supply to the femoral head⁴³ (Figure 1.4 on the next page). Accumulation of haematoma within the joint capsule may further reduce blood supply by impairing the microvasculature⁴⁵. As a consequence, the fracture risks progression to painful non-union and/or osteonecrosis of the femoral head if fixed *in situ*^{37,43}. The re-operation rates for internal fixation in such cases ranges from 10 to 48%, which may reflect differences in fracture displacement and anatomical variations in hip vasculature. One meta-analysis found that revision following internal fixation is due to non-union in 19% of cases and osteonecrosis in 10%⁴⁶.



Figure 1.3: A left-sided displaced intracapsular hip fracture

Most guidelines recommend treating displaced intracapsular hip fractures by replacing the femoral head (“arthroplasty”)^{42,48,49}. There are two commonly used arthroplasty techniques: hemiarthroplasty (HA) in which only the femoral head is replaced and THA in which both the femoral head and acetabulum are replaced

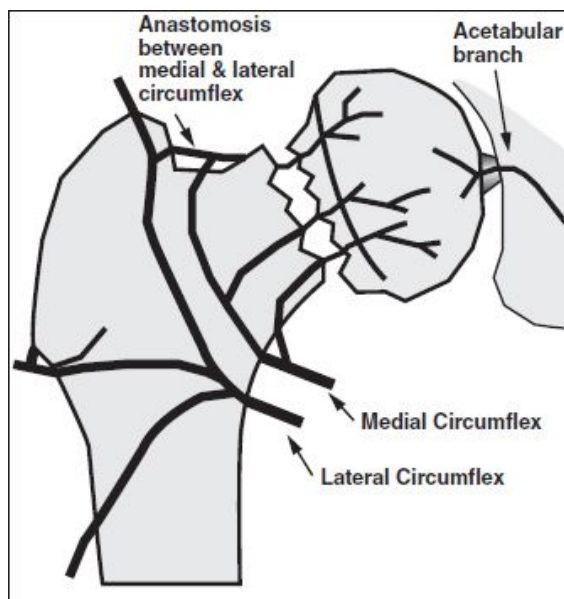


Figure 1.4: A displaced intracapsular hip fracture showing nutrient vessels. *Reproduced from Johnson 2004⁴⁷ with permission of Elsevier Inc.*

(Figure 1.5 on the following page). HA is a quicker operation and leads to less blood loss⁵⁰, although THA is associated with better health-related quality of life^{51–54}. The HA revision rate is thought to be higher but THA is associated with dislocation (Figure 1.6 on the next page), which may require closed reduction and sedation in an emergency department (ED) or operating theatre⁵⁰. The advantages and disadvantages of these two operations are described in Section 5.1.1 on page 124 and considered more fully in Chapter 6 on page 148. In general, THA is reserved for the fittest and most active patients⁵⁵.

There are two techniques for implanting an arthroplasty prosthesis. In an uncemented HA, the femoral shaft is prepared and the prosthesis gently hammered into the canal. In a cemented HA, the bone is prepared and then covered with cement before the prosthesis is press fit into the canal.

There is a concern that cementing the femoral canal may lead to cardiovascular complications, the so-called “bone cement implantation syndrome”⁵⁶. This is believed to occur as a result of embolization caused by increased intramedullary pressure while the femoral canal is instrumented and cemented. The syndrome is poorly understood but characterized by hypoxia, hypotension, arrhythmias, and sometimes

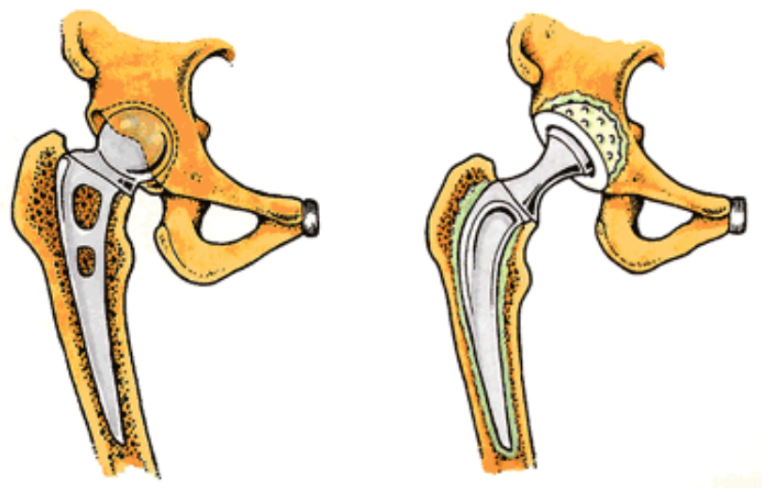


Figure 1.5: Schematic showing HA (left, with only femoral component) and THA (showing both femoral and acetabular components)

cardiac arrest⁵⁷. However, uncemented HA is associated with more post-operative pain, higher risk of peri-prosthetic fracture (both intra- and post-operatively), and higher rates of re-operation^{58,59}. In addition, the risk of bone cement implantation syndrome is thought to be low and potentially mitigated by careful peri-operative resuscitation^{43,59}.



Figure 1.6: Plain radiograph showing a dislocated right THA prosthesis. *Reproduced from Saleeb 2017⁶⁰ with permission from Elsevier Inc.*

1.5 Components of high-quality hip fracture care

The hip fracture population is characterized by advanced age, frailty and multi-morbidity. The median age of patients in the NHFD is 81 and the median superspell (i.e. total time in National Health Service (NHS) care) is 28 days⁶¹. Chapter 3 on page 72 describes the co-morbidity burden of older adults with hip fractures in England.

Historically, these fragile patients did not receive optimal care and hip fracture was considered a devastating injury. A saying attributed to Sir Reginald Watson Jones (1902-1972) quips that “we come to the world under the brim of the pelvis and go out through the neck of the femur”⁶². In 1955, Charles Heck described hip fracture care as “characterized by sandbags, long periods of recumbency, bedsores, low morale, a high mortality rate and often an extremity unfit for satisfactory function”⁶³. As late as 1999, the National Confidential Enquiry into Patient Outcome and Death (NCEPOD) inspectors expressed concern that “there were still a significant number of [hip fracture operations] which were performed by inappropriately junior trainees, often outside normal working hours”⁶⁴. However, there is growing recognition that care of these fragile patients is complex and requires careful coordination from a multidisciplinary team, including orthopaedic surgeons, anaesthetists, geriatricians, specialist nurses, physiotherapists, and occupational therapists^{37,43}.

Sahota and Currie have identified three elements of high-quality hip fracture care, which are “(1) high-quality acute and rehabilitation care delivered through coordinated multi-disciplinary teams; (2) High-quality secondary prevention of fragility fractures - bone protection and multi-disciplinary falls risk assessment... and (3) high-quality data collection and using audit standards to provide feedback to units, allowing them to monitor and benchmark what they do, and thus, to improve the hip fracture care and secondary prevention that they provide”⁶¹. These components will each be considered in the following subsections.

1.5.1 Acute and multi-disciplinary rehabilitation care

A number of factors related to acute care have been identified as likely determinants of hip fracture outcome. The most prominent of these are early surgery, early mobilization, and involvement of a specialist orthogeriatrician.

Early surgery

There is consistent evidence that surgical delay is associated with worse functional outcomes, more medical complications, and higher mortality^{65–69}. This is usually explained as a consequence of excessive immobility leading to complications of bed rest (e.g. VTE and decubitus ulcers), pain (e.g. pneumonia caused by atelectasis and cough inhibition due to hip pain), and rapid loss of muscle tone⁷⁰. Although the National Institute for Health and Care Excellence (NICE) recommends surgery within 36 hours⁵⁵, a number of studies have proposed that <24 hours is the critical threshold after which outcomes worsen^{67,68}.

Importantly, RTs testing early versus delayed hip fracture treatment are difficult to undertake and so all existing evidence comes from observational studies. This is problematic as there are many reasons why medically unwell hip fracture patients might undergo surgery later. For example, they may require physiological optimization (e.g. intravenous fluids), pre-operative investigations (e.g. echocardiogram), specific treatments (e.g. vitamin K to reverse warfarin anticoagulation), or the availability of a senior anaesthetist^{71–73}. It is likely that all previous observational studies addressing this question were biased by residual confounding. Nevertheless, the reasoning underlying early surgery is both intuitive and consistent with the available evidence. Untreated hip fractures are also painful⁷⁰ and so there is a plain humanitarian reason for delivering early surgery.

Early mobilization

Early mobilization is desirable for the same reason as early surgery (Subsection 1.5.1), i.e. to reduce the risk of medical complications. The aim in most cases of hip frac-

ture is to perform an operation that allows the patient to fully weight-bear post-operatively. Immediate post-operative weight-bearing is associated with shorter length of stay, greater likelihood of discharge to the patient's own home, better functional recovery, and reduced mortality^{74,75}. Similarly, delays to ambulation have been associated with worse outcomes in observational studies^{76,77}, as has lack of physiotherapy services at weekends⁷⁸. NICE currently recommends physiotherapy assessment and mobilization (unless clinically contraindicated) the day after surgery followed by daily mobilization until discharge⁵⁵.

Involvement of a specialist orthogeriatrician

Early collaborations between orthopaedic surgeons and geriatricians evolved into a medical sub-specialty dedicated to the care of older adults with fragility fractures⁷⁹. However, the first evidence in support of geriatrician involvement with orthopaedic inpatients only appeared in the 1990s, and then almost exclusively in the form of simple before-after observational studies^{80–82}. These early studies were followed by a NCEPOD report that highlighted the co-morbidity burden of hip fracture patients and recommended that orthopaedic surgeons “establish where there is sufficient expertise available within their team to manage the complex medical problems of these patients, or whether local guidelines for shared care should be developed”⁶⁴. In 2003, the British Orthopaedic Association (BOA) and British Geriatric Society (BGS) collaboratively authored *The Care of Fragility Fracture Patients* (henceforth *The Blue Book*), which set out clinical standards to include formal orthogeriatric services. RTs have since shown that orthogeriatric involvement reduces hip fracture mortality, complications, length of stay, and healthcare costs^{83–87}. One meta-analysis suggested that, when considered as a single group, orthogeriatric services reduce in-hospital (RR 0.60, 95% CI 0.43 to 0.84) and 12-month-mortality (0.83, 0.74 to 0.94)⁸⁸.

There are a number of different orthogeriatric models of care (Table 1.1 on page 31) and it is not yet known which of these achieves the best results for older

adults with hip fractures⁸⁹. However, formal orthogeriatric services are now recommended by most hip fracture guidelines in the developed world^{42,49,55,90}.

Table 1.1: Orthogeriatric models of care

Model	Description
Traditional care	Orthopaedic care with geriatrician consultation on request
Geriatric orthopaedic rehabilitation unit	Orthopaedic care until transfer for rehab under geriatricians
Orthogeriatric routine review	Orthopaedic ward with consistent geriatrician consultation
Orthopaedic routine review	Geriatric ward with consistent orthopaedic consultation
Integrated hip fracture service	Shared responsibility between orthopaedics and geriatricians
<i>Reproduced from Middleton 2018⁹¹ with permission from MDPI (Basel, Switzerland)</i>	

1.5.2 Secondary prevention of fragility fractures

NICE recommends that patients receive a formal hip fracture programme that includes liaison with falls prevention and bone health teams⁵⁵. They also provide specific guidance for secondary prevention of fragility fractures⁹², which is usually implemented at ward-level by an orthogeriatrician (Subsection 1.1 on the previous page). Multi-disciplinary Fracture Liaison Services (FLSs) have also been established to identify patients that may benefit from specific anti-fracture interventions and there is observational data to support the view that such services reduce long-term mortality following hip fracture⁹³. There is good RT evidence to support use of pharmacological agents (such as bisphosphonates) for secondary prevention of hip fractures^{94,95}. However, the range of interventions available for, and the evidence underlying, secondary prevention of hip fractures is beyond the scope of this thesis.

1.5.3 Data collection and clinical audit

The UK hip fracture community accepted the challenge of collecting high-quality audit data with which to feedback performance to hospitals against defined clinical standards. This took the form of the NHFD in 2007, which collects data from hospitals in England, Wales, and Northern Ireland. This NHFD is of central importance to this thesis and is described further in Section 1.6 and Subsection 1.7.1 on page 38.

1.6 Improving healthcare quality through audit

It has been argued that everyone in healthcare should “recognize that they have two jobs when they come to work every day: doing the work and improving it”⁹⁶.

In 1854, Florence Nightingale (1820-1910) arrived as a nurse at Scutari (now Uskudar, Istanbul, Turkey) during the Crimean War⁹⁷. She found the military hospital in disarray with a 42% mortality rate among hospital inpatients⁹⁸. The overwhelming cause of death was preventable infection rather than bullet, shell, or sabre. Having been tutored as a child by the eminent mathematician, J. J.

Sylvester⁹⁹, Nightingale began meticulously recording data about standards of care and patient outcomes. She drove a host of quality improvements at Scutari (e.g. hand washing, stopped sharing of bed linen between patients) and developed a new method (the polar area diagram) that allowed her to visualize changes over time⁹⁷. By the end of the Crimean War, her efforts were widely credited with reducing inpatient mortality from 42% to 2%⁹⁸.

Other pioneers of clinical audit included the Boston orthopaedic surgeon Ernest Codman (1869-1940) who proposed that every patient discharged from hospital should be followed up to learn the “end result” of their treatment¹⁰⁰. In Codman’s view, this would allow surgeons to learn which treatments worked and which did not, and so continuously improve their practice. This idea was not well received and Codman lost his privileges to practice at Massachusetts General Hospital¹⁰¹. He founded a new establishment (the “End Result Hospital”), which was the first to publish all outcomes in an annual report and declared errors in the care of a third of its patients¹⁰⁰.

The concept of data-driven quality improvement made little progress in the 19th and early 20th centuries beyond the influence of a few pioneers. However, audit has gradually become a central component of quality improvement in healthcare since the 1980s. In its modern form, clinical audit has been defined as a “quality improvement process that seeks to improve patient care and outcomes through systematic review of care against explicit criteria and the implementation of change”¹⁰².

Within the NHS, clinical audit was formally introduced in the government White Paper *Working for Patients* (1989)¹⁰³. The concept of clinical governance emerged in *The New NHS* (1997)¹⁰⁴, which also identified clinical audit as a key foundation on which to base healthcare quality improvements. A number of public inquiries following healthcare failings have recommended engagement with both local audit and national clinical audit programmes^{105,106}. The GMC now requires all registered medical practitioners to “take part in systems of quality assurance and quality improvement to promote safety. . . [including]. . . regular reviews and audits”¹⁰⁷.

1.6.1 Why is audit necessary?

There is considerable evidence for unexplained variation in hip fracture care and clinical treatment that falls short of recommended practice. Such variations might be explained by differences in resource allocation, culture, organization, or individual clinical performance. Another explanation is that individual clinicians – even subspecialists – cannot keep up with the fast pace of developments in any given field¹⁰⁸. However, this challenge may be mitigated by parallel advances in the techniques used to synthesize evidence and develop best-practice guidelines. Audit and feedback have therefore been proposed as a vehicle with which to drive adherence to best practice guidelines. There is consistent evidence to show that clinicians are not the best judges of their own performance and that self-assessment is less likely to drive improvements than external audit programmes using systematic data collection methods¹⁰⁹.

1.6.2 Can audit improve performance?

Most studies have found that participation in a formal clinical audit programme drives change more effectively than self-assessment¹⁰⁹ and self-monitoring¹¹⁰. A Cochrane review (including data from 140 RTs) reported that audit can modestly improve health professionals' compliance with clinical standards¹¹¹. Unsurprisingly, this review found gross between-study heterogeneity across interventions, clinical contexts, and outcomes as well as poor descriptions of the interventions themselves. The authors reported a median adjusted risk difference of 4.3% (interquartile range (IQR) 0.5% to 16.0%) absolute increase in compliance with clinical guidelines. However, the review also found that a quarter of interventions had large positive effects on quality of care (i.e. 16.0% absolute improvement) whereas another quarter had either a negative or null effect. This suggests that clinical audit *can* drive improvement but that establishing an audit programme does not guarantee success.

Two systematic reviews specifically considered whether or not feedback from clinical registries can improve processes and outcomes^{112,113}. However, they also

reported few rigorous studies and high between-study heterogeneity. Both reviews found a high number of studies that reported associations between feedback from clinical registries and improved quality-of-care processes. Neither reported consistent evidence in support of clinical registries having a positive effect on patient outcomes. The factors that appeared to influence the effect of interventions included engagement of recipients, quality of the audit data, and the degree of organizational support for improving services¹¹³.

1.6.3 Components of a successful audit programme

There has been little meaningful progress in terms of identifying the elements of a successful audit programme. This is partly because the interventions subject to RT evidence have not been designed within a common framework that would facilitate meaningful pooling of data and meta-analysis. However, Iver (2012)¹¹¹ did use meta-regression techniques to explore associations between features of clinical audit programmes and their reported effects. The key findings of this systematic review are summarized below:

Format of the feedback Feedback presented in written and verbal form resulted in a significantly larger effect than when feedback was only conveyed verbally (adjusted risk difference 8%, $p=0.020$)

Source of the feedback Feedback presented by a supervisor or senior colleague was associated with a larger effect than feedback compiled by the investigators themselves (adjusted risk difference 11%, $p<0.001$).

Regularity Interventions utilising monthly feedback were associated with the largest positive effect size. The adjusted risk difference when comparing monthly feedback with feedback on only one occasion was 11%, $p<0.001$.

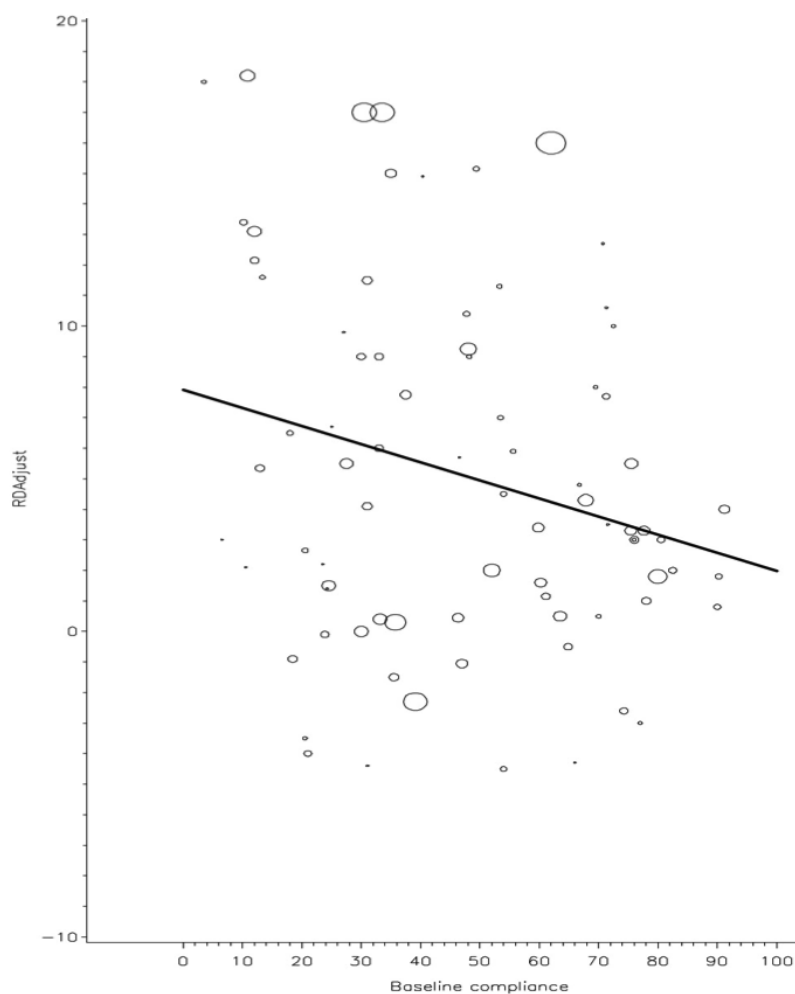


Figure 1.7: Bubble plot the association between adjusted risk difference and baseline compliance. *Reproduced from Iver 2012¹¹ with permission from John Wiley & Sons Inc.*

Instructions for improvement Interventions that include a measurable target and clear action plan were associated with the largest positive effects: adjusted risk difference 5% ($p < 0.001$) when compared with those that included neither.

Direction of change required Interventions recommending a decrease in current behaviour were associated with larger positive effects than those requiring an increase (adjusted risk difference 6%, $p < 0.001$)

Baseline performance Low baseline performance was positively associated with greater effects of the intervention (Figure 1.7).

Profession of the recipient Whether or not the recipient was a physician was not associated with estimated effect size.

1.7 National clinical audit in England

One classification system has distinguished three types of audit: external, internal, and clinical¹¹⁴. In this system, externally driven audits exist predominantly for the purpose of quality assurance, e.g. accreditation. By contrast, clinical audits are driven by healthcare professionals and have a greater focus on quality improvement. However, this system does not readily describe *national* clinical audit, which is a hybrid category but a growing feature of clinical governance across the NHS.

The first major national clinical audit in England was the NCEPOD, which reported in 1989¹¹⁵. The NCEPOD was commissioned to review standards of surgery and anaesthesia, which it addressed by sending inspectors to hospitals to physically inspect procedures and medical records. Importantly, the first NCEPOD report concluded that “data systems in the NHS are inadequate. Rates of events (admissions, operations and deaths) cannot be calculated because contemporary data are not available. Thus valid comparisons between hospitals, districts or regions cannot be made promptly enough to influence clinical practice”¹¹⁵.

The Bristol Public Inquiry in 2001 was followed by calls to establish national clinical audits¹⁰⁵ and ultimately prompted the Society of Cardiothoracic Surgeons of Great Britain and Ireland to publish surgeon-level mortality statistics for the first time¹¹⁶. In December 2012, HQIP was commissioned by NHS England to establish national audits of outcomes across all the major surgical specialties¹¹⁷. There are now 30 audit programmes managed by HQIP. The FFFAP is one of these programmes and incorporates the NHFD, National Audit of Inpatient Falls, and the Fracture Liaison Service Database¹¹⁸. The contract for day-to-day administration of the FFFAP is currently held by the Royal College of Physicians of London. The NHFD dataset forms the foundation of this thesis and is described in Subsection 1.7.1 on the next page.

1.7.1 The National Hip Fracture Database

The NHFD was launched in 2007 and receives data from all NHS hospitals in England, Wales, and Northern Ireland that treat patients with hip fractures. It was established independently of the Scottish Hip Fracture Audit, which ran from 1993 to 2008¹¹⁹.

Hospitals are incentivized to report cases to the NHFD by payments under the Hip Fracture Best Practice Tariff (BPT)¹²⁰. Specialist nurses within individual trusts typically upload data through a dedicated electronic portal. The NHFD captures approximately 98% of hip fractures from all participating acute hospitals and has recorded over 500,000 cases since 2007¹²¹. All adults with hip fractures are eligible for inclusion within the NHFD except for those aged less than 60 years or undergoing non-operative treatment. Hospital performance can be visualized using an online dashboard and inter-hospital comparisons are published in an annual report. There is a systematic process for managing hospitals that are found to be performance outliers across any one of the HQIP national clinical audits¹²².

The NHFD was established for the purposes of audit but has gradually been adopted as a resource by researchers interested in hip fracture care and outcomes. Although it is increasingly used for research purposes, the absolute number of published projects remains small (Figure 1.8 on the following page).

1.8 Thesis aim and structure

The overarching aim of the work in this thesis was to explore ways in which a national clinical audit can improve quality of care and guide clinical decision making.

Each chapter was designed to address a discrete research question, either to guide development of the NHFD or better understand an aspect of hip fracture care. In addition, I designed each project with the aim of gaining experience across a suite of methods that could form the basis of a research career using “real world” data. Throughout these projects I gained experience of information governance (partic-

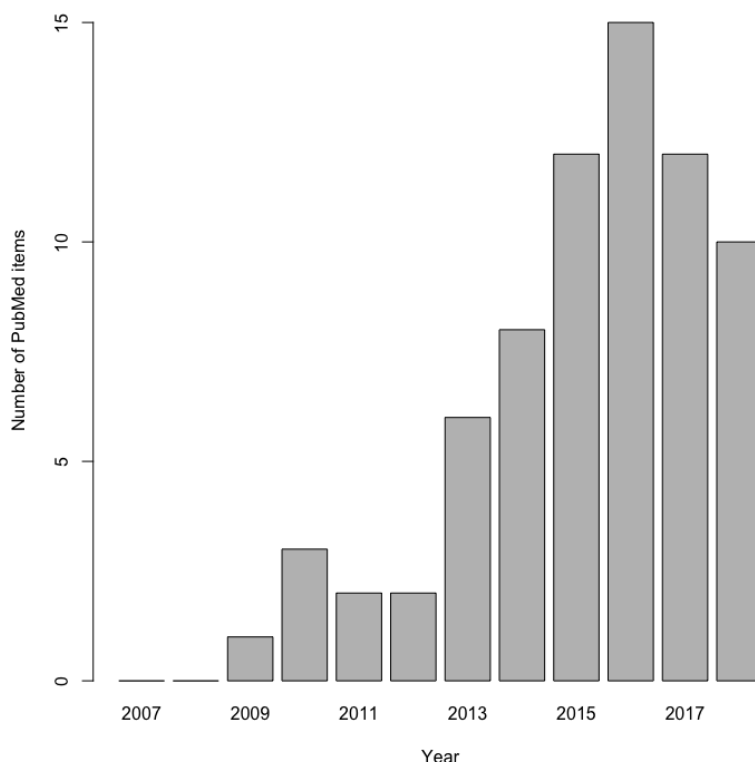


Figure 1.8: Number of PubMed items retrieved when searching for "National Hip Fracture Database" 2007-2018

ularly overcoming barriers to data access) as well as data management principles and familiarity with a range of different software packages (predominantly Stata but also R, SPSS, RevMan, and GradePro), as and when these were required for specific tasks. I also learned how to access and work with a range of national datasets, including the NHFD, Hospital Episode Statistics (HES), SMR01, and civil death registration data.

The core chapters in this thesis are each outlined below together with the research question that they addressed and the specific methods that were used.

1.8.1 Core chapters

Impact of public release of performance data on quality of care and patient outcomes

- **Research question:** Does publication of performance data change the behaviour of healthcare providers, quality of care, and patient outcomes.

- **Study design:** Systematic review of randomized and non-randomized studies.
- **Methods:** Systematic review, risk of bias assessments, data re-analysis using difference-in-differences (DID) and AutoRegressive Integrated Moving Average (ARIMA) analyses, standardized language statements, Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessments, and constructing a summary of findings table.

Risk adjustment in the National Hip Fracture Database

- **Research question:** Would the NHFD benefit from routine linkage to an administrative dataset and is the identification of mortality outliers sensitive to the choice of co-morbidity summary measure?
- **Study design:** Observational study using NHFD data and statistical model evaluation.
- **Methods:** Regression model selection (including tests for calibration and discrimination), multiple imputation, fitting a fractional polynomial, area under the receiver operating characteristic (AUROC) curves, calibration belt plots, and funnel plots.

Pay-for-performance and hip fracture outcomes

- **Research question:** Was introduction of the NHFD and the subsequent Hip Fracture BPT in England associated with improved hip fracture outcomes?
- **Study design:** Natural experiment with interrupted time series analysis (ITSA) and DID analyses using the Hospital Episode Statistics Admitted Patient Care (HES APC) and SMR01 datasets.
- **Methods:** ITSA, DID, projection modelling.

Inequalities in the use of total hip arthroplasty for hip fracture

- **Research question:** Is use of THA among individuals with a displaced intracapsular hip fracture based on national guidelines or are there systematic

inequalities?

- **Study design:** Observational cohort study using the NHFD.
- **Methods:** Recursive partitioning, multivariable logistic regression.

Total hip arthroplasty versus hemiarthroplasty for intracapsular hip fractures

- **Research question:** Are THA or HA associated with clinical outcomes for independently mobile older adults with intracapsular hip fractures?
- **Study design:** Systematic review of RTs compared with propensity score matched “real world” data from the NHFD.
- **Methods:** Systematic review, random-effects meta-analysis, propensity score matching, Rosenbaum sensitivity analyses, survival analysis (Kaplan-Meier plots, Cox regression, frailty models, and competing risks regression), multivariable regression), logistic regression and generalized linear models.

Chapter 2

Impact of public release of performance data on the quality of care and patient outcomes

2.1 Introduction

The NHFD began releasing annual reports to the public in 2007¹²³. The first report provided data that was sufficient to illustrate variation in outcomes between hospitals but the identity of individual providers was anonymized¹²³. This was changed for the 2010 report, which identified hospitals by name, ranked them by performance against the BPT indicators, and publicly identified outliers for standardised 30-day mortality¹²⁴. This approach is consistent with that taken by other national clinical audits, such as the National Joint Registry (NJR)¹²⁵ and the Trauma Audit & Research Network (TARN)¹²⁶¹.

It is often assumed that releasing performance data into the public domain will influence the behaviours of various stakeholders, and so ultimately lead to health system improvements^{127–129}. One study has conceptualized public reporting of performance data as (1) supporting patient choice, (2) improving accountability, and

¹Chapter published as Metcalfe D, Rios Diaz AJ, Olufajo OA, Massa MS, Ketelaar N, Flottorp SA, Perry DC. Impact of public release of performance data on the behaviour of healthcare consumers and providers. *Cochrane Database Sys Rev.* 2018;9:CD004538.

(3) allowing providers to benchmark their performance against others¹³⁰.

Although patient choice might help drive improvements in other disease areas, this is less likely amongst the hip fracture population. Hip fracture typically occurs amongst frail older adults (who are less likely to access performance data)^{131–133} and patients will rarely be in a position to influence the hospital to which they are conveyed. As hip fracture treatment is rarely planned, there is little opportunity for patients to research the performance of individual hospitals before injury^{134,135}.

Improved accountability may however be achieved by encouraging healthcare providers to focus on quality issues, as they know that performance measures will be published^{136,137}. This, in turn, may stimulate quality improvements, particularly as providers can quantify their own performance against that of other clinicians and hospitals.

Other proposed goals for performance measurements have been linked to controlling costs^{138,139}, regulating the overall healthcare system^{140,141}, and influencing the decisions of healthcare purchasers^{142–144}.

Professional concerns about public release of performance data often relate to the validity of both the performance measures themselves, and comparisons between health providers^{145–147}. There are concerns that failure to adequately adjust for case mix differences might lead to providers that treat higher-risk patients being labelled as poor performers, or to providers preferentially selecting lower-risk patients^{147–150}. In healthcare systems where providers charge for their services, “better” performing providers might feel empowered to increase charges, thereby restricting access to the highest quality care¹⁴⁴. An additional risk is that publication of performance data may lead to improved reporting, without necessarily improving performance. It has been argued that the care processes that are easiest to measure are often those that are least important in a quality improvement context, and can result in the de-prioritization of more important tasks¹⁵¹.

Earlier systematic reviews have suggested positive effects of publicly releasing performance data, but included a broad range of study designs^{136,152–154}. This chap-

ter aimed to review only the best available evidence for the impact of publicly releasing performance data.

The aim of this review was to estimate the effects of publicly releasing performance data on changing the behaviour of healthcare providers (professionals and organizations), quality of care, and patient outcomes.

2.2 Methods

A systematic review was undertaken under the auspices of the Cochrane Collaboration and following guidance developed by the Cochrane Effective Practice and Organisation of Care (EPOC) Group^{155,156}.

2.2.1 Study inclusion criteria

Types of studies

- Non-randomized trials, including cluster-randomized trials (cRTs).
- Non-randomized trials, including cluster-non-randomized trials (cNRTs), which use non-random methods of allocation, such as alternation or allocation by case note number.
- Controlled before-after studies, with at least two intervention sites and two control sites that were chosen for similarity of main outcome measures at baseline.
- Interrupted time series studies, with at least three data points before and three data points after the intervention.

Non-randomized studies were included in anticipation of a lack of RTs, but also because some interventions might not be appropriate for a trial (e.g. randomising participants to not receive important information that might affect their healthcare choices), and others might have a variable effect over time that is best observed by an alternative study design, such as an interrupted time series (ITS).

Types of participants

Patients or other healthcare consumers and healthcare providers (professionals and organizations) without any restriction by type of healthcare professional, provider, setting, or purchaser.

Types of interventions

Interventions were included that contained the following elements:

- Performance data about any aspect of the healthcare organizations or individuals, including process measures (e.g. waiting times), healthcare outcomes (e.g. mortality), structure measures (e.g. presence of waiting rooms), consumer or patient experiences (e.g. survey data), with or without expert or peer assessed measures, e.g. certification, accreditation, and quality ratings given by colleagues. Performance data were included if prepared and released by any organization, such as the government, insurers, consumer organizations, or providers.
- The release of performance data into the public domain in written or electronic form without regard to any minimum degree of accessibility. For example, this could include a report available in a publicly accessible library, as well as active dissemination directly to consumers through personal mailings.

Types of outcome measures

Primary outcomes

1. Changes in healthcare decisions taken by healthcare providers.
 - Objective measures of changing healthcare provider behaviour, such as changes to drug prescribing.
2. Changes in provider performance
 - Objective changes, such as reaching the correct diagnosis or time to treatment.

3. Changes in patient outcome.

- Objective changes, such as mortality or patient reported outcome measures.

Secondary outcomes The secondary outcomes were objective measures of changing purchaser behaviour (such as increased or decreased funding for services) unintended effects or harms, and any potential impact on equity (such as differential effects between advantaged and disadvantaged populations).

Search methods for identification of studies

Electronic searches Literature searches were undertaken using the Database of Abstracts of Reviews of Effects (DARE) for primary studies included in related systematic reviews. The following databases were also searched on 26th June 2017: Cochrane Central Register of Controlled Trials (CENTRAL; 2017, Issue 5) in the Cochrane Library; MEDLINE Ovid; Embase Ovid. The search strategies were developed together with a Cochrane EPOC Information Specialist. The search strategies included both keywords and controlled vocabulary terms. No language or time limits were applied.

Trial registries

- International Clinical Trials Registry Platform (ICTRP), WHO: www.who.int/ictrp/en/ (searched 26th June 2017)
- ClinicalTrials.gov, US National Institutes for Health (NIH): clinicaltrials.gov/ (searched 26th June 2017)

Other resources The reference lists of all included studies were searched manually. The search strategies used have not been reproduced in this thesis but are available in the published review¹⁵⁷.

Data collection and analysis

Selection of studies All titles and abstracts retrieved in the electronic search were downloaded to a reference management database. Duplicates were removed and two researchers then independently examined the remaining references. Assessments were recorded with points: “0” for exclusion, “1” for doubtful and “2” for inclusion. Two researchers independently rated each abstract and so scores could range from zero to four. Abstracts with a combined score of zero or one were excluded. Studies with a combined score of three or four were included. Two researchers resolved the fate of studies with a combined score of two by discussion. A third researcher adjudicated any disagreements that remained unresolved. Figure 2.1 on page 52 shows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram that accounts for exclusion of all items retrieved by the search strategy.

Data extraction and management Two researchers independently extracted data about the study design, patient and provider characteristics, interventions, outcome measures, and healthcare choices to a form specially designed for the review. Disagreements were resolved by discussion, and the judgement of a third researcher was accepted in the event of continued disagreement.

Assessment of risk of bias in included studies Risk of bias was assessed using guidance from the *Cochrane Handbook for Systematic Reviews of Interventions* (henceforth the *Cochrane Handbook*), which recommends using the following items: (i) adequate sequence generation, (ii) concealment of allocation, (iii) blinding, (iv) incomplete outcome data, (v) selective reporting, and (vi) no risk of bias from other sources¹⁵⁸. However, three additional criteria were also used following guidance from the EPOC Group: (vii) baseline characteristic similarity, (viii) baseline outcome similarity, and (ix) adequate protection against contamination¹⁵⁵. These nine criteria were used for RTs, non-randomized trials (NRTs), and controlled before-after (CBA) studies. Seven different criteria for ITS studies were used following EPOC

guidance: (i) the intervention is independent of other changes, (ii) the shape of the intervention effect is pre-specified, (iii) the intervention is unlikely to affect data collection, (iv) knowledge of the allocated interventions is adequately prevented during the study, (v) the outcome data are complete, (vi) reporting is not selective, and (vii) there is no risk of bias from other sources¹⁵⁵. Two researchers independently reached judgements about risk of bias using the guidance provided by the *Cochrane Handbook*¹⁵⁸ and EPOC 2013¹⁵⁵. Disagreements were resolved through discussion with a third researcher.

Measures of treatment effect In order to standardize reporting of effect sizes, data were re-analysed from individual studies to ensure that RT and CBA studies could be reported as relative effects. ITS data were reported as change in level and change in slope. The methods used for re-analysing and presenting these data are described in Section 2.2.1 on the following page.

Unit of analysis issues It was noted whether trials randomized patients or healthcare providers. If an analysis did not allow for clustering of patients within healthcare providers, a unit of analysis error was recorded, because such analyses tend to overestimate the precision of the treatment effect. In the event of a unit of analysis error and insufficient data to account for clustering, p-values and confidence intervals were not reported.

Dealing with missing data In the event of important missing data, study authors were contacted. As described in Section 2.2.1 on the next page, missing ITS data that was not available from study authors were extracted electronically when presented in graphs.

Assessment of heterogeneity There were substantial differences between the policies and interventions described. There were also differences between the settings in terms of both culture and health system delivery. Although some studies

evaluated similar interventions, there were still important clinical and methodological differences. As statistical tests for heterogeneity lack power when few studies are included, no attempt was made to calculate average effects across studies or to estimate statistical heterogeneity¹⁵⁹.

Assessment of reporting biases Funnel plots were not presented as a meta-analysis was not undertaken and there were fewer than 10 studies contributing to any individual analysis¹⁵⁸.

Data synthesis The EPOC recommendations were followed with regard to analysing data from individual studies and meta-analysis¹⁵⁵. Findings from CBA studies were expressed as relative effects. To achieve this, continuous variables were reported as relative change in outcome measures, adjusted for baseline differences.

Absolute DID analyses were undertaken and adjusted for differences in the post-intervention control group using: $((A2 - B2) - (A1 - B1))/A2$ where $A2$ represents the post-intervention intervention group, $B2$ the post-intervention control group, $A1$ the pre-intervention intervention group, and $B1$ the pre-intervention control group. For ease of comparison with the findings of CBA studies, the findings of RTs and NRTs were reported using the same DID analysis.

ITS are typically reported using regression analysis, such as ARIMA analysis. Pursuant to the EPOC recommendations, outcomes were presented along two dimensions: change in level and change in slope¹⁵⁵. The former represents the immediate effect of the intervention as measured by the difference between the fitted value for the first post-intervention time point and the predicted outcome at the same point, based only on an extrapolation of the pre-intervention slope. Change in slope is an expression of any longer-term effect of the intervention.

In the event that appropriate ITS data were not reported or available from authors but were presented graphically, values were extracted from graphs using Plot Digitizer v2.6.8¹⁶⁰. Plot Digitizer has been validated and shown to be more reliable than manually reading from graphs when compared with source data¹⁶¹.

True data points were extracted from all studies and lines of best fit only used in the event that true points were not available. A segmented time series model ($Y(t) = (B0 + B1) \times (preslope + (B2 \times postslope) + (B3 \times intervention) + e(t))$) was specified, in which $Y(t)$ was the outcome in month t . *Preslope* is a continuous variable that indicates time from the beginning of the study until the end of the pre-intervention phase, after which it was coded as a constant. *Postslope* is assigned the value 0 until after the intervention takes place, after which it is coded sequentially from 1 (i.e. 1, 2, 3). *Intervention* is assigned the value 0 pre-intervention and 1 in the post-intervention time period. In this model, $B1$ estimates the pre-intervention slope, $B2$ the post-intervention slope, and $B3$ the change in level, i.e. the difference between the first post-intervention time point and the extrapolated first post-intervention time point had the pre-intervention line continued into the post-intervention period. The difference in slope was determined using $B2 - B1$.

Effects were reported at 3, 6, 9, 12, and 24 months post-intervention when sufficient data were available. Given the substantial degree of clinical and methodological heterogeneity between the studies, findings were presented using a structured format but no meta-analysis was undertaken.

Summary of findings

The findings of the main interventions are summarized in Table 2.4 on page 63 to illustrate the certainty of the evidence. One researcher categorized the certainty of the evidence as high, moderate, low, or very low, using the five GRADE domains: study limitations, consistency of effect, imprecision, indirectness, and publication bias¹⁶². This was undertaken using Chapter 12 of the *Cochrane Handbook* and worksheets created by EPOC^{155,158}. These judgments were checked by all other researchers and disagreements resolved through discussion. When ratings were up- or down-graded, this was justified in Appendix C on page 250. Standardized statements for reporting effects and certainty of evidence were selected, based on the GRADE assessments for each outcome, and used throughout the review¹⁵⁶.

2.3 Results

The included studies are summarized in Table 2.1 on page 53 and described fully in Appendix A on page 230.

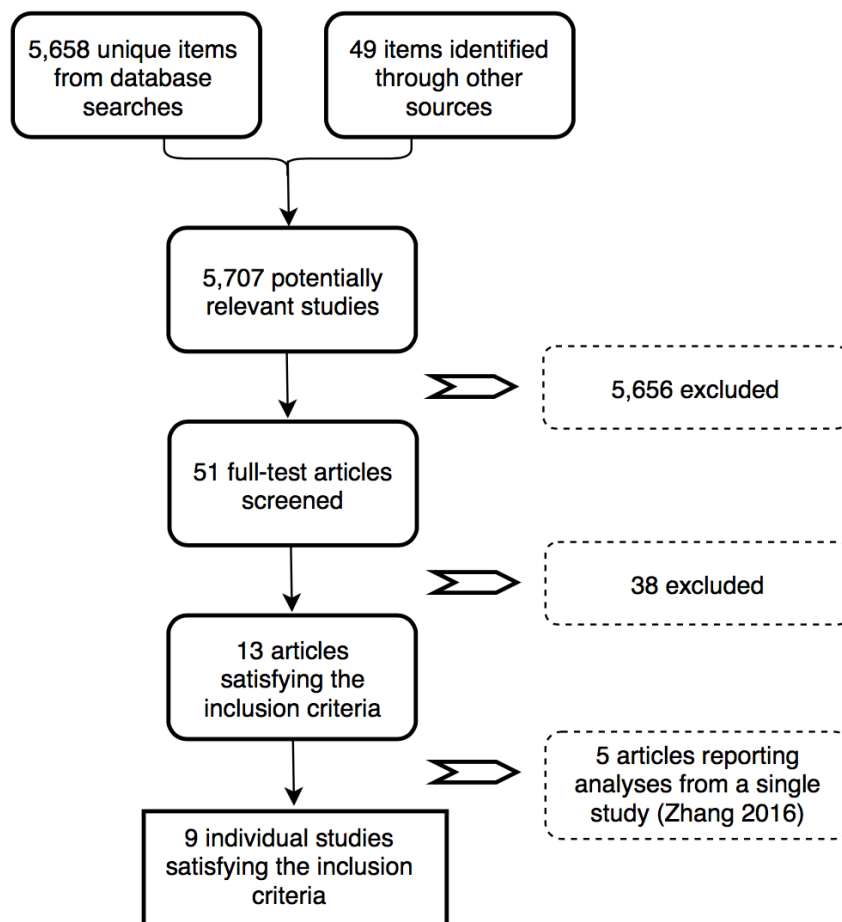


Figure 2.1: A PRISMA flow diagram illustrating the selection of studies satisfying the review inclusion criteria

Table 2.1: Summary of included studies

Study	Setting	Participants	Intervention	Outcome
Jang 2011 ¹⁶³ ITS	South Korea	Hospitals	Online press releases	Caesarean section rates
Flett 2015 ¹⁶⁴ ITS	USA	Paediatric hospitals	State-based mandatory public reporting of healthcare-associated infections	Blood culture tests Antibiotic prescribing
DeVore 2016 ¹⁶⁵ ITS	USA	Medicare enrollees	Hospital re-admission rates published on Hospital Compare	30-day re-admission 30-day outpatient visits 30-day ED visits
Joynt 2016 ¹⁶⁶ ITS	USA	Medicare enrollees with specific diagnoses	Hospital mortality rates published on Hospital Compare	30-day mortality
Liu 2017 ¹⁶⁷ ITS	USA	Acute hospitals	State-based mandatory public reporting of healthcare-associated infections	CLABSI
Tu 2009 ¹⁶⁸ cRT	Canada	Acute hospitals	Report cards released online and to media	Range of AMI and CHF quality indicators
Ikkersheim 2013 ¹⁶⁹ cRT	Netherlands	GPs	Report cards sent to GPs by post	Choice of hospital when making referrals
Zhang 2016 ¹⁷⁰ cRT	China	Primary care institutions	Display of antibiotic prescribing data in outpatient waiting areas	Antibiotic prescriptions Intravenous antibiotic prescriptions Average expenditure per prescription
Rinke 2015 ¹⁷¹ CBA	USA	Hospitals by state	State-based mandatory public reporting of healthcare-associated infections	Healthcare-associated infections

2.3.1 Results of the search

The electronic searches for this chapter retrieved 5,658 individual items and a further 49 were identified from other sources, e.g. manual searching of reference lists. 5,656 items were excluded because the titles and abstracts did not meet the inclusion criteria. Full-text versions of the remaining 51 articles were retrieved; 38 did not satisfy the inclusion criteria. Five of the remaining 13 articles reported separate analyses of a single cRT, and so these were treated as a single study for the purposes of the review¹⁷⁰. The final review therefore included nine studies (Figure 2.1 on page 52).

2.3.2 Included studies

The nine studies comprized more than 7,570 providers (e.g. professionals and organizations) and a further 3,333,386 clinical encounters (e.g. patient referrals, prescriptions). There were three cRTs^{168–170}, five ITS studies^{163–167}, and one CBA study¹⁷¹. Five were conducted in the United States of America (USA)^{164–167,171}, and one each in Canada¹⁶⁸, the Netherlands¹⁶⁹, Korea¹⁶³, and China¹⁷⁰.

Four studies focused on changes in the behaviour of providers^{163,164,169,170} and none of purchasers. Two studies reported data on changes to provider performance^{168,171} and five on patient outcomes^{164–168}. No study explicitly reported adverse events as a separate outcome, or gave particular consideration to effects on equitable health care.

Three US studies examined the effect of a single suite of interventions (i.e. laws mandating public reporting of healthcare-associated infections in the United States), which were introduced by some state legislatures between 2006 and 2009^{164,167,171}. Liu 2017¹⁶⁷ examined the effect of mandatory reporting on central line-associated blood stream infection (CLABSI) rates in adult intensive care units. They undertook an ITS study using data from hospitals contributing to the National Healthcare Safety Network between 2006 and 2012. States that did not introduce mandatory reporting were used to control for secular trends through a difference-in-difference

analysis. The other two studies focused their analyses on healthcare-associated infections in paediatric inpatients^{164,171}. Rinke 2015¹⁷¹ sought to determine whether mandatory CLABSI public reporting was associated with a reduction in a specific paediatric safety indicator (PDI12, i.e. selected infections due to medical care), which is defined using diagnosis codes on hospital discharge. They undertook a CBA study using the Kids' Inpatient Database, which is one of a suite of administrative healthcare databases coordinated by the Healthcare Cost and Utilization Project (HCUP) at the US Agency for Healthcare Research and Quality (AHRQ). Flett 2015¹⁶⁴ did not examine patient outcomes, but aimed to test the hypothesis that clinicians in hospitals that are required to report CLABSIs would modify their behaviour by sending fewer blood culture tests or prescribing longer courses of antibiotics. They undertook an ITS using data from the Pediatric Health Information System, which is a collaborative venture between children's hospitals that is used for clinical audit and quality improvement. The data were analysed using generalized linear mixed-effects models with auto-correlated residuals to compare CLABSI adjusted rate ratios before and after implementation of mandatory reporting laws.

The other two US studies each examined the impact of different public reporting initiatives on patient outcomes^{165,166}. They both used Medicare claims data, and so confined their analyses to the Medicare population, i.e. those aged 65 years or older. DeVore 2016¹⁶⁵ undertook an ITS to study the effect on 30-day re-admissions of publicly reporting risk-adjusted hospital re-admission rates for patients with selected conditions (acute myocardial infarction (AMI), heart failure, and pneumonia) on the *Hospital Compare* website. Joynt 2016¹⁶⁶ reported an ITS with a similar study design to DeVore 2016¹⁶⁵ but examined the impact on mortality rates of public reporting of mortality (for patients with the same three selected conditions) on *Hospital Compare*. They used hierarchical modelling to compare 30-day mortality in the pre- and post-reporting periods.

There were three cRTs outside the USA; one each in Canada¹⁶⁸, the Netherlands¹⁶⁹, and China¹⁷⁰. In Canada, Tu 2009¹⁶⁸ evaluated the public release of per-

formance data about 12 care quality indicators for AMI and six for congestive heart failure (CHF) in 86 hospitals. Participating hospitals were randomized to either early (January 2004) or delayed (September 2005) publication of performance report cards. The performance data were provided to individual hospitals and then publicized, both online and through popular media, with coverage achieved through television, radio, and newspapers. The outcomes reported by this study were any change in hospital performance, as measured using the 18 care quality indicators.

The cRT in the Netherlands randomized 26 General Practitioners (GPs) to receive either individualized hospital report cards (65.4%), or to a control group (34.6%) that did not receive this information¹⁶⁹. The study then captured individual patient referrals (for breast cancer, cataract surgery, and hip or knee replacement) to one of four hospitals in the region, using an electronic referral system.

Zhang 2016¹⁷⁰ undertook a cRT in Hubei Province, south central China. They matched 20 primary care providers within a single city, based on similar organizational characteristics. In this matched-pair cRT, half the providers were randomized to public reporting of injection prescribing, by way of league tables that were posted on outpatient bulletin boards. Performance data were also disseminated to both local health authorities and the leaders of hospitals in the intervention group. The outcomes were the percentage of prescriptions requiring antibiotics, percentage requiring intravenous antibiotics, and the average expenditure per prescription.

Finally, a single ITS study was undertaken in Seoul, South Korea by Jang 2011¹⁶³. In this study, the intervention was public release of data (online and in media releases) about caesarean section rates for 1,194 institutions across the country. These rates were publicized as part of a series of public releases, which were not described in detail. The outcome was change in risk-adjusted institutional caesarean section rates over the whole study period, and after each public release of data.

2.3.3 Excluded studies

In total, 35 studies were excluded after assessing full copies of the papers. The main reasons for exclusion were: ineligible study design (18), interventions did not contain process measures, health care outcomes, structure measures, consumer or patient experiences, expert- or peer-assessed measures (11), no objective outcome data were recorded or available for one or both arms (3), or the study was about hypothetical choices (3).

2.3.4 Risk of bias in included studies

The included studies were rated on different risk of bias items as appropriate for each study design (RT, NRT, CBA, or ITS), as described in Section 2.2.1 on page 47. The results of these assessments are shown in Table 2.2 on the following page and Table 2.3 on the next page.

Allocation

The extent of possible selection bias due to the random sequence generation process was unclear in one study because the precise method of random sequence generation was not described¹⁶⁹. Rinke 2015¹⁷¹ was at high risk as they used a CBA study design. Risk of bias was judged to be low for Zhang 2016¹⁷⁰ who “flipped a coin to randomly assign” paired primary care institutions and Tu 2009¹⁶⁸ who employed a dedicated study statistician to implement a stratified randomization process. The same judgements were made for allocation concealment as for random sequence generation, except for Zhang 2016¹⁷⁰, which was judged to be at high risk for allocation concealment given their use of a coin flip.

Blinding

Although hospitals and healthcare providers could not be blinded to their allocated groups, individual participants were unlikely to have been aware that a study was taking place. No study explicitly contacted individual patients or members of the

Table 2.2: Risk of bias assessments for RCT and CBA studies

	Tu 2009	Rinke 2015	Ikkersheim 2016	Zhang 2016
Random sequence generation	+	-	?	+
Allocation concealment	+	-	?	-
Adequate blinding	-	-	-	-
Incomplete outcome data	-	+	+	+
Selecting reporting	+	+	+	+
Similar baseline characteristics	+	+	+	+
Similar baseline outcomes	+	+	+	+
Protection against contamination	-	-	+	?
Other bias	+	+	+	+

Table 2.3: Risk of bias assessments for ITS studies

	Jang 2011	Flett 2015	Joynt 2016	DeVore 2016	Liu 2017
Incomplete outcome data	+	+	+	+	+
Selective reporting	+	+	+	+	+
Intervention independent	?	+	+	?	+
Shape of effect pre-specified	+	-	-	-	-
Knowledge of interventions	+	+	+	+	+
Effect on data collection	+	+	+	+	+
Other bias	?	?	+	+	+

public to inform them about the research question, intervention, or measured outcomes. Four studies were at high risk, because providers were likely to know that a study was taking place, and it was not possible to blind them to their group allocation^{168–171}.

Incomplete outcome data

Eight of the nine included studies were judged to be at low risk of attrition bias because their outcomes were based on routinely collected administrative data, e.g. electronic prescriptions or hospital referrals. Only Tu 2009¹⁶⁸ was judged to be at high risk of bias, because five randomized hospitals withdrew due to resource constraints; one after randomization and four during follow-up. Although only a small proportion (5.8%) of the hospitals randomized in this cRT withdrew, it is plausible that poorly performing institutions would be more likely to withdraw than those with average or high performance.

Selective reporting

Only Tu 2009¹⁶⁸ registered a trial protocol with ClinicalTrials.gov (NCT00187460) in advance of undertaking the study. All outcomes described in the protocol were presented in the final report, which also included all-cause mortality as an additional outcome. It was therefore judged to be at low risk of reporting bias. Although Zhang 2016¹⁷⁰ presented a trial protocol, this was published in March 2015, i.e. 18 months after the intervention began in October 2013. None of the remaining 10 studies registered a protocol in advance of randomization (RTs and NRTs) or data analysis (ITS and CBA studies).

Other potential sources of bias

As outlined in Section 2.2.1 on page 47, the three cRTs^{168–170} and one CBA study¹⁷¹ were assessed for bias in terms of baseline characteristics, baseline outcome measures, and protection against contamination. In addition, four additional sources of bias were assessed for the ITS studies: intervention is independent of other changes, shape of the intervention is pre-specified, intervention is unlikely to affect data collection, and knowledge of the allocated interventions is adequately prevented during the study^{163–167}.

Baseline characteristics

Four studies were judged to be at low risk of bias for baseline characteristics because the intervention and control groups were shown to be similar^{168–171}.

Baseline outcome measures

All five ITS studies presented baseline outcome measures that differed between the intervention and control groups. However, they also used appropriate statistical techniques, including multivariable regression^{168–171}, and DID techniques^{168,170,171} to account for differences in baseline between the groups. They were therefore all considered to be at low risk of bias from this source.

Protection against contamination

One study was judged to be at low risk of contamination because they randomized healthcare professionals¹⁶⁹ who were not clearly in contact with one another.

Two studies were judged to be at high risk. The authors of Tu 2009¹⁶⁸ stated that several hospitals in the delayed feedback group reported that they also initiated quality improvement activities after becoming aware that performance measures were due to be released publicly. As this was not quantified, it was difficult to determine the degree to which hospitals in the control group modified their activities in anticipation of having to publicly release performance data. Rinke 2015¹⁷¹ was assessed to be at high risk because hospitals in states that did not mandate healthcare-associated infection reporting might still have modified their practice, given that such laws were being introduced elsewhere in the USA.

Zhang 2016¹⁷⁰ was judged to be at unclear risk, because no specific efforts were taken to protect against contamination. However, it is not likely that their intervention (posters on bulletin boards in outpatient areas of intervention organizations) would necessarily have influenced behaviour in control institutions.

Intervention independent of other changes

In two ITS studies, it was unclear whether the intervention occurred independently of other changes over time, or whether the outcome was confounded by other events during the study period^{163,165}. The remaining three ITS studies were judged to be at low risk of bias. In the two studies that examined public reporting of CLABSIs, this was because they analysed data from a number of states that introduced legislation at different times^{164,167}. Joynt 2016¹⁶⁶ was judged to be at low risk because they did not demonstrate a substantial change in the post-intervention period, and so this was unlikely to be attributable to other factors.

Shape of intervention effect pre-specified

One ITS study pre-specified the shape of the intervention effect and so was assessed to be at low risk of bias in this domain¹⁶³. The remaining four ITS studies did not, and were judged to be at high risk.

Knowledge of the allocated interventions adequately prevented during the study

All five ITS studies reported objective outcome measures and so were judged to be at low risk of bias for this domain.

Intervention unlikely to affect data collection

The intervention was unlikely to affect data collection in any of the five ITS studies, as all were undertaken retrospectively, using routinely collected data. In all cases, the methods of data collection were the same before and after the intervention. Therefore, all five studies were judged to be at low risk of bias.

Effects of interventions

The studies included in this review used a wide range of different interventions, which are described in Table 2.1 on page 53. The core findings are described in Table 2.4 on page 63 and the judgements about levels of certainty are justified in Appendix C on page 250.

2.3.5 Primary outcomes**Changes in healthcare decisions taken by healthcare providers**

This review provides some indication of the likely effect of public release of performance data on decision-making by healthcare professionals. There was low-certainty evidence from four studies that public release of performance data may make little or no difference to decisions taken by healthcare professionals. These studies included

three million births¹⁶³, and 67 healthcare providers^{164,169,170}. Two studies reported modest effects on some outcomes^{169,170}. Ikkersheim 2013¹⁶⁹ did not find any clear affect on referral patterns following public release of data about cataract surgery, or hip and knee replacement. However, there was a small effect on referrals for breast cancer, with GPs in the intervention group referring 1.0% more cases ($p = 0.01$) to hospitals per incremental percentage point on the report card scale of medical effectiveness. Similarly, Zhang 2016¹⁷⁰ found that the effect of displaying prescription performance data in outpatient areas varied across outcomes and disease groups. Public release of performance data did not change the number of prescriptions containing antibiotics in the bronchitis group, two or more antibiotics in the gastritis group, injections in the hypertension group, or antibiotic injections in the bronchitis and hypertension groups. Similarly, the average prescription cost did not change for patients with hypertension. However, public release of performance data did appear to reduce prescriptions containing antibiotics for gastritis (intervention effect -12.7%, $p < 0.001$), two or more antibiotics for gastritis (-3.8%, $p = 0.005$), injections for gastritis (-10.6%, $p < 0.001$), and antibiotic injections for gastritis (-10.7%, $p < 0.001$). Average antibiotic prescription cost fell for patients with bronchitis (-7.9%, $p < 0.001$) and gastritis (-5.7%, $p = 0.005$). These mixed findings were also complicated by evidence that public release of prescribing data increased prescriptions containing antibiotics for patients with hypertension (intervention effect 2.0%, $p = 0.08$), and injections for bronchitis (2.0%, $p = 0.012$).

One study found that the first public release of hospital caesarean section rate data may have slightly reduced the number of patients undergoing this procedure (-0.8%, $p < 0.01$), and that this persisted until the end of the study, 20 months later¹⁶³. However, further public releases of data did not exhibit any further effect on caesarean section rates.

Finally, Flett 2015¹⁶⁴ did not find any evidence that mandatory public reporting of CLABSI had any effect on blood culture testing or antibiotic utilization in paediatric and neonatal intensive care units in the United States.

Table 2.4: Summary of findings table

Outcome	Impact	Clinical encounters (studies)	Certainty (GRADE)
Changes in healthcare decisions taken by healthcare providers (professionals and organizations)	Little or no difference to decisions taken by healthcare professionals. Two studies (2 cuts) found that some decisions might be affected. One study (ITS) found that decisions might be influenced by the initial release of data, but that subsequent releases might have less impact	3,000,000 births and 67 providers (4: 2 cRTs, 2 ITS)	Low
Changes in the healthcare utilization decisions of purchasers	No studies reported this outcome	-	-
Changes in provider performance	Little or no difference to objective measures of provider performance	82 healthcare providers (1: 1cRT)	Low
Changes in patient outcome	May slightly improve patient outcomes	315,092 hospitalizations and 7,503 healthcare providers (5: 1 cRT, 3 ITS, 1 CBA)	Low
Adverse effects	No studies reported this outcome	-	-
Impact on equity	No studies reported this outcome	-	-

Changes in provider performance

This review provides some indication of the likely effect of public release of performance data on healthcare provider performance. There was low-certainty evidence from one study that public release of performance data may make little or no difference to objective measures of provider performance. Tu 2009¹⁶⁸ included data from 82 healthcare providers and found that a media campaign and release of hospital performance data online had no effect on 11 of 12 AMI process-of-care quality indicators. The 12th AMI quality indicator (fibrinolytics given prior to transfer to the coronary care unit (CCU) or intensive care unit (ICU)) increased by 5.8% ($p = 0.02$), although no statistical correction was made for multiple hypothesis testing. Similarly, public release of performance data did not clearly affect five of six congestive CHF quality indicators, although the sixth (angiotensin-converting enzyme inhibitor (ACEi) or angiotensin receptor blocker (ARB) for left ventricular dysfunction (LVF)) increased by 5.9% ($P = 0.02$). Neither the AMI nor CHF composite process-of-care quality indicators improved following the public release of performance data.

The main outcomes in two studies described above, are sometimes considered evidence of provider performance^{163,170}. However, as these outcomes (caesarean section and antibiotic prescribing) may be appropriate clinical decisions, they are not direct evidence of poor performance and so these were instead considered in Section 2.3.5 on page 61.

Changes in patient outcome

Low-certainty evidence from five studies suggested that public release of performance data may slightly improve patient outcomes. The certainty of evidence was graded as low because the evidence was mixed, with two studies reporting improvements^{167,168} and three finding no evidence of improved patient outcomes^{165,166,171}. These five studies included 7,503 healthcare providers and 315,092 hospitalizations. Two studies reported that patient outcomes were not changed by publication of

hospital-level quality metrics on *Hospital Compare*, which is a website run by the Centers for Medicare & Medicaid Services (CMS)^{165,166}. DeVore 2016¹⁶⁵ did not find any evidence that publication of hospital re-admission rates had an effect on 30-day re-admissions for patients with AMI, heart failure, or pneumonia. Similarly, Joynt 2016¹⁶⁶ reported a very small slowing in a pre-existing trend (change 0.13% per quarter; 95% confidence interval (CI) 0.12% to 0.14%) towards reduced 30-day mortality following publication of mortality rates on *Hospital Compare*.

Rinke 2015¹⁷¹ did not find any evidence that mandatory hospital reporting of CLABSI affected the rate of paediatric CLABSIs. However, Liu 2017¹⁶⁷ reported a 34% reduction (incidence rate ratio 0.66, $p < 0.001$) in adult CLABSIs after mandatory reporting, when compared with the 25-month period before each state introduced legislation. This discrepancy between the findings of Rinke 2015¹⁷¹ and Liu 2017¹⁶⁷ might reflect a genuine difference in terms of impact on children and adult CLABSI rates. Importantly, both studies found that CLABSI rates declined across the USA during their study period, including in states that did not introduce mandatory reporting. It is unclear whether public release of performance data in some states contributed to this national decline, even within states that did not introduce mandatory reporting. Tu 2009¹⁶⁸ found that public release of hospital performance data online and through the media was associated with a 2.5% reduction in 30-day mortality ($p = 0.045$) for patients with AMI, although no such effect was observed in patients with CHF.

2.3.6 Secondary outcomes

No included studies reported data that could be used to determine whether public release of performance data can change the behaviour of purchasers, lead to unintended adverse effects, or impact on equity.

2.4 Discussion

2.4.1 Summary of main results

Changes in healthcare decisions taken by healthcare providers

There was low-certainty evidence with mixed findings from four studies, which reported either modest effects^{163,169,170}, or no effect¹⁶⁴, on healthcare decisions taken by healthcare providers. Two studies found evidence that public release of performance data had modest effects on some of the healthcare decisions taken by healthcare providers, but not all of the decisions measured^{169,170}. One study found that the first public release of data had a small but sustained effect on caesarean rates, although subsequent releases did not affect the rate any further¹⁶³.

Changes in provider performance

There was low-certainty evidence from one study that informed conclusions about the effect of public release of performance data on provider performance. A single RT addressed this question, and found that 2/18 (11.1%) of measured processes appeared to improve in the intervention hospitals¹⁶⁸. However, as no correction was made for multiple hypotheses testing¹⁷², this did not provide convincing evidence that provider performance was affected by public release of performance data.

Changes in patient outcome

Low-certainty evidence showed that five studies that included patient outcomes had inconsistent findings, with two reporting improvements^{167,168} and three reporting no difference^{165,166,171}.

Secondary outcomes

There were no studies that considered the effect of public release of performance data on changes in the healthcare utilization decisions of purchasers, unintended effects or harms, or on healthcare equity.

2.4.2 Overall completeness and applicability of evidence

There are many systems around the world that use public release of performance data. However, only a small proportion were represented in this review and so it is likely that most have either not been evaluated or were subject only to low-quality evaluations. It is notable that some interventions have been evaluated more robustly than others, with two studies in this review considering the CMS website *Hospital Care*^{165,166}, and three, the introduction of state-based mandatory reporting of CLABSIs^{164,167,171}. Similarly, the majority of the studies included in this review (6/9, 67%) were based in North America, with no representation at all from South America, Africa, or Australasia. It is therefore likely that a small number of initiatives have attracted a disproportionate number of studies, and there is clearly work that needs to be done to robustly evaluate similar interventions in other settings. There was also insufficient evidence to draw any conclusions about the healthcare utilization decisions of purchasers, adverse effects, or impact on equity.

One new study was included from Canada¹⁶⁸, which was published after the latest systematic reviews by Fung 2008¹³⁶, Shekelle 2008¹⁵³, and Faber 2009¹⁵⁴. The included studies evaluated interventions that used data that might have been originally collected for a purpose other than influencing behavior or improving outcomes. It is possible that custom-made interventions, using data collected for the specific purpose of influencing behaviour or improving outcomes, would have a greater impact. However, the lack of such interventions in the literature highlighted the fact that their delivery may be excessively resource intensive, and that future initiatives aimed at public release of performance data will continue to draw on data initially collected for a different purpose.

2.4.3 Certainty of the evidence

The certainty of the evidence that examined the effect of public release of performance data was low. Only 3/9 included studies (33%) were RTs and so the evidence for these outcomes was partly informed by non-randomized study designs. The use

of EPOC study design criteria ensured that all included observational studies took steps to minimize the risk of bias¹⁵⁵.

However the criteria proposed by the Cochrane EPOC Group includes ITS studies with three data points before and after the intervention, which is a pragmatic recommendation. There is emerging evidence to suggest that the number of data points alone should not be used to determine whether or not a ITS study is adequately powered¹⁷³. It is therefore possible that some of the ITS studies included in this review were underpowered to detect an intervention effect. There was also considerable heterogeneity across the settings, outcomes, and modes of public release; and inconsistent effects reported between studies. In terms of secondary outcomes, there were no studies that set out to consider behavior change on the part of purchasers, unintended adverse effects, or impact on equity.

2.4.4 Potential biases in the review process

Although the search was comprehensive, it is possible that relevant studies were missed. However, this risk was minimized by asking an Information Specialist to help design and implement the search strategy, and ensured that two researchers independently examined all items retrieved from the search. Data extraction and “risk of bias” assessments were independently undertaken by two researchers. Although the GRADE assessments were determined by a single researcher, these were checked by all members of the team, and disagreements resolved through discussion. These steps ensured that potential biases in the review processes were mitigated as much as possible. However, this stringent approach to study collection also meant excluding most of the studies that have evaluated public release of performance data in other settings. It is possible that this approach biased the review against settings that were less likely to deliver studies that satisfied the EPOC inclusion criteria, and this might have accounted for the over-representation of studies from North America, Europe, and Asia. It might also have led to the exclusion of studies (e.g. those utilising qualitative designs) that contained important information about the impact of

public release of performance data. However, it was necessary to limit the review to studies that were at the lowest possible risk of bias, to maximize the certainty of its findings. There may nevertheless be scope for future reviews to synthesize evidence from studies using a broader range of designs.

2.4.5 Comparison with other studies or reviews

The systematic literature search identified two relevant systematic reviews^{129,154}. The findings in this chapter agreed with these earlier publications that previous studies were limited by risk of bias, inconsistent findings, and heterogeneity of interventions, healthcare settings, and outcomes.

Campanella 2016¹²⁹ attempted a meta-analysis of data from 10 studies, and reported improved mortality (risk ratio 0.85, 95% CI 0.79 to 0.92). However, this finding was reported in the context of very high heterogeneity ($p < 0.0001$; $I^2 = 100\%$). The authors limited their meta-analysis to studies that reported sufficient data, and excluded those with inappropriate study designs, or those that were at high risk of bias. In contrast, the review presented in this chapter only considered studies that proffered the highest certainty of evidence, and did not consider a meta-analysis appropriate in view of the considerable degree of heterogeneity between studies (see *Assessment of heterogeneity*, Section 2.2.1 on page 48). Instead, the findings in this chapter were consistent with those of Fung 2008¹³⁶, which concluded, “studies of the effect of public reporting on outcomes provide mixed signals, and the usefulness of public reporting in improving patient safety and patient centeredness remains unknown, because few studies assessed these end points”.

2.4.6 Conclusions

This review found that the existing evidence base is inadequate to directly inform practice.

In order to understand the effectiveness of the public release of performance data, there is a need for more longitudinal studies with robust evaluation designs. In

particular, the evidence base would benefit from more studies that consider whether public release of performance data can improve patient outcomes, rather than simply healthcare processes. Further work should also specifically consider whether public release of performance data might result in adverse effects or harms.

Unfortunately, most studies were unable to guarantee that disseminated performance data actually reached its intended audience, i.e. that lack of effect was not simply a result of failed exposure to the intervention.

Therefore, future studies should consider carefully how they might maximize the number of people exposed to their intervention, and whether this can be quantified. However, the effect of public release of performance information in the “real world” is likely to be limited by difficulties in reaching its intended audience^{130–135,174}. Therefore, the need to ensure that performance data reach those who are intended to be influenced, needs to be balanced against the risk of reducing study validity by creating artificial conditions that cannot be replicated when the intervention is used in practice.

Further work is clearly required to delineate whether, how, and when public release of performance data affects patient outcomes. This will form part of the research question addressed by Chapter 4 on page 101 in relation to publication of NHFD data.

Although this review did not find evidence in support of publishing performance data, it also could not show that such interventions do *not* achieve their intended aims. Importantly, there was not any evidence of unintended consequences or harms. In an era of increasing transparency and accountability, it is likely that public release of performance data will become more commonplace in the absence of clear evidence that it does not work or is harmful. For example, the new online platform *My NHS* publishes a range of hospital performance measures, including at the level of individual consultants, for all major surgical specialties¹⁷⁵. Within orthopaedic surgery, this includes data for patients undergoing elective arthroplasty (ankle, elbow and shoulder, hip, and knee) as well as presenting with major trauma and hip

fracture.

As discussed above in Section 2.1 on page 42, the NHFD has publicly released performance metrics (including adjusted 30-day mortality) at the level of individual hospitals since 2010¹²⁴. These data appear in the NHFD annual reports¹²³ as well as online through *My NHS*¹⁷⁵. If publication of hospital-level outcome data is to positively improve performance, it is clearly important that outliers are correctly identified, which requires properly taking into account differences in case mix. Confidence in this process is also necessary to overcome professional objections to public reporting of performance data^{147–150}. This task will form the focus of Chapter 3 on page 72.

Chapter 3

Optimising case mix adjustment in the NHFD

3.1 Introduction

Chapter 2 on page 42 could not find any evidence to refute the claim that public release of performance data can improve healthcare performance and patient outcomes. A number of national clinical audits now publish outcomes at the level of NHS organizations or even individual clinicians¹¹⁶. One criticism of such initiatives is that differences in patient outcome may be attributable to case mix^{145–147}. A number of risks flow from inappropriately labelling healthcare providers as delivering poor care. First, hospitals that are actually delivering poor care (but that benefit from a correspondingly favourable case mix) may slip through the net and fail to be identified as performance outliers¹⁷⁶. Second, hospitals with an unfavourable case mix may have resources withdrawn (e.g. under pay-for-performance initiatives – see Chapter 4 on page 101) or be inappropriately labelled as “failing”, which has consequences for staff morale and recruitment^{177,178}. Finally, healthcare providers may choose to disregard audit findings and so explain their outlier status as an artefact that is unrelated to quality of care¹⁷⁹¹.

¹Pilot work from this chapter has been published as Metcalfe D, Masters J, Delmestri A, Judge A, Perry DC, Zogg CK, Gabbe BJ, Costa ML. Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. *BMC Med Res Methodol.* 2019;19(1):115.

3.1.1 Risk adjustment in the NHFD

Case mix adjustment is particularly important in the hip fracture population as these patients are typically characterized by advanced age, frailty, and multi-morbidity (see Section 1.5 on page 27). The first NHFD reports were based on data analyses undertaken by Quantics Consulting Ltd using a classification and regression tree approach for risk adjustment¹⁸⁰. However, the 2014 annual report was based on work from the Clinical Effectiveness Unit (CEU) at the Royal College of Surgeons of England. The CEU tested a number of different approaches and selected a logistic regression model using the variables age, sex, admission source, walking ability indoors, fracture type, and American Society of Anesthesiologists (ASA) grade¹⁸⁰. The last of these variables is the only measure of co-morbidity included within the NHFD risk adjustment model. Although Abbreviated Mental Test Score (AMTS) is collected by the NHFD and could act as a further measure of co-morbidity (by highlighting cognitive impairment), it is not currently used for risk adjustment.

3.1.2 ASA Physical Status Classification System

The ASA Physical Status Classification System was first developed in 1941 as a means of describing the physical status of patients preparing to undergo surgery¹⁸¹. This ranges from 1 (“normal healthy patient”) to 5 (“moribund patient unlikely to survive 24 hours”):

1. A normally healthy patient
2. A patient with mild systemic disease
3. A patient with severe systemic disease that is not incapacitating
4. A patient with incapacitating severe disease that is a constant threat to life
5. A moribund patient who is not expected to survive for 24 hours with or without an operation

Incorporation of ASA grade into risk scores

ASA physical status forms a central part of a number of risk scores aimed at predicting surgical mortality. These include the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) surgical risk calculator, which is available for 1,557 separate operations and intended to support consent conversations between surgeons and patients¹⁸². ASA physical status is also included within the National Emergency Laparotomy Audit (NELA) score, which is used by one of the national audits run by HQIP¹⁸³ (see Section 1.7 on page 37). It is also incorporated into measures such as the Surgical Risk Scale¹⁸⁴ and Surgical Outcome Risk Tool¹⁸⁵, which are intended to support audits of surgical outcomes.

Limitations of ASA

There are a number of problems with the ASA system¹⁸⁶. First, there are no clear distinctions between the categories, and grades are therefore assigned subjectively. Unsurprisingly, most studies have reported considerable inter-observer variability between anaesthetists when selecting ASA grades^{187,188}. Second, it is a simple system that does not readily accommodate complexity¹⁸⁶. For example, a patient with chronic obstructive pulmonary disease (COPD) that does not limit their mobility may be assigned ASA grade 3. However, a patient with many such co-morbidities (e.g. COPD, heart failure, diabetes, and metastatic cancer) might be assigned the same grade if their diseases were not incapacitating. Finally, most versions of ASA only include five grades, and so lack precision when used as a tool for comparing risk between many different patients¹⁸⁸.

3.1.3 Alternative co-morbidity measures

A number of co-morbidity summary measures have been developed to help classify patients according to their overall disease burden^{189–192}.

The most commonly used summary measure is the Charlson co-morbidity index (CCI)¹⁹¹. Charlson et al identified 17 diseases that optimally predict one-year mor-

tality when assigned a weight between 1 (e.g. peripheral vascular disease) and 6 (e.g. metastatic cancer)¹⁸⁹. Although the CCI is commonly used¹⁹¹ and has been widely validated¹⁹³, it was developed in the 1980s and has been criticized as outdated¹⁹⁴. A number of meta-analyses have found that an alternative summary measure proposed by Elixhauser et al¹⁹⁰ has superior predictive properties^{191,192}. In particular, the Elixhauser co-morbidity index (ECI) predicts mortality more effectively than CCI amongst patients with fractures of the cervical spine¹⁹⁵ and proximal humerus¹⁹⁵. However, although older adults with hip fractures have a high co-morbid disease burden, it is unclear which summary measure optimally predicts mortality in this population. The ECI is similar to the CCI (nine categories overlap the two measures: diabetes [uncomplicated and complicated], congestive heart failure, human immunodeficiency virus (HIV), metastatic cancer, renal disease, chronic pulmonary disease, rheumatic disease, and peripheral vascular disease) but includes almost twice as many diagnostic categories¹⁹⁶.

However, both the CCI and ECI were developed in general inpatient populations^{190,191}. One concept that might have been poorly captured by previous co-morbidity summary measures is frailty, which has been defined as a “state of increased vulnerability to poor resolution of homeostasis following a stress, which increases the risk of adverse outcomes including falls, delirium and disability”¹⁹⁷. There are a number of efforts afoot to define “frailty” in administrative datasets using diagnostic codes. A recent study described a Hospital Frailty Risk Score (HFRS), which is based on International Classification of Diseases, 10th Revision (ICD-10) codes and specifically developed for use in older adult inpatient populations¹⁹⁸.

Logistics of using alternative co-morbidity measures

ASA is typically collected as part of the routine pre-operative assessment of all patients with hip fractures. It is therefore relatively straightforward to include this single number as part of each NHFD submission. Alternative co-morbidity measures are more cumbersome as they require additional effort on behalf of clinical coders,

e.g. to screen notes for evidence of 30 separate Elixhauser co-morbidities¹⁹⁹. However, such data could potentially be incorporated into the NHFD risk adjustment models if a routine linkage was established between the NHFD and an administrative dataset, such as HES.

3.1.4 Aims

The aims of this study were to:

1. Determine whether incorporation of CCI, ECI, HFRS could improve the performance of the current NHFD risk adjustment model.
2. Determine whether AMTS could enhance performance of the existing NHFD model.
3. Determine whether the ability of the NHFD to identify performance outliers is sensitive to choice of summary co-morbidity measure and inclusion of AMTS.

3.2 Methods

The study protocol was approved by HQIP prior to data release but research ethics committee approval was not sought for secondary analysis of administrative data in line with Governance Arrangements for Research Ethics Committees (GafREC) guidelines²⁰⁰.

3.2.1 Data source

In this study, data from the NHFD was linked to records in the HES APC and civil registration death records.

The NHFD

The NHFD has already been described in Section 3 on page 72.

Hospital Episode Statistics

The HES APC dataset is managed by NHS Digital and collects data on all admissions to NHS hospitals as well as those treated in private hospitals but funded by the NHS²⁰¹. Approximately 98-99% of hospital activity in England is funded by the NHS²⁰². It is unlikely that many older adults with hip fractures were treated in the private sector during the study period. The HES APC dataset does not include information regarding ED attendances that do not lead to admission²⁰¹.

Office for National Statistics

The ONS holds data on all deaths registered in England and Wales. All English deaths should be captured, although registration could be delayed in cases referred to a coroner for *post mortem* or inquest. In 2016, upwards of 96% of deaths were registered within the year that they occurred²⁰³.

3.2.2 Inclusion criteria

This study included all patients aged ≥ 60 years that presented to hospital between 1st January 2016 and 31st December 2016 with a hip fracture. Patients treated at hospitals outside England were excluded as these could not be reliably linked to records in HES. Similarly, patients treated non-operatively were excluded as these are treated separately by NHFD annual reports.

3.2.3 Variables and outcomes

The variables extracted from the NHFD were those currently used in the existing risk adjustment model, i.e. age, sex, admission source, walking ability indoors, fracture type, and ASA grade¹⁸⁰.

ICD-10 diagnostic codes from HES APC were used to determine CCI, ECI, and HFRS. The codes used for CCI are found in Table G.1 on page 266, ECI in a paper by Quan et al¹⁹⁶, and HFRS in a paper by Gilbert et al¹⁹⁸. Age and co-morbidity summary scores were treated as continuous variables throughout²⁰⁴.

The primary outcome was 30-day mortality, which is the same as that used in the NHFD annual reports²⁰⁵.

3.2.4 Statistical analysis

The distribution of continuous variables was assessed visually using kernel density plots. Age was not linearly associated with outcome and so a fractional polynomial was fitted using the *fracpoly* module – and assessed visually using the *fracplot* command – in Stata.

Missing variables and imputation

Variable missingness was explored visually using cross-tabulation and the missing completely at random (MCAR) assumption formally evaluated using Little’s Chi-square test, which tests the null hypothesis that the pattern of missingness does not depend on the data values²⁰⁶. The distribution of missing values for components of the risk adjustment model are shown in Table 3.1 on page 82. There were no missing data for age or sex. As the MCAR assumption was not satisfied for the variables with missing data, multiple imputation was used in preference to listwise deletion²⁰⁷.

Multiple imputation was undertaken using sequential chained equations based on age, sex, and 30-day mortality. This technique operates under the assumption that data are missing at random (MAR), i.e. that missingness depends only on observed rather than unobserved values. Under this assumption, it should be possible to control for known values as any residual missingness will be MCAR²⁰⁸. Ordered logistic regression models were fitted for the ordinal variables (i.e. admission source, walking ability indoors, and ASA grade). An augmented regression approach was specified in the event of perfect prediction being detected, as recommended by White et al²⁰⁹.

Comparing prediction models

Logistic regression models were fitted with 30-day mortality as the dependent variable. The co-variables were those included in the NHFD risk adjustment model, as described in Subsection 3.1.1 on page 73. The models that included alternative co-morbidity summary measures were then layered on top of this base model. Model discrimination and calibration were then assessed. In this context, discrimination is a measure of how well a model can distinguish between those patients that survived and those that died within 30 days. Calibration is the ability of the model to correctly estimate the probability of future events, i.e. the extent to which predicted and observed proportions of 30-day mortality are the same²¹⁰.

Model discrimination was assessed using tests of equality for receiver operating characteristic (ROC) areas, which were undertaken using the *roccomp*²¹¹ module in Stata. The AUROC technique plots randomly drawn pairs of observations with different outcomes and describes the proportion of times in which a patient that died had a higher predicted risk of mortality than their paired survivor²¹⁰. The conventional thresholds for assessing discrimination using AUROC were adopted - 0.7-0.8 “acceptable”, 0.8-0.9 “excellent” and >0.9 “outstanding”²¹².

Model calibration was assessed using the Hosmer-Lemeshow test with ten predicted risk groups. This test evaluates whether or not observed and expected event rates match across a defined number of sub-groups within the population based on categories of risk²¹³. However, the Hosmer-Lemeshow test has been criticized as being sensitive to the arbitrary choice of risk categories and for only producing a p-value²¹⁴. Calibration belts were therefore plotted, which visualize the relationship between predicted and observed mortality rates across the range of probabilities together with confidence limits²¹⁵.

Elixhauser et al originally proposed including all co-morbidities as separate independent variables within regression models¹⁹⁰. However, this approach is unlikely to be practical for the NHFD, which benchmarks outcomes across many hospitals, including small units that report few deaths each year. An approach that incor-

porates 30 additional independent variables into the risk adjustment model would likely be under-powered to detect excess deaths in some hospitals²¹⁶. Elixhauser co-morbidities were therefore included as a single index using the weights proposed by van Walraven et al²¹⁶.

Identifying performance outliers

Performance outliers were identified using the risk adjustment model currently used by the NHFD. This study aimed to compare this list of outliers to those identified by the best performing model to determine what changes might be expected if a routine NHFD-HES linkage were established to inform future annual reports. The distribution of observed and adjusted high mortality outliers was visualized using funnel plots with 95% and 99.8% confidence limits. As specific permission was not sought to name individual hospitals, these have been anonymized throughout the thesis.

Information governance

Ethical approval was not sought in line with the latest GafREC guidance²¹⁷. Personal data was processed under Articles 6(1)(f) and 9(1)(f) of the General Data Protection Regulation (EU 2016/679) (GDPR).

3.3 Results

There were 56,191 cases of hip fracture within the NHFD that satisfied the inclusion criteria described in Subsection 3.2.2 on page 77.

3.3.1 Distribution and missingness across key variables

Table 3.1 on page 82 shows the proportion of missingness across variables that may be suitable for risk adjustment within the NHFD. Overall, levels of missing data were low with the most affected variable (AMTS) only missing in 2.1% of cases.

There was not any missing data for either age or sex.

All three co-morbidity measures (CCI, ECI, HFRS) were distributed similarly with a pronounced leftward skew (Figure 3.1 for the HFRS but also Figure D.1 on page 252 for CCI and Figure D.2 on page 253 for ECI). The median scores for each measure were: CCI 1 (IQR 1-3), ECI 2 (1-4), and HFRS 10.6 (5.3-17.9).

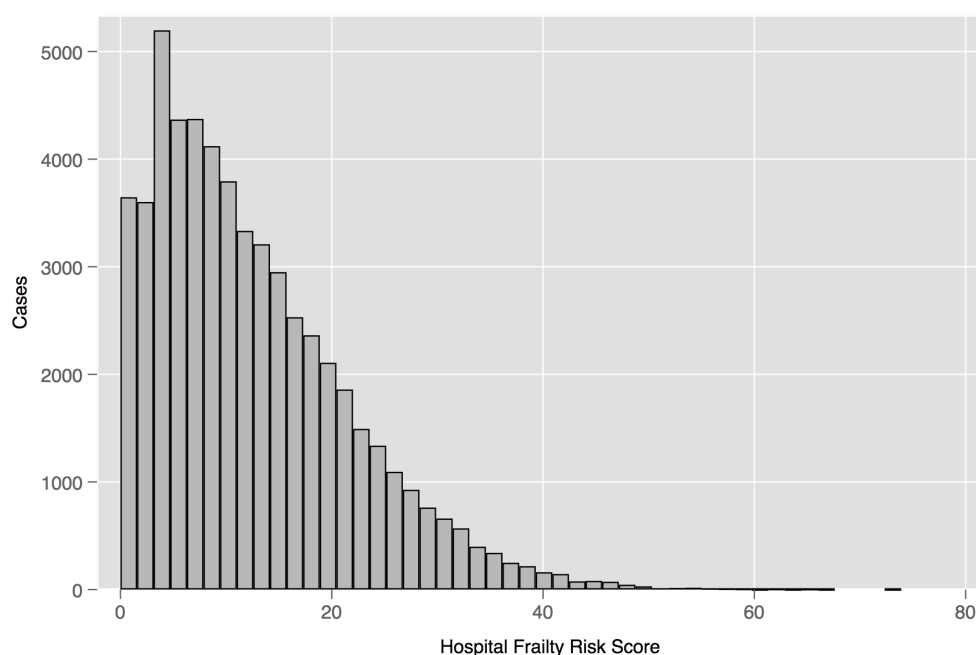


Figure 3.1: Distribution of the Hospital Frailty Risk Score in the NHFD.

3.3.2 Current NHFD risk adjustment model

The risk adjustment model currently used by the NHFD exhibited acceptable discrimination (0.752, 95% CI 0.474-0.758), although the Hosmer-Lemeshow test suggested that it may be miscalibrated ($p < 0.001$). Figure 3.2 on page 83 shows the ROC curve for the current base model and Figure 3.3 on page 83 the calibration belt plot. The latter shows that the existing model under-estimates risk for patients with probabilities of mortality in the ranges 0.00-0.03 and 0.22-0.69 and over-estimates those for patients in the probability range 0.05-0.14. However, as the distribution of estimated probabilities under the model was left-skewed (Figure 3.4 on page 84), miscalibration affecting lower probability ranges involved far more patients than those in the higher probability ranges. The current NHFD model under-estimated

Table 3.1: Key variables available from the NHFD

	Missing (%)	Distribution	
Sex	0 (0.0%)	Female	40,059 (71.3%)
		Male	16,132 (28.7%)
Admission source	24 (0.04%)	Own home	45,516 (81.0%)
		Residential care	6,456 (11.5%)
		Nursing care	4,195 (7.5%)
Walking ability indoors	490 (0.9%)	Freely mobile	20,571 (36.9%)
		Mobile indoors with one aid	12,234 (22.0%)
		Mobile outdoors with two aids	8,294 (14.9%)
		Indoor mobility but housebound	13,963 (25.1%)
		No functional mobility	639 (1.2%)
Fracture type	433 (0.8%)	Intertrochanteric (A1/A2)	16,732 (30.0%)
		Intertrochanteric (A3)	1,427 (2.6%)
		Intertrochanteric (unknown)	1,622 (2.9%)
		Intracapsular (undisplaced)	5,030 (9.0%)
		Intracapsular (displaced)	27,607 (49.5%)
		Subtrochanteric	3,340 (6.0%)
Age	0 (0.0%)	Median 84 (IQR 78-89)	
ASA	1,135 (2.0%)	Median 3 (IQR 2-3)	
AMTS	1,196 (2.1%)	Median 9 (IQR 5-10)	

risk for 19,861 individual patients (35.3% of the total NHFD) and over-estimated risk for 21,123 (37.6%).

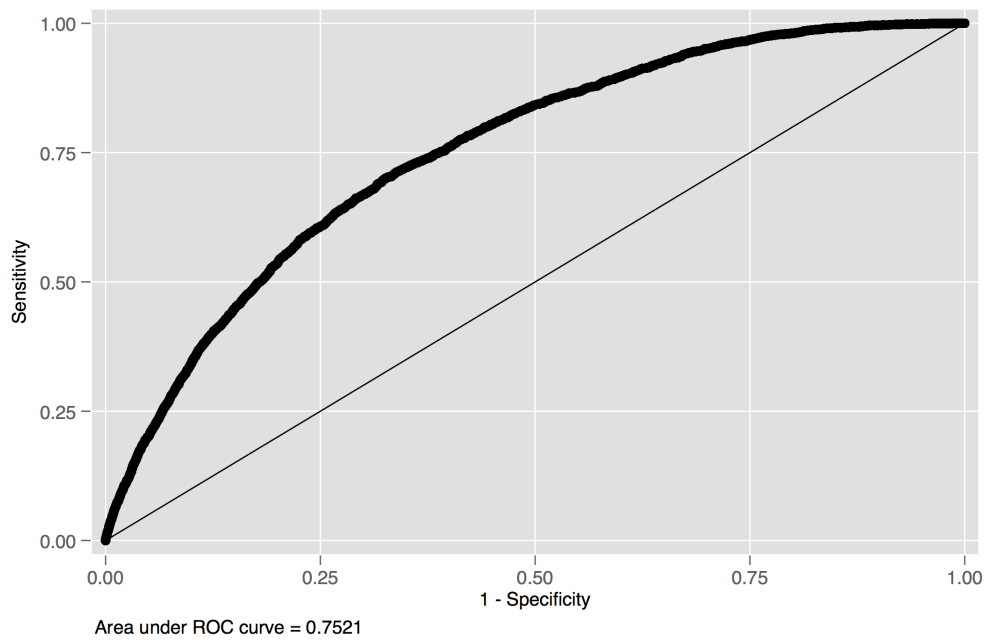


Figure 3.2: Receiver Operating Characteristic curve for the existing NHFD risk adjustment model.

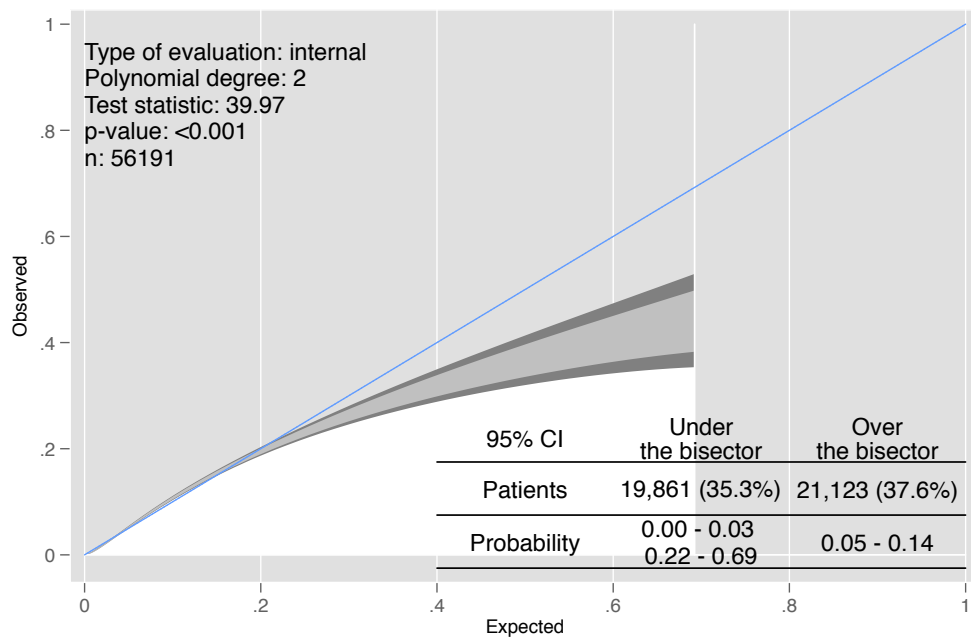


Figure 3.3: Calibration belt plot for the existing NHFD risk adjustment model.

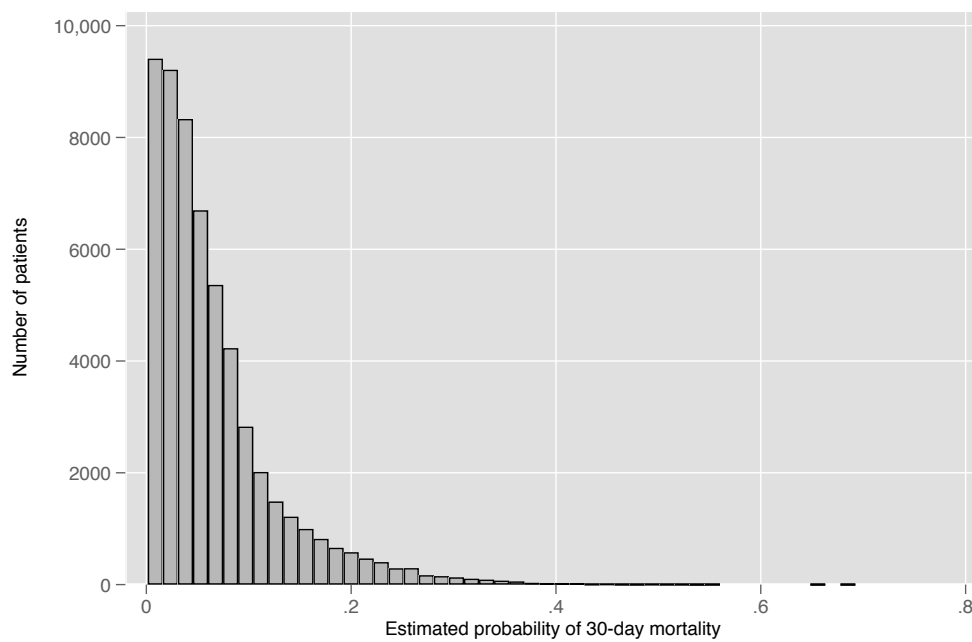


Figure 3.4: Histogram showing the distribution of estimated probabilities of 30-day mortality using the existing NHFD risk adjustment model.

Table 3.2: Performance characteristics for risk adjustment models

Model	Discrimination		Calibration	
	AUROC**	ROC curve	H-L test***	Calibration belt
Base*	0.752 (0.474-0.758)	Figure 3.2 on page 83	<0.001	Figure 3.3 on page 83
Base -ASA +CCI	0.722 (0.716-0.727)	Figure D.3 on page 253	<0.0001	Figure D.4 on page 254
Base -ASA +ECI	0.711 (0.706-0.717)	Figure D.5 on page 254	<0.0001	Figure D.6 on page 255
Base -ASA +HFRS	0.715 (0.709-0.721)	Figure 3.5 on the following page	0.0176	Figure 3.6 on page 87
Base +ASA +CCI	0.758 (0.751-0.765)	Figure D.7 on page 255	<0.0001	Figure D.8 on page 256
Base +ASA +ECI	0.753 (0.746-0.761)	Figure D.9 on page 256	<0.0001	Figure D.10 on page 257
Base +ASA +HFRS	0.764 (0.757-0.772)	Figure 3.7 on page 88	0.0444	Figure 3.8 on page 89
Base +ASA +CCI +HFRS	0.774 (0.767-0.781)	Figure 3.9 on page 90	0.0142	Figure 3.10 on page 90
Base +ASA +ECI +HFRS	0.769 (0.762-0.777)	Figure D.11 on page 257	0.0147	Figure D.12 on page 258
Base +ASA +AMTS	0.751 (0.744-0.759)	Figure D.13 on page 258	<0.0001	Figure D.14 on page 259
Base +ASA +ECI +HFRS +AMTS	0.781 (0.774-0.788)	Figure 3.11 on page 91	0.0011	Figure 3.12 on page 91

*Current NHFD risk adjustment model; **AUROC (95% confidence intervals); ***Hosmer-Lemeshow p -value

3.3.3 Alternative co-morbidity summary measures

ASA substituted for alternative co-morbidity measures

These comparisons are shown in Table 3.2 on the preceding page. All three alternative co-morbidity measures led to reduced discrimination of the model when compared with ASA. However, replacing ASA with the HFRS resulted in the best calibrated model presented in this chapter. The ROC curves and calibration belt plots for the model using HFRS are shown in Figure 3.5 and Figure 3.6 on the following page respectively. The plots not shown in this section are available in Appendix D on page 252.

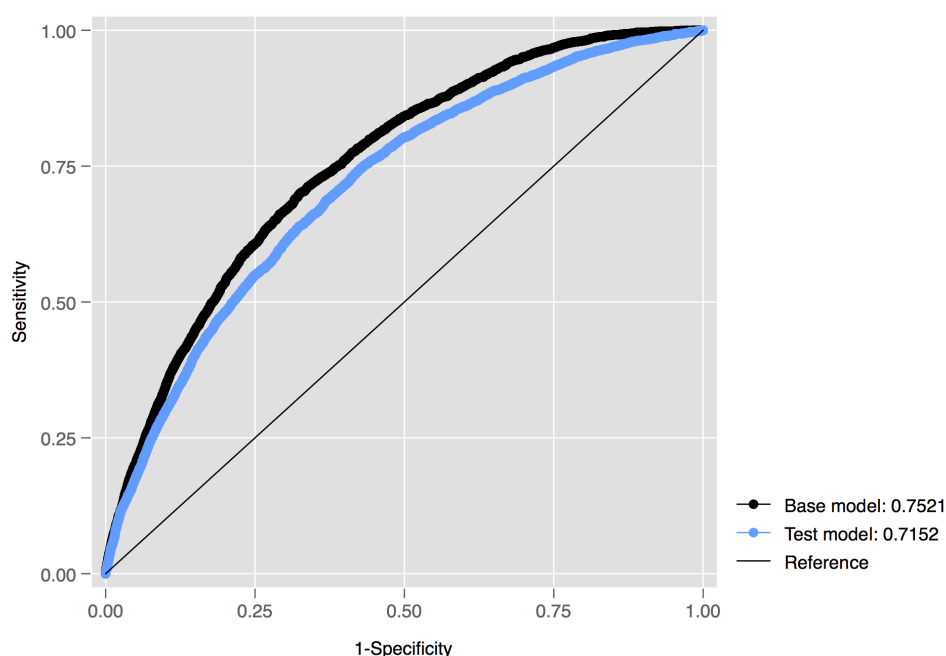


Figure 3.5: Receiver Operating Characteristic curves showing the NHFD model versus a model in which ASA has been replaced by the Hospital Frailty Risk Score.

ASA supplemented with additional co-morbidity measures

The best performing alternative co-morbidity measure was the HFRS when used in conjunction with the full NHFD base model (0.764 [0.757-0.772] versus 0.752 [0.745-0.760]). The ROC curves are shown in Figure 3.7 on page 88 and calibration belt plot in Figure 3.8 on page 89. The ROC curves and calibration belt plots not shown in this chapter are available in Appendix D on page 252.

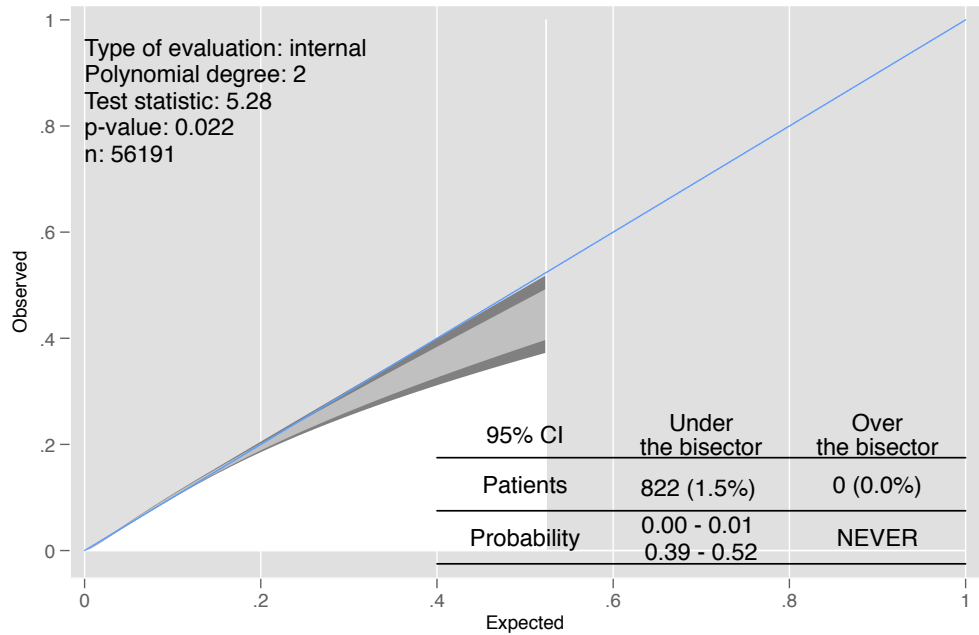


Figure 3.6: Calibration belt plot for a variation on the existing NHFD model in which ASA has been replaced by the Hospital Frailty Risk Score.

Using combinations of variables

The best performing model in terms of discrimination was the NHFD base model together with CCI and HFRS (0.774 [95% CI 0.767-0.781]). The ROC curves are shown in Figure 3.9 on page 90 and the calibration belt plot in Figure 3.10 on page 90. The plots for ECI and HFRS are available in Appendix D on page 252.

3.3.4 Incorporating AMTS

Including AMTS did not significantly increase the discrimination of the base model (0.751 [0.744-0.759] versus 0.755 [0.747-0.762] - see Figure D.13 on page 258 and Figure D.14 on page 259). However, when combined with the best performing model identified in Subsection 3.3.3, AMTS further improved discrimination (0.781 [0.774-0.788] versus 0.774 [0.767-0.781]). The properties of this model are shown in Figure 3.11 on page 91 and Figure 3.12 on page 91.

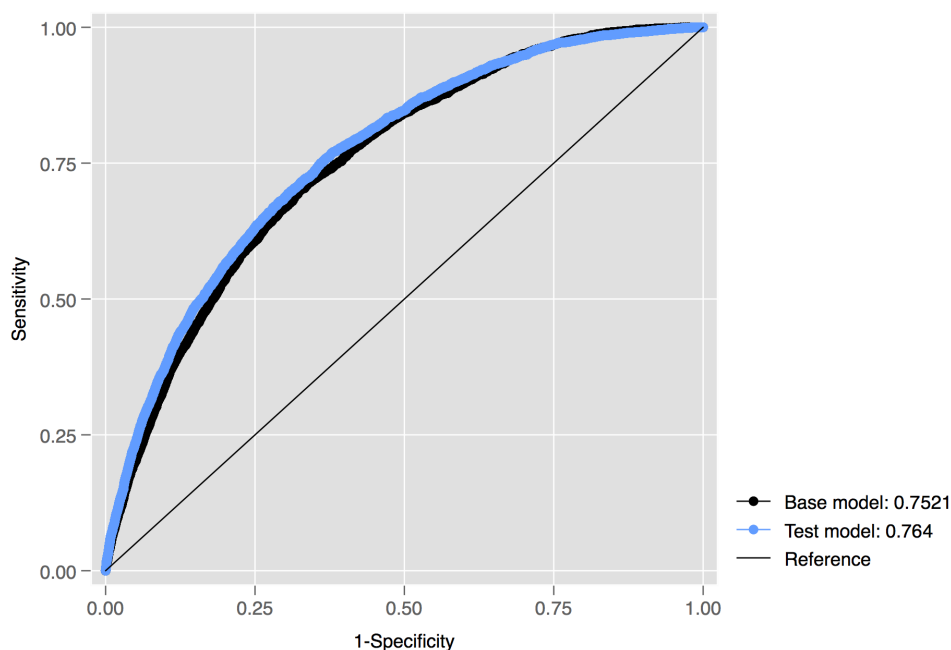


Figure 3.7: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with the Hospital Frailty Risk Score.

3.3.5 Identification of high mortality outliers

Three models were selected for further evaluation to determine whether or not choice of model affected identification of high mortality outliers.

1. Existing NHFD risk adjustment model.
2. Base model with ASA substituted for HFRS, which had acceptable discrimination (AUROC 0.715) and was the best calibrated model tested.
3. Base model (including ASA) supplemented with CCI, HFRS, and AMTS. This exhibited the best discrimination (AUROC 0.781) but - in common with all of the models - was miscalibrated.

Table 3.3 on page 92 shows the observed and expected proportions of 30-day mortality for the 12 hospitals identified as having crude mortality exceeding the 95% and 99.8% control limits.

The funnel plot for the existing NHFD model is shown in Figure 3.13 on page 93, the base model with ASA substituted for HFRS in Figure 3.14 on page 94, and the combined model (base plus CCI, HFRS, and AMTS) in Figure 3.15 on page 95.

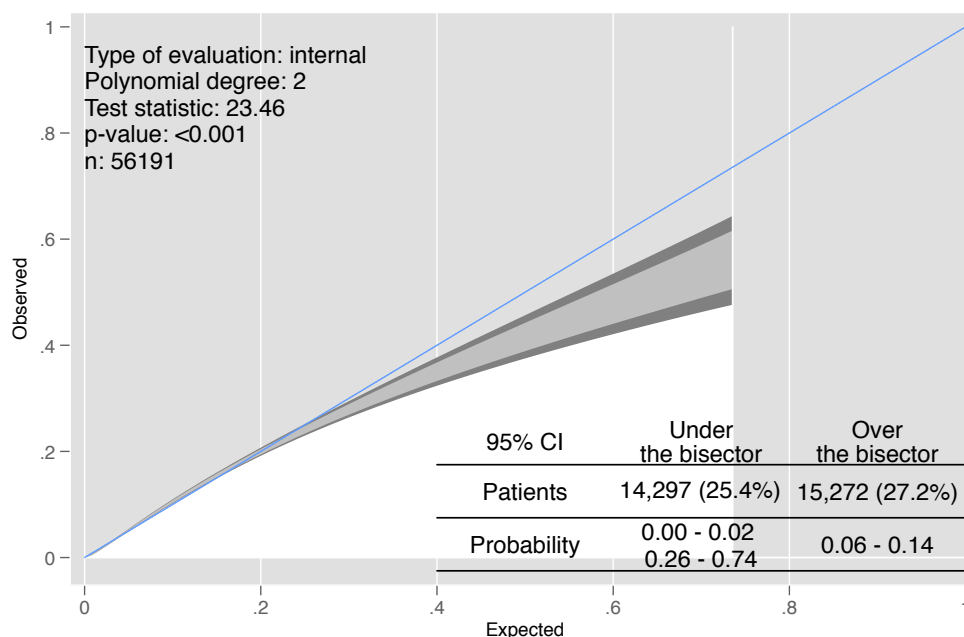


Figure 3.8: Calibration belt plot for the existing NHFD model supplemented by Hospital Frailty Risk Score.

Table 3.4 on page 96 shows the outlier status in each model for all hospitals identified as a 99.8% high mortality outlier in any of the models tested in this section. Four hospitals (3, 41, 144, and 161) were consistently identified as high mortality outliers across all three models but they were all also outliers based on *unadjusted* mortality. Two hospitals (6 and 65) were reported as high mortality outliers using the statistical models based on linked HES data but not when using the existing NHFD model. One hospital (155) was only identified as an outlier by the combined model (NHFD plus CCI, HFRS, and AMTS) and another (158) only by the HFRS model (NHFD with ASA substituted for HFRS). In all eight cases, these hospitals had unadjusted 30-day mortality rates that exceeded the 95% confidence limits (Table 3.3 on page 92).

Figure 3.16 on page 97 shows a funnel plot using the existing NHFD risk adjustment model with all eight outliers from Table 3.4 on page 96 identified as red triangles to indicate where these would fall in a typical NHFD annual report.

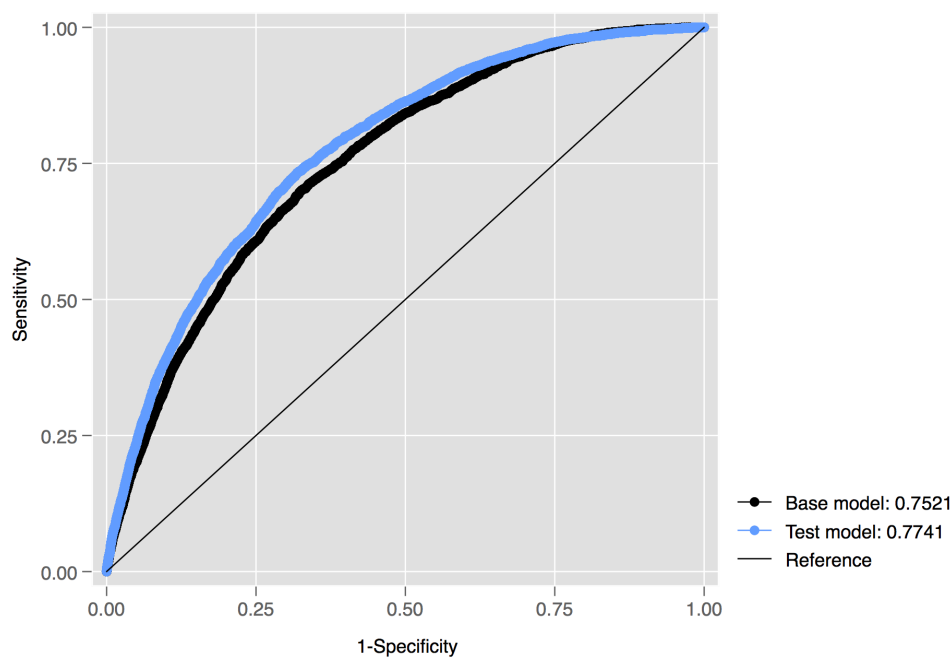


Figure 3.9: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with Charlson Co-morbidity Index and Hospital Frailty Risk Score

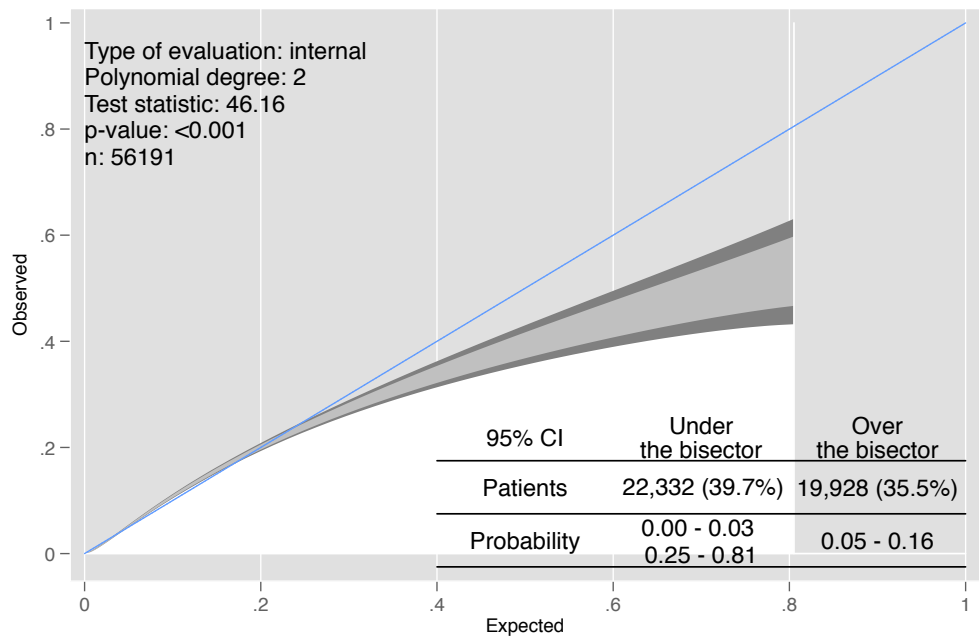


Figure 3.10: Calibration belt plot for the existing NHFD model supplemented with the Charlson Co-morbidity Index and Hospital Frailty Risk Score.

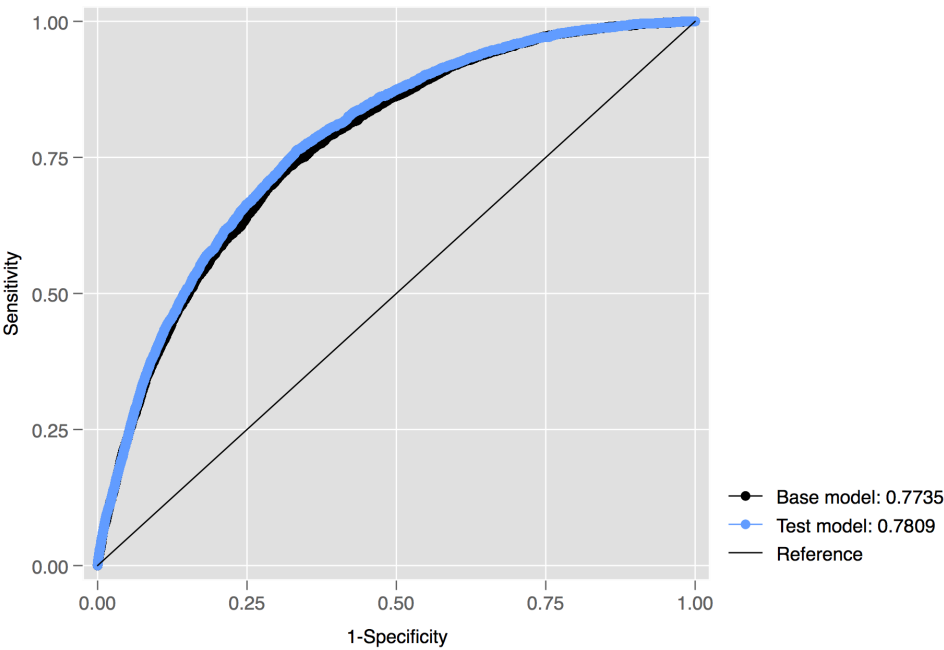


Figure 3.11: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with Charlson Co-morbidity Index, Hospital Frailty Risk Score, and Abbreviated Mental Test Score

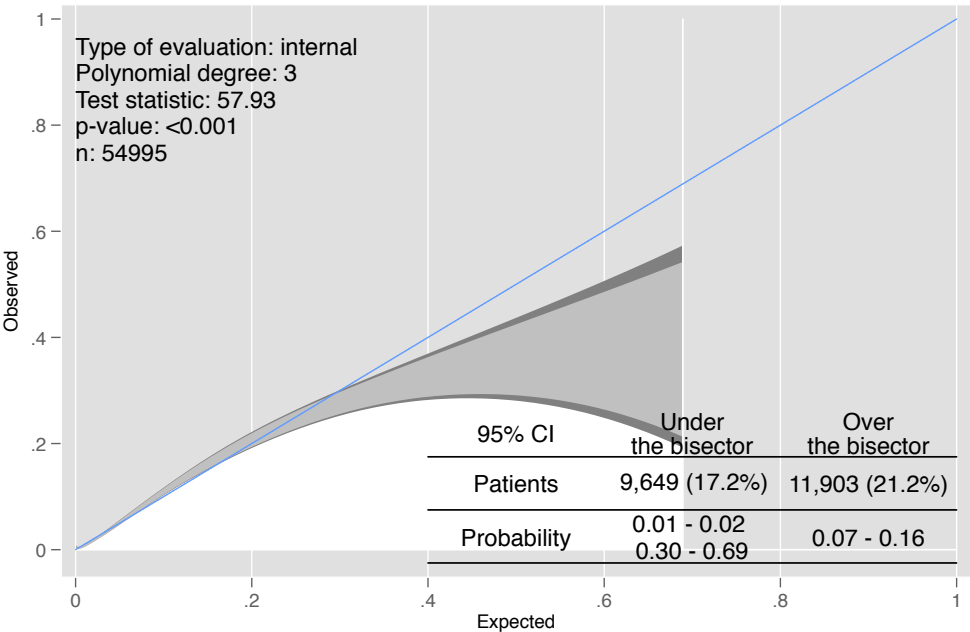


Figure 3.12: Calibration belt plot for the existing NHFD model supplemented with the Charlson Co-morbidity Index, Hospital Frailty Risk Score, and Abbreviated Mental Test Score.

Table 3.3: Hospitals identified as outliers for unadjusted 30-day mortality

		99.8% confidence limits		95% confidence limits	
Hospital	Observed	Lower	Upper	Lower	Upper
99.8% outliers	3	12.63	2.19	11.01	3.66
	6	11.50	2.75	10.45	4.03
	41	11.76	1.59	11.61	3.26
	144	10.17	3.13	10.06	4.29
	155	10.26	3.00	10.19	4.20
	161	10.79	2.95	10.24	4.17
95% outliers	2	10.26	2.31	10.88	3.74
	13	10.13	2.34	10.85	3.76
	56	9.60	2.85	10.34	4.10
	65	8.61	3.82	9.37	4.75
	87	12.12	0.00	14.08	1.61
	158	9.20	2.88	10.31	4.12

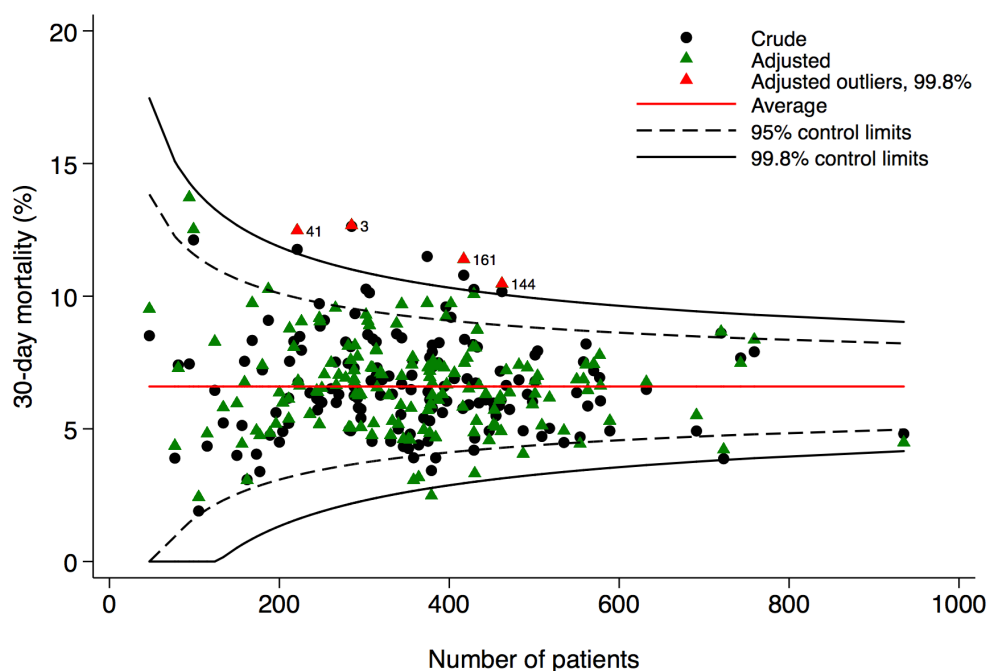


Figure 3.13: Funnel plot identifying hospital outliers for adjusted 30-day mortality using the existing NHFD model.

3.4 Discussion

This study showed that the existing NHFD risk adjustment model exhibits acceptable discrimination but is poorly calibrated. In particular, the model appears to underestimate the predicted risk of patients with the highest observed mortality.

3.4.1 ASA substituted for alternative co-morbidity measures

As discussed in Section 3.1.1 on page 73, the only variable that attempts to capture co-morbidity in the NHFD is ASA. This is a subjective classification system^{186–188} that allocates patients into one of only five categories and so may lack precision¹⁸⁸. In addition, it was developed 80 years ago as a means of estimating peri-operative risk in the general population¹⁸¹ and so was not specifically designed for use in the frail, older adult hip fracture population.

One possible means of improving the existing risk adjustment model is to seek

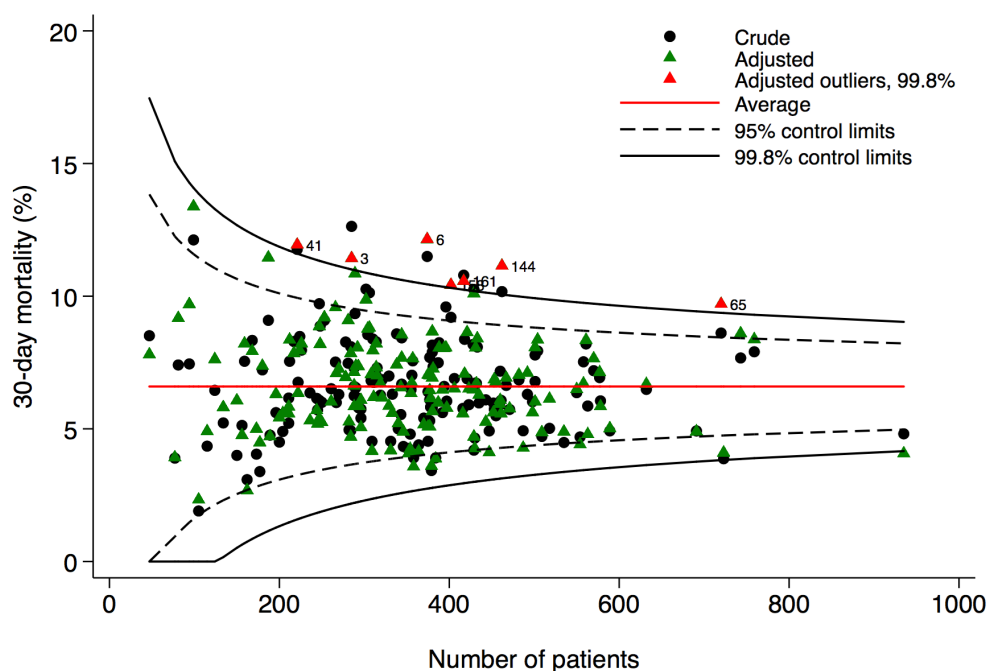


Figure 3.14: Funnel plot identifying hospital outliers for adjusted 30-day mortality using the NHFD model with ASA substituted for the Hospital Frailty Risk Score.

a routine linkage to alternative datasets, such as the HES APC. This would permit analysts to draw on data from patients' previous hospital admissions without further burdening clinical staff tasked with submitting data to the NHFD. This approach could also reduce the likelihood of hospitals “gaming” mortality figures, e.g. by increased coding of co-morbidities or over-estimating ASA²¹⁸.

A number of co-morbidity summary measures based on administrative data are well-validated (e.g. CCI and ECI) or have been developed specifically for the frail older adult inpatient population (e.g. HFRS). However, in this study, replacing ASA with any of these measures led to reduced model discrimination. Only the HFRS noticeably improved calibration, which may be because it was the only summary measure tested that was developed for use amongst older inpatients¹⁹⁸.

There are three possible explanations for the apparent primacy of ASA. First, it is a global assessment of peri-operative risk that is usually undertaken shortly before surgery by a senior anaesthetist. It is also possible that the ASA is not applied strictly and sometimes represents an heuristic judgement (on a scale of 1-5)

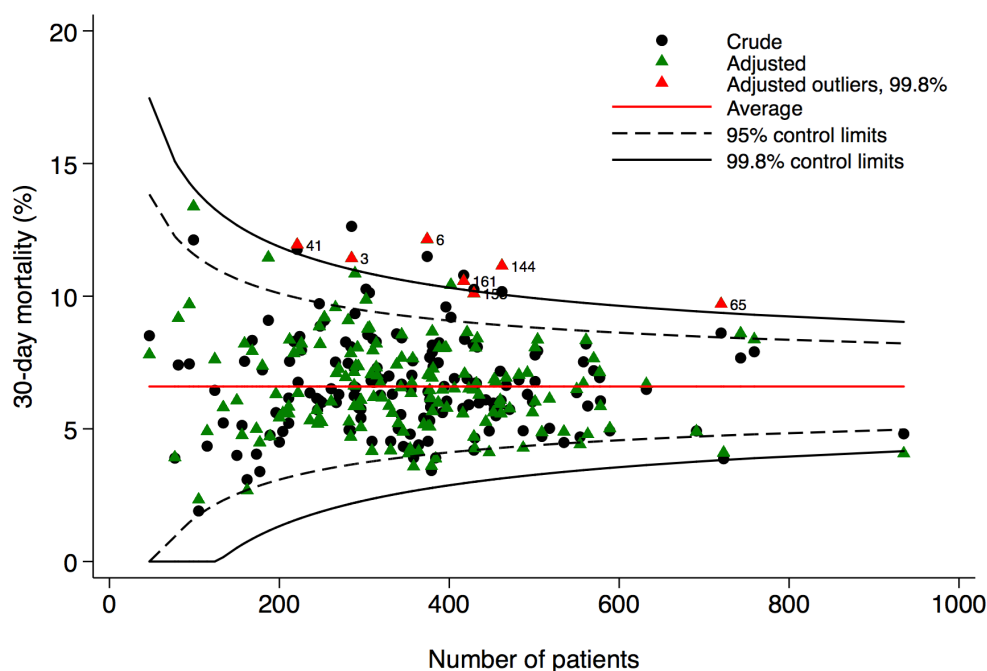


Figure 3.15: Funnel plot identifying hospital outliers for adjusted 30-day mortality using the NHFD model supplemented with Charlson Co-morbidity Index, Hospital Frailty Risk Score, and Abbreviated Mental Test Score.

about the patient’s likelihood of peri-operative death. By contrast, the other co-morbidity measures are principally based on diagnoses recorded during other clinical encounters by coders who are remote from the patient. Second, ASA is based on functional impairment rather than diagnoses, which may facilitate greater precision. For example, a patient with COPD may score 1 point on the CCI but be graded as anything from ASA 2 (“a patient with mild systemic disease”) in a patient with minimal symptoms to 5 (“a moribund patient who is not expected to survive”) in a patient with *cor pulmonale*¹⁸¹.

Finally, it is possible that HES does not capture all possible co-morbidities. In this study, a missing value in the ASA variable was imputed (see Subsection 3.2.4 on page 78) but a missing value for any of the co-morbidity scores based on linked data would necessarily have been interpreted as “CCI = 0”. This is a well-documented limitation of administrative datasets when compared against clinical registries²⁰¹.

Table 3.4: 99.8% outlier status for 30-day mortality based on choice of risk adjustment model

Hospital	30-day mortality			
	Unadjusted	Adjusted - model 1	Adjusted - model 2	Adjusted - model 3
3	X	X	X	X
6	X		X	X
41	X	X	X	X
65			X	X
144	X	X	X	X
155	X			X
158			X	
161	X	X	X	X

Model 1: Existing NHFD model.

Model 2: NHFD model with ASA substituted for HFRS.

Model 3: NHFD supplemented with CCI, HFRS, and AMTS.

3.4.2 ASA supplemented with additional co-morbidity measures

All three additional co-morbidity measures improved discrimination and calibration of the existing NHFD model when used in addition to ASA, but these gains were modest. The explanations for these modest improvements are likely to be the same as discussed in Subsection 3.4.1 on page 93. The best performing measure when used to supplement the NHFD model was the HFRS. In one sense this is surprising given that the HFRS was developed to predict frailty rather than clinical outcome, although it has also been shown to predict 30-day mortality in a validation cohort of older adult inpatients¹⁹⁸ and was developed for use in this population (Subsection 3.4.1 on page 93).

It is also possible that this finding simply reflects the greater range of scores assigned by the HFRS (0-73.9 in this cohort) versus those observed for the CCI (0-16) and ECI (0-13). Alternatively, the superiority of the HFRS may simply reflect the importance of frailty over absolute co-morbidity burden in the hip fracture population²¹⁹.

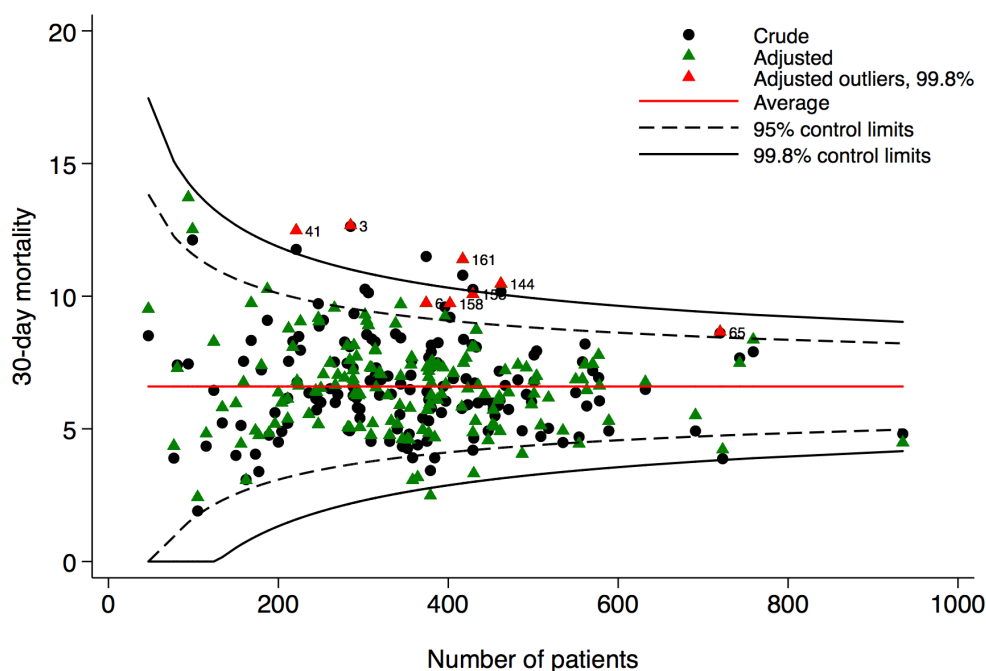


Figure 3.16: Funnel plot using the existing NHFD risk adjustment model but highlighting outliers from Table 3.4 on the previous page, which were identified by any one of the models evaluated in this section

3.4.3 Incorporating AMTS

The AMTS is a measure of cognitive impairment that is associated with delirium, acute illness (e.g. pneumonia) frailty, and dementia²²⁰. It might therefore be expected to predict 30-day mortality amongst patients in the NHFD. AMTS was initially left out of the existing NHFD model because this variable was frequently missing in the historical dataset used for the model development work in 2014¹⁸⁰. However, the data presented in this chapter suggests that overall levels of AMTS completion in the 2016 dataset was high (97.9%). This improvement in data completion may be related to the second iteration of the Hip Fracture BPT, which included documentation of AMTS as a quality standard for patients admitted from April 2012 (see Appendix E on page 260). However, as the BPT only relates to hospitals in England, it will be necessary to ensure that similar completion levels are observed for data submitted by hospitals in Wales and Northern Ireland before incorporating AMTS into the NHFD risk adjustment model. In any event, gains achieved by

including AMTS were minimal and unlikely worth jeopardising the ability of the national clinical audit to accurately benchmark hospital performance across England, Wales, and Northern Ireland.

3.4.4 Identification of high mortality outliers

Importantly, the choice of model did not make a substantial difference to the hospitals identified as being negative outliers for 30-day mortality. Although risk adjustment using linked HES data increased the number of hospitals found to exceed the 99.8% confidence limit for 30-day mortality, all such cases were identified by the existing NHFD model as exceeding the 95% confidence limit. These were also readily identifiable from analysis of unadjusted mortality data when the lower threshold (i.e. 95% confidence limit) was applied. It therefore seems unlikely that this function of the national clinical audit would be enhanced by pursuing a routine HES linkage, which would increase costs²⁰¹ and risk delays to the NHFD annual reports^{221,222}. This is particularly true given that HES only collects data from hospitals in England and so equivalent data linkages (e.g. the Patient Episode Database for Wales (PEDW)) would need to be pursued for Wales and Northern Ireland.

3.4.5 Limitations

There are a number of limitations to the study presented in this chapter. First, it used only a single year of NHFD data to compare models. It is likely that patterns of NHFD data collection and hip fracture outcomes will change over time, which may affect the performance of each model. Second, it was not possible to evaluate the Nottingham Hip Fracture Score (NHFS), which is the most extensively validated outcome prediction tool for older adults with hip fractures^{223–225}. This was because the NHFS includes pre-operative haemoglobin, which was not available from either the NHFD or HES²²⁶. However, a recent study showed that the performance of the existing NHFD risk adjustment model is similar to that of the NHFS²²⁷. Third, as discussed in Subsection 3.4.1 on page 93, it is likely that the true co-morbidity

burden is higher than recorded by administrative datasets such as HES²²⁸. It has also been shown that diagnoses are frequently miscoded in HES, which would likely affect the predictive performance of models drawing on this data^{201,229}. It is possible that this limitation might have been mitigated by access to a richer primary care database, such as the Clinical Practice Research Datalink (CPRD)²³⁰ or The Health Improvement Network (THIN)²³¹. Such resources are likely to record a more comprehensive range of patient co-morbidities than HES, which is limited to those recorded during hospital admissions. It is possible that linkage to richer - or even just different - datasets might have led to different findings. Primary care records would also have permitted evaluation of a wider range of frailty scores, such as the Rockwood Scale²³² and Electronic Frailty Index (eFI)²³³.

Finally, the quality of data in the NHFD has not yet been rigorously evaluated. Although this chapter found showed that levels of missing data are low, authors from one NHS organization have raised concerns that their own NHFD submissions contain inaccuracies^{234,235}. The NHFD reports have also suggested that some hospitals may be reported as outliers because of low quality data rather than because they are delivering poor care²⁰⁵. It is difficult to fully accept this explanation as it is possible that poor care and poor data quality may cluster within the same organizations. However, a national study that compared fracture classification in the NHFD with the type of operation found that few combinations were implausible²³⁶. This suggests that there is no widespread failure of coding data for NHFD submissions. The other key variables used for NHFD risk adjustment are likely to be readily available to coders from electronic patient records (e.g. age and sex) or are routinely documented as part of the medical clerking for patients with hip fractures (e.g. pre-injury mobility status and AMTS)²³⁷. Understanding the quality of NHFD data submitted by individual hospitals should however be prioritized by the national clinical audit and may be readily organized through orthopaedic trainee research networks, such as the Collaborative Orthopaedic Research Network (CORNET).

3.4.6 Conclusions

The existing risk adjustment model within the NHFD does an adequate job of consistently identifying high mortality outliers. Although improvements are possible, outlier hospitals can be readily identified by lowering the threshold at which alerts are triggered. This underlines the importance of national clinical audits operating a tiered alert process so that hospitals receive a different response depending on their estimated outlier status¹²².

The identification of outliers is however only one component of the overall aim of the national clinical audits to improve quality of care and patient outcomes. In Chapter 4 on page 101, this thesis will consider whether or not there is evidence for the NHFD improving outcomes, either alone as a voluntary audit of clinical standards, or as a vehicle for varying hospital remuneration based on the quality of care provided.

Chapter 4

Pay-for-performance and hip fracture outcomes

4.1 Introduction

Chapter 2 on page 42 showed that the existing evidence base is insufficient to conclude that public release of performance data affects healthcare performance or patient outcomes. There is clearly a need for further studies aimed at addressing this question (Subsection 2.4.6 on page 69). However, national clinical audits could also drive quality improvements through mechanisms that are not directly related to public release of performance data¹.

Although it was initially established for the purposes of national clinical audit^{205,238}, the availability of the NHFD as a data collection vehicle was one reason behind the choice of hip fractures as a target population for the BPT initiative in 2010^{239,240}.

¹Chapter published as Metcalfe D, Zogg CK, Judge A, Perry DC, Gabbe BJ, Willett K, Costa ML. Pay-for-performance and hip fracture outcomes: an interrupted time series and difference-in-differences analysis in England and Scotland. *Bone Joint J.* 2019;101-B:1015-1023.

4.1.1 Pay-for-performance in the NHS

Pay-for-performance initiatives are increasingly used as a mechanism for improving patient outcomes^{241–243}. These schemes link health payments to defined quality metrics in order to incentivize health providers to improve the quality or efficiency of care²⁴⁴. In the NHS, pay-for-performance has taken the form of the Payment-by-Results (PbR) programme, which was a flagship health policy of the 1997 Labour Government^{245,246}.

In the years before PbR, NHS hospitals were typically paid by block contracts, which were fixed sums for delivering services. Block contracts were local arrangements and permitted hospitals to negotiate higher payments based on their own cost circumstances. In *The NHS Plan* (2000), the Blair government argued that the NHS was suffering from “a lack of clear incentives and levers to improve performance”²⁴⁷. This view was developed in *Delivering the NHS Plan* (2002), which introduced the need for a system of “payment by results... [to offer] incentives to reward good performance... and to make the best use of available capacity”²⁴⁸. Under PbR, NHS organizations receive a fixed national payment that depends on their level of activity and the Healthcare Resource Group (HRG) to which cases are allocated. These fixed prices were intended to drive efficiency by removing the ability of individual hospitals to plead for higher payments. Surplus funds could be kept by the hospital and used to support other services at their discretion²⁴⁵.

In *High Quality Care For All* (2008), the Department of Health committed to “universalize best practice by... [paying] prices that reflect the cost of best practice rather than average cost... [which would]... be enabled through the Best Practice Tariffs programme”²⁴⁹. Similarly, in *Equity and Excellence* (2010), the Coalition Government declared that “providers will be paid according to their performance... payment should reflect outcomes, not just activity, and provide an incentive for better quality”²⁵⁰.

These commitments to best practice were implemented by the 2010/11 national tariff, which introduced a BPT across four high-volume clinical areas: cataract

surgery, cholecystectomy, stroke, and hip fracture. These were selected for having (1) significant unexplained variation in practice, (2) strong clinical consensus in support of best practice standards, and (3) a mechanism for collecting data about achievement of best practice²⁴⁰.

The 2010/11 national tariff also included a need for organizations to make efficiency savings, which continued in subsequent years. For example, the 2012/13 tariff prices were reduced by a further -1.8%. In the meantime, the BPT programme expanded with new tariffs published for major trauma care, emergency surgery, and outpatient procedures. As the base tariffs reduced, the disparity between these sums and BPT payments increased. For example, the difference between achieving the Hip Fracture BPT was worth £445 to hospitals in 2010/11, £890 in 2011/12, £1,335 in 2012/13, and £1,353 in 2016/17²⁴⁰. Responsibility for organising the BPTs moved from the Department of Health to Monitor and the NHS Commissioning Board from 2013/14²⁴⁶.

4.1.2 The Hip Fracture Best Practice Tariff

The Hip Fracture BPT paid hospitals an enhanced sum for each patient whose care satisfied all defined national clinical standards, e.g. surgery within 36 hours²⁵¹. Cases satisfying these standards, which have evolved over time (Appendix E on page 260), are identified from data submissions to the NHFD. Importantly, Scottish hospitals do not participate in the NHFD and are not subject to this BPT.

Evidence for the hip fracture BPT

There have been few formal evaluations of the Hip Fracture BPT in England. A small number of NHS organizations have reported improved compliance with the BPT criteria^{237,252–257}. This is consistent with evidence from annual NHFD reports^{205,238,258}. There is mixed evidence from single centre studies about the impact of the BPT on patient outcomes, with one reporting that BPT-driven quality improvements were associated with reduced mortality and others that satisfying the

BPT criteria is not a predictor of hip fracture mortality^{120,259}.

4.1.3 Objectives

This study aimed to (1) quantify any effect of the NHFD and BPT on hip fracture outcomes in England by using data from Scotland to control for secular trends and (2) estimate the effect of implementing similar initiatives in Scotland.

4.2 Methods

4.2.1 Study design

Natural experiment using interrupted time series²⁶⁰ and DID analysis²⁶¹.

4.2.2 Data sources

The study relied on national data from two sources in order to conduct quasi-experimental modeling of temporal trends. Changes in England (where the NHFD/BPT was introduced) were analysed as an “exposed” group and those in Scotland as a “control”. Given the countries’ geographic proximity, cultural similarities, and common political union within a United Kingdom, it was anticipated that secular changes in Scotland would closely mimic those in England had the NHFD/BPT not been implemented²⁶².

Data for England were abstracted from the HES APC dataset linked to ONS death certificate registrations. Data for Scotland were abstracted from SMR01.

Hospital Episode Statistics

The HES APC dataset has been described in Subsection 3.2.1 on page 77. In brief, this is an administrative dataset that collects data on all hospital admissions funded by the NHS in England²⁰¹.

Office for National Statistics

The civil death registration dataset has already been described in Subsection 3.2.1 on page 77. It holds data on all deaths registered in England and Wales with the exception of a small number of that require referral to a coroner (e.g. for *post mortem* or inquest) before registration.

Scottish Morbidity Records

SMR01 collects administrative data on episodes of inpatient care provided by all hospitals in Scotland. It is managed by the ISD for NHS National Services Scotland and is linked directly to Scottish death certificate data. ISD Scotland estimate that the SMR01 captures 99% of admissions to hospitals in Scotland²⁶³.

4.2.3 Setting and population

The UK is a unitary state composed of four countries: England, Scotland, Wales, and Northern Ireland. Comprehensive publicly-funded healthcare is freely available throughout the UK under the NHS. Provision of healthcare under the auspices of NHS Scotland was devolved to the Scottish Parliament by the Scotland Act 1998²⁶⁴.

Inclusion criteria

All adults aged ≥ 60 years were included in the analysis if they were treated for hip fracture in England or Scotland with inpatient admission dates between 1st January 2000 to 31st December 2016 and complete follow-up information for a period of 1-year following inpatient admission (2000-2017). No additional exclusion criteria were applied. Patients had to present with a primary ICD-10 diagnosis code on admission consistent with: S72.0 (“fracture of neck of femur”), S71.1 (“pertrochanteric fracture”), or S72.2 (“subtrochanteric fracture”). These represented all ICD-10 diagnostic codes that were compatible with the definition of “hip fracture” used in the NHFD²⁰⁵.

4.2.4 Intervention

The principal aim of this study was to determine the effect of introducing a pay-for-performance initiative on outcomes for older adults with hip fractures. However, the Hip Fracture BPT was only feasible once a framework had been established for capturing high quality clinical audit data. This framework was provided by the NHFD, which was launched three years previously. The analyses therefore examined the effect of (1) the introduction of the NHFD on 1st January 2007, (2) the introduction of the Hip Fracture BPT on 1st April 2010, and (3) the combined effect of the NHFD/BPT intervention.

National Hip Fracture Database

The NHFD was launched in 2007 and captures data on all adults aged ≥ 60 years with a hip fracture treated within England, Wales, and Northern Ireland (see Subsection 1.7.1 on page 38). In addition to publicly reporting hospital-level outcomes in an annual report, the NHFD provides an online platform through which clinical teams can visualize their outcomes and performance against the national clinical standards²⁵¹. These standards have changed over time and are shown in Appendix E on page 260.

Hip Fracture Best Practice Tariff

All NHS hospitals are reimbursed by a system of tariffs based on an adjusted formula applied to the “reference costs” returned by NHS organizations that estimate the cost of treating cases the previous year. In order to achieve the Hip Fracture BPT, hospitals must satisfy all of the criteria shown in Appendix E on page 260. The NHFD reports patient-level compliance with the national standards to the local Clinical Commissioning Group (CCG), which makes a quarterly uplift correction payment to individual hospitals.

4.2.5 Outcomes

The primary outcome was 30-day mortality. Secondary outcomes included 60-, 90-, and 365-day mortality as well as 30-, 60-, and 90-day re-admission, time to operation (defined as binary early time to theatre: <2 or >2 days), and acute index hospital length of stay (LOS) in days.

4.2.6 Statistical analysis

Secular trends

Differences in demographic and clinical variables were compared between countries before-and-after NHFD/BPT roll-out in 2007 to 2010 respectively in order to visualize potential differences between groups. The “pre-intervention” period was defined as cases presenting to hospital between 1st January 2000 and 31st December 2006. The “post-intervention” period was 1st May 2010 until 1st February 2018. Patients admitted in the interim period between 1st January 2007 and 31st March 2010 were incorporated into step-wise analyses examining changes before-and-after establishment of the NHFD in 2007 and the BPT in 2010. Co-variate information, presented in Table 4.1 on page 110, was largely consistent between groups with subtle time-consistent differences in admitted hip fracture patients in England and Scotland.

Before-and-after analysis

Changes in hip fracture outcomes in England and Scotland were first visualized graphically by month to inspect for obvious changes and ensure the existence of pre-intervention parallel trends, which is a requirement of quasi-experimental DID analysis. These visualizations included scatter plots with locally weighted scatterplot smoothing (LOWESS) lines. Quantitative assessment of before-and-after changes was also undertaken for English data using ITSA. ITSA was used to contextualize the main mortality DID results and to account for changes in English outcomes not reported in Scottish data (e.g. time to operation). The ITSA tech-

nique functions by fitting linear regression models to observations from the pre- and post-intervention periods. For the purposes of ITSA, longitudinal patient-level data were aggregated into monthly bins for each month of the calendar year and plotted by month as the proportion (or indicated quantile of index LOS) of each outcome of interest. Analyses were based on 84 pre-intervention points (patients admitted January 2000-December 2006) and 81 post-intervention points (patients admitted April 2010-December 2016). Models estimated the pre-intervention trend (“pre-intervention annual change”), change in level immediately following the intervention (“instant change”), and change in post-intervention trend (“post-intervention annual change”). The presence of auto-correlation was tested using the Durbin-Watson test. The extent to which the intercept of the post-intervention model deviates from the anticipated pre-intervention trend is assumed to represent an instantaneous causal effect of the intervention taking place over the same time span. Ongoing changes during the post-intervention period, the post-intervention slope, can sometimes be observed as a marked and maintained change from pre-intervention trends.

Difference-in-differences

Differences in mortality outcomes for England and Scotland were further compared using DID regression. This quasi-experimental technique functions by fitting linear models to temporally-aggregated data from the pre- and post-intervention periods. It includes coefficients for intervention group (i.e. England versus Scotland), time period (i.e. pre- versus post-intervention), and an interaction term between intervention group and time period. The magnitude and direction of the interaction term (the so-called “difference-in-differences” between temporal changes within each country) is assumed to represent the causal effect²⁶².

Software

StataIC v.15.0 was used for all statistical analyses. Panel data for ITSA were constructed from the admission-level master dataset using collapse commands. They

were analyzed using the ITSA module²⁶⁵ in Stata. The DIFF module²⁶⁶ was used for linear DID regression in order to obtain p-values and country-specific and time period-specific tabulations for tables. 95% confidence intervals were obtained by manually-fitted versions of the same models.

Information governance

Use of HES APC data for this project was approved by the NHS Digital Independent Group Advising on the Release of Data (IGARD). NHS Digital undertook the linkage to ONS data and created mortality flags at defined time points. Pseudo-anonymized data were then transmitted to researchers at the University of Oxford. The ISD of National Services Scotland provided pseudo-anonymized records from SMR01. Research Ethics Committee approval was not sought in line with GAfREC guidance²¹⁷. Personal data were processed under Articles 6(1)(f) and 9(2)(f) of the General Data Protection Regulation (EU) 2016/679.

4.3 Results

A total of 1,037,860 adults aged ≥ 60 years were admitted between 2000-2016 with a hip fracture in England, and 116,594 in Scotland. The demographic characteristics of these groups are presented in Table 4.1 on the next page, which shows a marked increase in recorded co-morbidities between the pre- and post-intervention periods in England (e.g. 5.5% CCI ≥ 3 rising to 18.9%) that was not evident in Scotland. Table 4.2 on page 111 presents ITSA results of English data whereas Table 4.3 on page 112 and Table 4.4 on page 112 show the more detailed DID analyses for mortality that include Scotland as a control region.

Table 4.1: Demographic differences between hip fracture populations in England and Scotland before (1st January 2000 to 31st December 2006) and after (1st April 2010 to 31st December 2016) implementation of the NHFD/BPT in England.

	Overall				Pre-intervention				Post-intervention			
	England		Scotland		England		Scotland		England		Scotland	
Number	1,037,860	89.9%	116,594	10.1%	391,697	89.4%	46,404	10.6%	446,098	90.3%	47,730	9.7%
Age												
60-64	30,454	2.9%	5,071	4.3%	10,225	2.6%	1,889	4.1%	13,692	3.1%	2,115	4.4%
65-69	49,178	4.7%	7,646	6.6%	17,336	4.4%	2,978	6.4%	23,178	5.2%	3,176	6.7%
70-74	83,750	8.1%	12,308	10.6%	33,538	8.6%	5,236	11.3%	34,752	7.8%	4,767	10.0%
75-79	152,778	14.7%	19,502	16.7%	63,205	16.1%	8,170	17.6%	60,507	13.6%	7,605	15.9%
80-84	240,553	23.2%	26,397	22.6%	96,160	24.5%	10,642	22.9%	97,993	22.0%	10,691	22.4%
85-89	259,565	25.0%	25,767	22.1%	92,498	23.6%	9,813	21.1%	113,854	25.5%	10,825	22.7%
>90	221,582	21.3%	19,903	17.1%	78,735	20.1%	7,646	16.5%	102,122	22.9%	8,551	17.9%
Sex												
Male	256,703	24.7%	29,141	25.0%	84,411	21.6%	10,426	22.5%	112,404	25.2%	12,927	27.1%
Female	781,157	75.3%	87,453	75.0%	307,286	78.4%	35,978	77.5%	323,694	72.6%	34,803	72.9%
Deprivation												
Least deprived 20%	199,962	19.2%	-	-	71,492	18.3%	-	-	89,156	20.0%	-	-
Less deprived 21-40%	222,711	21.5%	-	-	82,406	21.0%	-	-	97,321	21.8%	-	-
Middle quantile	222,714	21.9%	-	-	85,883	21.9%	-	-	98,569	22.1%	-	-
More deprived 21-40%	204,222	19.7%	-	-	78,449	20.0%	-	-	86,248	19.3%	-	-
Most deprived 20%	183,521	17.1%	-	-	73,467	18.8%	-	-	74,804	16.8%	-	-
Charlson Index												
0	465,976	44.9%	102,874	88.2%	229,389	58.6%	40,917	88.2%	143,451	32.2%	42,178	88.4%
1	309,067	29.8%	9,331	8.0%	104,123	26.6%	3,805	8.2%	140,969	31.6%	3,774	7.9%
2	139,411	13.4%	3,583	3.1%	36,722	9.4%	1,387	3.0%	77,399	17.4%	1,444	3.0%
≥ 3	123,406	11.9%	806	0.7%	21,463	5.5%	295	0.6%	84,279	18.9%	334	0.7%

Table 4.2: ITSA compared differences in English temporal trends before (1st January 2000 - 31st December 2006) and after (1st April 2010 - 31st December 2016) combined NHFD/BPT policy implementation in 2007-2010.

	Before	95% CI	Instant change	95% CI	p-value	After	95% CI
Mortality							
30-day	0.0%	-0.1%	-2.6%	-3.4%	<0.001	-0.2%	-0.2%
60-day	0.1%	0.0%	-4.3%	-5.5%	<0.001	-0.2%	-0.4%
90-day	0.2%	0.0%	-5.4%	-6.8%	<0.001	-0.2%	-0.4%
365-day	0.2%	0.1%	-5.3%	-6.3%	<0.001	-0.1%	-0.2%
Re-admission							
30-day	0.4%	0.3%	-1.3%	-2.2%	0.003	-0.2%	-0.4%
60-day	0.7%	0.5%	-1.4%	-2.3%	0.002	-0.1%	-0.1%
90-day	0.8%	0.6%	-1.2%	-2.1%	0.015	0.0%	-0.1%
Length of stay (days)							
50th percentile	-0.1	-0.2	-2.8	-3.5	<0.001	-0.4	-0.5
60th percentile	-0.1	-0.2	-3.8	-4.6	<0.001	-0.5	-0.7
70th percentile	-0.1	-0.3	-5.4	-6.4	<0.001	-0.7	-0.8
80th percentile	-0.3	-0.5	-7.3	-8.8	<0.001	-1.0	-1.2
Time to operation							
<<2 days	-0.6%	-0.9%	15.4%	13.7%	<0.001	0.7%	0.5%
							0.8%

Table 4.3: DID before and after implementation of the NHFD and BPT. DID in temporal mortality before (1st January 2000 to 31st December 2006) and after (1st January 2007 to 31st March 2010) implementation of the NHFD and before (1st January 2008 to 31st March 2010) and after (1st April 2010 to 31st December 2016) implementation of the BPT.

Scotland				England				NHFD			BPT		
2000-2006		2007-2009	2010-2016	2000-2006	2007-2009	2010-2016	2010-2016	DID	95% CI	p-value	DID	95% CI	p-value
30-day	9.8%	8.8%	8.5%	9.8%	8.7%	6.8%	6.8%	-0.1%	-0.5%	0.5%	-1.6%	-2.1%	<0.001
60-day	15.4%	13.8%	13.1%	15.6%	14.0%	11.4%	11.4%	0.0%	-0.6%	0.6%	-1.9%	-2.5%	<0.001
90-day	18.9%	17.3%	16.4%	19.4%	17.6%	14.7%	14.7%	-0.2%	-0.9%	0.5%	-2.0%	-2.6%	<0.001
365-day	32.0%	30.4%	29.8%	31.9%	30.2%	27.8%	27.8%	-0.1%	-1.0%	0.6%	-1.8%	-2.5%	<0.001

Table 4.4: DID before and after implementation of the combined NHFD/BPT intervention. DID differences in temporal mortality before (1st January 2000 - 31st December 2006) and after (1st April 2010 - 31st December 2016) combined NHFD/BPT implementation between 1st April 2010 in England versus Scotland

	DID	95% CI	p-value
30-day	-1.7%	-2.0%	<0.001
60-day	-1.9%	-2.3%	<0.001
90-day	-2.2%	-2.6%	<0.001
365-day	-1.9%	-2.5%	<0.001

4.3.1 30-day mortality

Figure 4.1 on the next page shows that the pre-intervention trends in 30-day mortality were the same in England and Scotland. 30-day mortality trended downwards in both countries after the launch of the NHFD, although the decline was more pronounced in England. The diverging lines became more obvious after April 2010 and this continued until the data were censored after 31st December 2016 (ITSA instant change following combined policy implementation -2.6 percentage-points [95% CI -3.4 to -1.7]; annual trend post-implementation -0.2 [-0.2 to -0.1]). DID analysis corroborated these findings, suggesting an overall reduction in 30-day mortality in England relative to Scotland of -1.7 percentage-points (95% CI -2.0 to -1.2). When stratified by each component of the intervention alone, the results suggest no reduction in 30-day mortality following NHFD introduction but a significant change of -1.6 percentage-points (-2.1 to -1.2) following introduction of the BPT. Between 2010-2016 in England, Figure 4.1 on the following page shows that there were 7,600 fewer deaths than expected within 30 days following implementation of the BPT.

4.3.2 60-day mortality

The effects on 60-day mortality exhibited the same direction and magnitude as on 30-day mortality.

4.3.3 90-day mortality

Figure 4.2 on page 115 shows similar findings for 90-day mortality, although the effect of the BPT was more apparent at this time point. DID analysis suggested that the combined intervention was associated with a change of -2.2 percentage-points (95% CI -2.6 to -1.6). However, this appeared to be driven entirely by the BPT: NHFD DID -0.2 percentage-points [95% CI -0.9 to -0.5] and BPT -2.0 [-2.6 to -1.3]).

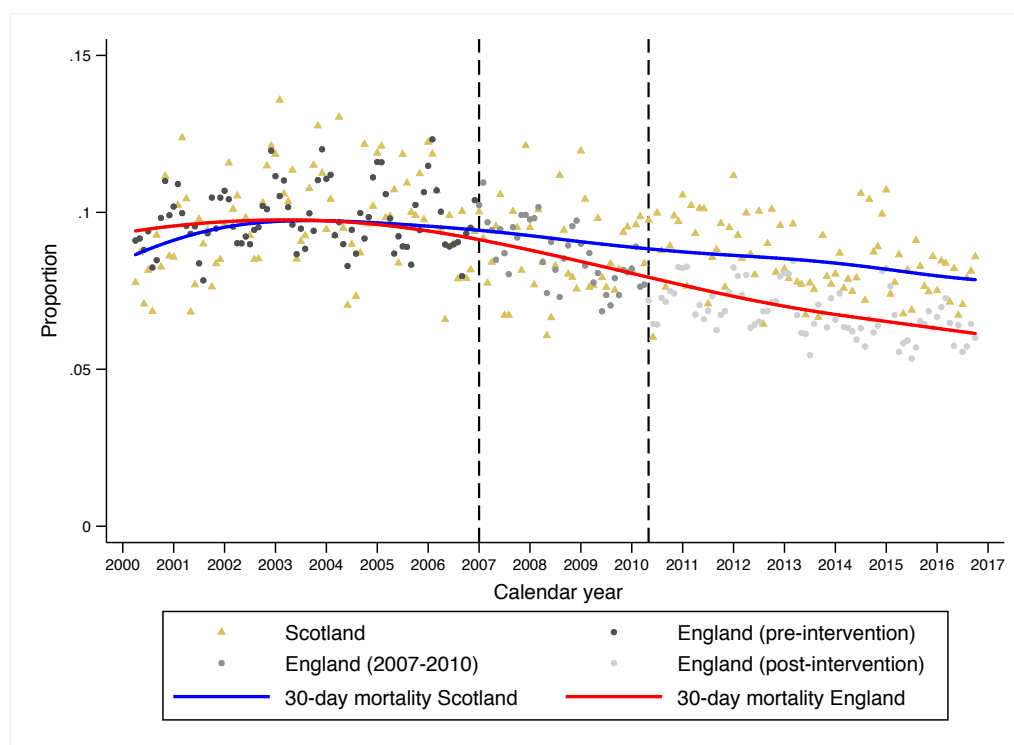


Figure 4.1: Monthly changes in 30-day mortality among adults aged ≥ 60 years, 2000-2016. Vertical dashed lines represent the implementation of the NHFD on 1st January 2007 and the BPT on 1st April 2010. There were 7,600 fewer deaths than expected in England following implementation of the BPT.

4.3.4 365-day mortality

Figure 4.3 on page 116 shows that the effect on mortality at 365 days was similar to that for mortality at 30- and 60-days. Although a small (non-significant) improvement was observed when the NHFD was introduced (DID -0.1 percentage points, 95% CI -1.0 to 0.6), the BPT was associated with a significant fall in 365-day mortality (-1.8, -2.5 to -1.0). The effect of the combined intervention on 365-day mortality was a change of -1.9 percentage-points (95% CI -2.5 to -1.3). Projection modelling presented in Figure 4.4 on page 117 suggests that, were the BPT to be implemented in Scotland in 2019, upwards of 115 annual deaths could be prevented each year – a number totaling more than 1,377 deaths by 2030.

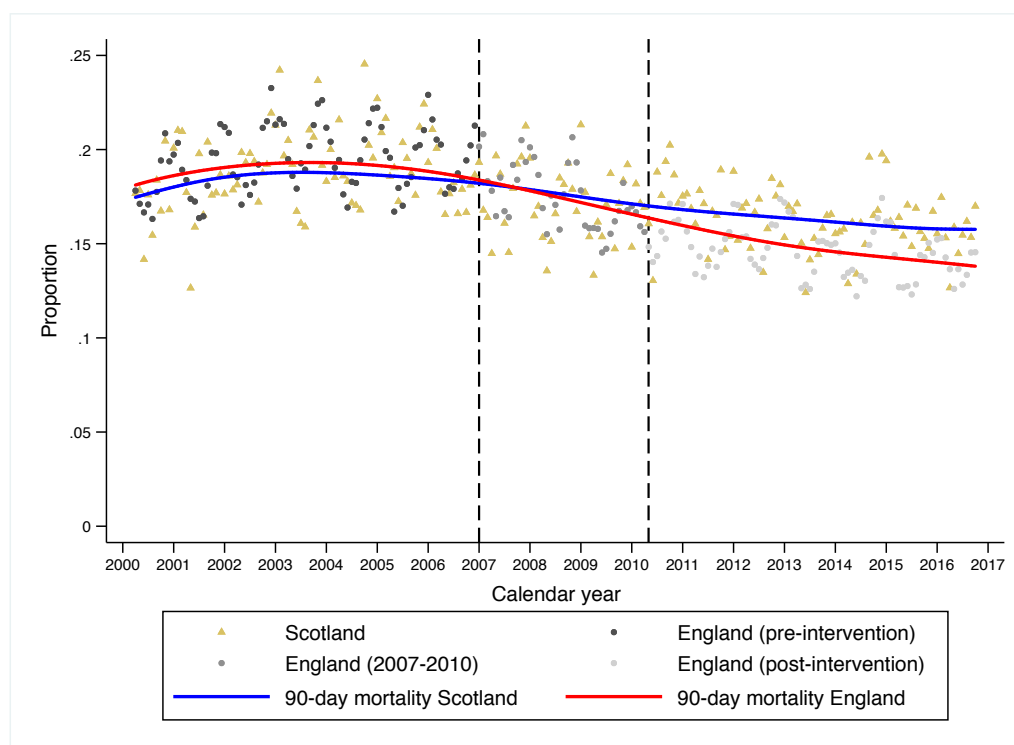


Figure 4.2: Monthly changes in 90-day mortality among adults aged ≥ 60 years, 2000-2016. Vertical dashed lines represent the implementation of the NHFD on 1st January 2007 and the BPT on 1st April 2010.

4.3.5 Re-admissions

Table 4.2 on page 111 shows that re-admissions at all time points (30-, 60-, and 90-days) were increasing steadily in England in the pre-implementation phase. The annual trend towards increasing 30-day re-admissions (0.4 percentage-points, 95% CI 0.3 to 0.5) was however reversed on implementation of the BPT (instant change -1.3 percentage-points, 95% CI -2.2 to -0.5) and this decline continued each year subsequently (annual trend post-implementation -0.2 percentage-points, 95% CI -0.4 to -0.1). Similar findings were observed for 60- and 90-day re-admissions, although the annual trend post-implementation did not change after the sudden fall associated with the BPT at these time points. Projection modeling presented in Figure 4.5 on page 118 suggests that were the BPT to be implemented in Scotland starting in 2019, upwards of 220 fewer re-admissions would be expected among Scottish hip fracture patients by 2030 within 30 days.

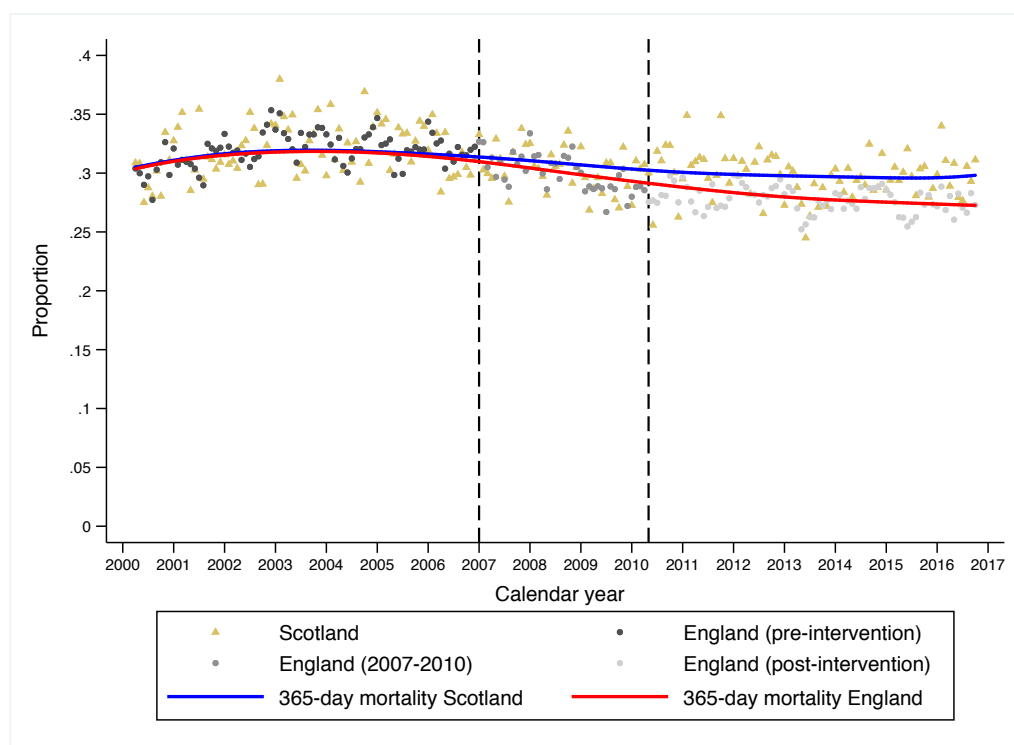


Figure 4.3: Monthly changes in 365-day mortality among adults aged ≥ 60 years, 2000-2016. Vertical dashed lines represent the implementation of the NHFD on 1st January 2007 and the BPT on 1st April 2010.

4.3.6 Time to operation

There was an annual trend towards fewer patients being operated within 36 hours in the pre-intervention period (annual trend -0.6 percentage-points, 95% CI -0.9 to -0.4). However, in the year following introduction of the NHFD/BPT, the proportion of patients reaching theatre within this time-frame increased by an absolute value of 15.4 percentage-points (95% CI 13.7 to 17.0) (Table 4.2 on page 111). This positive trend continued to increase by 0.7 percentage-points (95% CI 0.5 to 0.8) each year thereafter.

4.3.7 Length of stay

Median LOS was declining modestly (annual trend -0.6 days, 95% CI -0.2 to 0.0 days) in the pre-intervention period. This reduction increased following implementation of the NHFD/BPT (instant change following policy implementation: -2.8 days, 95% CI -3.5 to -2.1; annual trend post-policy implementation: -0.4 days, 95% CI -0.5

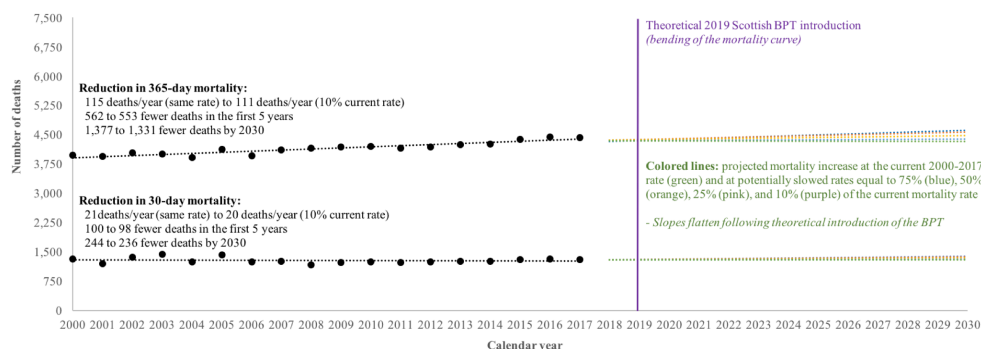


Figure 4.4: Projection model, which shows that 115 hip fracture deaths could be saved per year if a BPT was implemented and had the same effect in Scotland.

to -0.3) The magnitude of these reductions increased in a step-wise manner with each ascending quantile (e.g. 60th, 70th, and 80th) so that the largest reductions were observed amongst the patients with greatest index LOS (80th percentile instant change: -7.3 days, 95% CI -8.8 to -5.8; annual trend post-implementation: -1.0 days, 95% CI -1.2 to -0.8).

4.4 Discussion

This study provides strong evidence that the BPT drove changes in practice that reduced mortality for older adults with hip fractures in England by as many as 7,600 fewer deaths within 30 days between 2010 and 2016. It also suggests that the BPT increased the proportion of patients receiving an operation within 36 hours, shortened acute hospital LOS, and reduced re-admissions within 30, 60, and 90 days. There was evidence to suggest that some changes began before implementation of the BPT, which might be attributable to the NHFD.

A number of small studies have reported improved compliance with process mea-

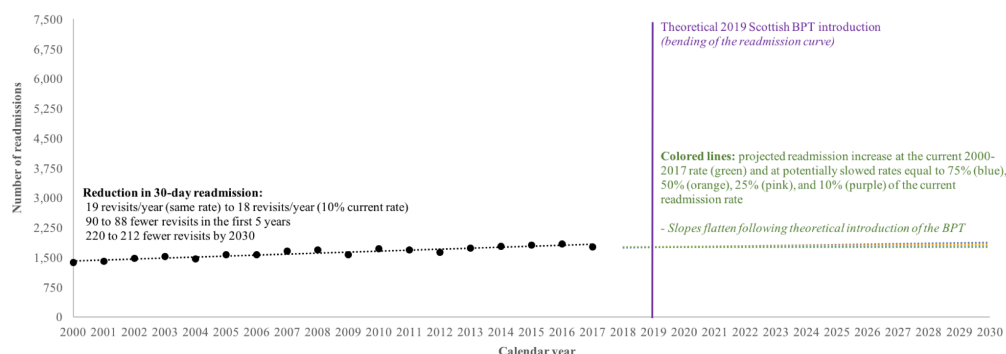


Figure 4.5: Projection model, which shows that 19 re-admissions could be prevented per year if a BPT was implemented and had the same effect in Scotland.

asures that were associated with introduction of the NHFD²⁶⁷ and BPT^{120,237,255}. The NHFD annual reports have also shown that English hospitals are increasingly achieving the hip fracture national clinical standards^{205,238,258}. One national time series study reported that mortality fell from 10.9% before the NHFD was launched to 8.5% afterwards²⁵¹. However, this study did not include a control population or analyze data after the BPT came into effect. The data in this chapter suggest that there was a gradual trend towards reduced mortality between 2007 and 2010 but that this was also apparent (albeit to a lesser extent) in Scotland, which did not participate in the NHFD. DID results restricted to the influence of the NHFD suggest that the differential trend between the two countries was not statistically significant following introduction of the NHFD in isolation. There was, however, a significant change leading up to full BPT implementation in 2010. DID results comparing BPT implementation alone between England and Scotland revealed a -2.0 percentage-point reduction in 90-day mortality (8.7% to 6.8% in England) that accounted for 90.9% of the overall effect (DID: -2.2 percentage-points, 9.8% to 6.8% in England).

A number of changes could account for improved outcomes over time across the UK, including publication of the BOA/BGS guidelines²⁶⁸, increasing recognition of the need for early surgery and post-operative rehabilitation²⁶⁹, and the emergence of orthogeriatrics as a medical sub-specialty dedicated to caring for older adults with fractures^{61,93}. It therefore seems unlikely that the NHFD alone accounted for the fall in mortality reported by Neuberger et al²⁵¹. Although this study suggests that the NHFD might have had a small positive effect on English hip fracture outcomes, this was not statistically significant. However, implementation of the BPT was associated with a marked and sustained improvement in hip fracture outcomes. It is nevertheless worth noting that the NHFD was a prerequisite for the choice of hip fracture outcomes as a target for pay-for-performance in England and so these two interventions are fundamentally linked^{239,240}. However, these data suggest that a system for rewarding best practice can improve outcomes beyond that of a voluntary audit of national clinical standards.

The improvements in LOS and re-admission suggest substantial resource savings attributable to the BPT in addition to reduced mortality^{242,243}. Importantly, the BPT itself did not require a substantial investment. It was initially set up as a payment of £445, which was based on an estimate of the cost that an average hospital would likely incur to provide additional operating capacity. However, the base tariff was reduced initially to adjust for compliance with the BPT criteria that was already present across the NHS. As a consequence of the falling base tariff, the BPT has accounted for a greater proportion of the overall payment to hospitals each year – from £445 in 2010/11 to £890 in 2011/12, £1,335 in 2012/13, and £1,353 in 2016/17²⁴⁰. As 100% compliance with the standards has not been achieved, the overall payment nationally by CCGs changed little over the first three years. Although a formal health economic analysis has not been undertaken, it is likely that the BPT delivered improved hip fracture care at reduced cost to NHS commissioners. There is a rectification process in the NHS of hospital trusts returning reference costs (for delivery of care) to ensure alignment between tariff price and average delivery

cost.

Evaluations of comparable pay-for-performance initiatives have reported mixed findings^{242,243}. Many of these earlier studies focussed on value-based purchasing (VBP), which is a strategy used by the CMS in the United States. The VBP programme withholds 2% of annual Medicare payments and allocates these to hospitals based on quality of care, compliance with best clinical practice, and patient experience²⁷⁰. However, few studies have been able to demonstrate improvements in mortality or re-admissions that are plausibly attributable to VBP^{271–273}. A number of explanations have been proposed for this finding²⁷⁴. First, the financial impact of the VBP is small (average \$213,000 bonus and \$1,200,000 penalty per hospital in 2015)²⁷⁵, which might be insufficient to motivate clinical pathway changes given that only a proportion of US patients are funded through CMS. Second, there are 21 individual measures and improving these in isolation is unlikely to achieve a hospital's overall score²⁷⁶. Third, the financial reward for improvement is unclear until the end of the performance period because the scheme is designed to be cost neutral and to allocate payments from low- to high-performers²⁷⁴. By contrast, the Hip Fracture BPT overcomes many of these criticisms in that it has a simple design, focuses on a small number of high-value measures, and carries a financial incentive that may be sufficient to motivate pathway change²⁷⁷. An alternative explanation for the success of the Hip Fracture BPT is that it was part of a more complex intervention that began with the national clinical audit. There is evidence that some hospitals engaged with the NHFD from 2007 by designing quality improvement processes aimed at improving their performance using data provided through online visual dashboards and in publicly accessible reports^{205,267}. An earlier qualitative study reported that “many participants reported that they had already made changes before the introduction of the BPT” and that such changes were driven by the NHFD²⁷⁸. It is however possible that the BPT provided the necessary additional impetus to help clinicians and hospital leaders create business cases that justified local investment in hip fracture services. This is also consistent with earlier qualitative findings: “The

NHFD was seen by interviewees as extremely important in terms of underpinning the BPT... however, this was not, on its own, sufficient to ensure the delivery of best practice... [the BPT was] seen as raising the profile of fragility hip fracture, with financial incentives encouraging... efforts on delivering care in accordance with best practice”²⁷⁸. The data in this chapter suggest that – despite concerns about the success of other schemes – it is possible to improve hip fracture outcomes through pay-for-performance. Further work should aim to identify the features that distinguish programmes that can demonstrably improve patient outcomes.

The apparent success of the Hip Fracture BPT in England could have policy implications for a number of countries. First, although there are 30 national clinical audits operated by the HQIP²⁵⁸ in England, there are only 21 BPTs used by NHS England to refine healthcare payments²⁷⁹. This suggests that there are further opportunities to extend pay-for-performance to other patient populations in England. Second, these data raise the possibility that introduction of a comparable pay-for-performance initiative might reduce hip fracture mortality in Scotland. Finally, this study might encourage policy makers outside the UK to consider implementing pay-for-performance to improve hip fractures outcomes. For example, although the CMS coordinates health payments for most patients aged >65 years in the United States, the VBP programme does not yet extend to hip fractures. There are over 200,000 hip fractures in the USA each year^{280,281} with a reported mortality of 5.2% at 30-days²⁸⁰. If the estimated benefits of the BPT in England were generalizable to the USA, the CMS expansion of pay-for-performance to older adults with hip fractures could prevent as many as 3,600 deaths per year.

4.4.1 Strengths and limitations

The strengths of this study are the use of comprehensive national cohorts linked to death certificate registrations and a “control” region, which overcame the limitations of earlier before-and-after studies. The ecological study design overcame issues arising from changes in coding patient characteristics over time. For example, there

was a marked increase in recorded co-morbidities in HES (that was not evident in SMR01) between the two periods, which might have been driven by the introduction of PbR (see Subsection 4.1.1 on page 102). This could have implications for studies using before-and-after designs that rely on co-morbidities recorded in administrative data to adjust patient outcomes. It may be particularly important for researchers that include patient records with variable “look back” periods given the apparent change in co-morbidity coding in England over time.

There are, however, a number of possible limitations to the approach adopted in this chapter. First, this study would still be vulnerable to confounding if another event had occurred at the same time as the NHFD/BPT but only affected outcomes in either England or Scotland²⁶². However, the factors that are thought to have driven recent trends towards improved hip fracture outcomes (such as the rise of orthogeriatrics as a medical sub-specialty⁶¹) would be expected to have applied across the UK. Second, some variables (e.g. LOS) were not available in Scottish data and so were limited to undertaking ITSA without a control region for this outcome. As discussed above, the absence of a control can result in erroneous attribution of secular change to a single intervention²⁶². It is, however, reassuring that the findings from ITSA for the other outcomes were consistent with those of the DID analyses. Third, outcomes were selected that could be readily quantified using administrative data. Although mortality and re-admissions are important quality metrics, other outcomes such as pain, mobility, and health-related quality of life might be more important to patients²⁸².

Finally, the focus of this study was on hip fracture outcomes and so it could not determine whether the BPT had an effect on other patient groups. Unintended consequences of the BPT could include the deprioritization of other older patients with lower limb injuries (e.g. distal femur or ankle) who share many of the same vulnerabilities as those with hip fractures^{283,284}. Alternatively, further benefits might extend to such patients (a so-called “halo effect”) as hospitals are likely to have invested in orthogeriatricians and dedicated trauma theatres in order to achieve the

BPT. The effect of pay-for-performance on related patient groups should be a focus for future research.

4.4.2 Conclusions

This study provides evidence that marrying a BPT to the existing national clinical audit improved hip fracture outcomes in England. It is therefore possible that this model could improve outcomes – and reduce costs – across other disease groups. Policy makers and clinicians should support the controlled expansion of the BPT model to other clinical areas and health policy environments.

The BPT is currently paid for cases that satisfy all seven standards of care. These have changed every few years (Appendix E on page 260) but clearly cannot simultaneously encapsulate all elements of hip fracture care. It is even possible that a hospital trust might perform well against the BPT standards while providing sub-optimal care in other domains. It has been suggested that one risk of a target is that it distorts priorities and may lead to under-investment in other areas²⁸⁵. Chapter 5 on page 124 will consider performance against a national hip fracture recommendation that is not yet a component of the Hip Fracture BPT.

Chapter 5

Inequalities in the use of total hip arthroplasty for hip fracture

5.1 Introduction

Chapter 4 on page 101 showed that the quality of hip fracture care and patient outcomes both improved when a BPT was integrated into the hip fracture national clinical audit. As discussed in Subsection 1.6.1 on page 34, the aims of national clinical audit include standardising treatment and reducing unwarranted variation. In this chapter, the NHFD was used to explore compliance with a national hip fracture recommendation that is not included amongst the BPT care standards¹.

5.1.1 Arthroplasty for displaced intracapsular hip fractures

Displaced intracapsular hip fractures are at high risk of disrupted blood supply to the femoral head, which can lead to osteonecrosis and painful non-union (Section 1.4 on page 22). As the femoral head is at risk of osteonecrosis, it is typically replaced rather than fixed *in situ*^{48,58,286}. There is good evidence that replacing the femoral head (“arthroplasty”) is superior to internal fixation, both in terms quality of life

¹Previous iteration of this study published as Perry DC, Metcalfe D, Griffin XL, Costa ML. Inequalities in access to total hip arthroplasty for hip fracture: population-based study. *BMJ*. 2016;353:i2021.

and less frequent need for revision surgery²⁸⁷. However, there are two arthroplasty procedures that are commonly used to treat these fractures: HA and THA.

During a THA, both the femoral head and acetabulum are replaced but, during a HA, only the femoral is replaced. Both procedures facilitate immediate weight-bearing. In general, HA is a quicker and simpler procedure, which is within the skill set of most general orthopaedic surgeons. Hemiarthroplasty has good short-term results but there are long-term risks of acetabular wear amongst the most active patients, and revision rates are higher than for THA²⁸⁸. However, the increased short-term risks of THA are balanced against its apparent long-term benefits. Although the risk-benefit profiles vary between these two operations, it has been shown that hip fracture patients undergoing THA have better function as well as less need for revision surgery than HA^{58,286,289,290}. However, THA prostheses are at higher risk of dislocation than patients receiving HA and so may require closed reduction under sedation in hospital.

One obvious solution to balancing these considerations is to reserve THA for the fittest and most active patients²⁹¹. The rationale underlying this approach is that patients in this group are most likely to tolerate a bigger operation and are also at greatest risk of acetabular wear if undergoing HA. As a group, they may have a greater life expectancy and so stand to benefit from the higher quality of life associated with THA for a longer period. It therefore seems likely that the long-term benefits of THA will stack up more favourably against the short-term risks for the fittest and most active patients.

NICE adopted this view in June 2011 when it issued Clinical Guideline 124 (CG124), which recommended that THA be offered to patients with a displaced intracapsular hip fracture who are “(a) able to walk independently out of doors with no more than the use of a stick (b) not cognitively impaired and (c) medically fit for anaesthesia and the procedure”⁴⁸.

There are concerns that some hospitals - particularly smaller units - may not have sufficient expertise to guarantee prompt THA for all eligible patients²⁹¹. Unlike

HA, the need to correctly site an acetabular component (with attendant risk of recurrent dislocation if sub-optimally positioned) means that THA is only performed by a proportion of orthopaedic surgeons^{291,292}. Two local audits have reported poor compliance with NICE CG124 and under-provision of THA for hip fracture^{293,294}. One of these studies also reported that, although use of THA increased between 2012 and 2013, almost half of the patients undergoing THA did not actually meet the NICE criteria for this operation²⁹⁴.

At the time that the study in this chapter was undertaken, THA provision was not included as a quality indicator within the NHFD and so the extent to which surgeons comply with this recommendation was unknown. This study therefore aimed to identify whether the national use of THA is consistent with NICE CG124 and whether there are systematic inequalities with regards to the use of THA for hip fracture.

5.2 Methods

A retrospective observational study was performed using data collected prospectively by the NHFD. The study protocol was approved by the HQIP prior to data release but research ethics committee approval was not sought for secondary analysis of administrative data in line with GAfREC guidelines²⁰⁰.

5.2.1 Data source

The NHFD has already been described in detail in Section 3 on page 72.

5.2.2 Inclusion criteria

This study included all patients aged ≥ 60 years that presented to a hospital in England between 1st April 2011 and 31st December 2016 with a displaced intracapsular hip fracture. Patients were excluded if their fracture was coded as “pathological”, as this may represent a heterogeneous group that includes patients with disseminated

cancer.

5.2.3 Variables and outcomes

Data cleaning involved several steps. Two patients had ages recorded as >115 years (both >1000 years), which were re-coded to exclude this variable. In 32 (0.01%) cases, the AMTS was not recorded as an integer and so scores were rounded to the nearest integer. On 1st April 2014 the NHFD data collection tool was updated to record mobility differently within the revised database. Earlier data were therefore mapped onto the new version using an algorithm presented in Appendix F on page 262. In the event of hospital trust reconfiguration (e.g. closure or merger), the hospital code at the time of data entry was used. As a consequence, a small number of hospitals only contributed data for a few months prior to reconfiguration.

The variables extracted from the NHFD were age (whole years), sex, lower layer super output area (LSOA), date of admission, treating hospital, pre-morbid mobility, ASA physical status classification score, and AMTS (see Subsection 1.7.1 on page 38). As described in Subsection 3.1.2 on page 73, the ASA score ranges between 1 (healthy patient) and 5 (moribund patient not expected to survive for 24 hours with or without surgery)¹⁸¹. The AMTS is a test of ten questions (e.g. “what is your age?”), which gives a score from 0 (zero answers correct) to 10 (all correct)²⁹⁵.

Deprivation scores for patients living in England were determined using the index of multiple deprivation (IMD) 2007. These scores reflect deprivation related to income, health and disability, employment, barriers to housing and services, living environment, education, and crime²⁹⁶. IMD scores were generated from LSOAs, which were then categorized into quintiles of deprivation based on the population of the UK.

Day of the week was determined from the date of admission. Hip fracture surgery in the UK usually takes place on the next available trauma operating list, which is the day following admission for most patients in the NHFD ($>65\%$). “Weekend” surgery was therefore identified by admission on a Friday or Saturday. Date of

surgery was categorized into deciles, which approximated 7-month periods across 69-month period between 1st April 2011 and 31st December 2016.

5.2.4 Statistical analysis

Guideline compliance was determined using a decision tree ordered to mirror the NICE recommendations, i.e. based on mobility (mobile outdoors with or without the use of a stick), cognition (defined as AMTS >8), and fitness for anaesthesia (defined as ASA 1 or 2). Although the cut-offs used for AMTS and ASA are not expressly stated in CG124, these have been used by NICE to monitor compliance with the guideline²⁹⁷. An AMTS score below 8 has previously been shown to identify cognitive impairment²⁹⁸ and adopted as a pragmatic threshold by the Royal College of Physicians²⁹⁹.

A method called recursive partitioning (RP) was used to determine the optimal decision tree that explains current practice. RP is a statistical technique for multivariable analysis that models how variables are best organized to predict a given outcome (e.g. THA). In RP, decision trees are built by identifying a variable that best splits the data into two groups. RP defines a cut-off (split) for continuous or ordinal variables, to enable the decision tree to correctly classify the maximum members of the population. Categorical variables are similarly grouped in RP, to build a tree with the least error. This process is then applied separately to each sub-group and continues recursively until either a maximum number of steps are reached or no further improvement is possible³⁰⁰. RP was undertaken using the “rpart” function in R. The tree was built using 10-fold cross validation and a negative complexity parameter to ensure that the maximum tree was built. Predictors included in the model were age, sex, mobility, AMTS, ASA, IMD quintile, and day of admission. The tree was pruned using the complexity (“cp”) function of the smallest tree within one standard error of the best functioning tree, i.e. the tree with the smallest error, which was confirmed graphically. A pragmatic approach was adopted to consider the tree complexity and efficiency related to clinical practice.

Individuals that fulfilled the NICE criteria were further analysed to explore factors associated with undergoing THA. A RP decision tree was constructed to model differentiating between THA and no-THA in this subgroup. The treating hospital was included as a factor variable, which allowed the partitioning algorithm to select optimal cut-off points for best fit within the model.

A mixed effects logistic regression model was fitted to explore factors associated with the use of THA amongst patients that fulfilled the NICE criteria. Age was included as a continuous variable, AMTS, ASA, admission date decile, and IMD quintile as ordinal variables, and weekend surgery as a categorical variable. Weekend admission was then substituted for day of the week to explore this predictor further in a second analysis. The unique hospital identifier was included as a centre level random effect.

Statistical analyses were performed using R and StataIC v.15.0.

5.3 Results

In the 69-month period between 1st April 2011 and 31st December 2016, there were 141,247 patients with non-pathological displaced intracapsular hip fractures recorded within the NHFD. Although 27,762 (19.7%) patients satisfied the NICE criteria to receive a THA, only 10,703 (38.6%) of these patients actually underwent this procedure. Amongst the 17,138 patients that underwent THA, 6,435 (37.5%) did not fulfil the NICE criteria.

5.3.1 Explaining the variation

The RP algorithm identified 7 terminal nodes (6 splits) as the optimal model (Figure 5.1 on the following page). The variable with the greatest importance was patient age, with a cut-off at 78 years defining the initial split (Figure 5.2 on the next page). A mobility split occurred between patients that ambulate independently and those that required the use of a stick. The other important predictive variables

were similar to those recommended by NICE, with splits occurring as predicted at $ASA \geq 3$ and $AMTS \geq 9$.

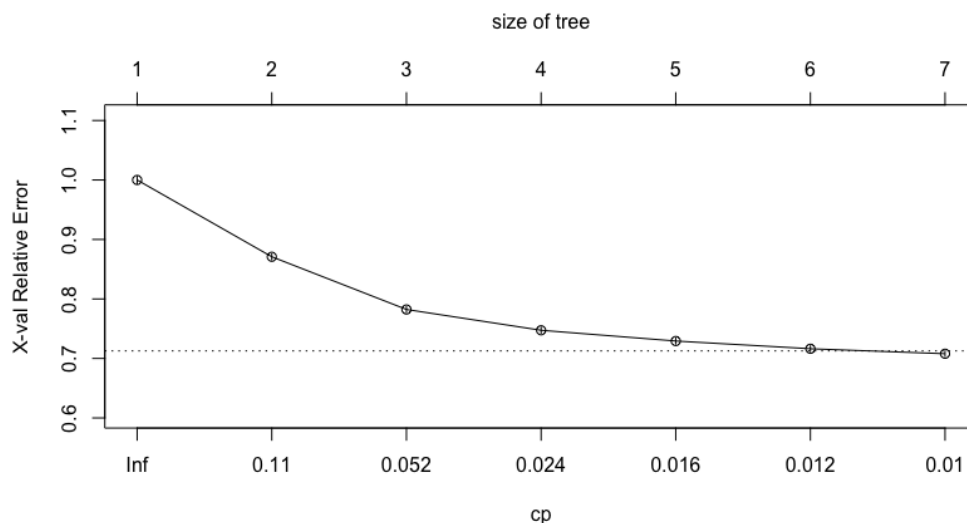


Figure 5.1: Graph showing limited improvement in the model beyond incorporating 7 terminal nodes.

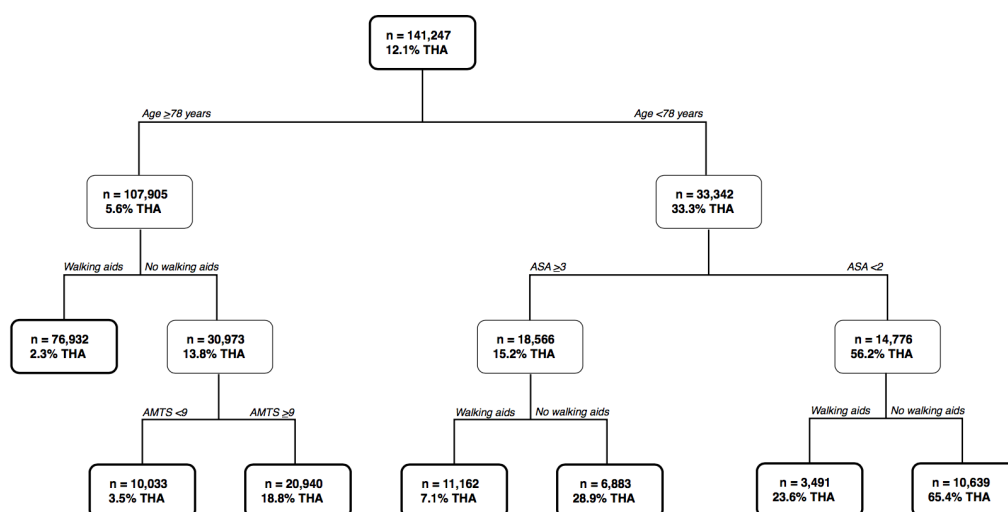


Figure 5.2: A decision tree for THA amongst all patients with displaced intracapsular hip fractures using a recursive partitioning algorithm.

Figure 5.4 on page 132 shows that, amongst the 27,762 patients fulfilling the NICE eligibility criteria, the RP algorithm identified 5 terminal nodes (4 splits) as the optimal model as more nodes increased the complexity of the tree with little associated gain in efficiency. Age was the most significant predictor, with 79 years identifying the splitting point (Figure 5.5 on page 132). For patients aged 79 and

above, there was a further age threshold at 85. Amongst the patients aged below 79 years, pre-injury mobility (with or without the use of a stick) was the most important predictor. Those mobilising without aids exhibited a further threshold at age 71 but the majority of these patients underwent THA regardless of age category: 58.4% amongst those aged ≥ 71 and 79.6% amongst those aged < 71 . Hospital variation amongst individuals fulfilling the NICE guidelines was considerable and ranged from 7.4% to 69.0% (Figure 5.3).

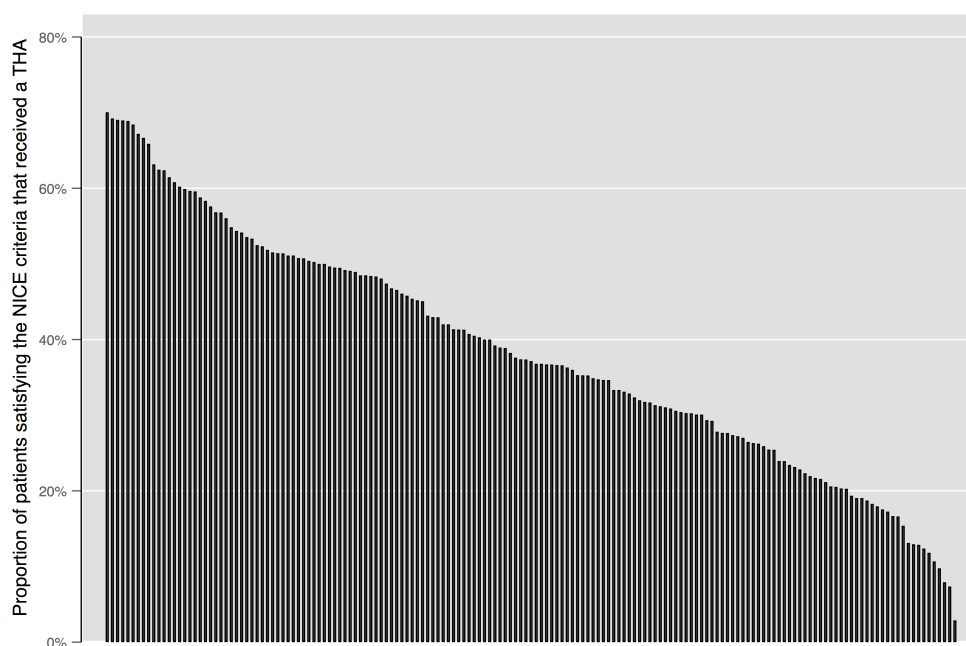


Figure 5.3: Bar chart showing variation in compliance with NICE CG124 by hospital contributing data to the NHFD

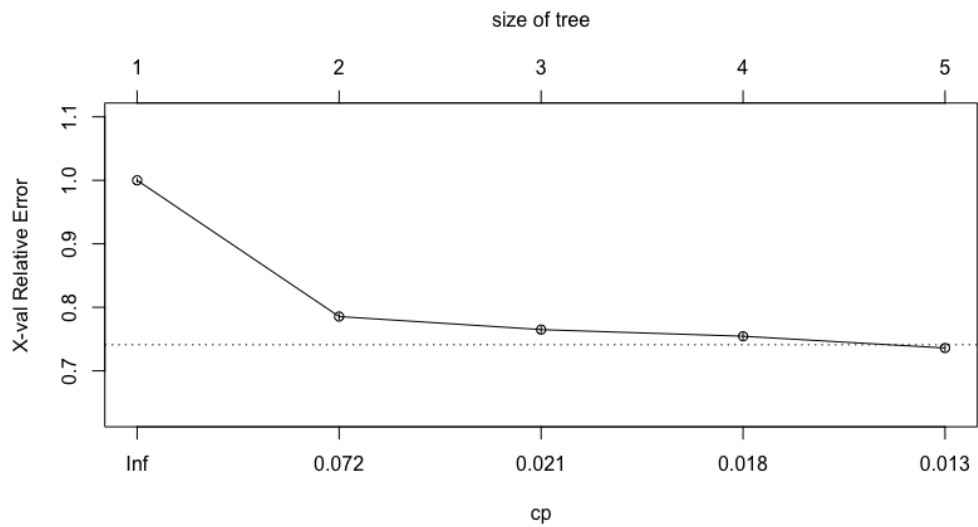


Figure 5.4: Graph showing limited improvement in the model beyond incorporating 5 terminal nodes.

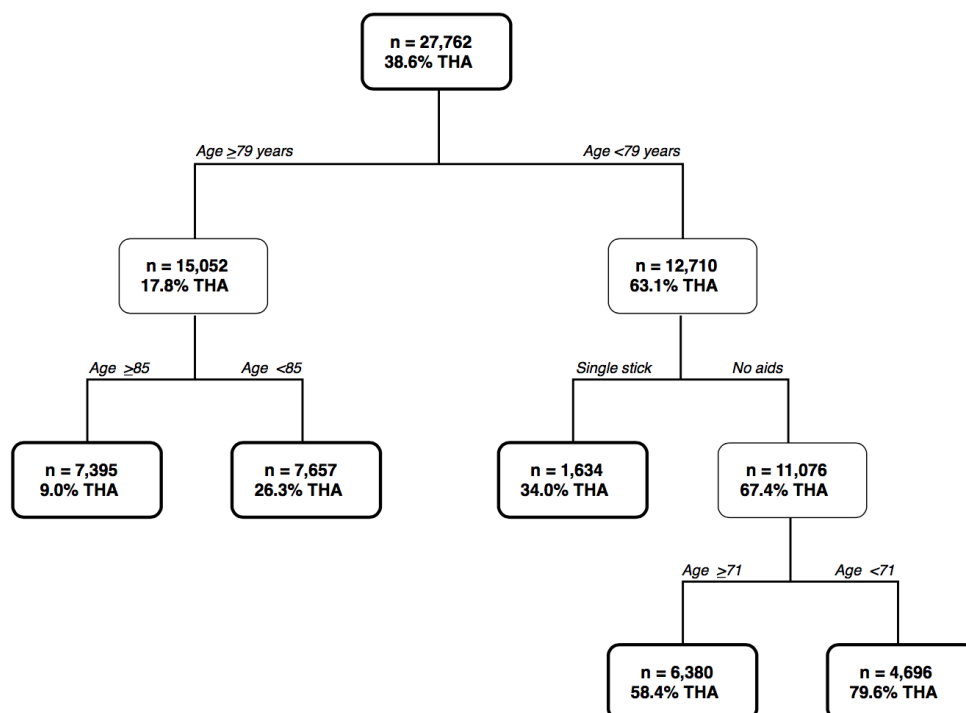


Figure 5.5: A decision tree for THA amongst patients with displaced intracapsular hip fractures that are eligible according to NICE guidelines using a recursive partitioning algorithm.

Table 5.1: The proportion of NICE-eligible patients undergoing THA by time period

	Patients undergoing THA	Patients fulfilling NICE criteria	Proportion of eligible patients undergoing THA
1st Apr 2011 - 24th Nov 2011	590	2,181	22.3%
25th Nov 2011 - 28th Jun 2012	795	1,830	30.3%
29th Jun 2012 - 3rd Feb 2013	909	1,768	34.0%
4th Feb 2013 - 30th Aug 2013	1,030	1,650	38.4%
31st Aug 2013 - 19th Mar 2014	1,020	1,644	38.3%
20th Mar 2014 - 16th Oct 2014	1,139	1,710	40.0%
17th Oct 2014 - 30th Apr 2015	1,186	1,650	41.8%
31st Apr 2015 - 23rd Nov 2015	1,260	1,646	43.4%
24th Nov 2015 - 8th Jun 2016	1,317	1,584	45.5%
9th Jun 2016 - 31st Dec 2016	1,457	1,396	51.1%

5.3.2 Predictors of receiving THA if NICE-eligible

There was a progressive increase in the provision of THA for eligible individuals across the study period (Table 5.1 on the previous page). The logistic regression model (Table 5.2 on the following page) showed that increasing age (odds ratio (OR) 0.84, 95% CI 0.84 to 0.85), using a stick to mobilize (0.25, 0.23 to 0.28), and operation on Saturday (0.61, 0.54 to 0.69) or Sunday (0.62, 0.54 to 0.71) were associated with reduced odds of undergoing THA. Similarly, the odds of undergoing THA reduced in a step-wise manner with each point reduction in AMTS (AMTS 8 0.43, 0.37 to 0.50), ASA (ASA 2 0.53, 0.47 to 0.59), and deprivation quintile (most deprived quintile 0.66, 0.59 to 0.74). Sex was not significantly associated with whether or not NICE-eligible patients received a THA.

5.3.3 Predictors of receiving THA if NICE-ineligible

A further analysis was conducted amongst individuals with a non-pathological displaced intracapsular fracture that did not fulfil the NICE eligibility criteria for THA (n=113,485). 6,435 (5.7%) such individuals underwent a THA. Using the same regression model, similar inequalities emerged. The receipt of THA outside the recommendations of NICE was least common amongst those from the poorest quintile of socioeconomic deprivation (OR 0.65, 95% CI 0.58 – 0.72), with a step-wise decrease from the richest quintile. Similarly, patients were least likely to receive a THA when admitted on Saturday (OR 0.65, 95% CI 0.58 to 0.74) or Sunday (0.60, 0.53 to 0.69) when compared with Wednesday. The odds of receiving a THA on other days of the week were similar to Wednesday: Monday (OR 1.08, 95% CI 0.97 to 1.20), Tuesday (0.97, 0.87 to 1.08), Thursday (0.97, 0.87 to 1.08), and Friday (1.01, 0.91 to 1.12). Increasing time decile was only weakly associated with provision of THA to patients that did not satisfy the NICE criteria (OR 1.03, 95% CI 1.02 to 1.05).

Table 5.2: Mixed effects logistic regression model showing the odds of undergoing THA amongst NICE-eligible patients

	Odds ratio	95% CI	p-value
Age (per year)	0.84	0.84 to 0.85	<0.001
AMTS			
10	1.00	Ref	Ref
9	0.65	0.59 to 0.72	<0.001
8	0.43	0.37 to 0.50	<0.001
ASA grade			
1	1.00	Ref	Ref
2	0.53	(0.47 to 0.59)	<0.001
Mobility			
Walk independently	1.00	Ref	Ref
Walk with one stick	0.25	0.23 to 0.28	<0.001
Sex			
Female	1.00	Ref	Ref
Male	0.93	0.86 to 1.01	0.100
Day of surgery			
Saturday	0.61	0.54 to 0.69	<0.001
Sunday	0.62	0.54 to 0.71	<0.001
Monday	1.06	0.94 to 1.19	0.313
Tuesday	0.97	0.82 to 0.86	1.09
Wednesday	1.00	Ref	Ref
Thursday	0.93	0.82 to 1.05	0.227
Friday	1.07	0.95 to 1.20	0.296
Deprivation quintile			
1 - Least deprived	1.00	Ref	Ref
2	0.94	0.88 to 0.85	0.283
3	0.88	0.79 to 0.98	0.019
4	0.76	0.68 to 0.85	<0.001
5 - Most deprived	0.66	0.59 to 0.74	<0.001
Date of surgery			
Per 7-month period	1.14	1.14 to 1.16	<0.001

5.4 Discussion

In June 2011, NICE recommended that THA should be offered to patients with a displaced intracapsular hip fracture who can walk independently outdoors (with no more than a single mobility aid), are cognitively intact, and are medically fit to undergo the operation⁴⁸. This guideline is consistent with a developing evidence base, which suggests that THA leads to better functional outcomes than HA following hip fracture^{58,286,289,290}.

This observational study used a large national audit dataset and demonstrated substantial unexplained variation (7.4 to 69.0% between hospitals) in the use of THA following a hip fracture. As patient-level predictors were unable to fully account for this variation, it is likely to reflect systematic differences in practice between centres. However, the provision of THA appeared to be influenced by a number of patient characteristics, including age, AMTS, ASA, socioeconomic status, and pre-fracture mobility. Other predictors of THA included the treating hospital and the day of admission. The use of THA amongst eligible patients increased steadily but remained low throughout the study period.

5.4.1 Barriers to increased THA provision

There are a number of possible obstacles to compliance with the THA recommendation in NICE CG124. These include guideline overload, lack of specialist hip surgeons, efforts to reduce surgical delay, and concerns about the underlying evidence base.

A 2011 survey of guidelines found that clinicians caring for a single hip fracture patient should comply with around 75 guideline and policy documents. Many of these were difficult to access, long, complex, and contradictory³⁰¹. Within NICE CG124 alone, the THA provision is only one amongst a total of 31 separate recommendations. The full version of the guideline runs to 658 pages⁵⁵, which can be read alongside a 280-page addendum³⁰². It is therefore possible that awareness of the recommendation did not immediately permeate the orthopaedic community.

A further potential obstacle to delivering THA for all eligible hip fracture patients is the availability of experienced hip surgeons. It is widely accepted that patients undergoing elective THA by a low-volume surgeon have greater risks of dislocation, need for revision surgery, post-operative complications, and death^{292,303–306}. For this reason, many orthopaedic surgeons do not perform THA for hip fracture if this procedure is not part of their routine elective practice. It is possible that the number of surgeons performing THA routinely will increase across the NHS if it becomes a “core” orthopaedic trauma operation in future.

There is consistent evidence from observational studies that early surgery is associated with reduced hip fracture mortality^{66,67}. This is thought to be related to prolonged immobility and subsequent complications, such as pneumonia, decubitus ulcers, and VTE⁶⁶. In the absence of someone to deliver THA, some surgeons may prefer to provide an operation (e.g. HA) promptly to ensure early mobility despite the THA criteria being satisfied. However, it is possible that the fittest patients (i.e. those satisfying the THA criteria) may be at less risk from a prolonged period of immobility than the general hip fracture population. A study exploring whether or not surgical delay is associated with higher mortality amongst the THA-eligible hip fracture population would be informative. It is also possible that surgeons might opt for non-compliant early surgery to maximize operating theatre efficiency and ensure high performance against the BPT indicators (Section 4.1.2 on page 103). There are many examples of targets in healthcare having subtle but unintended consequences for treatment decisions²⁸⁵. Whether or not THA is currently associated with surgical delay will be addressed in Chapter 6 on page 148.

One possibility for non-compliance with the NICE THA recommendation is that orthopaedic surgeons are unconvinced by the underlying evidence. Despite studies in support of THA, an international survey of orthopaedic surgeons found that 73% prefer to use HA for displaced intracapsular hip fractures³⁰⁷. Although there is evidence that orthopaedic surgeons are willing to change practice in the face of strong trial evidence, such changes are gradual and incomplete³⁰⁸. The strength of

the evidence base underlying CG124 is addressed in Chapter 6 on page 148 and may finally be resolved by the forthcoming Hip fracture Evaluation with ALternatives of Total hip arthroplasty versus Hemiarthroplasty (HEALTH) trial³⁰⁹.

5.4.2 Strengths and limitations of the study

The main strength of this study was its use of a dataset that captures almost every hip fracture treated in England, Wales, and Northern Ireland. There were variables that aligned closely with the NICE eligibility criteria, which permitted the recommended treatment algorithm to be mapped over the administrative data recorded within the NHFD. As the variables required were closely aligned to the principal functions of the NHFD, there was minimal missing data.

The main limitation was that the NHFD does not record individual patient co-morbidities and so it was not possible to determine whether specific co-morbid diseases were associated with differences in the use of THA. This study did not have permission from HQIP or NHS Digital to use linked HES APC data. Some of the variables in this analysis (e.g. age and deprivation) could therefore simply represent a tendency towards a greater co-morbidity burden. However, as shown in Chapter 3 on page 72 and previous studies, ASA score has equivalent or even greater predictive value for mortality and complications than standard co-morbidity measures based on administrative data, such as the Charlson Co-morbidity Index)^{310–312}. It is unlikely that patients assigned an ASA score <2 ($2 = \text{“mild systemic disease”}$) were medically unfit to undergo THA.

Although IMD is the most widely used index of deprivation, it is an aggregate measure based on LSOAs of approximately 1,500 people or 650 households³¹³. It is therefore lacks precision for identifying the extent to which any one individual is deprived. It has also been criticized because its different constituents parts are unequally weighted, e.g. “income” 22.5%, “health deprivation and disability” 13.5%)³¹⁴. The choice of weights is based on a mix of evidence and value judgement³¹⁴, and there is evidence that varying the weights can have substantial effects on the distri-

bution of scores³¹⁵.

Finally, the NHFD does not include sufficient detail to understand clinical decision-making at the level of individual patients. For example, it is possible that THA was fully discussed with some patients and HA chosen following a balanced risk-benefit discussion. However, such substantial inter-hospital variation in compliance with CG124 suggests that there are likely to be systematic differences in THA provision.

5.4.3 Implications of the study findings

The origins of NICE began in a policy white paper, *The New NHS: Modern, Dependable* (1997). NICE was established as a NHS special health authority to provide “a strong lead on clinical and cost-effectiveness, drawing up new guidelines and ensuring they reach all parts of the health service”¹⁰⁴. This was a response to the rapidly expanding landscape of available healthcare treatments and to reduce unexplained variation in care across the NHS, the so-called “postcode lottery”³¹⁶. The *NHS Constitution for England* (2015) subsequently guaranteed patients access to all “drugs and treatments that have been recommended by NICE for use in the NHS”³¹⁷.

Consequences of non-compliance with NICE guidelines

Recommendations arising from NICE technology appraisals and specialized technology appraisals impose a duty on NHS England to make approved treatments available within 3 months of publication. This duty has a statutory basis under s.7 and s.8 of The National Institute for Health and Care Excellence (Constitution and Functions) and the Health and Social Care Information Centre (Functions) Regulations 2013 (henceforth “The Regulation”).

However, NICE CG124, which included the THA recommendation, does not carry the weight of either a technology recommendation or specialized technology recommendation. The role of NICE in “giving advice and guidance” is instead governed by s.5 of the Regulation, which, unlike s.7 and s.8, does not explicitly impose a duty on NHS organizations to comply. Although such guidelines are often there-

fore assumed to be voluntary, their status is evolving in English case law towards a situation in which they too may bind commissioners, providers, and even individual clinicians.

The general principle that governs the behaviour of state organizations in relation to national guidance is set out in *R. v. North Derbyshire Health Authority, ex parte Fisher* (1997). In this case, which preceded the creation of NICE, Dyson J found that the decision of a public sector organization not to follow guidance from the Secretary of State could only be lawful when there is some “special factor which it considered exceptionally justified departure”. In *R. v. Thanet Clinical Commissioning Group, ex parte Rose* (2014), Justice J found that NICE guidance “has the same status as that of the Secretary of State in *ex parte Fisher*... [and] it would surely follow that the CCG could not disagree with NICE; it would need to find an exceptional basis for not following the NICE recommendation”. Providers and commissioners cannot, therefore, lawfully restrict access to treatments that are approved or recommended by NICE without exceptional justification.

The extent to which individual clinicians are bound by NICE is less clear. NICE clinical guidelines typically lead with the statement that recommendations “represent the view of NICE, arrived at after careful consideration of the evidence available... when exercising their judgement, healthcare professionals are expected to take these recommendations fully into account... however, the [recommendations do] not override the individual responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and/or guardian or carer”. Sir Michael Rawlins, while Chair of NICE, reiterated this point when writing “there appears to be confusion about the circumstances in which it is obligatory for... [doctors] to follow NICE guidance... the quick answer is ‘never’”³¹⁸. Sir Michael accepted that it is not possible for guidelines to regulate every clinical encounter and that treatments shown to be effective at a population level might not be best for individual patients. He also accepted that some guidelines are aspirational and so it might not be immediately possi-

ble for clinicians to provide NICE-compliant treatment to every eligible patient³¹⁸. There are nevertheless two ways in which NICE guidelines might be relevant to the decision-making of individual clinicians: defining the minimum standard of care and ensuring informed consent.

Liability for negligence in English law requires (1) a duty of care, (2) breach of that duty, (3) that the breach caused harm and (4) that the harm was not too remote a consequence of that breach³¹⁹. Whether or not a duty of care is breached depends on whether or not the plaintiff's actions fell below the relevant "standard of care". Historically, the minimum standard of care in English tort law has been determined by recourse to accepted medical practice and not written guidelines³²⁰. In *Bolam v. Friern Hospital Management Committee* (1957), McNair J found that a claim of negligence must fail if a clinician "acted in accordance with a practice accepted as proper by a responsible body of medical men skilled in that particular art". Under *Bolam*, a doctor could freely depart from national guidelines as long as she could find expert witnesses willing to state that this was acceptable behaviour³²⁰. However, *Bolam* has been revised over the last 20 years, most notably by *Bolitho v. City and Hackney Health Authority* (1996) in which the House of Lords found that the minimum standard of care is a question of law to be determined by the court rather than other doctors. According to Lord Brown-Wilkinson, "the court has to subject the expert medical evidence to scrutiny and to decide whether the practice is reasonable... the issue of reasonableness is for the court and not the medical profession". This move towards an objective test for determining the minimum standard of care has enhanced the significance of clinical guidelines. The courts have shown considerable willingness to follow national guidelines when determining the legal standard of care in *Re C* (1998), *Penney, Palmer and Canon v. East Kent Health Authority* (2000) and *Fotadar v. St George's Healthcare NHS Trust* (2005). The very low compliance with NICE CG124 should therefore be a concern for orthopaedic surgeons treating hip fractures. A patient eligible for THA who nevertheless underwent HA could succeed in a claim for negligence if they were

to go on to develop complications that are specific to HA, such as acetabular wear requiring revision. Although clinicians might have good reason to depart from NICE CG124, it is clear that the THA recommendation should be actively considered and any reasons for non-compliance documented carefully.

Although many hip fracture patients lack capacity to make informed treatment decisions and are treated according to their “best interests” under s.4 of the Mental Capacity Act 2005, a substantial proportion are able to consent for themselves. The median AMTS of patients recorded in the NHFD is 9 (IQR 5-10)³²¹. As the THA recommendation in NICE CG124 relates specifically to those without cognitive impairment, it is likely that a majority of these patients are in a position to consent for themselves. A surgeon that operates on a competent patient in the absence of consent risks liability in the torts of battery or negligence³²². A successful claim in the tort of battery is only likely where there was “a complete absence of consent” and, in cases where the consent process was alleged to be “deficient”, a claim is more likely to arise in the tort of negligence.

The standard of care required of a surgeon taking consent from a patient has historically been defined by *Bolam*; an approach re-affirmed by the House of Lords in *Sidaway v. Board of Governors of the Bethlem Royal Hospital* (1985). In *Sidaway*, Lord Diplock warned that a frank discussion with the patient about alternative treatments could risk “detering the patient from undergoing the treatment which in the expert opinion of the doctor it is in the patient’s interest to undergo... to decide what risks the existence of which a patient should be voluntarily warned... is as much an exercise of professional skill and judgment as any other part of the doctor’s comprehensive duty of care to the individual patient”. Perhaps unsurprisingly, the paternalistic approach in *Sidaway* has since been overruled by their lordships in *Montgomery v. Lanarkshire Health Board* (2015). In *Montgomery*, a pregnant woman of small stature and with diabetes was not warned about the risk of shoulder dystocia from a vaginal birth because her surgeon did not believe that the alternative mode of delivery (i.e. caesarean section) was justified. The delivery

was complicated by shoulder dystocia and - during instrumented delivery - the child suffered hypoxic injury and cerebral palsy. The claimant argued that she would have pursued a caesarean section had she been warned that there was even a small risk of this outcome from a vaginal delivery. In their judgement, Lords Kerr and Reid wrote that “the doctor is... under a duty to take reasonable care to ensure that the patient is aware of any material risks involved in any recommended treatment, and of any reasonable alternative or variant treatments”. They added that the test of “materiality” was “whether, in the circumstances of the particular case, a reasonable person in the patient’s position would be likely to attach a significant risk, or the doctor is or should be aware that the particular patient would be likely to attach significance to it”. Although this definition of materiality was formulated in relation to risk, it seems likely that the same test would be used to determine whether or not the patient should have been informed about alternative treatments³²³.

THA is clearly an alternative treatment to HA that should be discussed with competent patients with a displaced intracapsular hip fracture as part of obtaining informed consent. It would be difficult to justify not including THA as part of the consent conversation in the patient group for whom this procedure is recommended by NICE. Although surgeons might be reluctant to discuss THA because of logistical concerns (e.g. lack of a specialist hip surgeon or the possibility of introducing delay), it is unlikely that such a justification would satisfy the principles laid down in *Montgomery*.

Socioeconomic inequalities

This study found that deprivation quintile was inversely associated with use of THA, both amongst patients that satisfied the NICE criteria and those that did not.

Health inequalities and barriers to accessing healthcare are well-documented in private insurance-based healthcare systems, such as in the US³²⁴. The NHS, by contrast, has provided universal and comprehensive healthcare “free at the point of delivery” since 1948³²⁵. Most commentators recognize that equity of access to

healthcare is an important component of healthcare quality. However, this goal is formalized in the NHS by s.23(1)(13G) of the Health and Social Care Act 2012, which imposes a duty on commissioners to “have regard to the need to... reduce inequalities between patients with respect to their ability to access health services”. There is nevertheless a substantial literature that documents inequality of outcomes within the UK by patient characteristics, such as age³²⁶, sex³²⁷, ethnicity³²⁸, and socioeconomic status^{329,330}.

A number of studies have reported that hip fracture mortality in the UK is associated with socioeconomic deprivation^{65,331,332}. The largest study analysed administrative data from England and Wales (HES and PEDW respectively) and reported that socioeconomic deprivation was associated with 30-day mortality (most deprived quintile OR 1.19, 95% CI 1.15-1.23), which persisted at least as long as 365-days³³². Globally, worse hip fracture outcomes have also been associated with socioeconomic deprivation in Denmark³³³, Italy³³⁴, Taiwan³³⁵, and the US³³⁶.

Most explanations for such disparities are complex and multi-factorial³³⁷ but one possibility is that poorer hip fracture patients receive lower quality care. Although perhaps surprising in a system providing universal healthcare, studies across a range of conditions have reported that richer patients receive slightly better quality of care across the NHS^{325,338–340}. A number of explanations have been suggested for this finding, including patient choice, under-provision of high-quality healthcare services in deprived areas, and logistical barriers to accessing care, such as difficulties travelling long distances for specialist care and taking time away from work^{340,341}. However, some studies have raised the possibility that inequalities might be driven - at least in part - by variations in clinician decision making.

There are many potential explanations for the observation in this chapter that socioeconomic deprivation is inversely associated with receiving NICE-compliant treatment. These include variations in confounding factors (e.g. co-morbidity burden not adequately captured by ASA) and differences in patient preferences. Another possibility is that affluent patients are more likely to ask questions and advocate for

themselves³⁴². However, it is also possible that heuristic judgments about which patients are sufficiently “independent” to benefit from THA could be influenced by implicit surgeon bias.

Studies using approaches such as the implicit association test (IAT) consistently report that medical students and doctors exhibit implicit preferences for patients with high socioeconomic status^{343,344}. There is also evidence to suggest that clinicians use socioeconomic status as a surrogate for other personality characteristics, such as a patient’s normal level of physical activity³⁴⁵ and desire for information about alternative treatment options³⁴². It is easy to imagine how this could lead to reduced provision of THA for poorer patients. However, the evidence as to whether or not such biases influence treatment decisions is mixed and less certain^{344,346}. Although the data in this chapter suggest an inverse association between socioeconomic deprivation and provision of NICE-compliant treatment, it cannot show that compliance with CG124 is influenced by surgeon implicit bias. However, the inverse association documented in this chapter does risk exacerbating health inequalities and provides further reason to promote clear, evidence-based, national guidelines.

A hip fracture weekend effect

The limited availability of suitably experienced hip surgeons discussed previously (Section 5.4.1 on page 136) might account for the reduced provision of THA observed at weekends. This finding is important in the context of recent proposals to develop seven-day services across the NHS³⁴⁷.

The existing discussion about weekend outcomes has been principally framed around increased weekend mortality^{348,349}. Although individual centres have previously reported worse hip fracture mortality at weekends³⁵⁰, this was not found by two independent analyses of data from the NHFD^{68,351}. However, the focus on weekend mortality risks masking other concerns about variable hip fracture treatment at weekends. For example, a previous study using NHFD data reported that hip fracture patients are less likely to receive a prompt operation or to be assessed

by a geriatrician when presenting at the weekend³⁵¹. The finding in this chapter that access to THA may be restricted outside normal working hours adds additional nuance to the concern about variable NHS care at weekends.

Regionalization of hip fracture services might be one means of ensuring equitable access to THA by providing a critical mass of specialist hip surgeons able to support such a service every day. Dedicated hip fracture centres have already been successfully piloted in Germany^{352,353}. However, the potential benefits of regionalization would need to be weighed against competing considerations such as the desire of older adults to be treated close to their homes³⁵⁴.

Opportunities to learn from non-compliance

This study implicitly treated NICE CG124 as the “gold standard” for defining high quality hip fracture treatment. However, although there is good evidence for THA as a treatment for hip fracture^{58,286,289,290}, its precise indications are not well-defined. The optimal RP model suggested that surgeons might consider factors that could be relevant even if not strictly included within the NICE guidelines. For example, older patients were less likely to undergo THA, as were those that mobilized using a stick compared to those mobilising independently without aids. This model offers a glimpse into the collective judgment of orthopaedic surgeons and could be used to help inform the development of future NICE guidelines in the absence of higher-level evidence.

5.4.4 Impact of the study

This study highlighted substantial non-compliance with the THA recommendation in NICE CG124. An early iteration of this study was published in *TheBMJ*³⁵⁵ together with a linked editorial³⁵⁶ and reported by popular media outlets, such as *The Daily Mirror*³⁵⁷. It was subsequently used as a justification by NICE for updating the THA recommendation: “a decision to update this part of the guideline was made after topic experts noted a population based study that reported a low level of com-

pliance (around 30% nationally) with the NICE CG124 recommendation”³⁰². In a 280-page addendum to the guideline, NICE undertook a full review of the evidence around clinical and cost effectiveness of arthroplasty interventions for displaced intracapsular hip fractures³⁰². As a consequence of this review, NICE re-asserted its support for THA in selected patients with this fracture type and announced a new quality standard, which states that “adults with displaced intracapsular hip fracture receive cemented hemiarthroplasty or, if they are assessed as clinically eligible, a total hip replacement”. In the published paper, we recommended that the NHFD “report data on THA provision at the hospital-level to help achieve greater consistency across the NHS”. This recommendation was adopted by the NHFD, which now audits individual hospitals against this standard on an annual basis²⁰⁵.

5.4.5 Conclusion

Compliance with the NICE recommendation on THA for hip fracture is improving but remains poor. There continues to be substantial inter-hospital variation in practice, which is not readily explained by patient-level differences. The limited use of THA amongst patients from deprived areas, the inappropriately high use amongst patients from more affluent areas and inequalities in the provision of treatment at the weekend are particular concerns. Despite clear national guidelines, systematic differences in use of THA for hip fracture persists throughout the NHS. There are now multiple efforts afoot - both from NICE and the FFFAP - to further standardize care for this patient population.

Although a number of barriers to THA provision have been discussed in Section 5.4.1 on page 136, one that could be readily addressed is the professional consensus around the evidence for THA. A large multi-centre RT is in progress³⁰⁹ but - in the meantime - there are opportunities to learn from previous RTs as well as “real world” data from the NHFD. This is the challenge that forms the focus of Chapter 6 on page 148.

Chapter 6

Total hip arthroplasty versus hemiarthroplasty for intracapsular hip fractures

6.1 Introduction

The data in Chapter 5 on page 124 reported unexplained variation in provision of THA to patients that satisfy the criteria recommended by NICE. One possible explanation for non-adherence to NICE CG124 is that surgeons are not yet convinced by the evidence underlying this guideline. This chapter aims to review the RT evidence underlying CG124 as well as consider the impact of increasing provision of THA in the “real world”, i.e. outside the controlled environment of clinical trials¹.

Although the NHFD was established for the purposes of national clinical audit, it also represents a comprehensive cohort study of older adults with hip fractures in England, Wales, and Northern Ireland. In previous chapters, this thesis has framed the NHFD as a vehicle for improving outcomes through publishing performance data (Chapter 2 on page 42), effectively identifying performance outliers (Chapter 3 on page 72), and administering performance-based remuneration to healthcare

¹Chapter published as Metcalfe D, Judge A, Perry DC, Gabbe B, Zogg CK, Costa ML. Total hip arthroplasty versus hemiarthroplasty for independently mobile older adults with hip fractures. *BMC Musculoskelet Disord.* 2019;20:226.

providers (Chapter 4 on page 101) as well as auditing compliance with evidence-based guidelines and tracking healthcare equity (Chapter 5 on page 124). However, as discussed in Subsection 1.7.1 on page 38, the NHFD also provides an opportunity to embed studies aimed at learning from the outcomes of hip fracture patients and creating generalizable knowledge that can be used to inform treatment decisions. It is this element of the NHFD that will be exploited by the present chapter.

6.1.1 HA versus THA

The treatment of displaced intracapsular hip fractures has been discussed in Section 1.4 on page 22 and Subsection 5.1.1 on page 124. Although HA is performed more frequently, a number of organizations (such as the American Academy of Orthopaedic Surgeons (AAOS)⁴⁹) and the UK NICE⁵⁵ recommend offering THA to selected hip fracture patients owing to perceived functional benefits. NICE recommends offering THA to patients that (1) could walk independently before the fracture (2) are not cognitively impaired and (3) are medically fit for both anaesthesia and the procedure⁵⁵. Despite this recommendation, an international survey of orthopaedic surgeons found that 73% favour HA³⁰⁷, with studies demonstrating less than a third of eligible patients actually receive THA³⁵⁵. One explanation for this discrepancy is that the evidence in support of THA is mixed. A number of small RTs have suggested that THA is associated with better functional outcomes, fewer wound infections, and reduced need for secondary procedures^{51–54}. However, THA is also a more complex procedure that requires longer surgical time, is associated with greater blood loss, and has a higher risk of subsequent dislocation⁵⁰.

It is also uncertain whether the reported benefits for THA over HA^{51–54} can be replicated beyond the controlled environment of clinical trials. For example, there is a clear association between THA outcome and surgeon volume³⁵⁸ and it is likely that patients will be preferentially recruited to THA trials by experienced arthroplasty surgeons. It has been suggested that increasing the number of generalist surgeons providing THA will offset the benefits of this intervention for patients with hip frac-

tures⁴⁹. Similarly, there are concerns that the unavailability of appropriately trained arthroplasty surgeons might delay operative treatment, and delays are thought to worsen outcomes for this vulnerable patient group^{269,359}. It is for these reasons that the “real world” effect of increasing use of THA in the hip fracture setting has been identified as a hip fracture research recommendation by the AAOS⁴⁹.

The objectives of this chapter were to undertake an updated meta-analysis of RTs and use data from a comprehensive national cohort of hip fractures to provide “real world” context to the existing trial literature. The overall aim was to compare the outcomes between HA and THA for independently mobile older adults with hip fractures.

6.2 Methods

6.2.1 Systematic review and meta-analysis

A scoping review identified a number of previous systematic reviews that compared HA and THA for patients with displaced intracapsular hip fractures. A simplified search strategy was undertaken using a modification of the method first proposed by Sampson 2008 et al³⁶⁰, which has been shown to be highly sensitive (median sensitivity 100%) for identifying RTs when applied to systematic reviews with clinically focused research questions³⁶¹. A broad search strategy (fracture* AND (“total hip” OR hemiarthroplasty) AND “systematic review”) was used to search three databases (Medline 1966-, EMBASE 1947-, and CINAHL 1982-) on 1st August 2018 to identify previous reviews comparing HA and THA. The reference lists of all reviews were searched and the forward citation facility in PubMed used to identify trials published after each systematic review. Trial reference lists and citations were also searched for further studies. No language restrictions were applied. The full texts of all RTs were then screened by two authors to identify those satisfying the following inclusion criteria. A single author evaluated studies published in Chinese with help from a Chinese-speaking health economist with experience of hip fracture research.

The inclusion criteria were:

- A RT or NRT.
- Including patients predominantly aged >60 years with displaced intracapsular hip fractures.
- Excluding patients that had cognitive impairment or limited mobility before injury.
- Reporting dislocation, revision, mortality, unplanned re-admission, functional outcomes or health-related quality of life (using any validated scale).

Study characteristics and outcome data were extracted by one author and checked by a second. The intention was to report all outcomes at 12-months for consistency. Two authors independently determined risk of bias using criteria recommended by the *Cochrane Handbook*¹⁵⁸ and resolved disagreements by consensus. These data were presented to guide judgements about the certainty of the evidence and not to determine eligibility for inclusion within meta-analyses. Data were pooled to estimate risk ratios (for categorical outcomes) and mean differences (for continuous outcomes) using the DerSimonian and Laird method for random-effects meta-analysis³⁶² as high levels of between-study heterogeneity were anticipated when pooling trials from different patient populations and healthcare settings. Standardized mean differences were reported when studies reported the same outcome measured on difference scales. When studies did not provide standard deviations necessary to inform confidence intervals, these were calculated from absolute p-values¹⁵⁸. Meta-analyses were undertaken using RevMan v.5.0 (Cochrane Collaboration, Vienna, Austria). The systematic review was reported in line with the PRISMA statement³⁶³ and the protocol registered prospectively in the PROSPERO database with reference CRD42018109415.

6.2.2 Observational cohort study

An observational study was undertaken using a comprehensive national cohort of older adults with displaced intracapsular hip fractures to extend and contextualize

the existing RT literature. Propensity score matching was used to mimic randomization as far as is possible using observational data.

Data sources

The cohort was defined using the NHFD, and patient records linked both to the HES APC dataset and ONS civil death registration data.

National Hip Fracture Database The NHFD has already been described comprehensively in Section 3 on page 72.

Hospital Episode Statistics The HES APC dataset has been described in Section 3.2.1 on page 77. In brief, it includes data on all admissions to NHS hospitals and to independent sector providers that are funded by the NHS²⁰¹, which accounts for approximately 99% of hospital activity in England²⁰¹.

Office for National Statistics Civil registration death records previously held by the ONS have been described in Section 3.2.1 on page 77. These data should be complete except for the small number of cases referred to a coroner, which cannot be registered until coronial enquiries are complete and a death certificate has been issued.

Study population

The study period was 28th March 2011 until 4th January 2017. The start date was the earliest point at which the NHFD captured unique patient identifiers that could facilitate linkage to other datasets and the end date was chosen to facilitate 12 months follow-up. Only patients treated at hospitals in England were included. The inclusion criteria were otherwise those recommended by NICE⁵⁵ and described in Section 5.1 on page 124:

- All adults aged >60 .

- Displaced fracture of the femoral neck that was deemed unsuitable for internal fixation.
- Independently mobile or using a single stick before injury.
- Medically fit to undergo hip arthroplasty, defined as an ASA grade $<2^{355}$.
- Patients without substantial cognitive impairment, defined as an AMTS $>8^{355}$.

Patients that presented to hospitals in Wales, Northern Ireland, and the Isle of Man were excluded as HES only captures data from hospitals in England. Cases were also excluded if they could not be positively matched to records within HES based on their NHS number, sex, date of birth, and full post-code.

Outcomes

The primary outcomes were dislocation, revision, and mortality within 12-months. The secondary outcomes were surgical delay, length of stay, discharge to own home, and re-admission within 30 days. Surgical delay, length of stay, and discharge destination were available directly from the NHFD. Revision operations were identified from HES and defined by Office of Population Censuses and Surveys (OPCS) v4 (OPCS4) procedure codes previously used in other studies and incorporating codes recommended for this purpose by the United Kingdom (UK) NJR¹²⁵. Dislocation OPCS4 codes were identified by manual searches using *disloc**, *manipula**, and *reduc**. The diagnostic codes used to define dislocation and revision are reproduced in Table G.2 on page 267 and Table G.3 on page 268 respectively.

Statistical analysis

Matching Propensity scores were calculated that represented the estimated probability of each patient undergoing THA based on characteristics that are known to be associated with outcome in this population: age, sex, pre-injury mobility status, admission source, ASA physical status grade, CCI, AMTS, and IMD³⁶⁴. CCI was determined using diagnostic codes reproduced in Table G.1 on page 266. The model was otherwise specified iteratively to achieve the best possible match, as judged

by visual inspection of the distribution of propensity scores after matching and plots of co-variables against propensity scores by treatment status. Post-estimation statistical checks³⁶⁵ were also undertaken, which included t-tests for differences in means and confirmation that the standardized mean difference for each co-variable between the groups was $<1\%$ ³⁶⁶. The final model utilized 1:1 nearest neighbour matching with a 0.02 calliper (as recommended by Austin³⁶⁷), no replacement, and the common support restriction. All subsequent descriptive, regression, and survival analyses were confined to the propensity score matched groups.

Matching sensitivity analyses In the absence of randomization, it could not usually be assumed that the probability of undergoing THA would be equal across the HA and THA groups. However, this *may* be true of a matched cohort *if there are no unobserved confounders*, which could then permit causal inferences to be drawn³⁶⁸. As it is not possible to determine the magnitude of any such hidden biases from observational data, this study employed the bounding technique proposed by Rosenbaum³⁶⁹. This approach employs sensitivity analyses for each primary outcome to estimate the extent to which claims about the significance of an association depend on the assumption that there are no unobserved confounding variables³⁷⁰. All three primary outcomes (see Subsection 6.2.2 on the previous page) were binary variables. Rosenbaum’s approach uses a modification of McNemar’s test, which is based on a simple 2x2 cross-tabulation. In this study, McNemar’s statistic simply represents the difference between the total number of patients undergoing THA and the total number of outcomes (e.g. dislocation) in the THA group. The statistic has a Chi-square distribution and an associated p-value, which is valid under the assumption that there is not any unobserved confounding³⁶⁸. In Rosenbaum’s sensitivity analysis, upper and lower bounds on McNemar’s statistic are calculated based on values of “gamma”, which are calculated using a formula described elsewhere³⁶⁹. For $\gamma = 2$, any given individual in a matched pair could have been twice as likely to receive THA due to an unobserved confounder and the association with the outcome being tested would still have been significant at the threshold $p < 0.05$ ³⁶⁸.

Descriptive statistics Categorical variables were compared using Chi-square tests and non-normally distributed continuous variables using the Kruskal-Wallis one-way analysis of variance test. Length of stay data were only analysed for the proportion of patients that were discharged alive from hospital to prevent left skew caused by early deaths.

Survival analysis Kaplan-Meier estimates were plotted with 95% confidence intervals for cumulative survival free from unplanned secondary procedures. The proportional hazards assumption was tested by visual examination and statistical assessment of the relationship between event time and Schoenfeld residuals. The proportional hazards assumption was satisfied and so Cox regression models were fitted with the primary outcome (dislocation and/or revision) as the independent variable. Mortality is high in this population and so a sensitivity analysis was undertaken using Fine & Gray competing risks regression models with death specified as the competing risk³⁷¹. Competing risks regression models were also fitted for dislocation and revision as individual events. The co-variables for all regression models were those described in Paragraph 6.2.2 on page 153 as the basis for propensity score matching, which include five of the six used routinely in the NHFD for case mix adjustment²²⁷. The sixth NHFD case mix co-variable (i.e. fracture type) was not used because only patients with displaced intracapsular hip fractures were included in this study. Year of fracture was included as an ordinal variable within regression models to account for the possibility of changing outcomes over time.

Multivariable regression Multivariable logistic regression was used to adjust for residual imbalance between the two groups in respect of discharge to own home and 30-day re-admission. The co-variables were as specified above. Length of stay data conformed to a gamma distribution and so were adjusted using a generalized linear model (GLM) together with post-estimation calculations of average marginal effects to yield predicted mean differences and 95% CI. Logistic regression and GLMs utilized cluster-robust standard errors and robust variance estimators³⁷² to account

for the lack of independence between matched records³⁷³.

Propensity score matching was achieved using the MatchIt application³⁷⁴ and Rosenbaum bounds estimated using the rbounds application³⁷⁵ for R (R Foundation for Statistical Computing, Vienna, Austria). All subsequent analyses were undertaken using StataIC v.15 (StataCorp, College Station, TX, USA). Two tailed $p < 0.05$ was adopted *a priori* as the threshold for statistical significance.

Information governance

Linkage of NHFD data was supported by the HQIP Data Access Request Group (DARG) and the Confidentiality Advisory Group (CAG) on behalf of the Secretary of State for Health and Social Care under s.251 NHS Act 2006 (CAG ref 8-03(PR11)/2013). Access to HES was approved by the NHS Digital IGARD and civil registration mortality data by the ONS Micro Panel Release Panel (MRP) under s.39(5) Statistics and Registration Service Act 2007. Formal research ethics committee approval was not required for secondary analysis of pseudonymized data in line with the NHS Health Research Authority guidelines²⁰⁰.

6.3 Results

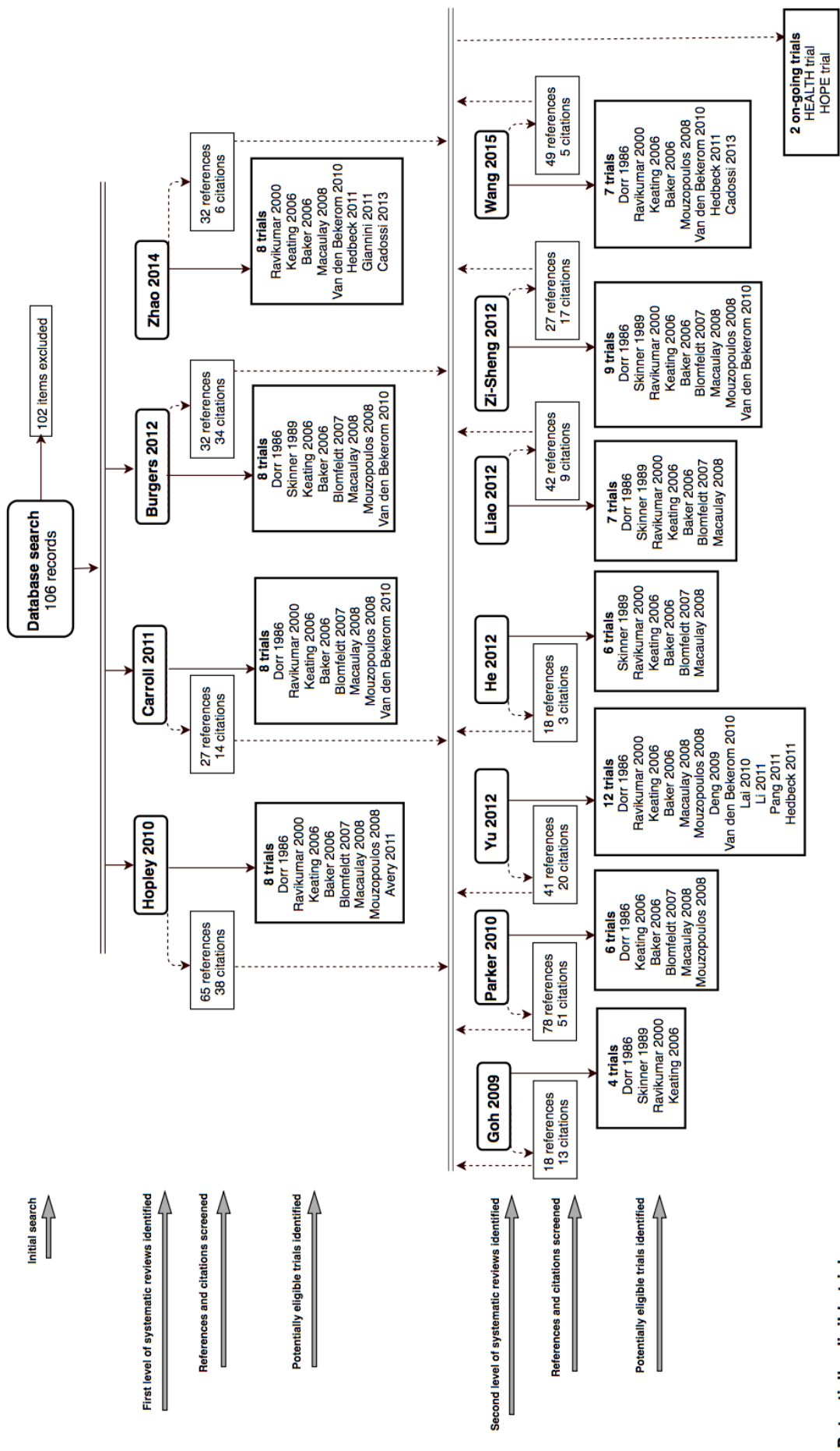
6.3.1 Meta-analysis of randomized trials

There were 11 previous systematic reviews^{50,54,58,286,290,376–381} but none reported analyses limited to patients that were cognitively intact and independently mobile before injury (Figure 6.1 on page 158). The 11 earlier reviews included 16 RT reports, which presented data from 14 individual RTs. Table H.1 on page 271 provides detailed explanations for excluding trial reports. Eight RTs did not satisfy the restricted inclusion criteria of this systematic review, e.g. they did not exclude patients with cognitive impairment or limited mobility. One study could not be retrieved despite extensive attempts.

Five RTs satisfied the eligibility criteria for this review and these are described in Table H.2 on page 272. Two were undertaken in the UK^{382,383} and one each in Sweden³⁸⁴, Italy³⁸⁵, and the USA³⁸⁶. A further eligible RT is on-going³⁰⁹. Characteristics of the RTs and risk of bias assessments are described in Table H.2 on page 272 and Table H.3 on page 273, respectively. All the RTs used adequate random sequence generation techniques and were judged to be at low risk of attrition bias as loss to follow-up was low. However, no RT sought to blind patients, personnel, or outcome assessors.

6.3.2 Observational cohort study

There were 143,871 patients with displaced intracapsular hip fractures that underwent HA or THA and could be matched to a record within HES (Figure 6.2 on page 159). 28,099 (19.5%) satisfied the pre-specified inclusion criteria, i.e. ASA <2, AMTS >8, and independently mobile. Table I.1 on page 275 shows that the groups initially varied considerably in terms of baseline characteristics. After propensity score matching, 12,290 cases were selected for further analysis, and Table 6.1 on page 160 shows that the baseline characteristics of the matched groups were similar. The distribution of propensity scores also improved after matching (see Figure 6.3 on page 163, Figure I.1 on page 276, Figure I.2 on page 277, and Figure I.3 on page 278).



Dorr 1986, Skinner 1989, Ravikumar 2000, Keating 2006, Baker 2006, Macaulay 2008, Mouzopoulos 2008, Deng 2009, Van den Bekerom 2010, Lai 2010, Li 2011, Pang 2011, Avery 2011, Hedbeck 2011, Giannini 2011, Cadossi 2013.

Figure 6.1: PRISMA flow diagram showing randomized and quasi-randomized controlled trials from previous systematic reviews

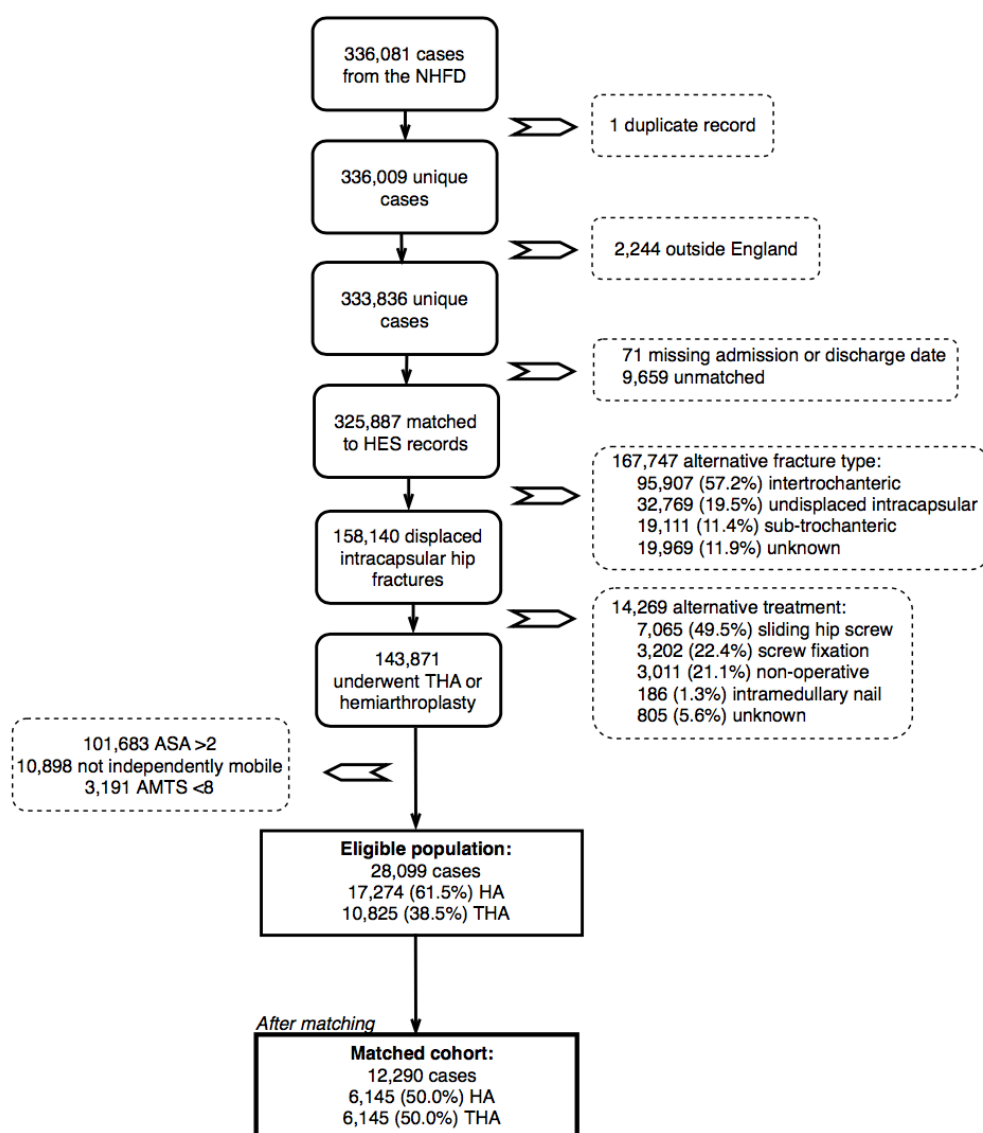


Figure 6.2: A flow diagram showing inclusion of cases within the study.

Table 6.1: Characteristics of the matched population

	Hemiarthroplasty	Total hip arthroplasty	Total
Age*	77 (72-81)	77 (73-81)	77 (73-81)
Sex**			
Male	1,347 (21.9%)	1,321 (21.5%)	2,668 (21.7%)
Female	4,798 (78.1%)	4,824 (78.5%)	9,622 (78.3%)
ASA*	2 (2-2)	2 (2-2)	2 (2-2)
Pre-injury mobility**			
Independently mobile	5,308 (86.7%)	5,326 (86.7%)	10,634 (86.7%)
Mobile indoors with one aid	837 (13.6%)	819 (13.3%)	1,656 (13.5%)
AMTS*	10 (10-10)	10 (10-10)	10 (10-10)
Admission source**			
Own home	6,071 (98.8%)	6,092 (99.1%)	12,163 (99.0%)
Rehabilitation unit	8 (0.1%)	2 (0.0%)	10 (0.1%)
Residential/nursing home	37 (0.6%)	18 (0.3%)	55 (0.5%)
Acute hospital	29 (0.5%)	33 (0.5%)	62 (0.5%)

*Median (interquartile range)

**Number (percentage)

^aKruskall-Wallis one-way analysis of variance

^bChi-square test

The primary and secondary outcomes from the propensity score matched study are summarized in Table 6.2 on the following page.

Table 6.2: Primary and secondary outcomes from the propensity score matched study

	Hemiarthroplasty	Total hip arthroplasty	P
Dislocation (12-months)	57 (0.9%)	96 (1.6%)	0.002*
Revision (12-months)	THA sub-distribution hazard ratio 1.73 (95% CI 1.24 to 2.41)	67 (1.1%)	b
Mortality (12-months)	106 (1.7%)	159 (2.6%)	<0.001*
	THA sub-distribution hazard ratio 0.66 (95% CI 0.48 to 0.90)		**
	338 (5.4%)		<0.001*
	THA hazard ratio 0.45 (95% CI 0.37 to 0.54)		**
Surgical delay (hours)	22.2 (17.8-39.0)	23.9 (18.9-40.6)	<0.001
Length of stay (days)	10 (7-15)	9 (7-13)	<0.001
	THA predicted mean difference -1.92 (95% CI -2.30 to -1.55)		daysf
Discharge home	5,017 (80.7%)	5,519 (88.6%)	<0.001*
	THA adjusted odds ratio 1.77 (95% CI 1.58 to 1.99)		^
Re-admission (30-days)	361 (5.9%)	356 (5.8%)	0.847
	THA adjusted odds ratio 0.96 (95% CI 0.82 to 1.11)		^

*Chi2 test; **Competing risks regression model; ***Royston-Parmar flexible parametric model;
§Median (interquartile range); §§Kruskall-Wallis one-way analysis of variance; §§§Generalized linear model; ^Multivariable logistic regression model.

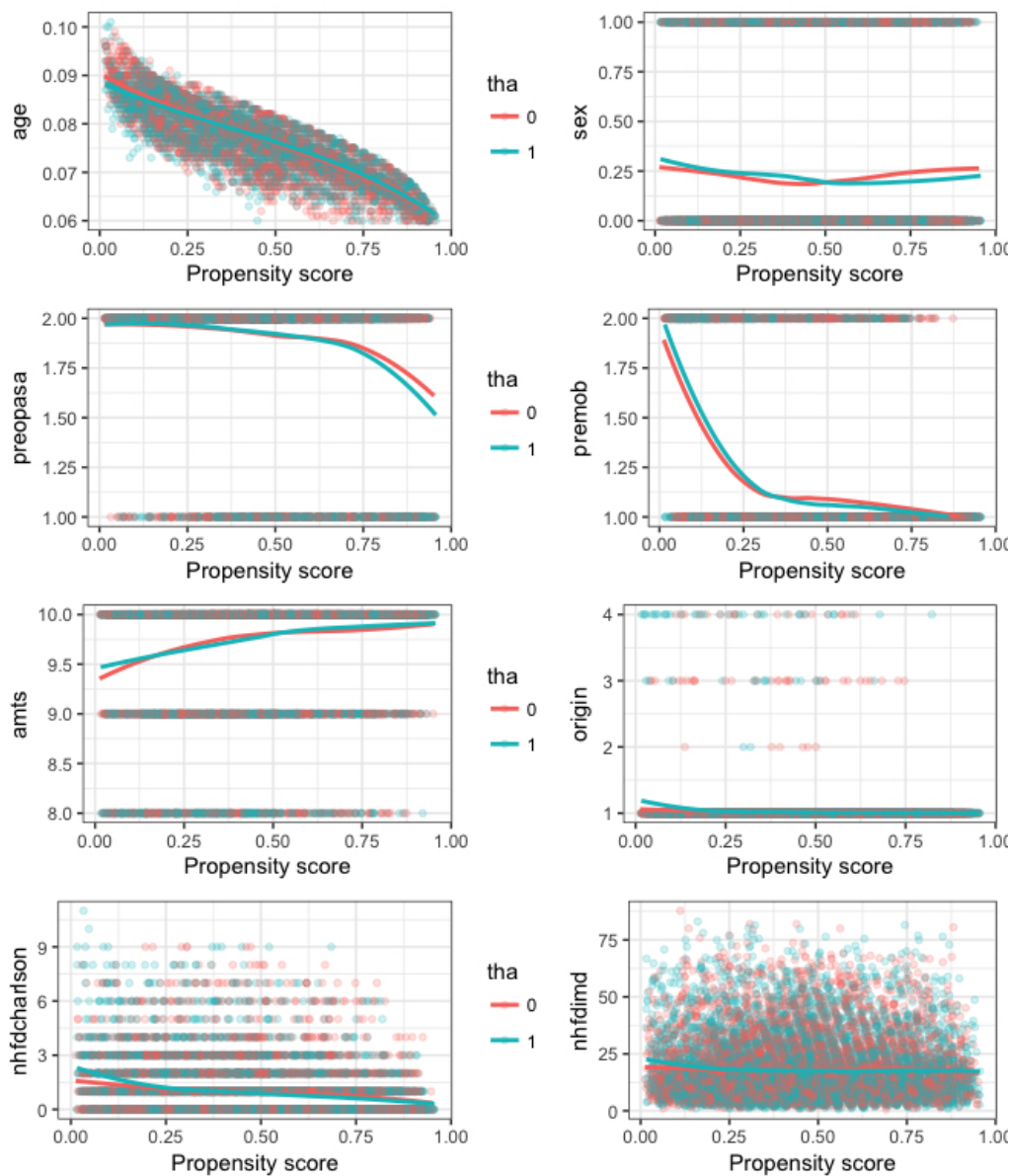


Figure 6.3: Co-variables plotted against propensity scores by treatment status. If the distributions are identical, this indicates that the groups have the same mean for each value of the propensity score and so are well matched. "tha 0" represents patients in the HA group and "tha 1" those in the THA group

Sensitivity analyses

Rosenbaum sensitivity analyses were undertaken (as described in Paragraph 6.2.2 on page 153) to estimate the magnitude of effect that an unobserved confounding variable would need to exert in order to undermine the statistical significance of a reported association. These are shown in Table 6.3 on the following page and described in the relevant paragraphs in Section 6.3.3 on page 165.

Table 6.3: Rosenbaum sensitivity tests

	Outcome						
	Dislocation		Revision		Mortality		
Gamma	Lower	Upper	Lower	Upper	Gamma	Lower	Upper
1.0	0	0.0001	0	0.0000	1.0	0	0.0000
1.5	0	0.0008	0	0.0000	1.1	0	0.0000
2.0	0	0.0031	0	0.0002	1.2	0	0.0000
2.5	0	0.0072	0	0.0006	1.3	0	0.0001
3.0	0	0.0127	0	0.0016	1.4	0	0.0013
3.5	0	0.0192	0	0.0032	1.5	0	0.0122
4.0	0	0.0264	0	0.0054	1.6	0	0.0616
4.5	0	0.0339	0	0.0081	1.7	0	0.1898
5.0	0	0.0416	0	0.0113	1.8	0	0.3997
5.5	0	0.0493	0	0.0148	1.9	0	0.6317
6.0	0	0.0569	0	0.0187	2.0	0	0.8151
6.5	0	0.0644	0	0.0228	2.1	0	0.9237
7.0	0	0.0716	0	0.0270	2.2	0	0.9738
7.5	0	0.0786	0	0.0313	2.3	0	0.9924
8.0	0	0.0855	0	0.0357	2.4	0	0.9981
8.5	0	0.0920	0	0.0401	2.5	0	0.9996
9.0	0	0.0984	0	0.0446	2.6	0	0.9999
9.5	0	0.1045	0	0.0490	2.7	0	1.0000
10.0	0	0.1103	0	0.0534	2.8	0	1.0000
Gamma represents the odds of differential assignment to treatment due to unobserved factors. Cells in bold font represent the highest p-value <0.05 and so the critical gamma for each outcome.							

6.3.3 Primary outcomes

Dislocation

All 5 RTs reported risk of dislocation. Although the pooled effect estimate suggested higher risk of dislocation amongst those undergoing THA, this was not significant (THA 9/233 [3.9%] versus HA 2/234 [0.9%], risk ratio (RR) 2.77 [95% 0.81 to 9.48], Figure 6.4). Within the propensity score matched cohort, those undergoing THA were significantly more likely to dislocate than those with HA (1.6% versus 0.9%, X^2 $p=0.002$). This finding persisted when adjusting for co-variables in a competing risks regression model (THA sub-distribution hazard ratio [standardized hazard ratio (SHR)] 1.73, 95% CI 1.24 to 2.41). The Rosenbaum sensitivity analysis reported a critical gamma of 5.5 (Table 6.3 on the previous page), which suggests that any given individual in a matched pair could have been 5.5 times as likely to receive THA due to an unobserved confounder and the association with revision surgery would still have been significant at the threshold $p<0.05$.

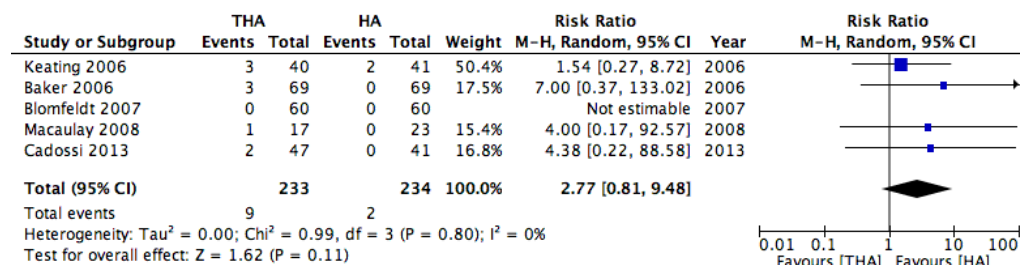


Figure 6.4: A forest plot showing risk of 12-month dislocation within eligible clinical trials.

Revision

All five RTs reported risk of revision^{52,382–385}. The pooled effect estimate was initially in favour of HA, although this association was not statistically significant (HA 8/234 [3.4%] versus 15/233 [6.4%], RR 1.52 [95% CI 0.56 to 4.14], Figure 6.5 on the following page). The association also diminished when the data reported by Cadossi et al³⁸⁵ were excluded as these authors had trialled a non-standard THA prosthesis and reported an unusually high revision rate (HA 8/193 [4.1%] versus 9/186 [4.8%],

RR 1.16 [95% CI 0.46 to 2.91]). However, within the propensity score matched cohort, a greater proportion of HA patients underwent revision surgery within the subsequent 12 months than THA (1.7% versus 1.1%, X^2 $p < 0.001$). This finding persisted when adjusting for co-variables in a competing risks regression model (THA SHR 0.66, 95% CI 0.48 to 0.90). The Rosenbaum sensitivity analysis reported a critical gamma of 9.5 (Table 6.3 on page 164), which suggests that this analysis was very insensitive to unobserved confounding.

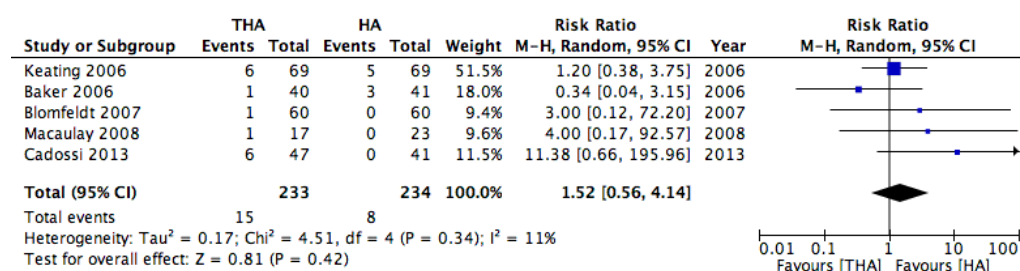


Figure 6.5: A forest plot showing risk of 12-month revision within eligible clinical trials.

Mortality

Four RTs reported mortality at 12 months and one at 6 months. A higher proportion of patients undergoing HA died (36/234, 15.4%) than those in the THA group (21/233, 9.0%; RR 0.63, 95% CI 0.38 to 1.04, Figure 6.6 on the following page). Within the propensity score matched cohort, 12-month mortality was higher in the HA group (5.4% versus 2.6%, X^2 $p < 0.001$) and this persisted within a multi-level flexible parametric survival model (hazard ratio 0.45, 95% CI 0.37 to 0.54). Twelve-month mortality within the observational cohort is illustrated by a Kaplan-Meier plot in Figure 6.7 on the next page. However, it is important to note that the mortality analysis was more sensitive to unobserved confounding than the other two primary outcomes (Table 6.3 on page 164). As gamma was 1.5, an unobserved confounder that increased an individual's likelihood of undergoing THA by 1.6 times could be sufficient to undermine the association with mortality.

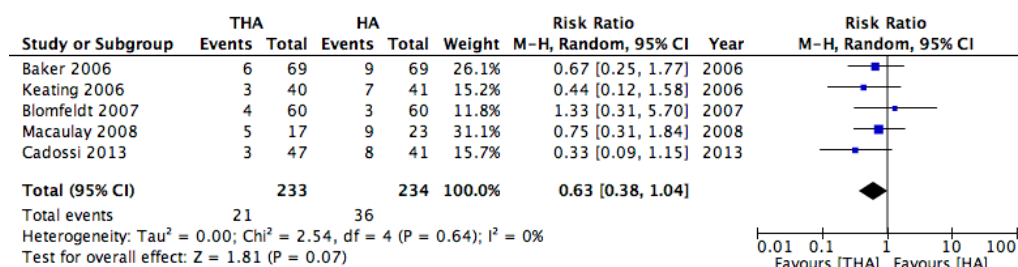


Figure 6.6: A forest plot showing risk of 12-month mortality within eligible clinical trials.

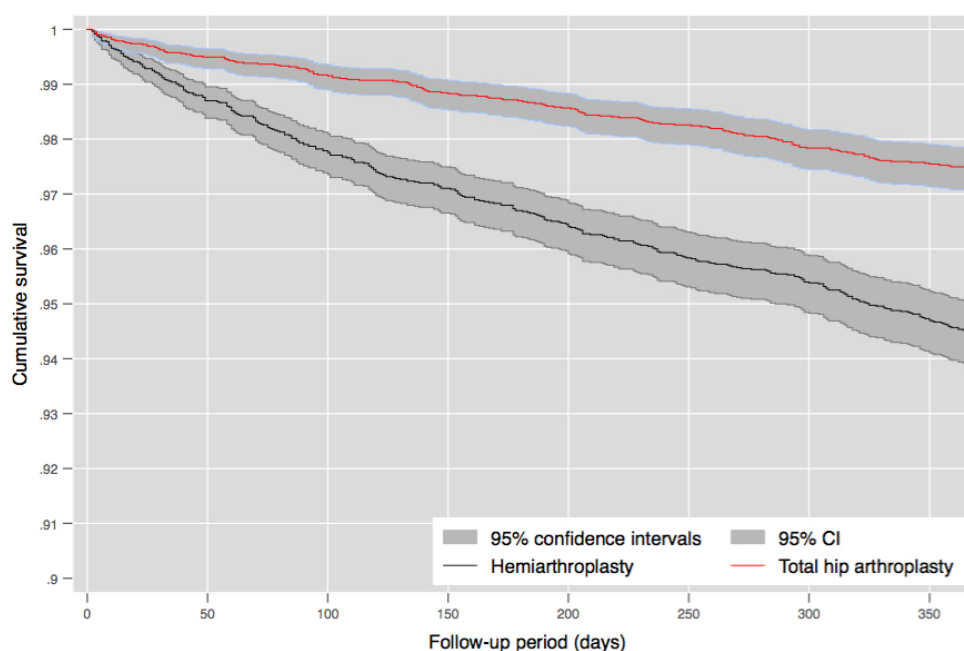


Figure 6.7: Kaplan-Meier plot showing mortality for patients in the propensity score matched cohort.

6.3.4 Secondary outcomes

Time to surgery

Two RTs^{382,385} (164 patients) reported no difference in time to surgery between THA and HA (THA mean difference -0.44 [95% CI -0.93 to 0.05]). Within the propensity score matched cohort, patients underwent HA more promptly than THA (median 22.2 [IQR 17.8-29.0] versus 23.9 [18.9-40.6] hours, Kruskal-Wallis $p < 0.001$).

Duration of surgery

All five RTs (462 patients) reported surgical duration. Although THA took longer than HA, and this difference was statistically significant, the absolute effect was small (mean difference 15.0 [95% CI 6.4 to 23.7] minutes). Duration of surgery was not available from the propensity score matched cohort.

Length of stay

Two RTs (123 patients) reported length of stay and there was no intervention effect on this outcome (THA mean difference 1.50 [95% CI 0.00 to 3.00] days). In the propensity score matched cohort, patients undergoing HA stayed in hospital longer than those undergoing THA (median 10 [IQR 7-15] versus 9 [7-13] days, Kruskal-Wallis, $p < 0.001$). When adjusting for co-variables within a GLM, patients undergoing THA experienced a shorter length of stay (predicted mean difference -1.92 [95% CI -2.30 to -1.55] days).

Discharge destination

No RT reported discharge destination as an outcome. Within the propensity score matched cohort, a smaller proportion of patients undergoing HA were discharged to their own home than THA (80.7% versus 88.6%, X^2 $p < 0.001$). Within a multi-variable logistic regression model, those undergoing THA also had higher odds of being discharged to their own home (adjusted Odds Ratio (aOR) 1.77, 95% CI 1.58 to 1.99).

30-day re-admission

No RT reported unplanned re-admission to hospital as an outcome. Within the propensity score matched cohort, there was no statistically significant difference in 30-day re-admission between the two groups (HA 5.9% versus THA 5.8%, X^2 $p = 0.847$), and this finding persisted within a multivariable logistic regression model (aOR 0.96, 95% CI 0.82 to 1.11).

Hip functional outcomes

All five RTs reported joint-specific functional outcomes measured at 12-months. Three studies used the Harris Hip Score^{52,384,385} (234 patients) and one each used the Oxford Hip Score³⁸² (81 patients), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)⁵² (40 patients), and a bespoke hip questionnaire³⁸³ (138 patients). Higher scores on all of these measures reflect better outcomes except for the Oxford Hip Score in which a higher score represents worse function. There were no differences in terms of total score (THA standardized mean difference [SMD] 0.17 [95% CI -0.20 to 0.53]) or either the pain (-0.01 [-0.49 to 0.48]) or function (0.18 [-0.03 to 0.39]) domains. However, there was a difference in favour of THA once the data from Cadossi et al³⁸⁵ were excluded (THA mean difference 7.25 [95% CI 3.01 to 11.49]). The only study using the Oxford Hip Score reported a difference between the groups in favour of THA (THA mean difference -3.50 [95% CI -6.66 to -0.34]). However, the only study reporting data from a timed up and go (TUG) test⁵² (40 patients) – which measures the time that it takes a patient to rise from a chair, walk three metres, turn around, walk back to the chair, and sit down – did not find a difference between the groups (THA mean difference -0.70 [95% CI -8.01 to 6.61] seconds).

Health-related quality of life

Two studies^{52,382} (121 patients) reported components of the Short-Form 36 (SF36) and one the EuroQol Group 5D (EQ-5D)³⁸³ (183 patients). There were no differences in the mental (THA mean difference 2.30 [95% CI -8.57 to 13.18]) or physical (2.98 [-0.89 to 6.85]) component summary scores of the SF36 or the EQ-5D (0.10 [0.00 to 0.20]) utility score.

6.4 Discussion

This study identified five RTs that compared HA and THA amongst independently mobile older adults with displaced intracapsular hip fractures^{52,382–385}. These trials were typically small (median 89 patients) single-centre studies that were limited by few events (pooled totals 11/467 [2.4%] dislocations, 23/467 [4.9%] revisions, and 57/467 [12.2%] deaths). No individual trial reported differences in outcomes and it is even possible that the pooled analyses were underpowered to detect important differences between the groups. Data were therefore analysed from the largest available cohort of hip fracture patients and propensity score matching used to replicate randomization as far as is possible using observational data.

6.4.1 Primary outcomes

The observational data confirmed the non-significant trend reported by RTs that THA has a higher risk of 12-month dislocation. However, this study also found a 33% lower risk of 12-month revision for THA patients, which is contrary to the RT finding of “no difference” between the groups.

Dislocation

Previous systematic reviews comparing HA and THA (typically in unselected hip fracture populations) have reported increased risk of dislocation following THA, which is consistent with the meta-analysis presented in this chapter^{50,54,58,286,290,376–381}. Although the risk of dislocation did not reach the threshold for statistical significance, the pooled analysis was based on small numbers (ten dislocation events across five RTs) and so likely underpowered. This chapter was therefore supplemented with a national cohort study of hip fractures, which confirmed the signal from pooled RT data that dislocation is significantly associated with THA. Sensitivity analyses suggested that this finding was relatively insensitive to the effect of unobserved confounding.

The NHFD does not collect data about surgical approach, which is known to be associated with dislocation³⁸⁷ and so could be a further source of confounding.

As discussed in Section 6.1 on page 148, the AAOS have expressed concern that outcomes (such as dislocation rates) might be worse outside the controlled environment of RTs. In this chapter, the dislocation point estimate for HA was the same (0.9%) between the cohort study and meta-analysis whereas the estimate for THA was *lower* outside the RT environment (3.9% versus 0.9%). One possibility is that this represents differences in the effectiveness of follow-up between the two study designs. It is likely that some dislocations (e.g. a small number reduced and discharged home directly from the ED) might not have been recorded as an inpatient admission and so missed by HES whereas the RTs might be expected to achieve more robust patient follow-up. However, the dislocation rates reported in this cohort study are similar to those based on administrative datasets in the US³⁸⁸ and Canada³⁸⁹, as well as the UK³⁹⁰. Although frail patients may be more likely to be admitted after a dislocation (because of difficulty coping at home after such an event), the effect of such a bias would increase the capture of HA dislocations as this is the group that ought to be less mobile if residual confounding persists within the cohort. In any event, the data presented here do not support the AAOS concern that THA dislocation is higher in the “real world” than within RTs.

Revision

A more confusing picture emerged from the analyses using revision surgery as the outcome. Previous systematic reviews have reported higher rates of revision amongst patients undergoing HA^{50,54,58,286,290,376–381}, which is contrary to the meta-analysis presented in this chapter. One possibility is that HAs undertaken in the fittest patients (i.e. those included in the present meta-analysis) are less likely to require revision than those in the general hip fracture population. For example, patients with multiple co-morbidities might be more prone to infection, which is the cause for almost a quarter of early revisions following HA^{391,392}. However, the observational

cohort study data in this chapter supported the previous meta-analyses in finding higher revision rates following HA. This analysis was estimated to be very insensitive to the potential effects of residual confounding.

Unfortunately, the reasons for revision were not readily discernible from the data. One possibility is that increased revision surgery is a response to complications that are specific to HA, such as acetabular wear (Section 6.1 on page 148). However, it is also possible that revision of HA prostheses is more common because there is an obvious salvage procedure, i.e. THA.

As in the case of dislocation (Subsection 6.4.1 on page 170), absolute rates of revision surgery were lower in the cohort study than the meta-analysis, which may represent more effective follow-up by RTs. However, the rates were similar to those reported by other observational cohort studies^{388–390}. Previous work in other surgical settings has found that OPCS codes in HES can reliably be used to identify some operations, although this varies substantially between procedures³⁹³. However, a range of codes had to be used to define “revision surgery” and it is possible that some cases were not captured.

Mortality

Importantly, this cohort study identified a 58% lower risk of 12-month mortality for patients undergoing THA. This may represent the effect of residual confounding as there is likely to be a strong selection pressure on surgeons to select the fittest patients for THA. Importantly, the Rosenbaum sensitivity analyses for the mortality outcome showed that this analysis was much more sensitive to the potential impact of unobserved confounding than those for dislocation and revision. If unobserved confounding had the effect of increasing an individual's likelihood of receiving THA by 1.6 times, this would be sufficient to undermine the reported association with mortality. It is very plausible that a more accurate or precise measure of co-morbidity than ASA or an entirely unquantified variable (such as frailty) would surpass this threshold. For example, the analyses in Chapter 5 on page 124 showed that being

independently mobile pre-injury was highly associated with THA (OR 3.99, 95% CI 3.63 to 4.39) when compared with using a single stick. Although the study in this chapter accounted for mobility status (both by limiting the cohort to the most mobile categories and through matching), there is clearly potential for variables to have strong associations with selection for THA. The data in Chapter 5 on page 124 certainly suggest that surgeons employ a heuristic decision making when selecting candidates for THA, which is not readily captured by variables in the NHFD. However, although this finding from the observational cohort study may reflect residual confounding, a similar association was observed in the meta-analysis of data from all 5 trials. One possibility is that the increased power available from the observational cohort has confirmed an association first suggested in the RT data. This finding would however need to be replicated in further studies before it could be used to guide surgical decision-making.

6.4.2 Secondary outcomes

Hospital stay and disposition

This chapter also presented data that has not previously been reported by RTs, including time to surgery, length of stay, discharge destination, and 30-day re-admission. The study found that patients undergoing THA waited longer for an operation (approximately 1.7 hours), although this delay is unlikely to be clinically significant. Although the AAOS have expressed concern that increased provision of THA might lead to operative delays⁴⁹, data in this chapter suggest that hospitals in England are providing THA within a timeframe that is comparable to HA. The data also suggest that THA was associated with a shorter length of stay (by approximately 1.9 days) and increased odds of discharge home. However, there was no difference between the groups in terms of 30-day re-admission.

Functional outcomes and health-related quality of life

There was mixed evidence from the RTs as to whether or not functional outcomes or health-related quality of life vary between the groups at 12-months. The meta-analyses did not identify any statistically significant differences, although one study reported significantly better Oxford Hip Scores in the THA group³⁸². There is however evidence that the functional and health-related quality of life benefits of THA only become apparent after a number of years⁵³ and so it is however possible that these meta-analyses understated the functional benefits of THA in this population.

6.4.3 Use of propensity score matching

The principal advantage of propensity score matching is that it combines co-variables into a single score, which can increase balance between groups without discarding large numbers of observations. Some alternative matching techniques suffer from the curse of dimensionality, i.e. the likelihood of finding exact matches falls as more co-variables are included in the model³⁹⁴. However, a number of criticisms have been recently levelled at propensity score matching. These include the claim that propensity score matching may actually increase imbalance between groups³⁹⁵ and magnify the effects of hidden confounders³⁹⁶.

One particular concern for this study is that propensity score matching often excludes a high number of patients in any given dataset³⁹⁶. In this study, over 56% of the 28,099 eligible patient records were pruned from the final matched sample. As records were only preserved when a match was available, it is likely that patients with very high or low propensities to undergo THA were under-represented in the final sample. This may explain why there were differences between the demographic characteristics of the matched and unmatched populations. For example, a greater proportion of patients in the matched cohort were independently mobile without aids pre-injury than in the unmatched group (86.7% versus 75.5%). However, the groups were identical across some characteristics (e.g. ASA and AMTS) and similar across the others (e.g. age, sex, and admission source). It is however possible that

the matched cohort was not fully representative of cases in the parent dataset.

6.4.4 Conclusion

There is one on-going RT³⁰⁹ that might – either in isolation or when combined with data from previous trials – report sufficient events to identify differences between the two operations. However, the AAOS has expressed concern that the benefits of THA might not be generalizable beyond the controlled environment of clinical trials⁴⁹. The RTs identified in this study were all based in large academic centres and two^{384,385} specified that operations were only performed by experienced arthroplasty surgeons.

Observational datasets can provide important context for RT findings as they reflect “real world” practice in which operations may also be performed in smaller orthopaedic units, by generalist orthopaedic surgeons, and by trainees. It is therefore reassuring that, although the propensity score matched cohort mirrored the RT participants in terms of HA dislocation rate (both 0.9%), the THA dislocation rate was lower in the observational cohort than reported by trials. There were also fewer revisions identified in the propensity score matched cohort than were reported by the RTs.

Although it is possible that some dislocation and revision events were not captured by this cohort study, these findings were similar both in magnitude and direction (THA dislocation 1.9% versus 0.8%; revision 0.4% versus 2.3%) to a recent study from Canada³⁸⁹. It is therefore possible that contemporary prostheses perform better (in terms of major hip complications) than those used in trials undertaken between 2006 and 2013.

This study confirmed previous findings that dislocation and revision rates are higher for hip fracture patients undergoing THA. However, it did not find evidence to support the view that provision of THA leads to clinically significant delays or worse outcomes for older adults with hip fractures. Similarly, there was not any evidence that dislocation or revision rates are higher in England outside the context

of clinical trials. The finding of increased mortality amongst patients undergoing HA requires urgent further study to determine whether or not this can be replicated in other balanced populations.

Chapter 7

Conclusion

This chapter draws together the findings of studies presented in this thesis and makes recommendations for policy change and further research.

7.1 Review of core chapters

7.1.1 Impact of public release of performance data on quality of care and patient outcomes

Research question: Does publication of performance data change the behaviour of healthcare providers, quality of care, and patient outcomes.

Study design: Systematic review of RT and selected (i.e. high-quality CBA and ITS) observational study designs.

Findings: The existing evidence base was inadequate to directly inform practice. There was low-certainty evidence with mixed findings from four studies that measured effects on healthcare decisions taken by healthcare providers with three reporting modest effects and one no effect. There was low-certainty evidence from one study that measured effects on performance, which reported that a small number of quality-of-care processes improved in intervention hospitals. However, there was no correction for multiple hypotheses testing in this study, which was therefore

only considered to provide weak evidence. The five studies that measured patient outcomes reported inconsistent findings, with two reporting improvements and three reporting no difference.

Outputs: Metcalfe D, Rios Diaz AJ, Olufajo OA, Massa MS, Ketelaar N, Flottorp SA, Perry DC. Impact of public release of performance data on the behaviour of healthcare consumers and providers. *Cochrane Database Sys Rev.* 2018;9:CD004538.

7.1.2 Risk adjustment in the National Hip Fracture Database

Research question: Would the NHFD benefit from routine linkage to an administrative dataset and is the identification of mortality outliers sensitive to the choice of co-morbidity summary measure?

Study design: Observational cohort studying using a single year of NHFD data to evaluate discrimination and calibration of alternative risk adjustment models.

Findings: The existing risk adjustment model within the NHFD does an acceptable job of identifying high mortality outliers. Outlier hospitals can be identified by lowering the threshold at which alerts are triggered, which may be preferable to seeking a routine linkage to HES.

7.1.3 Pay-for-performance and hip fracture outcomes

Research question: Was introduction of the NHFD and the subsequent Hip Fracture BPT in England associated with improved hip fracture outcomes?

Study design: Natural experiment with ITSA and DID analyses using HES APC and SMR01.

Findings: There was little change in quality of care or patient outcomes that could be directly attributed to the NHFD. However, against the backdrop of the NHFD,

the BPT appeared to have driven changes in practice that reduced mortality for older adults with hip fractures in England. The BPT was also associated with more patients receiving an operation within 36 hours, shortened acute hospital LOS, and reduced re-admissions.

Output: Metcalfe D, Zogg CK, Judge A, Perry D, Gabbe B, Costa ML. Impact of the Hip Fracture Best Practice Tariff in England. “Best of the Best” oral presentation. Orthopaedic Trauma Society Annual Meeting, Burton-on-Trent, U.K. 10th January 2018.

Output: Metcalfe D, Zogg CK, Judge A, Perry DC, Gabbe BJ, Willett K, Costa ML. Pay-for-performance and hip fracture outcomes: an interrupted time series and difference-in-differences analysis in England and Scotland. *Bone Joint J.* 2019;101-B:1015-1023.

7.1.4 Inequalities in the use of THA for hip fracture

Research question: Is use of THA among individuals with a displaced intracapsular hip fracture based on national guidelines or are there systematic inequalities?

Study design: Observational cohort study using the NHFD.

Findings: There was substantial unexplained variation (7.4 to 69.0% between hospitals) in the use of THA following a hip fracture. Patient-level predictors were unable to fully account for this variation and so it is likely to reflect systematic differences in practice between centres. However, the provision of THA did appear to be influenced by some patient characteristics, including age, AMTS, ASA, pre-fracture mobility, and socioeconomic status. Other predictors of THA included the treating hospital and the day of admission. The use of THA amongst eligible patients increased steadily but remained low between 2011 and 2016.

Output: Perry DC, Metcalfe D, Costa ML. Inequalities in access to total hip arthroplasty for hip fracture: a population-based study. *Lancet*. 2016;387(S1);S81.

Output: Perry DC, Metcalfe D, Griffin XL, Costa ML. Inequalities in access to total hip arthroplasty for hip fracture: population-based study. *BMJ*. 2016;353:i2021.

7.1.5 THA versus HA for intracapsular hip fractures

Research question: Are THA or HA associated with clinical outcomes for independently mobile older adults with intracapsular hip fractures?

Study design: Systematic review of RTs compared with propensity score matched “real world” data from the NHFD.

Findings: Five small RTs (median 89 patients) were included in the meta-analysis and it is possible that even the pooled analyses were under-powered to detect important differences. The RT data showed a non-significant association between THA and 12-month dislocation, although this observation was reproduced and statistically significant in the cohort study. Although the RT data suggested “no difference” between HA and THA in terms of revision, there was an estimated 33% lower risk of 12-month revision for THA patients in the cohort study. Both the RT and cohort study data suggested a lower risk of mortality for patients undergoing THA. There was not any evidence for worse THA outcomes outwith RT and patients undergoing THA in the NHFD waited only 1.7 hours longer for their operation, which is unlikely to be clinically meaningful.

Output: Metcalfe D, Judge A, Perry DC, Gabbe B, Zogg CK, Costa ML. Total hip arthroplasty versus hemiarthroplasty for independently mobile older adults with hip fractures. *BMC Musculoskelet Disord*. 2019;20:226.

7.2 Themes and recommendations

7.2.1 Public release of performance data

There is a lack of high quality studies evaluating interventions in which performance data is released to the public. This is surprising given that there is clearly enthusiasm for such interventions from policy makers, purchasers, and the public (see Section 1.7 on page 37). It is particularly important as there are real concerns that public release of performance data may have unintended consequences, such as “gaming” by hospitals, selection of low-risk cases, and increasing healthcare inequity by allowing the best resourced and most mobile patients to preferentially select the “best” healthcare providers^{144,147–151}. Chapter 2 on page 42 found that there was a particular lack of studies evaluating potential effects on patient outcomes rather than process measures (Subsection 2.4.6 on page 69). In the absence of high quality data, it is not possible to base policy recommendations on this systematic review, although it is important to note that policies aimed at releasing performance data to the public have advanced some way beyond the underlying evidence base.

7.2.2 Risk adjustment in the NHFD

In order to achieve its clinical audit function, the NHFD must correctly identify mortality outliers. The evidence in Chapter 3 on page 72 suggests that there may be room for improvement of the risk adjustment model currently used by the NHFD. However, this study also found that routine linkage to administrative datasets such as HES may not lead to substantial improvements in model performance. Instead, the FFFAP should consider undertaking a formal evaluation of the quality of data submitted to the NHFD, which is likely to be better and easier to improve than variables harvested from administrative data.

7.2.3 Pay-for-performance and hip fracture outcomes

The findings of studies evaluating pay-for-performance programmes in other countries have been largely disappointing (Section 4.4 on page 117). However, there is anecdotal evidence to suggest that hip fracture pathways in England were transformed by the linked initiatives of the NHFD and Hip Fracture BPT⁶¹. The combination of defined quality standards and straightforward access to performance data made hip fracture care a target for clinicians undertaking local audits and quality improvement projects^{237,252–257,267}. There is therefore reason to think that the NHFD/BPT might affect care differently from initiatives that are determined by clinical coding remote from clinicians, such as PbR (see Subsection 4.1.1 on page 102).

The evidence in Chapter 4 on page 101 suggests that the BPT (and possibly - to a lesser extent - the NHFD) drove improvements in both performance measures and clinical outcomes. This should be considered as an exemplar model for engaging clinical staff with quality improvement and driving positive service re-organization. Policy makers in Wales, Scotland, and Northern Ireland should also consider whether their hip fracture outcomes could be improved by capitalising on the findings from this chapter.

7.2.4 Total hip arthroplasty for hip fracture

Chapter 5 on page 124 documented substantial variation across the country in terms of access to THA for eligible hip fracture patients. This is important in the context of improved quality of hip fracture care shown in Chapter 4 on page 101. Taken together, this suggests that healthcare providers may have focused their efforts and attention on improving performance against the BPT clinical standards. It is also important because this variation appears to be inter-related with structural disparities in access to THA, e.g. for patients living in poor areas and requiring a hip fracture operation at weekends. Guaranteeing consistent access to care across all days of the week is a healthcare priority for the current government³⁹⁷.

Chapter 6 on page 148 showed that there is reasonable evidence for offering THA, particularly when considered against the weak evidence base supporting many other orthopaedic trauma procedures^{398–400}. In particular, concerns about whether or not the clinical trial evidence can be generalized to routine clinical practice were not substantiated by the evidence in this chapter. Alarming, both the meta-analysis of trial data and the “real world” observational evidence suggested a mortality benefit in favour of THA amongst the fittest patients. The forthcoming HEALTH trial adopted unplanned secondary procedures as its primary outcome but has collected mortality data as a secondary outcome measure³⁰⁹. These data should be published urgently once available in order to confirm or refute the mortality findings in this chapter.

Even in the absence of a mortality difference, Chapter 6 on page 148 suggests that there is good evidence for using THA amongst the fittest patients across a range of outcome measures. Hip fracture units should therefore consider how best they can guarantee availability of THA to all patients regardless of day-of-the-week or the particular sub-specialty of the surgeon on duty. Individual surgeons should be mindful that THA may now have become the “standard of care” for the fittest hip fracture patients, even if clinical practice does not yet routinely meet this standard (Subsection 5.4.3 on page 139). As patients in the group recommended to receive THA by NICE do not have cognitive impairment, it is unlikely to be acceptable to perform HA without an explicit discussion with the patient about alternative treatments.

If the HEALTH trial provides further support for THA, it is likely that the number of patients undergoing this operation will increase. This will pose logistical problems for a number of hip fracture units, which may have to consider innovative organizational changes depending on local circumstances. These possibilities could include regionalising hip fracture care into few centres, offering a dedicated fracture service provided by a small number of surgeons who routinely perform THA, staffing a “hip surgeon” rota each day, or sharing hip surgeons between a small number of

hospitals^{352,353}.

As shown in Chapter 4 on page 101, hip fracture units *do* respond when quality processes are reported to the NHFD and particularly when they are tied to remuneration under the BPT. It therefore seems likely that the most effective mechanism for driving compliance with the THA recommendation would be to include this as a BPT standard. Some progress has already been made in this direction on the back of evidence submitted in this thesis.

The finding of low compliance with NICE CG124 (Chapter 5 on page 124) was reported in national print media, such as *The Daily Mirror*³⁵⁷, and prompted a 280-page addendum from NICE, which re-evaluated the clinical and cost effectiveness of interventions for intracapsular hip fractures³⁰². This addendum re-iterated support for THA and created a new quality standard, which states that “adults with displaced intracapsular hip fracture receive cemented hemiarthroplasty or, if they are assessed as clinically eligible, a total hip replacement”. The paper originally published from this chapter recommended that the THA “report data on THA provision at the hospital-level to help achieve greater consistency across the NHS”. This recommendation has since been adopted by the NHFD, which now audits individual hospitals against this standard on an annual basis²⁰⁵. This is clearly a step in the right direction and lays the necessary groundwork for incorporating this new quality standard into the BPT.

7.3 Access to data

One recurring theme of this DPhil was that of difficulties gaining access to linked data, which was despite working with experienced supervisors, in an institution with accredited information governance systems, and adequate funding.

Although there are legislative and organisational differences between England and the devolved nations, healthcare across the UK is provided under a single conceptual framework, i.e. the NHS. This presents an opportunity for studies using routinely collected data that might not be possible in healthcare systems with more

disparate funding arrangements⁴⁰¹. It is therefore disappointing that organizations in England have been so slow to harmonize data collection and establish routine linkages between datasets.

My initial plan was to base this thesis on a linkage between the NHFD, HES, CPRD, and ONS mortality data. Some of the most significant obstacles I encountered are documented in Appendix J on page 279 to give future researchers a flavour as to what they might expect when pursuing a bespoke linkage between datasets. In brief, my attempt to link NHFD and CPRD data first began in October 2015 and was finally abandoned in January 2018. My first application to link NHFD and HES data was submitted on 1st August 2016 and the linked dataset finally received on 26th February 2018, i.e. 574 days later.

The obstacles outlined in Appendix J on page 279 are not intended as criticism of individuals or even any particular organization. Much of the difficulties arose because of the need for multiple organizations to approve the protocol for a single project (see Figure 7.1 on page 189). Such approvals could not usually be sought concurrently and any significant change required by one organization potentially required the whole process to begin again. In most cases, approvals could only be granted by a committee, which only met at specific times in the year. Such lengthy, disjointed, and uncertain processes raise particular issues for researchers working to academic and/or funding deadlines.

The issues around coordinating data linkages in England are well-documented^{221,222,401–404} and have caused problems for student researchers²²¹, large-scale research projects^{401–403}, and national clinical audits^{222,403,404}. The processes in place at NHS Digital towards the end of my fellowship appeared more streamlined than they were at the beginning. However, there are clearly opportunities to further improve the data sharing landscape in England for the benefit of patients and the public.

In the meantime, future researchers may wish to ensure that proposed data linkages are not too ambitious (e.g. involving too many data controllers), applications are submitted early (ideally before doctoral work is scheduled to begin), and that a

back-up plan is available in case data linkages fail due to circumstances beyond the control of the research team.

7.4 Future work

This thesis has demonstrated that the NHFD can support a range of study designs aimed at guiding care for older adults with hip fractures. Although NHFD data are regularly used to monitor local quality processes and outcomes^{237,252–257,267}, it is currently under-utilized as a resource for hip fracture researchers (Subsection 1.7.1 on page 38). There is clearly scope to adapt the approach used in Chapter 6 on page 148 to other important hip fracture groups for whom there are controversies around surgical management, e.g. cannulated screw versus compression hip screw fixation for undisplaced intracapsular fractures and compression hip screw versus intramedullary nail fixation for intertrochanteric fractures^{43,44,405}.

The NHFD exists within a comprehensive universal health service and offers a rare opportunity to capture data on a national population of older adults with hip fractures. This thesis has shown that it is feasible to link the NHFD to administrative datasets in order to capture re-admissions, re-operations, and robust data about deaths registered anywhere within England and Wales. Future projects could exploit such linkages to incorporate health economic analyses and guide recommendations about surgical decision making across the NHS.

Although the focus of this thesis has been on surgical interventions, such a comprehensive cohort of older adults with hip fractures could also provide information that might help this group of patients in other ways. For example, linkage to primary care datasets (such as the CPRD²³⁰ or THIN²³¹) could be used to determine which medications patients were using before and after their hip fracture. This could provide information about both the safety and effectiveness of drugs aimed at primary and secondary prevention of hip fracture.

In particular, there is evidence that patients are at highest risk of re-fracture in the early period after a hip fracture⁴⁰⁶. This is presumably because these pa-

tients have identified themselves as having bone fragility and being at risk of falls (Section 1.2 on page 20). Other factors may include impaired mobility following the combined insults of injury, a large operation, and a hospital admission⁴⁰⁷. As traditional anti-resorptive therapies take many months to reduce bone fragility⁴⁰⁸, there is a window of risk that could be closed using short-term treatment with faster acting agents. The NHFD could be used to identify the characteristics of hip fracture patients who are at highest risk of early re-fracture and so might benefit from more intensive anti-osteoporosis treatment.

One problem with using administrative datasets alone for this purpose is that it is challenging to identify unique hip fractures events. For example, a hip fracture patient who is discharged from hospital but re-admitted with pneumonia 6 weeks later will likely receive a second HES record coded with the diagnosis “hip fracture”. A failure of fixation (e.g. a compression hip screw pulling out of the bone) would also likely result in re-admission with a “hip fracture” code. Both sets of circumstances may give the erroneous impression to researchers that the patient experienced a second hip fracture. However, this problem does not arise in clinical registries - such as the NHFD - as each record is known to represent a discrete event.

The gold standard for evaluating interventions is the RT. There are concerns that RTs cannot tell “the whole story” as they typically occur within a narrow range of high-performing hospitals and their findings may not be readily generalized to routine clinical practice. The data in Chapter 6 on page 148 did not support this as a particular concern for the RT literature around THA for hip fracture. However, it would be good practice for clinical trialists to plan large-scale observational studies using routine data to run alongside multi-centre RTs. Such studies could be designed to collect the same data and overlapping patient outcomes with a view to commenting on the generalizability of the RT findings. This approach could help reduce the temptation to overstate conclusions from RTs but also allay concerns about RT data that may deter surgeons from changing their clinical practice.

Although RTs and well-designed observational studies should be developed in

parallel, the obvious “next step” is to embed clinical trials within routinely collected healthcare data⁴⁰⁹. The advantages of this approach are obvious but include low cost, better generalizability, and long-term follow-up⁴¹⁰. A number of RTs have already been embedded within the multi-centre World Hip Trauma Evaluation (WHiTE) cohort study^{10,411–414}. In the future, it may be possible to embed such trials within national datasets (such as the NHFD), which could dramatically improve the recruitment and follow-up of trial patients. Although there is still a lack of high quality evidence to guide the care of older adults with hip fractures, the framework now exists to learn from these patients and continue to improve outcomes for this vulnerable population.

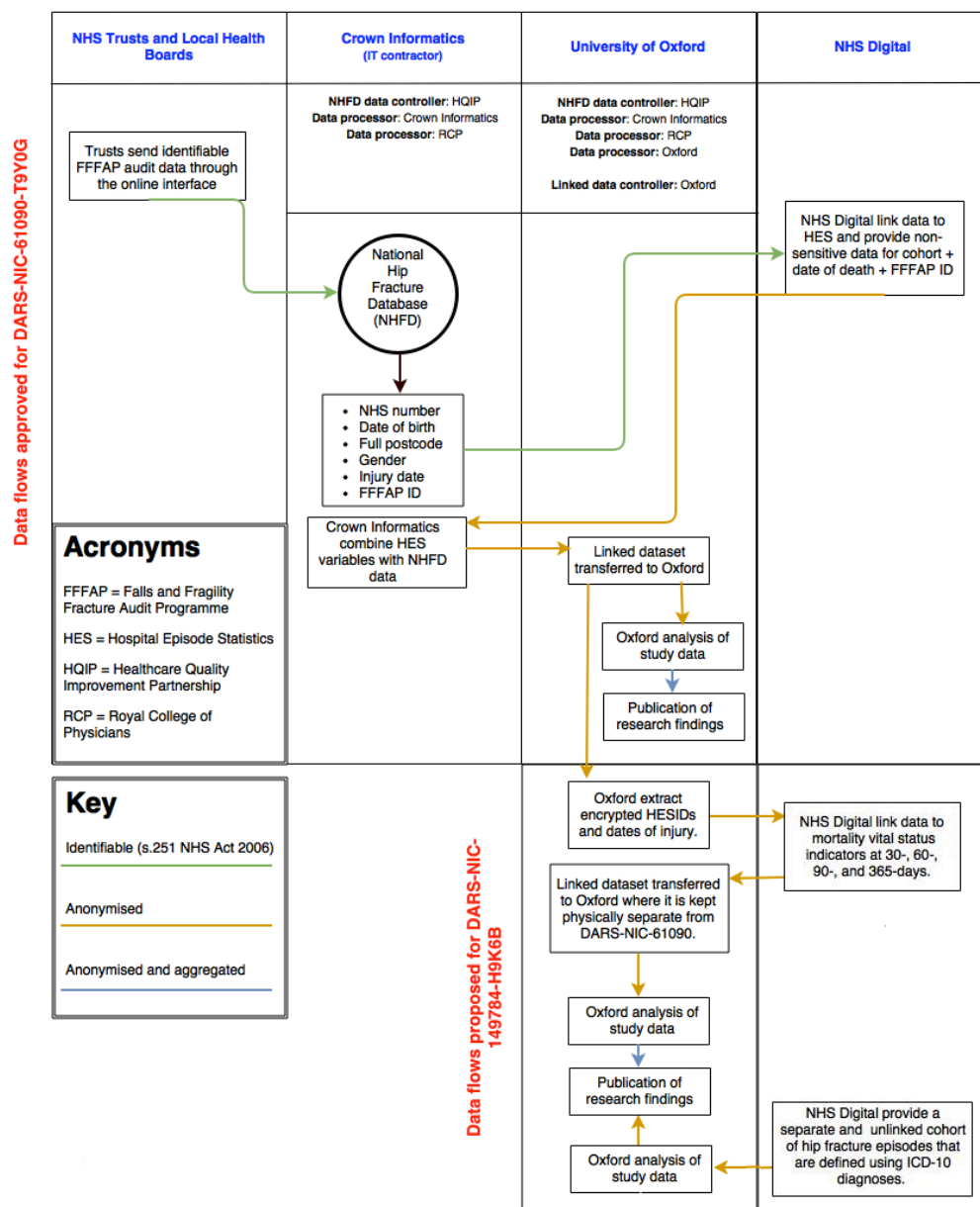


Figure 7.1: A diagram illustrating the complexity of data flows without including a CPRD linkage or consideration of legal bases under the European Union (EU) GDPR

Bibliography

- [1] Consensus Development Conference. “Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis”. *Am J Med* 94.6 (1993), pp. 646–50.
- [2] D. Metcalfe. “The pathophysiology of osteoporotic hip fracture”. *Mcgill J Med* 11.1 (2008), pp. 51–7.
- [3] Office of the Surgeon General (US). *Bone Health and Osteoporosis: A Report of the Surgeon General*. Rockville (MD), 2004.
- [4] E. Hernlund et al. “Osteoporosis in the European Union: medical management, epidemiology and economic burden. A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA)”. *Arch Osteoporos* 8 (2013), p. 136.
- [5] T. P. van Staa et al. “Epidemiology of fractures in England and Wales”. *Bone* 29.6 (2001), pp. 517–22.
- [6] E. M. Curtis et al. “The impact of fragility fracture and approaches to osteoporosis risk assessment worldwide”. *Bone* 104 (2017), pp. 29–38.
- [7] United Nations. *World Population Prospects*. Report. 2017.
- [8] C. Cooper et al. “Secular trends in the incidence of hip and other osteoporotic fractures”. *Osteoporos Int* 22.5 (2011), pp. 1277–88.

-
- [9] A. Papaioannou et al. “The impact of incident fractures on health-related quality of life: 5 years of data from the Canadian Multicentre Osteoporosis Study”. *Osteoporos Int* 20.5 (2009), pp. 703–14.
- [10] X. L. Griffin et al. “Recovery of health-related quality of life in a United Kingdom hip fracture population. The Warwick Hip Trauma Evaluation—a prospective cohort study”. *Bone Joint J* 97-B.3 (2015), pp. 372–82.
- [11] S. Borhan et al. “Incident Fragility Fractures Have a Long-Term Negative Impact on Health-Related Quality of Life of Older People: The Canadian Multicentre Osteoporosis Study”. *J Bone Miner Res* (2019), e3666.
- [12] R. Griffiths and M. Parker. “Bone cement implantation syndrome and proximal femoral fracture”. *Br J Anaesth* 114.1 (2015), pp. 6–7.
- [13] J. R. Center et al. “Mortality after all major types of osteoporotic fracture in men and women: an observational study”. *Lancet* 353.9156 (1999), pp. 878–82.
- [14] A. Papaioannou et al. “Lengthy hospitalization associated with vertebral fractures despite control for comorbid conditions”. *Osteoporos Int* 12.10 (2001), pp. 870–4.
- [15] C. L. Leibson et al. “Mortality, disability, and nursing home use for persons with and without hip fracture: a population-based study”. *J Am Geriatr Soc* 50.10 (2002), pp. 1644–50.
- [16] T. L. Jarvinen et al. “Overdiagnosis of bone fragility in the quest to prevent hip fracture”. *BMJ* 350 (2015), h2088.
- [17] J. A. Kanis. “Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report. WHO Study Group”. *Osteoporos Int* 4.6 (1994), pp. 368–81.
- [18] S. A. Wainwright et al. “Hip fracture in women without osteoporosis”. *J Clin Endocrinol Metab* 90.5 (2005), pp. 2787–93.

-
- [19] E. Sornay-Rendu et al. “Bone Microarchitecture Assessed by HR-pQCT as Predictor of Fracture Risk in Postmenopausal Women: The OFELY Study”. *J Bone Miner Res* 32.6 (2017), pp. 1243–1251.
- [20] J. A. Kanis et al. “The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women”. *Osteoporos Int* 18.8 (2007), pp. 1033–46.
- [21] J. Hippisley-Cox and C. Coupland. “Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study”. *BMJ* 344 (2012), e3427.
- [22] N. D. Nguyen et al. “Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks”. *Osteoporos Int* 19.10 (2008), pp. 1431–44.
- [23] K. L. Stone et al. “BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures”. *J Bone Miner Res* 18.11 (2003), pp. 1947–54.
- [24] O. Johnell et al. “Predictive value of BMD for hip and other fractures”. *J Bone Miner Res* 20.7 (2005), pp. 1185–94.
- [25] H. Wagner et al. “Simply ask them about their balance—future fracture risk in a nationwide cohort study of twins”. *Am J Epidemiol* 169.2 (2009), pp. 143–9.
- [26] J. Compston. “Overdiagnosis of osteoporosis: fact or fallacy?” *Osteoporos Int* 26.8 (2015), pp. 2051–4.
- [27] C. Todd and D. Skelton. *What are the main risk factors for falls amongst older people and what are the most effective interventions to prevent these falls?* Report. 2004.
- [28] M. I. Smith et al. “Predicting falls and when to intervene in older people: a multilevel logistical regression model and cost analysis”. *PLoS One* 11.7 (2016), e0159365.

-
- [29] R. L. Drake. *Gray's Anatomy for Students*. 2nd edition. Philadelphia, PA, USA: Elsevier Inc., 2009.
- [30] I. Gilligan, S. Chandraphak, and P. Mahakkanukrauh. "Femoral neck-shaft angle in humans: variation relating to climate, clothing, lifestyle, sex, age and side". *J Anat* 223.2 (2013), pp. 133–51.
- [31] J. B. Stiehl, D. Jacobson, and G. Carrera. "Morphological analysis of the proximal femur using quantitative computed tomography". *Int Orthop* 31.3 (2007), pp. 287–92.
- [32] F. V. Wagner et al. "Capsular ligaments of the hip: anatomic, histologic, and positional study in cadaveric specimens with MR arthrography". *Radiology* 263.1 (2012), pp. 189–98.
- [33] P. G. Collin et al. "Hip fractures in the elderly-: A Clinical Anatomy Review". *Clin Anat* 30.1 (2017), pp. 89–97.
- [34] A. W. Grose et al. "The surgical anatomy of the blood supply to the femoral head: description of the anastomosis between the medial femoral circumflex and inferior gluteal arteries at the hip". *J Bone Joint Surg Br* 90.10 (2008), pp. 1298–303.
- [35] M. Kalhor et al. "Anatomic variations in femoral head circulation". *Hip Int* 22.3 (2012), pp. 307–12.
- [36] V. Perumal, S. J. Woodley, and H. D. Nicholson. "Ligament of the head of femur: A comprehensive review of its anatomy, embryology, and potential function". *Clin Anat* 29.2 (2016), pp. 247–55.
- [37] M. Bhandari and M. Swiontkowski. "Management of Acute Hip Fracture". *N Engl J Med* 377.21 (2017), pp. 2053–2062.
- [38] H. H. Handoll and M. J. Parker. "Conservative versus operative treatment for hip fractures in adults". *Cochrane Database Syst Rev* 3 (2008), p. CD000337.
- [39] S. Chariyalertsak, P. Suriyawongpisal, and A. Thakkinstain. "Mortality after hip fractures in Thailand". *Int Orthop* 25.5 (2001), pp. 294–7.

-
- [40] R. Jain, A. Basinski, and H. J. Kreder. “Nonoperative treatment of hip fractures”. *Int Orthop* 27.1 (2003), pp. 11–7.
- [41] M. J. Parker et al. “Cost-benefit analysis of hip fracture treatment”. *J Bone Joint Surg Br* 74.2 (1992), pp. 261–4.
- [42] Scottish Intercollegiate Guidelines Network. *Management of hip fracture in older people*. Report. 2009.
- [43] M. A. Fernandez, X. L. Griffin, and M. L. Costa. “Management of hip fracture”. *Br Med Bull* 115.1 (2015), pp. 165–72.
- [44] Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) Investigators. “Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial”. *Lancet* 389.10078 (2017), pp. 1519–1527.
- [45] M. F. Swiontkowski. “Intracapsular fractures of the hip”. *J Bone Joint Surg Am* 76.1 (1994), pp. 129–38.
- [46] M. Bhandari et al. “Internal fixation compared with arthroplasty for displaced fractures of the femoral neck. A meta-analysis”. *J Bone Joint Surg Am* 85-A.9 (2003), pp. 1673–81.
- [47] E. O. Johnson, K. Soultanis, and P. N. Soucacos. “Vascular anatomy and microcirculation of skeletal zones vulnerable to osteonecrosis: vascularization of the femoral head”. *Orthop Clin North Am* 35.3 (2004), pp. 285–91, viii.
- [48] National Institute for Health and Care Excellence. *Hip fracture: the management of hip fracture in adults. NICE clinical guideline 124*. Report. 2011.
- [49] American Academy of Orthopaedic Surgeons. *Management of Hip Fractures in the Elderly*. <http://tiny.cc/otp05y>. 2015.
- [50] L. Liao et al. “A meta-analysis of total hip arthroplasty and hemiarthroplasty outcomes for displaced femoral neck fractures”. *Arch Orthop Trauma Surg* 132.7 (2012), pp. 1021–9.

-
- [51] J. F. Keating et al. “Displaced intracapsular hip fractures in fit, older people: a randomised comparison of reduction and fixation, bipolar hemiarthroplasty and total hip arthroplasty”. *Health Technol Assess* 9.41 (2005), pp. iii–iv, ix–x, 1–65.
- [52] W. Macaulay et al. “Prospective randomized clinical trial comparing hemiarthroplasty to total hip arthroplasty in the treatment of displaced femoral neck fractures: winner of the Dorr Award”. *J Arthroplasty* 23.6 Suppl 1 (2008), pp. 2–8.
- [53] C. J. Hedbeck et al. “Comparison of bipolar hemiarthroplasty with total hip arthroplasty for displaced femoral neck fractures: a concise four-year follow-up of a randomized trial”. *J Bone Joint Surg Am* 93.5 (2011), pp. 445–50.
- [54] P. T. Burgers et al. “Total hip arthroplasty versus hemiarthroplasty for displaced femoral neck fractures in the healthy elderly: a meta-analysis and systematic review of randomized trials”. *Int Orthop* 36.8 (2012), pp. 1549–60.
- [55] National Institute for Health and Care Excellence (NICE). *Hip fracture: management*. Report. NICE, 2017.
- [56] F. Griffiths et al. “Evaluating recovery following hip fracture: a qualitative interview study of what is important to patients”. *BMJ Open* 5.1 (2015), e005406.
- [57] A. J. Donaldson et al. “Bone cement implantation syndrome”. *Br J Anaesth* 102.1 (2009), pp. 12–22.
- [58] M. J. Parker, K. S. Gurusamy, and S. Azegami. “Arthroplasties (with and without bone cement) for proximal femoral fractures in adults”. *Cochrane Database Syst Rev* 6 (2010), p. CD001706.
- [59] E. Langslet et al. “Cemented versus uncemented hemiarthroplasty for displaced femoral neck fractures: 5-year followup of a randomized trial”. *Clin Orthop Relat Res* 472.4 (2014), pp. 1291–9.

-
- [60] H. Saleeb, R. Kanvinde, and T. Rahman. “Literature review and case report: Current concepts for concomitant intra and extracapsular fractures of neck of femur in elderly patients”. *Trauma Case Rep* 8 (2017), pp. 24–31.
- [61] O. Sahota and C. Currie. “Hip fracture care: all change”. *Age Ageing* 37.2 (2008), pp. 128–9.
- [62] Unknown. Author. “Fracture neck of femur - Still an unsolved issue”. *J Orthop* 13.1 (2016), A1–3.
- [63] C. V. Heck. “Management of hip fracture in the geriatric patient”. *J Am Geriatr Soc* 3.2 (1955), pp. 113–6.
- [64] K. G. Callum et al. *Extremes of age. The 1999 report of the National Confidential Enquiry into Perioperative Deaths*. Report. NCEPOD, 1999.
- [65] A. Bottle and P. Aylin. “Mortality associated with delay in operation after hip fracture: observational study”. *BMJ* 332.7547 (2006), pp. 947–51.
- [66] N. Simunovic et al. “Effect of early surgery after hip fracture on mortality and complications: systematic review and meta-analysis”. *CMAJ* 182.15 (2010), pp. 1609–16.
- [67] D. Pincus et al. “Association Between Wait Time and 30-Day Mortality in Adults Undergoing Hip Fracture Surgery”. *JAMA* 318.20 (2017), pp. 1994–2003.
- [68] A. Sayers et al. “The association between the day of the week of milestones in the care pathway of patients with hip fracture and 30-day mortality: findings from a prospective national registry - The National Hip Fracture Database of England and Wales”. *BMC Med* 15.1 (2017), p. 62.
- [69] B. Sobolev et al. “Mortality effects of timing alternatives for hip fracture surgery”. *CMAJ* 190.31 (2018), E923–E932.
- [70] F. Leung et al. “Does timing of surgery matter in fragility hip fractures?” *Osteoporos Int* 21.Suppl 4 (2010), S529–34.

-
- [71] T. Tran et al. “The impact of oral anticoagulation on time to surgery in patients hospitalized with hip fracture”. *Thromb Res* 136.5 (2015), pp. 962–5.
- [72] A. Aqil et al. “Achieving hip fracture surgery within 36 hours: an investigation of risk factors to surgical delay and recommendations for practice”. *J Orthop Traumatol* 17.3 (2016), pp. 207–13.
- [73] P. Guy et al. “Feasibility of using administrative data for identifying medical reasons to delay hip fracture surgery: a Canadian database study”. *BMJ Open* 7.10 (2017), e017869.
- [74] L. B. Oldmeadow et al. “No rest for the wounded: early ambulation after hip surgery accelerates recovery”. *ANZ J Surg* 76.7 (2006), pp. 607–11.
- [75] K. Singler et al. “A plea for an early mobilization after hip fractures. The geriatric point of view.” *European Geriatric Medicine* 4.1 (2013), pp. 40–42.
- [76] H. K. Kamel et al. “Time to ambulation after hip fracture surgery: relation to hospitalization outcomes”. *J Gerontol A Biol Sci Med Sci* 58.11 (2003), pp. 1042–5.
- [77] A. L. Siu et al. “Early ambulation after hip fracture: effects on function and mortality”. *Arch Intern Med* 166.7 (2006), pp. 766–71.
- [78] T. Frenkel Rutenberg et al. “Timing of physiotherapy following fragility hip fracture: delays cost lives”. *Arch Orthop Trauma Surg* 138.11 (2018), pp. 1519–1524.
- [79] R. E. Irvine and M. B. Devas. “The geriatric orthopaedic unit.” *J Bone Joint Surg Am* 49 (1963), pp. 186–187.
- [80] R. B. Lefroy. “Treatment of patients with fractured neck of the femur in a combined unit”. *Med J Aust* 2.12 (1980), pp. 669–70.
- [81] J. R. Elliot et al. “The added effectiveness of early geriatrician involvement on acute orthopaedic wards to orthogeriatric rehabilitation”. *N Z Med J* 109.1017 (1996), pp. 72–3.

-
- [82] M. Lundstrom et al. "Reorganization of nursing and medical care to reduce the incidence of postoperative delirium and improve rehabilitation outcome in elderly patients treated for femoral neck fractures". *Scand J Caring Sci* 13.3 (1999), pp. 193–200.
- [83] T. M. Huusko et al. "Randomised, clinically controlled trial of intensive geriatric rehabilitation in patients with hip fracture: subgroup analysis of patients with dementia". *BMJ* 321.7269 (2000), pp. 1107–11.
- [84] M. Vidan et al. "Efficacy of a comprehensive geriatric intervention in older patients hospitalized for hip fracture: a randomized, controlled trial". *J Am Geriatr Soc* 53.9 (2005), pp. 1476–82.
- [85] M. Stenvall et al. "Improved performance in activities of daily living and mobility after a multidisciplinary postoperative rehabilitation in older people with femoral neck fracture: a randomized controlled trial with 1-year follow-up". *J Rehabil Med* 39.3 (2007), pp. 232–8.
- [86] Y. I. Shyu et al. "Interdisciplinary intervention for hip fracture in older Taiwanese: benefits last for 1 year". *J Gerontol A Biol Sci Med Sci* 63.1 (2008), pp. 92–7.
- [87] J. I. Gonzalez-Montalvo et al. "The orthogeriatric unit for acute patients: a new model of care that improves efficiency in the management of patients with hip fracture". *Hip Int* 20.2 (2010), pp. 229–35.
- [88] K. V. Grigoryan, H. Javedan, and J. L. Rudolph. "Orthogeriatric care models and outcomes in hip fracture patients: a systematic review and meta-analysis". *J Orthop Trauma* 28.3 (2014), e49–55.
- [89] S. Sabharwal and H. Wilson. "Orthogeriatrics in the management of frail older patients with a fragility fracture". *Osteoporos Int* 26.10 (2015), pp. 2387–99.
- [90] Australian & New Zealand Society for Geriatric Medicine. *Position Statement* 5. Report. 2010.

-
- [91] M. Middleton. “Orthogeriatrics and hip fracture care in the UK: factor driving change to more integrated models of care”. *Geriatrics* 3.3 (2018), p. 55.
- [92] National Institute for Health and Care Excellence (NICE). *Raloxifene and teriparatide for the secondary prevention of osteoporotic fragility fractures in postmenopausal women*. Report. 2008.
- [93] S. Hawley et al. “Clinical effectiveness of orthogeriatric and fracture liaison service models of care for hip fracture patients: population-based longitudinal study”. *Age Ageing* 45.2 (2016), pp. 236–42.
- [94] K. W. Lyles et al. “Zoledronic acid and clinical fractures and mortality after hip fracture”. *N Engl J Med* 357.18 (2007), pp. 1799–809.
- [95] D. Prieto-Alhambra et al. “Fracture prevention in patients with cognitive impairment presenting with a hip fracture: secondary analysis of data from the HORIZON Recurrent Fracture Trial”. *Osteoporos Int* 25.1 (2014), pp. 77–83.
- [96] J. S. Myers and B. M. Wong. “Measuring outcomes in quality improvement education: success is in the eye of the beholder”. *BMJ Qual Saf* (2019).
- [97] C. J. Gill and G. C. Gill. “Nightingale in Scutari: her legacy reexamined”. *Clin Infect Dis* 40.12 (2005), pp. 1799–805.
- [98] B. M. Dossey. “Florence Nightingale: a 19th-century mystic”. *J Holist Nurs* 28.1 (2010), pp. 10–35.
- [99] S. Lipsey. “Mathematical education in the life of Florence Nightingale”. *Newsletter of the Association for Women in Mathematics* 23.4 (1993), pp. 11–12.
- [100] E. A. Codman. “The classic: A study in hospital efficiency: as demonstrated by the case report of first five years of private hospital”. *Clin Orthop Relat Res* 471.6 (2013), pp. 1778–83.
- [101] D. Neuhauser. “Ernest Amory Codman MD”. *Qual Saf Health Care* 11.1 (2002), pp. 104–5.

-
- [102] National Institute for Clinical Excellence (NICE). *Principles for best practice in clinical audit*. Oxford, U.K.: Radcliffe Publishing Ltd., 2002.
- [103] Department of Health. *Working For Patients*. Report. 1989.
- [104] Department of Health. *The New NHS. Modern. Dependable*. Report. 1997.
- [105] I. Kennedy. *The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995*. Report. 2001.
- [106] R. Francis. *Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry*. Report. 2013.
- [107] General Medical Council. *Good medical practice*. Report. 2013.
- [108] S. A. Flottorp et al. *Using audit and feedback to health professionals to improve the quality and safety of healthcare*. Report. World Health Organization, 2010.
- [109] D. A. Davis et al. "Accuracy of physician self-assessment compared with observed measures of competence: a systematic review". *JAMA* 296.9 (2006), pp. 1094–102.
- [110] A. M. Audet et al. "Measure, learn, and improve: physicians' involvement in quality improvement". *Health Aff (Millwood)* 24.3 (2005), pp. 843–53.
- [111] N. Ivers et al. "Audit and feedback: effects on professional practice and health-care outcomes". *Cochrane Database Syst Rev* 6 (2012), p. CD000259.
- [112] D. M. E. Hoque et al. "Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review". *PLoS One* 12.9 (2017), e0183667.
- [113] S. N. van der Veer et al. "Improving quality of care. A systematic review on how medical registries provide information feedback to health care providers". *Int J Med Inform* 79.5 (2010), pp. 305–23.

-
- [114] L. Hut-Mossel et al. “Understanding how and why audits work: protocol for a realist review of audit programmes to improve hospital care”. *BMJ Open* 7.6 (2017), e015121.
- [115] E. A. Campling, H. B. Devlin, and J. N. Lunn. *The Report of the National Confidential Enquiry into Perioperative Deaths*. Report. 1989.
- [116] B. Bridgewater et al. “Has the publication of cardiac surgery outcome data been associated with changes in practice in northwest England: an analysis of 25,730 patients undergoing CABG surgery under 30 surgeons over eight years”. *Heart* 93.6 (2007), pp. 744–8.
- [117] B. Bridgewater, D. Irvine, and B. Keogh. “NHS transparency”. *BMJ* 347 (2013), f4402.
- [118] *Falls and Fragility Fractures*. <http://tiny.cc/bwp05y>. 2019.
- [119] S. D. Hannah et al. “The changing case-mix of hip fractures in Scotland - evidence from the Scottish Hip Fracture Audit”. *Scott Med J* 62.4 (2017), pp. 142–146.
- [120] B. Oakley et al. “Does achieving the best practice tariff improve outcomes in hip fracture patients? An observational cohort study”. *BMJ Open* 7.2 (2017), e014190.
- [121] D. Metcalfe et al. “Quality of care for patients with a fracture of the hip in major trauma centres: a national observational study”. *Bone Joint J* 98-B.3 (2016), pp. 414–9.
- [122] Healthcare Quality Improvement Partnership (HQIP). *Detection and management of outliers for national clinical audits*. Report. 2017.
- [123] R. Wakeman et al. *The National Hip Fracture Database Preliminary National Report 2009*. Report. 2009.
- [124] C. Currie et al. *The National Hip Fracture Database National Report 2010*. Report. 2010.

-
- [125] *OPCS codes relevant to procedures recorded on the NJR*. <http://tiny.cc/4bq05y>. 2012.
 - [126] *Trauma Care in England and Wales*. <http://tiny.cc/bxp05y>. 2018.
 - [127] D. M. Berwick, B. James, and M. J. Coye. “Connections between quality measurement and improvement”. *Med Care* 41.1 Suppl (2003), pp. I30–8.
 - [128] PC. Smith et al. *Performance measurement for health system improvement. Experiences, Challenges and Prospects*. Cambridge, U.K.: Cambridge University Press, 2009.
 - [129] P. Campanella et al. “The impact of public reporting on clinical outcomes: a systematic review and meta-analysis”. *BMC Health Serv Res* 16 (2016), p. 296.
 - [130] J. Greenhalgh et al. “How do aggregated patient-reported outcome measures data stimulate health care improvement? A realist synthesis”. *J Health Serv Res Policy* 23.1 (2018), pp. 57–65.
 - [131] J. H. Hibbard et al. “Consumer competencies and the use of comparative quality information: it isn’t just about literacy”. *Med Care Res Rev* 64.4 (2007), pp. 379–94.
 - [132] R. Canaway et al. “Perceived barriers to effective implementation of public reporting of hospital performance data in Australia: a qualitative study”. *BMC Health Serv Res* 17.1 (2017), p. 391.
 - [133] R. Canaway et al. “"What is meant by public?": Stakeholder views on strengthening impacts of public reporting of hospital performance data”. *Soc Sci Med* 202 (2018), pp. 143–150.
 - [134] A. Aggarwal et al. “Patient Mobility for Elective Secondary Health Care Services in Response to Patient Choice Policies: A Systematic Review”. *Med Care Res Rev* 74.4 (2017), pp. 379–403.
 - [135] G. Moscelli et al. “Socioeconomic inequality of access to healthcare: Does choice explain the gradient?” *J Health Econ* 57 (2018), pp. 290–314.

-
- [136] C. H. Fung et al. “Systematic review: the evidence that publishing patient care performance data improves quality of care”. *Ann Intern Med* 148.2 (2008), pp. 111–23.
- [137] M. Hendriks et al. “Dutch healthcare reform: did it result in performance improvement of health plans? A comparison of consumer experiences over time”. *BMC Health Serv Res* 9 (2009), p. 167.
- [138] D. M. Berwick and D. L. Wald. “Hospital leaders’ opinions of the HCFA mortality data”. *JAMA* 263.2 (1990), pp. 247–9.
- [139] C. A. Sirio and J. L. McGee. “Public reporting of clinical outcomes—the data needs of health care stakeholders”. *Am J Med Qual* 11.1 (1996), S78–81.
- [140] G. E. Rosenthal et al. “Using hospital performance data in quality improvement: the Cleveland Health Quality Choice experience”. *Jt Comm J Qual Improv* 24.7 (1998), pp. 347–60.
- [141] F. T. Schut and W. P. Van de Ven. “Rationing and competition in the Dutch health-care system”. *Health Econ* 14.Suppl 1 (2005), S59–74.
- [142] R. H. Brook. “Health care reform is on the way: do we want to compete on quality?” *Ann Intern Med* 120.1 (1994), pp. 84–6.
- [143] J. H. Hibbard et al. “Choosing a health plan: do large employers use the data?” *Health Aff (Millwood)* 16.6 (1997), pp. 172–80.
- [144] D. B. Mukamel and A. I. Mushlin. “Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports”. *Med Care* 36.7 (1998), pp. 945–54.
- [145] K. L. Sherman et al. “Surgeons’ perceptions of public reporting of hospital and individual surgeon quality”. *Med Care* 51.12 (2013), pp. 1069–1075.
- [146] F. Kiernan and F. Rahman. “Measuring surgical performance: A risky game?” *Surgeon* 13.4 (2015), pp. 213–7.

-
- [147] E. M. Burns et al. “Understanding the strengths and weaknesses of public reporting of surgeon-specific outcome data”. *Health Aff (Millwood)* 35.3 (2016), pp. 415–21.
- [148] J. H. Wasfy et al. “Public reporting in cardiovascular medicine: accountability, unintended consequences, and promise for improvement”. *Circulation* 131.17 (2015), pp. 1518–27.
- [149] D. M. Shahian et al. “Risk aversion and public reporting. Part 1: observations from cardiac surgery and interventional cardiology”. *Ann Thorac Surg* 104.6 (2017), pp. 2093–2101.
- [150] R. K. Wadhera, J. D. Anderson, and R. W. Yeh. “High-risk percutaneous coronary intervention in public reporting states: the evidence, exclusion of critically ill patients, and implications”. *Curr Heart Fail Rep* 14.6 (2017), pp. 514–518.
- [151] J. M. Loeb. “The current state of performance measurement in health care”. *Int J Qual Health Care* 16 Suppl 1 (2004), pp. i5–9.
- [152] M. N. Marshall et al. “The public release of performance data: what do we expect to gain? A review of the evidence”. *JAMA* 283.14 (2000), pp. 1866–74.
- [153] P. G. Shekelle et al. *Does public release of performance results improve quality of care? A systematic review*. Report. 2008.
- [154] M. Faber et al. “Public reporting in health care: how do consumers use quality-of-care information? A systematic review”. *Med Care* 47.1 (2009), pp. 1–8.
- [155] Effective Practice Organisation of Care (EPOC). *Analysis in EPOC reviews. EPOC resources for review authors*. Report. 2013.
- [156] Effective Practice Organisation of Care (EPOC). *Reporting the effects of an intervention in EPOC reviews*. Report. 2018.

-
- [157] D. Metcalfe et al. “Impact of public release of performance data on the behaviour of healthcare consumers and providers”. *Cochrane Database Syst Rev* 9 (2018), p. CD004538.
- [158] J. Higgins and S. Green. *Cochrane Handbook for Systematic Reviews of Interventions*. London, U.K.: The Cochrane Collaboration, 2011.
- [159] J. B. Schroll, R. Moustgaard, and P. C. Gotzsche. “Dealing with substantial heterogeneity in Cochrane reviews. Cross-sectional study”. *BMC Med Res Methodol* 11 (2011), p. 22.
- [160] J. Huwaldt. *Plot Digitizer*: <http://plotdigitizer.sourceforge.net>. Computer Program. 2015.
- [161] A. Jelcic Kadic et al. “Extracting data from figures with software was faster, with higher interrater reliability than manual extraction”. *J Clin Epidemiol* 74 (2016), pp. 119–23.
- [162] G. Guyatt et al. “GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables”. *J Clin Epidemiol* 64.4 (2011), pp. 383–94.
- [163] W. M. Jang et al. “Effect of repeated public releases on cesarean section rates”. *J Prev Med Public Health* 44.1 (2011), pp. 2–8.
- [164] K. B. Flett et al. “Impact of Mandatory Public Reporting of Central Line-Associated Bloodstream Infections on Blood Culture and Antibiotic Utilization in Pediatric and Neonatal Intensive Care Units”. *Infect Control Hosp Epidemiol* 36.8 (2015), pp. 878–85.
- [165] A. D. DeVore et al. “Has Public Reporting of Hospital Readmission Rates Affected Patient Outcomes?: Analysis of Medicare Claims Data”. *J Am Coll Cardiol* 67.8 (2016), pp. 963–72.
- [166] K. E. Joynt et al. “Public Reporting of Mortality Rates for Hospitalized Medicare Patients and Trends in Mortality for Reported Conditions”. *Ann Intern Med* 165.3 (2016), pp. 153–60.

-
- [167] H. Liu et al. “Impact of State Reporting Laws on Central Line-Associated Bloodstream Infection Rates in U.S. Adult Intensive Care Units”. *Health Serv Res* 52.3 (2017), pp. 1079–1098.
- [168] J. V. Tu et al. “Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial”. *JAMA* 302.21 (2009), pp. 2330–7.
- [169] D. Ikkersheim and X. Koolman. “The use of quality information by general practitioners: does it alter choices? A randomized clustered study”. *BMC Fam Pract* 14 (2013), p. 95.
- [170] C. Liu et al. “Does public reporting influence antibiotic and injection prescribing to all patients? A cluster-randomized matched-pair trial in china”. *Medicine (Baltimore)* 95.26 (2016), e3965.
- [171] M. L. Rinke et al. “State-Mandated Hospital Infection Reporting Is Not Associated With Decreased Pediatric Health Care-Associated Infections”. *J Patient Saf* 11.3 (2015), pp. 123–34.
- [172] R. Bender and S. Lange. “Adjusting for multiple testing—when and how?” *J Clin Epidemiol* 54.4 (2001), pp. 343–9.
- [173] S. Hawley et al. “Sample size and power considerations for ordinary least squares interrupted time series analysis: a simulation study”. *Clin Epidemiol* 11 (2019), pp. 197–205.
- [174] O. C. Damman et al. “An international comparison of web-based reporting about health care quality: content analysis”. *J Med Internet Res* 12.2 (2010), e8.
- [175] *My NHS*. <http://tiny.cc/hvp05y>. 2019.
- [176] D. M. Shahian and S. L. Normand. “What is a performance outlier?” *BMJ Qual Saf* 24.2 (2015), pp. 95–9.

-
- [177] J. F. Figueroa, D. E. Wang, and A. K. Jha. “Characteristics of hospitals receiving the largest penalties by US pay-for-performance programmes”. *BMJ Qual Saf* 25.11 (2016), pp. 898–900.
- [178] E. B. Fos. “The unintended consequences of The Centers for Medicare and Medicaid Services pay-for-performance structures on safety-net hospitals and the low-income, medically vulnerable population”. *Health Serv Manage Res* 30.1 (2017), pp. 10–15.
- [179] J. Rumbold and S. Seaton. “Mid Staffs: disaster by numbers (or ‘how to create a drama out of a statistic’)”. *Journal of Medical Law and Ethics* 4.1 (2016), pp. 57–70.
- [180] C. Tsang and D. A. Cromwell. *Statistical methods developed for the National Hip Fracture Database annual report, 2014. A technical report*. Report. 2014.
- [181] M. Saklad. “Grading of patients for surgical procedures”. *Anaesthesiology* 2 (1941), pp. 281–284.
- [182] K. Y. Bilimoria et al. “Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons”. *J Am Coll Surg* 217.5 (2013), pp. 833–42.
- [183] O. Peacock et al. “Thirty-day mortality in patients undergoing laparotomy for small bowel obstruction”. *Br J Surg* 105.8 (2018), pp. 1006–1013.
- [184] R. Sutton et al. “The Surgical Risk Scale as an improved tool for risk-adjusted analysis in comparative surgical audit”. *Br J Surg* 89.6 (2002), pp. 763–8.
- [185] K. L. Protopapa et al. “Development and validation of the Surgical Outcome Risk Tool (SORT)”. *Br J Surg* 101.13 (2014), pp. 1774–83.
- [186] D. Mayhew, V. Mendonca, and B. V. S. Murthy. “A review of ASA physical status - historical perspectives and modern developments”. *Anaesthesia* 74.3 (2019), pp. 373–379.

-
- [187] W. L. Aronson, M. S. McAuliffe, and K. Miller. “Variability in the American Society of Anesthesiologists Physical Status Classification Scale”. *AANA J* 71.4 (2003), pp. 265–74.
- [188] R. Riley, C. Holman, and D. Fletcher. “Inter-rater reliability of the ASA physical status classification in a sample of anaesthetists in Western Australia”. *Anaesth Intensive Care* 42.5 (2014), pp. 614–8.
- [189] M. E. Charlson et al. “A new method of classifying prognostic comorbidity in longitudinal studies: development and validation”. *J Chronic Dis* 40.5 (1987), pp. 373–83.
- [190] A. Elixhauser et al. “Comorbidity measures for use with administrative data”. *Med Care* 36.1 (1998), pp. 8–27.
- [191] M. T. Sharabiani, P. Aylin, and A. Bottle. “Systematic review of comorbidity indices for administrative data”. *Med Care* 50.12 (2012), pp. 1109–18.
- [192] A. Molto and M. Dougados. “Comorbidity indices”. *Clin Exp Rheumatol* 32.5 Suppl 85 (2014), pp. 131–134.
- [193] J. Karres et al. “Predicting 30-day mortality following hip fracture surgery: evaluation of six risk prediction models”. *Injury* 46.2 (2015), pp. 371–7.
- [194] H. Quan et al. “Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries”. *Am J Epidemiol* 173.6 (2011), pp. 676–82.
- [195] M. E. Menendez et al. “Predicting in-hospital mortality in elderly patients With cervical spine fractures: a comparison of the Charlson and Elixhauser comorbidity measures”. *Spine (Phila Pa 1976)* 40.11 (2015), pp. 809–15.
- [196] H. Quan et al. “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data”. *Med Care* 43.11 (2005), pp. 1130–9.
- [197] A. Clegg et al. “Frailty in elderly people”. *Lancet* 381.9868 (2013), pp. 752–62.

-
- [198] T. Gilbert et al. “Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study”. *Lancet* 391.10132 (2018), pp. 1775–1782.
- [199] F. Aslam and N. A. Khan. “Tools for the assessment of comorbidity burden in rheumatoid arthritis”. *Front Med (Lausanne)* 5 (2018), p. 39.
- [200] Department of Health. *Governance arrangements for research ethics committees*. Report. Department of Health, 2011.
- [201] A. Herbert et al. “Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC)”. *Int J Epidemiol* 46.4 (2017), pp. 1093–1093i.
- [202] National Audit Office. *Healthcare across the UK: a comparison of the NHS in England, Scotland, Wales and Northern Ireland*. Report. 2012.
- [203] Office for National Statistics. *Impact of registration delays on mortality statistics: 2016*. Report. 2016.
- [204] D. G. Altman and P. Royston. “The cost of dichotomising continuous variables”. *BMJ* 332.7549 (2006), p. 1080.
- [205] Falls and Fragility Fracture Audit Programme (FFFAP). *National Hip Fracture Database (NHFD) Annual Report 2017*. Report. 2017.
- [206] R. J. A. Little. “A test of missing completely at random for multivariate data with missing values”. *Journal of the American Statistical Association* 83.404 (1988), pp. 1198–1202.
- [207] O. Harel et al. “Multiple Imputation for Incomplete Data in Epidemiologic Studies”. *Am J Epidemiol* 187.3 (2018), pp. 576–584.
- [208] J. B. Carlin et al. “Tools for analyzing multiple imputed datasets”. *Stata Journal* 3.3 (2003), pp. 226–244.
- [209] I. R. White, R. M. Daniel, and P. Royston. “Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical data.” *Computational Statistics and Data Analysis* 54 (2010), pp. 2267–2275.

-
- [210] N. Serrano. “Calibration strategies to validate predictive models: is new always better?” *Intensive Care Med* 38.8 (2012), pp. 1246–8.
- [211] M.A. Cleves. “From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models.” *The Stata Journal* 2.3 (2002), pp. 301–313.
- [212] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. 2nd edition. New York, USA: Wiley-Blackwell Ltd., 2000.
- [213] D. W. Hosmer and S. Lemeshow. “Goodness of fit tests for the multiple logistic regression model”. *Communications in Statistics - Theory and Methods* 9.10 (1980), pp. 1043–1069.
- [214] C. S. Crowson, E. J. Atkinson, and T. M. Therneau. “Assessing calibration of prognostic risk scores”. *Stat Methods Med Res* 25.4 (2016), pp. 1692–706.
- [215] G. Nattino, S. Finazzi, and G. Bertolini. “A new test and graphical tool to assess the goodness of fit of logistic regression models”. *Stat Med* 35.5 (2016), pp. 709–20.
- [216] C. van Walraven et al. “A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data”. *Med Care* 47.6 (2009), pp. 626–33.
- [217] NHS Health Research Authority. *Governance arrangements for research ethics committees: 2018 edition*. Report. 2018.
- [218] L. G. Glance et al. “Accuracy of hospital report cards based on administrative data”. *Health Serv Res* 41.4 Pt 1 (2006), pp. 1413–37.
- [219] L. P. Fried et al. “Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care”. *J Gerontol A Biol Sci Med Sci* 59.3 (2004), pp. 255–63.
- [220] Y. Yang et al. “Risk factors for postoperative delirium following hip fracture repair in elderly patients: a systematic review and meta-analysis”. *Aging Clin Exp Res* 29.2 (2017), pp. 115–126.

-
- [221] J. Filippou. “Slow and costly access to anonymised patient data impedes academic research”. *BMJ* 351 (2015), h5087.
- [222] K. Stewart, R. Buckingham, and F. Martin. “Delayed access to clinical audit data has risks for patient care”. *BMJ* 351 (2015), h5812.
- [223] M. D. Wiles et al. “Nottingham Hip Fracture Score as a predictor of one year mortality in patients undergoing surgical repair of fractured neck of femur”. *Br J Anaesth* 106.4 (2011), pp. 501–4.
- [224] T. C. Marufu et al. “Prediction of 30-day mortality after hip fracture surgery by the Nottingham Hip Fracture Score and the Surgical Outcome Risk Tool”. *Anaesthesia* 71.5 (2016), pp. 515–21.
- [225] K. Tilkeridis et al. “Validity of Nottingham Hip Fracture Score in different health systems and a new modified version validated to the Greek population”. *Med Sci Monit* 24 (2018), pp. 7665–7672.
- [226] N. Gunasekera et al. “Hip fracture audit: the Nottingham experience”. *Osteoporos Int* 21.Suppl 4 (2010), S647–53.
- [227] C. Tsang et al. “Predicting 30-day mortality after hip fracture surgery: Evaluation of the National Hip Fracture Database case-mix adjustment model”. *Bone Joint Res* 6.9 (2017), pp. 550–556.
- [228] T. Lloyd, S. R. Deeny, and A. Steventon. “Weekend admissions may be associated with poorer recording of long-term comorbidities: a prospective study of emergency admissions using administrative data”. *BMC Health Serv Res* 18.1 (2018), p. 863.
- [229] N. A. Heywood et al. “Improving accuracy of clinical coding in surgery: collaboration is key”. *J Surg Res* 204.2 (2016), pp. 490–495.
- [230] E. Herrett et al. “Data Resource Profile: Clinical Practice Research Datalink (CPRD)”. *Int J Epidemiol* 44.3 (2015), pp. 827–36.

-
- [231] B. T. Blak et al. “Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates”. *Inform Prim Care* 19.4 (2011), pp. 251–5.
- [232] K. Rockwood et al. “A global clinical measure of fitness and frailty in elderly people”. *CMAJ* 173.5 (2005), pp. 489–95.
- [233] L. N. Lansbury et al. “Use of the electronic Frailty Index to identify vulnerable patients: a pilot study in primary care”. *Br J Gen Pract* 67.664 (2017), e751–e756.
- [234] D. J. Cundall-Curry et al. “Data errors in the National Hip Fracture Database: a local validation study”. *Bone Joint J* 98-B.10 (2016), pp. 1406–1409.
- [235] J. E. Lawrence et al. “The use of an electronic health record system reduces errors in the National Hip Fracture Database”. *Age Ageing* doi:10.1093/ageing/afy177 (2018).
- [236] J. Masters et al. “Interpreting and reporting fracture classifications and operation type in hip fracture: implications for research studies and routine national audits”. *Bone Joint J* In press (2019).
- [237] M. Chamberlain and H. Pugh. “Improving inpatient care with the introduction of a hip fracture pathway”. *BMJ Qual Improv Rep* 4.1 (2015).
- [238] Healthcare Quality Improvement Partnership (HQIP). *The National Hip Fracture Database National Report 2011*. Report. 2011.
- [239] K. Grasic, A. R. Mason, and A. Street. “Paying for the quantity and quality of hospital care: the foundations and evolution of payment policy in England”. *Health Econ Rev* 5.1 (2015), p. 50.
- [240] B Gerschlick. *Country Background Note: United Kingdom (England)*. Report. 2016.
- [241] R. J. O’Connor and V. C. Neumann. “Payment by results or payment by outcome? The history of measuring medicine”. *J R Soc Med* 99.5 (2006), pp. 226–31.

-
- [242] A. Maynard. “The powers and pitfalls of payment for performance”. *Health Econ* 21.1 (2012), pp. 3–12.
- [243] Y. K. Ogundeji, J. M. Bland, and T. A. Sheldon. “The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes”. *Health Policy* 120.10 (2016), pp. 1141–1150.
- [244] L. Marshall, A. Charlesworth, and J. Hurst. *The NHS payment system: evolving policy and emerging evidence*. Report. 2014.
- [245] S. Farrar et al. “Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis”. *BMJ* 339 (2009), b3047.
- [246] Department of Health. *A Simple Guide to Payment by Results*. Report. 2011.
- [247] Department of Health. *The NHS Plan*. Report. 2000.
- [248] Department of Health. *Delivering the NHS Plan*. Report. 2002.
- [249] Department of Health. *High Quality Care For All - NHS Next Stage Review Final Report*. Report. 2008.
- [250] Department of Health. *Equity and Excellence: Liberating the NHS*. Report. 2010.
- [251] J. Neuburger et al. “The impact of a national clinician-led audit initiative on care and mortality after hip fracture in England: an external evaluation using time trends in non-audit data”. *Med Care* 53.8 (2015), pp. 686–91.
- [252] S. K. Khan et al. “The Best Practice Tariff helps improve management of neck of femur fractures: a completed audit loop”. *Br J Hosp Med (Lond)* 74.11 (2013), pp. 644–7.
- [253] M. Kommer, K. Gokaraju, and S. Singh. “Changing the consultant on calls from a daily to weekly rotation system reduces time to theater for patients with hip fracture to improve quality of care: a retrospective study of 2 cohorts of patients presenting with hip fracture”. *Geriatr Orthop Surg Rehabil* 5.2 (2014), pp. 69–72.

-
- [254] E. Britton and W. Nash. “The new neck of femur fracture target: experience in a district general hospital”. *Int J Health Care Qual Assur* 27.1 (2014), pp. 36–43.
 - [255] R. Lisk and K. Yeong. “Reducing mortality from hip fractures: a systematic quality improvement programme”. *BMJ Qual Improv Rep* 3.1 (2014).
 - [256] D. Hawkes et al. “Improving the care of patients with a hip fracture: a quality improvement report”. *BMJ Qual Saf* 24.8 (2015), pp. 532–8.
 - [257] M. Diamant et al. “"Early Trigger" Intravenous Vitamin K: Optimizing Target-Driven Care in Warfarinised Patients With Hip Fracture”. *Geriatr Orthop Surg Rehabil* 6.4 (2015), pp. 263–8.
 - [258] Healthcare Quality Improvement Partnership (HQIP). <https://www.hqip.org.uk/az-of-nca/>. <http://tiny.cc/sxp05y>. 2018.
 - [259] S. K. Khan et al. “Achieving best practice tariff may not reflect improved survival after hip fracture treatment”. *Clin Interv Aging* 9 (2014), pp. 2097–102.
 - [260] P. Craig et al. “Using natural experiments to evaluate population health interventions: new Medical Research Council guidance”. *J Epidemiol Community Health* 66.12 (2012), pp. 1182–6.
 - [261] E. Kontopantelis et al. “Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis”. *BMJ* 350 (2015), h2750.
 - [262] J. B. Dimick and A. M. Ryan. “Methods for evaluating changes in health care policy: the difference-in-differences approach”. *JAMA* 312.22 (2014), pp. 2401–2.
 - [263] Information Services Division Scotland. *SMR Completeness*. <http://tiny.cc/5pp05y>. 2018.
 - [264] K Robson. *The National Health Service in Scotland*. Report. 2016.

-
- [265] A. Linden. *ITSA: Stata module to perform interrupted time series analysis for single and multiple groups*. Computer Program. 2017.
- [266] J. M. Villa. *DIFF: Stata module to perform differences in differences estimation*. Computer Program. 2018.
- [267] N. K. Patel et al. “Implementing the National Hip Fracture Database: An audit of care”. *Injury* 44.12 (2013), pp. 1934–9.
- [268] British Orthopaedic Association and British Geriatrics Society. *The Care of Patients with Fragility Fracture*. London: BOA, 2007.
- [269] C. P. Bretherton and M. J. Parker. “Early surgery for patients with a fracture of the hip decreases 30-day mortality”. *Bone Joint J* 97-B.1 (2015), pp. 104–8.
- [270] Medicare Learning Network (MLN). *Hospital Value-Based Purchasing*. Report. 2017.
- [271] J. L. Lee et al. “Value-based insurance design: quality improvement but no cost savings”. *Health Aff (Millwood)* 32.7 (2013), pp. 1251–7.
- [272] T. T. Chee et al. “Current State of Value-Based Purchasing Programs”. *Circulation* 133.22 (2016), pp. 2197–205.
- [273] J. F. Figueroa et al. “Association between the Value-Based Purchasing pay for performance program and patient mortality in US hospitals: observational study”. *BMJ* 353 (2016), p. i2214.
- [274] M. D. Dalzell. “Now Might Be the Right Time To Kill Hospital Value-based Purchasing Program”. *Manag Care* 26.5 (2017), pp. 14–16.
- [275] J. Rau. *1,700 hospitals win quality bonuses from Medicare, but most will never collect*. <http://tiny.cc/b4p05y>. Jan. 2015.
- [276] Leavitt Partners. *Growth of population-based payments is not associated with a decrease in market-level cost growth, yet*. Report. 2018.

-
- [277] A. K. Jha. “Time to get serious about pay for performance”. *JAMA* 309.4 (2013), pp. 347–8.
- [278] R. McDonald et al. *A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs*. Report. 2012.
- [279] NHS England and NHS Improvement. *2017/18 and 2018/19 National Tariff Payment System*. Report. 2016.
- [280] C. A. Brauer et al. “Incidence and mortality of hip fractures in the United States”. *JAMA* 302.14 (2009), pp. 1573–9.
- [281] E. Michael Lewiecki et al. “Hip fracture trends in the United States, 2002 to 2015”. *Osteoporos Int* 29.3 (2018), pp. 717–722.
- [282] K. L. Haywood et al. “Developing a core outcome set for hip fracture trials”. *Bone Joint J* 96-B.8 (2014), pp. 1016–23.
- [283] H. E. Lester, K. L. Hannon, and S. M. Campbell. “Identifying unintended consequences of quality indicators: a qualitative study”. *BMJ Qual Saf* 20.12 (2011), pp. 1057–61.
- [284] J. R. Smith et al. “Distal femoral fractures: The need to review the standard of care”. *Injury* 46.6 (2015), pp. 1084–8.
- [285] K. G. Shojania. “Are increases in emergency use and hospitalisation always a bad thing? Reflections on unintended consequences and apparent backfires”. *BMJ Qual Saf* (2019).
- [286] C. Hopley et al. “Primary total hip arthroplasty versus hemiarthroplasty for displaced intracapsular hip fractures in older patients: systematic review”. *BMJ* 340 (2010), p. c2332.
- [287] F. Frihagen, L. Nordsletten, and J. E. Madsen. “Hemiarthroplasty or internal fixation for intracapsular displaced femoral neck fractures: randomised controlled trial”. *BMJ* 335.7632 (2007), pp. 1251–4.

-
- [288] M. Clayer and J. Bruckner. “The outcome of Austin-Moore hemiarthroplasty for fracture of the femoral neck”. *Am J Orthop (Belle Mead NJ)* 26.10 (1997), pp. 681–4.
- [289] P. P. Avery et al. “Total hip replacement and hemiarthroplasty in mobile, independent patients with a displaced intracapsular fracture of the femoral neck: a seven- to ten-year follow-up report of a prospective randomised controlled trial”. *J Bone Joint Surg Br* 93.8 (2011), pp. 1045–8.
- [290] L. Yu, Y. Wang, and J. Chen. “Total hip arthroplasty versus hemiarthroplasty for displaced femoral neck fractures: meta-analysis of randomized trials”. *Clin Orthop Relat Res* 470.8 (2012), pp. 2235–43.
- [291] R. H. Jay and D. Hipps. “Hip fracture—great steps forward but we still need better evidence. A commentary on NICE CG124 and QS16 on fractured neck of femur”. *Age Ageing* 47.5 (2018), pp. 630–632.
- [292] C. J. Lavernia and J. F. Guzman. “Relationship of surgical volume to short-term mortality, morbidity, and hospital charges in arthroplasty”. *J Arthroplasty* 10.2 (1995), pp. 133–40.
- [293] L. C. Walker et al. “Provision of total hip replacement for displaced intracapsular hip fracture and the outcomes: audit of local practice based on NICE guidelines”. *Hip Int* 26.2 (2016), pp. 153–7.
- [294] A. Fishlock, C. Scarsbrook, and R. Marsh. “Adherence to guidelines regarding total hip replacement for fractured neck of femur”. *Ann R Coll Surg Engl* 98.6 (2016), pp. 422–4.
- [295] H. M. Hodkinson. “Evaluation of a mental test score for assessment of mental impairment in the elderly”. *Age Ageing* 1.4 (1972), pp. 233–8.
- [296] D. McLennan et al. *The English Indices of Deprivation 2010*. Report. Department for Communities and Local Government, 2011.

-
- [297] National Institute for Health and Care Excellence (NICE). *Hip fracture: the management of hip fracture in adults [CG124]. Measuring the use of this guidance*. <http://tiny.cc/4up05y>. 2015.
- [298] D. M. MacKenzie et al. “Brief cognitive screening of the elderly: a comparison of the Mini-Mental State Examination (MMSE), Abbreviated Mental Test (AMT) and Mental Status Questionnaire (MSQ)”. *Psychol Med* 26.2 (1996), pp. 427–30.
- [299] L. Young and J. George. *Guidelines for the diagnosis and management of delirium in the elderly*. Report. 1997.
- [300] T. M. Therneau and E. J. Atkinson. *An introduction to recursive partitioning using the rpart routines*. Report. Mayo Foundation, 2018.
- [301] J. Carthey et al. “Breaking the rules: understanding non-compliance with policies and guidelines”. *BMJ* 343 (2011), p. d5283.
- [302] National Institute for Health and Care Excellence (NICE). *Addendum to Clinical Guideline 124, Hip fracture: management*. Report. 2017.
- [303] U. Hedlundh et al. “Surgical experience related to dislocations after total hip arthroplasty”. *J Bone Joint Surg Br* 78.2 (1996), pp. 206–9.
- [304] J. N. Katz et al. “Association between hospital and surgeon procedure volume and outcomes of total hip replacement in the United States medicare population”. *J Bone Joint Surg Am* 83-A.11 (2001), pp. 1622–9.
- [305] J. N. Katz et al. “Association of hospital and surgeon volume of total hip replacement with functional status and satisfaction three years following surgery”. *Arthritis Rheum* 48.2 (2003), pp. 560–8.
- [306] E. Losina et al. “Early failures of total hip replacement: effect of surgeon volume”. *Arthritis Rheum* 50.4 (2004), pp. 1338–43.
- [307] M. Bhandari et al. “Operative management of displaced femoral neck fractures in elderly patients. An international survey”. *J Bone Joint Surg Am* 87.9 (2005), pp. 2122–30.

-
- [308] M. L. Costa, S. S. Jameson, and M. R. Reed. “Do large pragmatic randomised trials change clinical practice?: assessing the impact of the Distal Radius Acute Fracture Fixation Trial (DRAFFT)”. *Bone Joint J* 98-B.3 (2016), pp. 410–3.
- [309] M. Bhandari et al. “Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial”. *BMJ Open* 5.2 (2015), e006263.
- [310] J. W. Dekker et al. “Use of different comorbidity scores for risk-adjustment in the evaluation of quality of colorectal cancer surgery: does it matter?” *Eur J Surg Oncol* 38.11 (2012), pp. 1071–8.
- [311] W. P. Tan et al. “American Society of Anesthesiologists class and Charlson’s comorbidity index as predictors of postoperative colorectal anastomotic leak: a single-institution experience”. *J Surg Res* 184.1 (2013), pp. 115–9.
- [312] R. G. Whitmore et al. “ASA grade and Charlson Comorbidity Index of spinal surgery patients: correlation with complications and societal costs”. *Spine J* 14.1 (2014), pp. 31–8.
- [313] Department for Communities and Local Government. *The English Indices of Deprivation 2015 - Frequently Asked Questions (FAQs)*. Report. 2016.
- [314] I. Deas et al. “Measuring neighbourhood deprivation: a critique of the Index of Multiple Deprivation”. *Environment and Planning C: Government and Policy* 21 (2003), pp. 883–903.
- [315] J. Saunders. “Weighted Census-based deprivation indices: their use in small areas”. *J Public Health Med* 20.3 (1998), pp. 253–60.
- [316] B. Vyawahare et al. “Impact of the National Institute for Health and Care Excellence (NICE) guidance on medical technology uptake: analysis of the uptake of spinal cord stimulation in England 2008-2012”. *BMJ Open* 4.1 (2014), e004182.

-
- [317] Department of Health and Social Care. *NHS Constitution for England*. Report. 2015.
- [318] M Rawlins. “NICE guidelines are crucial – but they are not compulsory”. *Pulse* (Oct. 2012).
- [319] S. D. Ferrara, Rafael Boscolo-Berto, and Guido Viel. *Malpractice and medical liability : European state of the art and guidelines*. Berlin: Springer, 2013, xxv, 368 pages.
- [320] B. Hurwitz. “Legal and political considerations of clinical practice guidelines”. *BMJ* 318.7184 (1999), pp. 661–4.
- [321] D. Metcalfe et al. “Validation of a prospective cohort study of older adults with hip fractures”. *Bone Joint J* In press (2019).
- [322] L. Johnston. “Informed consent and the lingering shadow of *Chester v Afshar*: Part 1”. *Scottish Law Times* 81 (2015), p. 86.
- [323] I. Hogarth. <https://clinicalnegligence.blog/2018/06/04/surgical-consent-case-report-materiality-of-risk-montgomery-vs-bolam/>. Blog. 2018.
- [324] S. L. Dickman, D. U. Himmelstein, and S. Woolhandler. “Inequality and the health-care system in the USA”. *Lancet* 389.10077 (2017), pp. 1431–1441.
- [325] M. Asaria et al. “How a universal health system reduces inequalities: lessons from England”. *J Epidemiol Community Health* 70.7 (2016), pp. 637–43.
- [326] R. Raine et al. “Sociodemographic variations in the contribution of secondary drug prevention to stroke survival at middle and older ages: cohort study”. *BMJ* 338 (2009), b1279.
- [327] R. Raine et al. “Social variations in access to hospital care for patients with colorectal, breast, and lung cancer between 1999 and 2006: retrospective analysis of hospital episode statistics”. *BMJ* 340 (2010), b5479.
- [328] A. Szczepura. “Access to health care for ethnic minority populations”. *Post-grad Med J* 81.953 (2005), pp. 141–7.

-
- [329] A. Dixon et al. “Is the British National Health Service equitable? The evidence on socioeconomic differences in utilization”. *J Health Serv Res Policy* 12.2 (2007), pp. 104–9.
- [330] C. Lejeune et al. “Socio-economic disparities in access to treatment and their impact on colorectal cancer survival”. *Int J Epidemiol* 39.3 (2010), pp. 710–7.
- [331] S. E. Roberts and M. J. Goldacre. “Time trends and demography of mortality after fractured neck of femur in an English population, 1968-98: database study”. *BMJ* 327.7418 (2003), pp. 771–5.
- [332] K. Thorne et al. “The impact of social deprivation on mortality following hip fracture in England and Wales: a record linkage study”. *Osteoporos Int* 27.9 (2016), pp. 2727–2737.
- [333] P. K. Kristensen et al. “Socioeconomic inequality in clinical outcome among hip fracture patients: a nationwide cohort study”. *Osteoporos Int* 28.4 (2017), pp. 1233–1243.
- [334] A. P. Barone et al. “Effects of socioeconomic position on 30-day mortality and wait for surgery after hip fracture”. *Int J Qual Health Care* 21.6 (2009), pp. 379–86.
- [335] I. L. Hsu et al. “Socioeconomic Inequality in One-Year Mortality of Elderly People with Hip Fracture in Taiwan”. *Int J Environ Res Public Health* 15.2 (2018).
- [336] C. J. Dy et al. “Racial and Socioeconomic Disparities in Hip Fracture Care”. *J Bone Joint Surg Am* 98.10 (2016), pp. 858–65.
- [337] M. Marmot. *Fair society, healthy lives: strategic review of health inequalities in England post-2010*. Report. 2010.
- [338] T. Doran et al. “Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework”. *Lancet* 372.9640 (2008), pp. 728–36.

-
- [339] J. Dixon et al. *Equity in the provision of palliative care in the UK: review of the evidence*. Report. 2015.
- [340] R. Cookson et al. “Socio-economic inequalities in health care in England”. *Fiscal Studies* 37.3-4 (2016), pp. 371–403.
- [341] M. Goddard et al. “Where did all the GPs go? Increasing supply and geographical equity in England and Scotland”. *J Health Serv Res Policy* 15.1 (2010), pp. 28–35.
- [342] S. Willems et al. “Socio-economic status of the patient and doctor-patient communication: does it make a difference?” *Patient Educ Couns* 56.2 (2005), pp. 139–46.
- [343] A. H. Haider et al. “Association of unconscious race and social class bias with vignette-based clinical assessments by medical students”. *JAMA* 306.9 (2011), pp. 942–51.
- [344] C. FitzGerald and S. Hurst. “Implicit bias in healthcare professionals: a systematic review”. *BMC Med Ethics* 18.1 (2017), p. 19.
- [345] M. van Ryn and J. Burke. “The effect of patient race and socio-economic status on physicians’ perceptions of patients”. *Soc Sci Med* 50.6 (2000), pp. 813–28.
- [346] A. H. Haider et al. “Unconscious race and social class bias among acute care surgical clinicians and clinical treatment decisions”. *JAMA Surg* 150.5 (2015), pp. 457–64.
- [347] A. Kleebauer and C. Comerford. “Government commits to seven-day NHS”. *Nurs Manag (Harrow)* 22.3 (2015), p. 6.
- [348] N. Freemantle et al. “Weekend hospitalization and additional risk of death: an analysis of inpatient data”. *J R Soc Med* 105.2 (2012), pp. 74–84.
- [349] B. Keogh. “Should the NHS work at weekends as it does in the week? Yes”. *BMJ* 346 (2013), f621.

-
- [350] C. J. Thomas et al. “The weekend effect: short-term mortality following admission with a hip fracture”. *Bone Joint J* 96-B.3 (2014), pp. 373–8.
 - [351] J. Neuburger et al. “Safe working in a 7-day service. Experience of hip fracture care as documented by the UK National Hip Fracture Database”. *Age Ageing* 47.5 (2018), pp. 741–745.
 - [352] T. W. Lau, C. Fang, and F. Leung. “The effectiveness of a geriatric hip fracture clinical pathway in reducing hospital and rehabilitation length of stay and improving short-term mortality rates”. *Geriatr Orthop Surg Rehabil* 4.1 (2013), pp. 3–9.
 - [353] M. Kelly and S. L. Kates. “Geriatric fracture centers-improved patient care and economic benefits : English Version”. *Unfallchirurg* (2015).
 - [354] B. J. Gabbe et al. “Patient perspectives of care in a regionalised trauma system: lessons from the Victorian State Trauma System”. *Med J Aust* 198.3 (2013), pp. 149–52.
 - [355] D. C. Perry et al. “Inequalities in use of total hip arthroplasty for hip fracture: population based study”. *BMJ* 353 (2016), p. i2021.
 - [356] H. Chaudhry. “Total hip arthroplasty after hip fracture”. *BMJ* 353 (2016), p. i2217.
 - [357] O. Lerche. *Poor patients less likely to have hip replacement surgery, say experts*. <http://tiny.cc/r3p05y>. Apr. 2016.
 - [358] B. Ravi et al. “Relation between surgeon volume and risk of complications after total hip arthroplasty: propensity score matched cohort study”. *BMJ* 348 (2014), g3284.
 - [359] M. C. Fu et al. “Surgery for a fracture of the hip within 24 hours of admission is independently associated with reduced short-term post-operative complications”. *Bone Joint J* 99-B.9 (2017), pp. 1216–1222.
 - [360] M. Sampson et al. “Surveillance search techniques identified the need to update systematic reviews”. *J Clin Epidemiol* 61.8 (2008), pp. 755–62.

-
- [361] M. Rice et al. “Testing the effectiveness of simplified search strategies for updating systematic reviews”. *J Clin Epidemiol* 88 (2017), pp. 148–153.
- [362] R. DerSimonian and N. Laird. “Meta-analysis in clinical trials”. *Control Clin Trials* 7.3 (1986), pp. 177–88.
- [363] D. Moher et al. “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement”. *BMJ* 339 (2009), b2535.
- [364] M. A. Brookhart et al. “Variable selection for propensity score models”. *Am J Epidemiol* 163.12 (2006), pp. 1149–56.
- [365] E. Leuven and B. Sianesi. *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphic, and covariate imbalance testing*. Computer Program. 2003.
- [366] P. C. Austin. “Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples”. *Stat Med* 28.25 (2009), pp. 3083–107.
- [367] P. C. Austin. “Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations”. *Biom J* 51.1 (2009), pp. 171–84.
- [368] L. J. Keele. *An overview of rbounds: an R package for Rosenbaum bounds sensitivity analysis with matched data*. Report. 2010.
- [369] P. R. Rosenbaum. *Observational Studies*. 2nd edition. New York, USA: Springer, 2002.
- [370] S. O. Becker and M. Caliendo. “Sensitivity analysis for average treatment effects”. *The Stata Journal* 7.1 (2007), pp. 71–83.
- [371] J. P. Fine and R. J. Gray. “A proportional hazards model for the subdistribution of a competing risk”. *Journal of the American Statistical Association* 94.446 (1999), pp. 496–509.

-
- [372] R. L. Williams. “A note on robust variance estimation for cluster-correlated data”. *Biometrics* 56.2 (2000), pp. 645–6.
- [373] P. C. Austin. “The performance of different propensity score methods for estimating marginal hazard ratios”. *Stat Med* 32.16 (2013), pp. 2837–49.
- [374] D. Ho et al. *MatchIt: nonparametric preprocessing for parametric causal inference*. Computer Program. 2018.
- [375] L. J. Keele. *rbounds: perform Rosenbaum bounds sensitivity tests for matched and unmatched data*. Computer Program. 2014.
- [376] S. K. Goh et al. “Meta-analysis comparing total hip arthroplasty with hemiarthroplasty in the treatment of displaced neck of femur fracture”. *J Arthroplasty* 24.3 (2009), pp. 400–6.
- [377] C. Carroll et al. “Hemiarthroplasty and total hip arthroplasty for treating primary intracapsular fracture of the hip: a systematic review and cost-effectiveness analysis”. *Health Technol Assess* 15.36 (2011), pp. 1–74.
- [378] A. Zi-Sheng et al. “Hemiarthroplasty vs primary total hip arthroplasty for displaced fractures of the femoral neck in the elderly: a meta-analysis”. *J Arthroplasty* 27.4 (2012), pp. 583–90.
- [379] J. H. He et al. “Meta-analysis comparing total hip arthroplasty with hemiarthroplasty in the treatment of displaced femoral neck fractures in patients over 70 years old”. *Chin J Traumatol* 15.4 (2012), pp. 195–200.
- [380] Y. Zhao et al. “Outcome of hemiarthroplasty and total hip replacement for active elderly patients with displaced femoral neck fractures: a meta-analysis of 8 randomized clinical trials”. *PLoS One* 9.5 (2014), e98071.
- [381] F. Wang et al. “Comparison of bipolar hemiarthroplasty and total hip arthroplasty for displaced femoral neck fractures in the healthy elderly: a meta-analysis”. *BMC Musculoskelet Disord* 16 (2015), p. 229.

-
- [382] R. P. Baker et al. “Total hip arthroplasty and hemiarthroplasty in mobile, independent patients with a displaced intracapsular fracture of the femoral neck. A randomized, controlled trial”. *J Bone Joint Surg Am* 88.12 (2006), pp. 2583–9.
- [383] J. F. Keating et al. “Randomized comparison of reduction and fixation, bipolar hemiarthroplasty, and total hip arthroplasty. Treatment of displaced intracapsular hip fractures in healthy older patients”. *J Bone Joint Surg Am* 88.2 (2006), pp. 249–60.
- [384] R. Blomfeldt et al. “A randomised controlled trial comparing bipolar hemiarthroplasty with total hip replacement for displaced intracapsular fractures of the femoral neck in elderly patients”. *J Bone Joint Surg Br* 89.2 (2007), pp. 160–5.
- [385] M. Cadossi et al. “A comparison of hemiarthroplasty with a novel polycarbonate-urethane acetabular component for displaced intracapsular fractures of the femoral neck: a randomised controlled trial in elderly patients”. *Bone Joint J* 95-B.5 (2013), pp. 609–15.
- [386] W. Macaulay et al. “Total hip arthroplasty is less painful at 12 months compared with hemiarthroplasty in treatment of displaced femoral neck fracture”. *HSS J* 4.1 (2008), pp. 48–54.
- [387] C. Rogmark and O. Leonardsson. “Hip arthroplasty for the treatment of displaced fractures of the femoral neck in elderly patients”. *Bone Joint J* 98-B.3 (2016), pp. 291–7.
- [388] Z. Wang and T. Bhattacharyya. “Outcomes of Hemiarthroplasty and Total Hip Arthroplasty for Femoral Neck Fracture: A Medicare Cohort Study”. *J Orthop Trauma* 31.5 (2017), pp. 260–263.
- [389] B. Ravi et al. “Comparing complications and costs of total hip arthroplasty versus hemiarthroplasty for femoral neck fracture: a propensity-score matched population-based study”. *In press* (2019).

-
- [390] S. S. Jameson et al. “Cemented hemiarthroplasty or hip replacement for intracapsular neck of femur fracture? A comparison of 7732 matched patients using national data”. *Injury* 44.12 (2013), pp. 1940–4.
 - [391] M. J. Grosso et al. “Hemiarthroplasty for Displaced Femoral Neck Fractures in the Elderly Has a Low Conversion Rate”. *J Arthroplasty* 32.1 (2017), pp. 150–154.
 - [392] K. Iamthanaporn, K. Chareancholvanich, and C. Pornrattanamaneewong. “Reasons for revision of failed hemiarthroplasty: Are there any differences between unipolar and bipolar?” *Eur J Orthop Surg Traumatol* 28.6 (2018), pp. 1117–1123.
 - [393] G. Bortolussi et al. “Identifying cardiac surgery operations in hospital episode statistics administrative database, with an OPCS-based classification of procedures, validated against clinical data”. *BMJ Open* 9.3 (2019), e023316.
 - [394] A. Abadie and G.W. Imbens. “Large sample properties of matching estimators for average treatment effects”. *Econometrica* 74 (2006), pp. 235–267.
 - [395] G. King and R. Nielsen. “Why propensity scores should not be used for matching” (2016), <http://j.mp/2ovYGsW>.
 - [396] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge, U.K.: Cambridge University Press, 2013.
 - [397] NHS England. *Delivering the Forward View: NHS Planning Guidance 2016/17 - 2020/21*. Report. 2015.
 - [398] M. L. Costa et al. “Percutaneous fixation with Kirschner wires versus volar locking plate fixation in adults with dorsally displaced fracture of distal radius: randomised controlled trial”. *BMJ* 349 (2014), g4807.
 - [399] A. Rangan et al. “Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus: the PROFHER randomized clinical trial”. *JAMA* 313.10 (2015), pp. 1037–47.

-
- [400] K. Willett et al. “Close contact casting vs surgery for initial treatment of unstable ankle fractures in older adults: a randomized clinical trial”. *JAMA* 316.14 (2016), pp. 1455–1463.
- [401] K. Harron et al. “Challenges in administrative data linkage for research”. *Big Data Soc* 4.2 (2017), p. 2053951717745678.
- [402] N. Dattani et al. “Accessing electronic administrative health data for research takes time”. *Arch Dis Child* 98.5 (2013), pp. 391–2.
- [403] Wellcome Trust. *Delays faced by researchers trying to access data from HSCIC*. Report. 2015.
- [404] R. C. Jones, K. G. Jones, and S. Gaduzo. *Re: Delayed access to clinical audit data has risks for patient care*. Web Page. 2015.
- [405] J. M. Queally et al. “Intramedullary nails for extracapsular hip fractures in adults”. *Cochrane Database Syst Rev* 9 (2014), p. CD004961.
- [406] J. Ryg et al. “Hip fracture patients at risk of second hip fracture: a nationwide population-based cohort study of 169,145 cases during 1977-2001”. *J Bone Miner Res* 24.7 (2009), pp. 1299–307.
- [407] A. Zisberg et al. “Low mobility during hospitalization and functional decline in older adults”. *J Am Geriatr Soc* 59.2 (2011), pp. 266–73.
- [408] G. Rodan et al. “Bone safety of long-term bisphosphonate treatment”. *Curr Med Res Opin* 20.8 (2004), pp. 1291–300.
- [409] D. C. Perry et al. “Designing clinical trials in trauma surgery: overcoming research barriers”. *Bone Joint Res* 3.4 (2014), pp. 123–9.
- [410] L. G. Hemkens. “How routinely collected data for randomized trials provide long-term randomized real-world evidence”. *JAMA Netw Open* 1.8 (2018), e186014.

-
- [411] X. L. Griffin et al. “the Targon femoral neck hip screw versus cannulated screws for internal fixation of intracapsular fractures of the hip: a randomised controlled trial”. *Bone Joint J* 96-B.5 (2014), pp. 652–7.
- [412] X. L. Griffin et al. “The Warwick Hip Trauma Evaluation One: a randomised pilot trial comparing the X-Bolt Dynamic Hip Plating System with sliding hip screw fixation in complex extracapsular hip fractures: WHiTE (One)”. *Bone Joint J* 98-B.5 (2016), pp. 686–9.
- [413] M. L. Costa et al. “World Hip Trauma Evaluation (WHiTE): framework for embedded comprehensive cohort studies”. *BMJ Open* 6.10 (2016), e011679.
- [414] A. L. Sims et al. “A randomized controlled trial comparing the Thompson hemiarthroplasty with the Exeter polished tapered stem and Unitrax modular head in the treatment of displaced intracapsular fractures of the hip: the WHiTE 3: HEMI Trial”. *Bone Joint J* 100-B.3 (2018), pp. 352–360.
- [415] S. H. Chung et al. “Changes in the cesarean section rate in Korea (1982–2012) and a review of the associated factors”. *J Korean Med Sci* 29.10 (2014), pp. 1341–52.
- [416] X. Wang et al. “Effect of publicly reporting performance data of medicine use on injection use: a quasi-experimental study”. *PLoS One* 9.10 (2014), e109594.
- [417] L. Yang et al. “Public reporting improves antibiotic prescribing for upper respiratory tract infections in primary care: a matched-pair cluster-randomized trial in China”. *Health Res Policy Syst* 12 (2014), p. 61.
- [418] C. Liu, X. Zhang, and J. Wan. “Public reporting influences antibiotic and injection prescription in primary care: a segmented regression analysis”. *J Eval Clin Pract* 21.4 (2015), pp. 597–603.
- [419] Y. Tang, C. Liu, and X. Zhang. “Public reporting as a prescriptions quality improvement measure in primary care settings in China: variations in effects associated with diagnoses”. *Sci Rep* 6 (2016), p. 39361.

Appendix A

Chapter 2 - Characteristics of included studies

Table A.1: Jang 2011 - Study characteristics

Methods	
Design:	ITS
Country:	South Korea
Care setting:	Paediatric and neonatal intensive care units
Duration:	2003 to 2007
Dataset:	HIRA National Quality Improvement database
Total participants:	Not stated; approximately 3,000,000 live births would have been included between January 2003 and May 2007 according to data from another study ⁴¹⁵
Unit of analysis:	Individual hospitals
Data analysis:	Time series ARIMA analysis
Participants	
Inclusion criteria:	All hospitals performing 100 or more deliveries per year
Hospital types:	Tertiary care hospitals (3.6%), general hospitals (13.1%), hospital (13.1%), clinic (35.4%)
Hospital regions:	Capital city (4.9%), metropolis (31.7%), satellite city (22.5%), city (24.5%), rural (16.3%)
Hospital ownership:	Public (3.2%), non-public (96.8%)
Hospital deliveries:	>700 (4.3%), 201 to 700 (26.4%), <200 (69.4%)
Interventions	
Intervention:	Repeated public release of information (online, press releases) on hospital caesarean rates
Duration:	Four distinct interventions (September 2005, January 2006, September 2006, January 2007)
Deliverer:	HIRA, South Korea
Funding:	HIRA, South Korea
Outcomes	
Main outcome:	Risk-adjusted institutional caesarean section rates

Table A.2: Flett 2015 - Study characteristics

Methods	
Design:	ITS (with non-intervention control hospitals)
Country:	USA
Care setting:	Paediatric and neonatal intensive care units
Duration:	2004 to 2012
Dataset:	Paediatric Health Information System (PHIS)
Total participants:	21 acute hospitals
Unit of analysis:	Individual hospitals; accounted for clustering within hospitals
Data analysis:	Generalised linear mixed-effects models with auto-correlated residuals
Participants	
Inclusion criteria:	Children's hospitals in US states that submitted data to the PHIS
Hospitals:	17 hospitals in 9 states that introduced public reporting of CLABSI rates, and 4 hospitals in 4 states without public reporting. Minimal data provided about the number or characteristics of individual patients treated within these hospitals
Interventions	
Intervention:	State-based mandatory public reporting of healthcare-associated infections
Duration:	Public reporting introduced between July 2005 and April 2010 (depending on state) and lagged behind legislation by 6 to 27 months
Deliverer:	Individual state legislatures
Funding:	Unclear
Outcomes	
Main outcomes:	Blood cultures per 1000 patient days; number of antibiotic days per 1000 patient days

Table A.3: DeVore 2016: Study characteristics

Methods	
Design:	ITS
Country:	USA
Care setting:	Acute hospitals
Duration:	1st July 2006 to 30th June 2012
Dataset:	5% nationally representative sample of Medicare beneficiaries
Total participants:	315,092 hospitalisations
Unit of analysis:	Individual hospitalisations; accounted for clustering within hospitals
Data analysis:	Regression models
Participants	
Inclusion criteria:	All patients enrolled with Medicare, i.e. predominantly those aged 65 years or older
Hospitals:	More than 4,100 hospitals in the USA
Participants:	315,092: 37,829 acute myocardial infarction (16.0%), 100,189 heart failure (42.5%), 17,907 diabetes (7.6%), 80,091 chronic obstructive pulmonary disease (33.9%)
Interventions	
Intervention:	Public reporting of risk-standardised hospital readmission rates on a public website, Hospital Compare
Duration:	June 2009 until the study end date in 2012
Deliverer:	CMS, US Department of Health and Human Services
Funding:	CMS (federal government funding)
Outcomes	
Main outcome:	30-day post-discharge re-admission to hospital
Secondary outcomes:	30-day post-discharge outpatients visits; 30-day post-discharge emergency department visits; 30-day post discharge observation stays without readmission

Table A.4: Joynt 2016 - Study characteristics

Methods	
Design:	ITS
Country:	USA
Care setting:	Acute hospitals
Duration:	January 2005 to November 2012
Dataset:	Medicare inpatient files
Total participants:	20,707,266
Unit of analysis:	Individual patients; accounted for clustering within hospitals
Data analysis:	Multivariable logistic regression
Participants	
Inclusion criteria:	All Medicare fee-for-service enrollees hospitalised with any of the 15 most common non-surgical discharge diagnoses. Medicare is predominantly composed of patients aged 65 years or older
Hospitals:	3,970 hospitals
Hospital types:	6.8% major teaching hospital, 18.3% minor teaching hospital, 74.9% non-teaching)
Participant characteristics:	Mean age 79 years, 41% male
Interventions	
Intervention:	Public release of hospital performance data (using 30-day mortality), published on a publicly accessible website. The intervention was the addition of 30-day mortality to publicly accessible hospital performance data in 2008. In the pre-intervention period, hospital performance data were available in the same format, but was limited to process metrics
Duration:	4 years
Deliverer:	Hospital Compare, which is maintained by the CMS
Funding:	CMS
Outcomes	
Main outcome:	Risk-adjusted 30-day mortality

Table A.5: Liu 2017 - Study characteristics

Methods	
Design:	ITS (with non-intervention control hospitals)
Country:	USA
Care setting:	Adult ICU
Duration:	2006 to 2012
Dataset:	Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN)
Total participants:	244 acute hospitals
Unit of analysis:	CLABSI; accounted for clustering within hospitals
Data analysis:	Multi-variable regression, using a DID approach from hospitals in states that did not introduce mandatory reporting
Participants	
Inclusion criteria:	All non-Veterans Affairs (VA) acute hospitals enrolled in the NHSN were eligible to participate
Hospitals:	244 hospitals with 475 ICU
Hospital teaching status:	control (469 ICU days, 59.1%), intervention (844, 76.2%)
ICU bed size:	Control (45 ICU days, 5.7%), intervention (68, 6.1%)
Patient days per year:	Control (mean 1384.1, standard deviation (SD) 2152.0), intervention (1855.4, SD 1447.6)
Participant characteristics:	No substantial case mix data provided
Interventions	
Intervention:	Mandatory public reporting of healthcare-associated infections
Duration:	Variable, depending on the state being studied
Deliverer:	Individual state legislatures
Funding:	Unclear
Outcomes	
Main outcome:	CLABSIs per 1000 patient days

Table A.6: Tu 2009 - Study characteristics

Methods	
Design:	cRT
Country:	Canada (Ontario)
Care setting:	Acute hospitals
Duration:	1 April 2004 to 31 March 2005
Dataset:	Prospective chart review by research nurses, and study linkage to the Ontario Registered Persons Vital Statistics Database for mortality outcomes
Total participants:	82 hospital organisations
Unit of allocation:	Hospital organisations
Unit of analysis:	Individual patients; accounted for clustering of patients within hospitals
Data analysis:	Multivariable logistic regression
Sample size calculation:	Quote: “The study had 84% power to detect 5% absolute difference on the composite quality indicators”
Participants	
Inclusion criteria:	Acute hospitals participating in Ontario, Canada that were identified from the Canadian Institute for Health Information hospital discharge administrative database 1999 to 2001 and treated more than 15 patients with AMI annually
Participants:	86 hospital corporations
Hospital characteristics:	12% teaching hospitals in the intervention group versus 10% in the control group; 74% community hospitals in the intervention versus 79% in the control group
AMI characteristics:	Median age 69 (IQR 57 to 78) both groups; female 35.4% versus 36.7%
CHF characteristics:	Median age 77 (IQR 70 to 84) versus 77 (69-84); female 51.3% versus 49.2%
Interventions	
Intervention:	Report cards with baseline performance data publicly released online and at a press conference
Control:	Report cards publicly released after data had been collected, i.e. a delayed release of data for the control group
Duration:	January to 1 April 2004
Deliverer:	The Canadian Cardiovascular Outcomes Research Team, which is a national team of cardiovascular outcomes researchers from across Canada
Funding:	Canadian Institutes of Health Research team grant in cardiovascular outcomes research
Outcome	
Main outcomes:	Composite AMI indicators; composite CHF indicators
Secondary outcomes:	12 AMI process-of-care indicators; 6 CHF process-of-care indicators; 30-day and 1-year mortality for patients in 5 following sub-groups (AMI, ST-elevation myocardial infarction (STEMI), non-ST-elevation myocardial infarction (NSTEMI), CHF, and CHF with left ventricular dysfunction (LVD))

Table A.7: Ikkersheim 2013 - Study characteristics

Methods	
Design:	cRT
Country:	The Netherlands (Eindhoven)
Care setting:	Primary care
Duration:	2009 to 2010
Dataset:	Prospective data collection by GPs
Total participants:	26 GPs (2:1 randomisation to intervention)
Unit of allocation:	Individual GPs; accounted for clustering of GPs within practices
Data analysis:	Multivariable logistic regression using a DID approach
Sample size calculation:	Not reported; statistical significance was assessed at the 0.05 level
Participants	
Inclusion criteria:	All GPs within the Eindhoven region
Participants:	26 GPs, with 17 in the intervention group and 9 in the control group
Participant characteristics:	male 41% (intervention) versus 44% (control), urban 35% (intervention) versus 33% (control)
Interventions	
Intervention:	Report cards sent by post to GPs that included a variety of quality indicators that depended on the specific condition (breast cancer, cataract surgery, hip or knee replacement)
Duration:	No details provided
Deliverer:	Research team
Funding:	The Netherlands Organisation for Health Research and Development (ZonMw), the Dutch organisation for health research and development
Outcome	
Main outcome:	Choice of hospital when making patient referrals

Table A.8: Zhang 2016 - Study characteristics

Methods	
Design:	cRT
Country:	China (Hubei Province)
Care setting:	Primary care
Duration:	2013 to 2014
Dataset:	Data collected from patient electronic health records
Total participants:	748,632 outpatient prescriptions from 20 primary healthcare institutions
Unit of allocation:	Primary healthcare institutions (paired and matched for similar characteristics)
Unit of analysis:	Individual prescriptions; accounted for clustering of prescriptions by individual prescribers
Data analysis:	Multivariable logistic regression using a DID approach
Sample size calculation:	Not reported; statistical significance was assessed at the 0.05 level
Participants	
Inclusion criteria:	Primary care institutions selected from within Qian Jiang City
Participants:	20 providers, 10 of which were in the intervention group, and 10 in the control group
Institution characteristics:	60 beds in the intervention group versus 66 in the control group; 28 versus 26 doctors, 50,199 versus 49,108 annual outpatient visits
Patient characteristics:	Mean age 37.5 years, 49.5% male
Interventions	
Intervention:	Public display of prescription information (percentage of prescriptions requiring antibiotics, percentage requiring injections, and average patient expenditure) on outpatient department bulletin boards in participating institutions
Duration:	1 October 2013 to 31 August 2014
Deliverer:	Outpatient departments of participating institutions
Funding:	National Natural Science Foundation of China
Outcome	
Main outcomes:	Percentage of prescriptions requiring antibiotics; percentage of prescriptions requiring combined antibiotics; percentage of prescriptions requiring injections; average expenditure per prescription
Other	
Note:	Zhang 2016 represents a single study that was reported in five articles ^{170,416–419} that individually satisfied our inclusion criteria. However, the senior author confirmed that these represented multiple analyses of a single cluster-RT ¹⁷⁰ . The cRT was therefore presented (as the original study design and higher level of evidence), rather than the designs (e.g. CBA and ITS) that were presented in the other articles

Table A.9: Rinke 2015 - Study characteristics

Methods	
Design:	CBA
Country:	USA
Care setting:	Acute hospitals
Duration:	2000 to 2009
Dataset:	HCUP Kids' Inpatient Database
Total participants:	4,705,857 paediatric hospital discharges
Unit of allocation:	Paediatric hospital discharges)
Unit of analysis:	Paediatric hospital discharges; accounted for clustering of discharges within hospitals and states
Data analysis:	Multivariable logistic regression
Sample size calculation:	Not reported; statistical significance was assessed at the 0.05 level
Participants	
Inclusion criteria:	All paediatric hospital discharges eligible for paediatric safety indicator (PDI12) (i.e. length of stay 2 or more days) in a state that was categorised as 'never reporters' (18 states), '2006 reporters' (2 states), or '2009 reporters' (7 states)
Hospitals:	3,207; 2,066 of which were never reporters, 135 were 2006 reporters, and 1,006 were 2009 reporters
Hospital teaching status:	Never reporters (52%), 2006 reporters (55%), 2009 reporters (58%)
Patients:	4,705,857 discharges, 2,580,621 of which were from never reporters, 179,322 from 2006 reporters, and 1,945,914 from '2009 reporters'
Patient age:	Never reporters (mean 3.5, SD 5.5), 2006 reporters (4.4, SD 6.0), 2009 reporters (3.6, SD 5.6)
Patient sex:	Never reporters (male 54%, female 46%), 2006 reporters (54% male, 46% female), 2009 reporters (55% male, 45% female)
Interventions	
Intervention:	Mandatory public reporting of healthcare-associated infections) on outpatient department bulletin boards in participating institutions
Control:	No mandatory reporting of healthcare-associated infections
Duration:	Mandatory CLABSI reporting introduced in 2006 or 2009
Deliverer:	Individual hospitals, as mandated by state legislatures
Funding:	Unclear
Outcome	
Main outcome:	PDI12, which was defined by the AHRQ as "selected infections due to medical care", and determined using discharge International Statistical Classification of Diseases 9 th Revision Clinical Modification (ICD-9-CM) codes

Appendix B

Chapter 2 - Risk of bias assessments

Table B.1: DeVore 2016 - Risk of bias

Bias	Risk	Explanation
Incomplete outcome data (attrition bias)	Low	Outcome data available for all patients
Selective reporting (reporting bias)	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Other bias	Low	No additional biases identified
Intervention is independent of other changes?	Unclear	Not stated whether there were other confounding events that might have changed performance over time.
Shape of intervention effect pre-specified?	High	Shape of intervention not pre-specified
Knowledge of the interventions adequately prevented during the study?	Low	Individuals would not have been aware of the study, as this was performed using routinely collected administrative data
Intervention unlikely to bias data collection?	Low	Routinely collected administrative data that were collected independently of the individuals at whom the public release of performance data were directed

Table B.2: Flett 2015 - Risk of bias

Bias	Risk	Explanation
Incomplete outcome data (attrition bias)	Low	Outcome data for all included hospitals, except for one that was excluded because of excessive missing data
Selective reporting (reporting bias)	Low	All outcomes and results outlined in the Method section are reported in tables, text, or both
Other bias	Unclear	No additional biases identified
Intervention is independent of other changes?	Low	Not stated whether there were other confounding events that might have changed performance over time. However, this was unlikely overall, as each state implemented mandatory reporting at different stages and using different regulatory mechanisms
Shape of intervention effect pre-specified?	High	Shape of intervention effect not pre-specified
Knowledge of the interventions adequately prevented during the study?	Low	Individuals would not have been aware of the study as this was performed retrospectively, using a clinical registry
Intervention unlikely to bias data collection?	Low	Routinely collected clinical data, so data collection was unlikely to be biased by the intervention

Table B.3: Jang 2011 - Risk of bias

Bias	Risk	Explanation
Incomplete outcome data (attrition bias)	Low	Outcome data available for all patients
Selective reporting (reporting bias)	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Other bias	Low	No additional biases identified
Intervention is independent of other changes?	Unclear	Not stated whether there were other confounding events that might have changed performance over time
Shape of intervention effect pre-specified?	Low	The authors pre-specified that repeated releases would decrease and that caesarean section rates of institutions with higher caesarean section rates in the period before repeated releases would decrease further after repeated releases than those with lower starting rates
Knowledge of the interventions adequately prevented during the study?	Low	Did not state explicitly that those responsible for data collection were informed that the publication of performance data was part of a study
Intervention unlikely to bias data collection?	Low	Routinely collected administrative data that were collected independently of the individuals at whom the public release of performance data were directed

Table B.4: Joynt 2016 - Risk of bias

Bias	Risk	Explanation
Incomplete outcome data (attrition bias)	Low	Outcome data available for all patients
Selective reporting (reporting bias)	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Other bias	Low	No additional biases identified
Intervention is independent of other changes?	Low	Not stated whether there were other confounding events that might have changed performance over time. However, this was unlikely, given that this study identified few changes in outcome after the intervention
Shape of intervention effect pre-specified?	High	Shape of intervention not pre-specified
Knowledge of the interventions adequately prevented during the study?	Low	Individuals would not have been aware of the study, as this was performed using routinely collected administrative data
Intervention unlikely to bias data collection?	Low	Routinely collected administrative data that were collected independently of the individuals at whom the public release of performance data were directed

Table B.5: Liu 2017 - Risk of bias

Bias	Risk	Explanation
Incomplete outcome data (attrition bias)	Low	Outcome data available for all patients
Selective reporting (reporting bias)	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Other bias	Low	No additional biases identified
Intervention is independent of other changes?	Low	Not stated whether there were other confounding events that might have changed performance over time. However, this was unlikely overall, as each state implemented mandatory reporting at different stages, and using different regulatory mechanisms
Shape of intervention effect pre-specified?	High	Shape of intervention not pre-specified
Knowledge of the interventions adequately prevented during the study?	Low	Individuals would not have been aware of the study, as this was performed using routinely collected administrative data
Intervention unlikely to bias data collection?	Low	Routinely collected administrative data that were collected independently of the individuals at whom the public release of performance data were directed

Table B.6: Ikkersheim 2013 - Risk of bias

Bias	Risk	Explanation
Random sequence generation (selection bias)	Unclear	Method of randomisation unclear
Allocation concealment (selection bias)	Unclear	No statement about allocation concealment
Adequate blinding of participants, personnel and outcome assessors?	High	No blinding of participants or personnel; the outcomes measured GP behaviour (i.e. referral patterns); individual GPs were not blinded to the group allocation
Incomplete outcome data (attrition bias)?	Low	Data from all participating GPs included
Selective reporting (reporting bias)?	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Baseline characteristics similar?	Low	Some baseline characteristics described (health professional sex and urban location), which suggested that the groups were balanced
Baseline outcomes similar?	Low	Baseline outcomes varied between hospitals, although multivariable logistic regression was used to adjust for baseline differences
Protection against contamination	Low	No specific safeguards against contamination, although it was unlikely that GPs shared hospital report cards amongst themselves when they knew these were the subject of a trial
Other bias	Low	No additional biases identified

Table B.7: Tu 2009 - Risk of bias

Bias	Risk	Explanation
Random sequence generation (selection bias)	Low	Method of randomisation not explicitly stated, but this was undertaken by a dedicated study statistician who used a stratified randomisation process
Allocation concealment (selection bias)	Low	Quote: “This random assignment was stratified by type of hospital and performed by a study statistician”
Adequate blinding of participants, personnel and outcome assessors?	High	Quote: “It was not possible to blind the hospitals to their status”
Incomplete outcome data (attrition bias)?	High	One hospital withdrew from the baseline phase after randomisation, and 4 withdrew from the follow-up phase, all due to resource constraints. No intention-to-treat analysis was performed. Additional exclusions of patients were not reported
Selective reporting (reporting bias)?	Low	A protocol was registered in advance of randomisation and all outcomes were reported in the final report, which also included a new outcome (all-cause mortality)
Baseline characteristics similar?	Low	Baseline characteristics of patients and hospitals between the intervention and control groups were similar
Baseline outcomes similar?	Low	Baseline outcomes presented and varied between hospitals, although results were presented as absolute change, and so accounted for baseline differences
Protection against contamination	High	Quote: “One unanticipated observation was that several hospitals in the delayed feedback group reported that they also initiated some quality improvement activities after becoming aware of the publicly released early feedback report cards, before receiving their own hospital-specific results”
Other bias	Low	No additional biases identified

Table B.8: Zhang 2016 - Risk of bias

Bias		Risk	Explanation
Random generation bias)	sequence (selection bias)	Low	Quote: “We flipped a coin to randomly assign one (primary care institution) into the intervention group and another into the control group”
Allocation concealment (selection bias)		High	Healthcare providers could not be blinded to the allocation
Adequate blinding of participants, personnel and outcome assessors?		High	It was not possible to blind personnel, who must have been aware of the group to which their primary care institution had been allocated
Incomplete data (attrition bias)?	outcome	Low	Injection prescribing data retrieved from a comprehensive administrative database
Selective reporting (reporting bias)?	reporting	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both. Although a protocol for the cRT was published, this appeared eighteen months after the trial reports stated that the intervention began
Baseline characteristics similar?	characteris-	Low	Some baseline characteristics described (e.g. age and sex), which suggested that the groups were balanced
Baseline outcomes similar?	outcomes	Low	Baseline outcomes presented and varied between hospitals. However, the hospitals were paired according to characteristics, and the results analysed using a DID approach and regression models to account for residual baseline differences
Protection against contamination	against	Unclear	No statement as to whether or not other events might have influenced performance over time
Other bias		Low	No additional biases identified

Table B.9: Rinke 2015 - Risk of bias

Bias	Risk	Explanation
Random sequence generation (selection bias)	High	CBA study, so did not use random sequence allocation
Allocation concealment (selection bias)	High	No allocation concealment as hospitals would have known whether or not their state mandated public reporting
Adequate blinding of participants, personnel and outcome assessors?	High	No blinding of participants, personnel, or outcome assessors, as all parties would have known whether or not their state mandated public reporting
Incomplete outcome data (attrition bias)?	Low	Outcome data for all included hospitals
Selective reporting (reporting bias)?	Low	All outcomes and results outlined in the Method section were reported in tables, text, or both
Baseline characteristics similar?	Low	Baseline characteristics differed, but were sufficiently similar to undertake the study using appropriate analyses
Baseline outcomes similar?	Low	Baseline outcomes differed, but were sufficiently similar to undertake the study using appropriate analyses (2.4 PDI12 per 1000 discharges in the never-reporting states, 2.6 in the 2006 reporter states, and 3.0 in the 2009 reporter states)
Protection against contamination	High	Unable to protect against contamination, as hospitals in states without mandatory reporting might have been influenced by states in which these laws were introduced
Other bias	Low	No additional biases identified

Appendix C

Chapter 2 - Certainty of the evidence

Table C.1: Certainty of the evidence

Studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Overall certainty
Outcome: Changes in healthcare decisions taken by healthcare providers (professionals and organisations)							
4	2 RT 2 ITS	Initial: 3 Final: 3	-1*	0	0	0	Low
Studies: Flett, Ikkersheim, Jang, Zhang * -1 for inconsistency as Zhang showed a change in behaviour, which was not consistently observed throughout the other 3 studies							
Outcome: Changes in provider performance							
1	1 RT		-2*	0	0	0	Low
Studies: Tu * -2 for risk of bias as there was attrition of participating hospitals, evidence of contamination of the intervention to control hospitals, and blinding was not possible given the nature of the intervention							
Outcome: Changes in patient outcome							
5	1 RT 3 ITS 1 CBA	0	-2*	0	0	0	Low
Studies: DeVore, Joynt, Liu, Rinke, Tu * -2 for inconsistency as there was marked disagreement between studies with 2 showing improvement in patient outcome (Tu, Liu) and 3 showing no such improvement (DeVore, Joynt, Rinke)							

No studies for the outcomes "Changes in healthcare decisions taken by healthcare purchasers", "Impact on equity", "Changes in staff morale", or "Adverse effects".

Appendix D

Chapter 3 - Supplementary plots

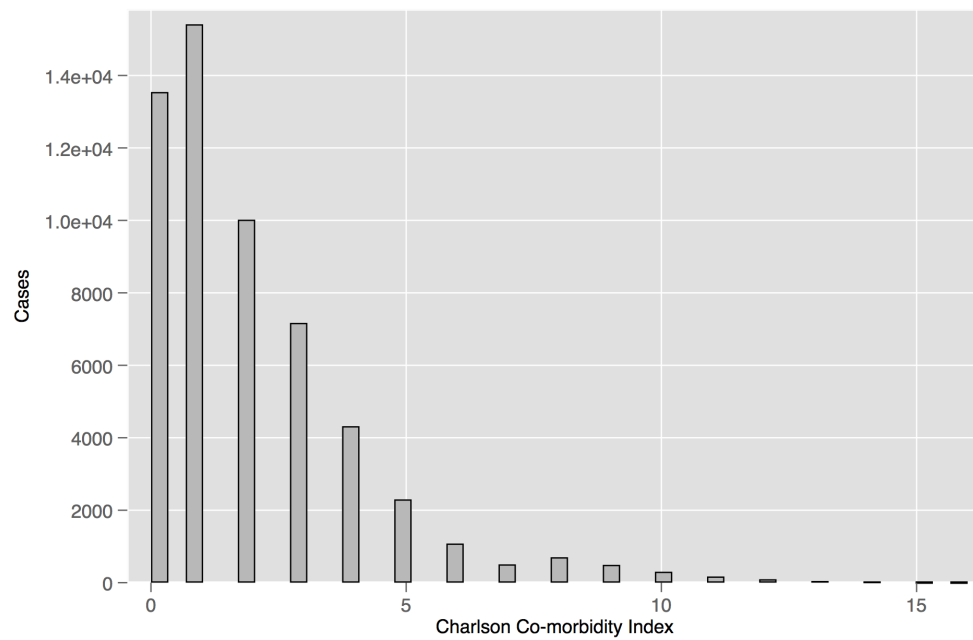


Figure D.1: Distribution of the Charlson Co-morbidity Index in the NHFD.

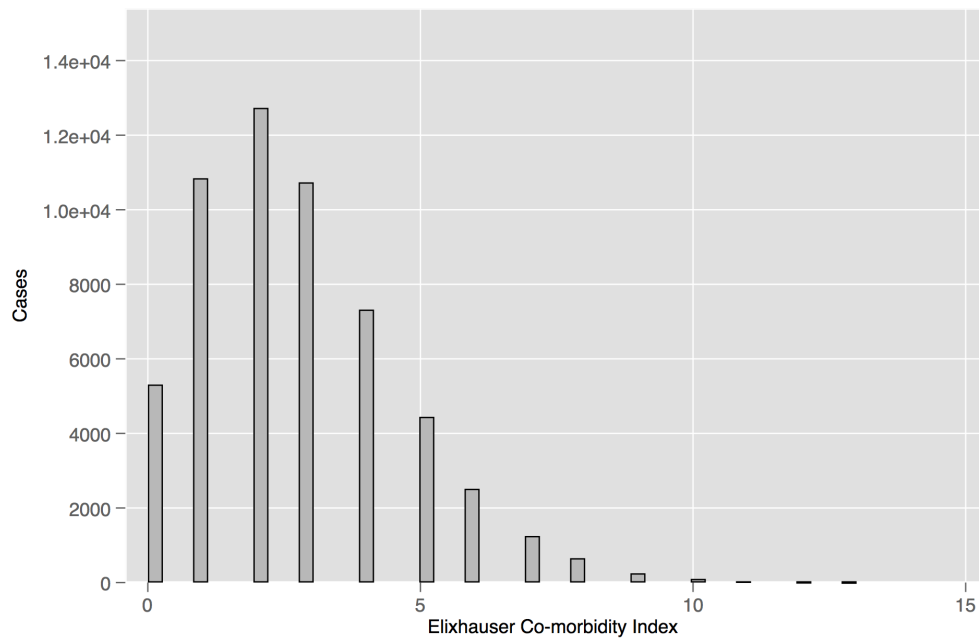


Figure D.2: Distribution of the Elixhauser Co-morbidity Index in the NHFD.

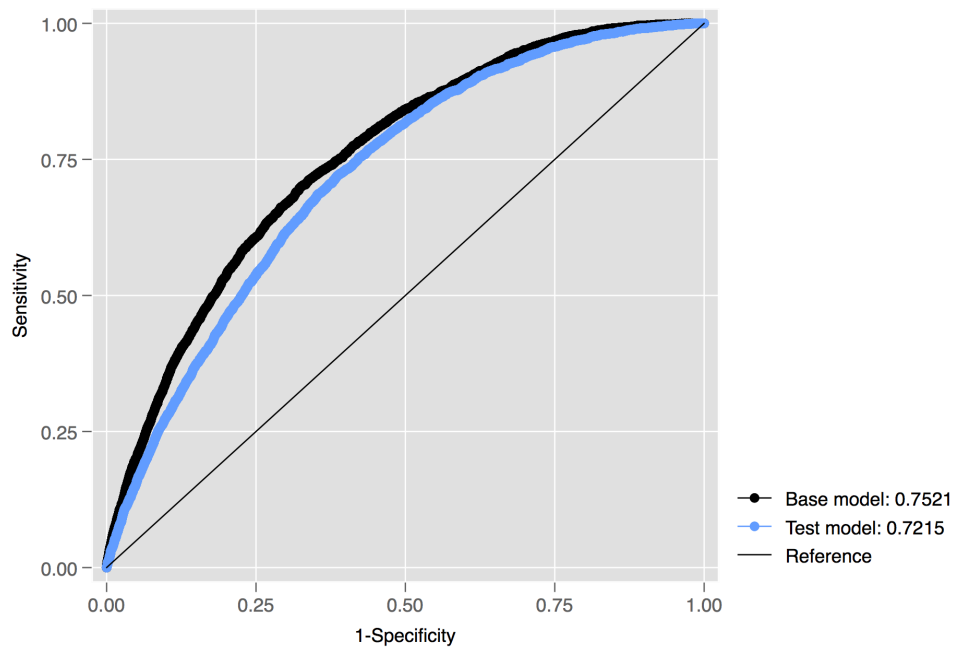


Figure D.3: Receiver Operating Characteristic curves showing the NHFD model versus the model with ASA substituted for Charlson Co-morbidity Index.

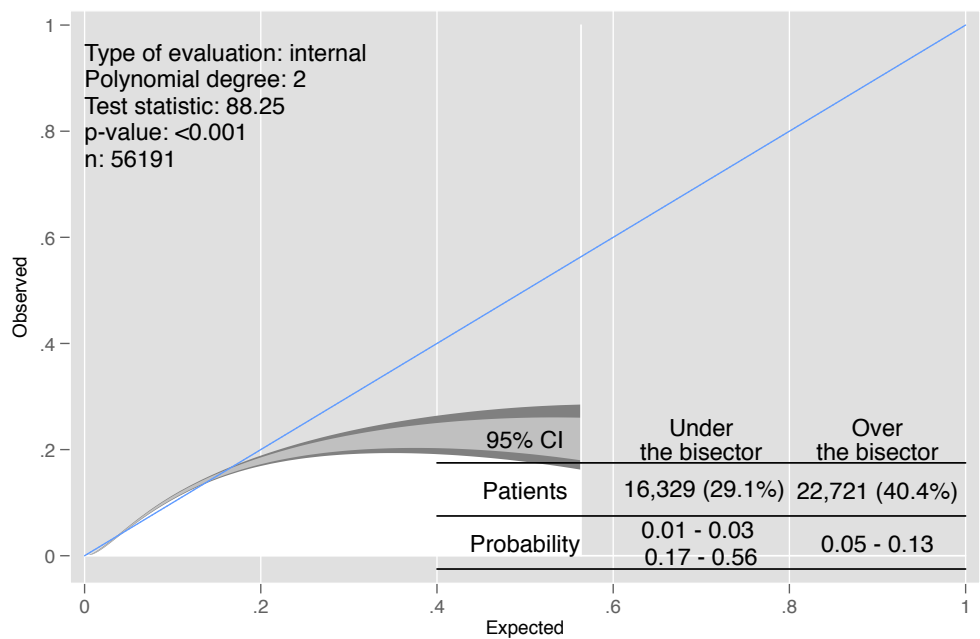


Figure D.4: Calibration belt plot for the existing NHFD model with ASA substituted for Charlson Co-morbidity Index.

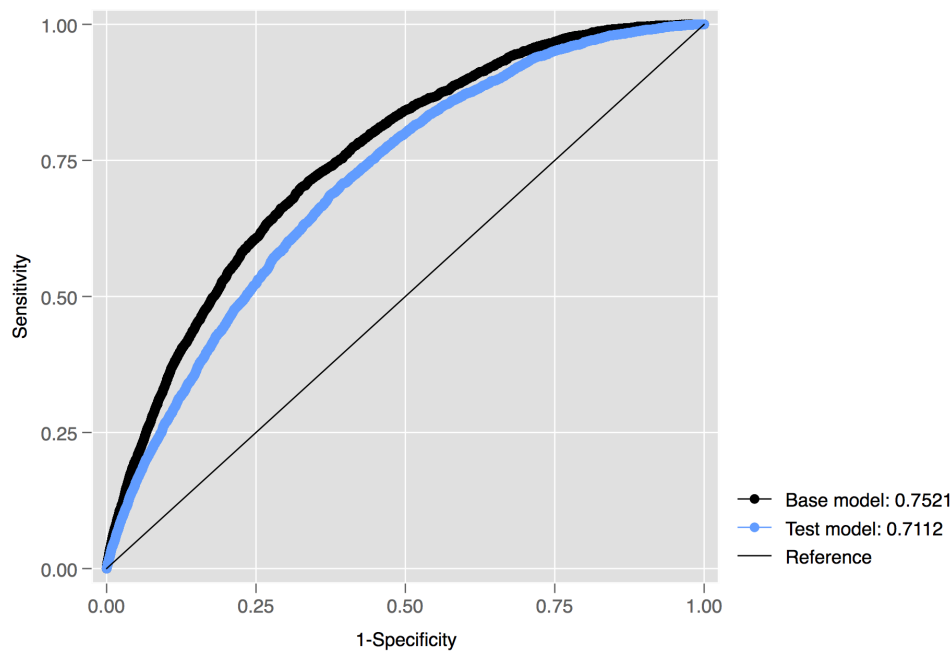


Figure D.5: Receiver Operating Characteristic curves showing the NHFD model versus the model with ASA substituted for Elixhauser Co-morbidity Index.

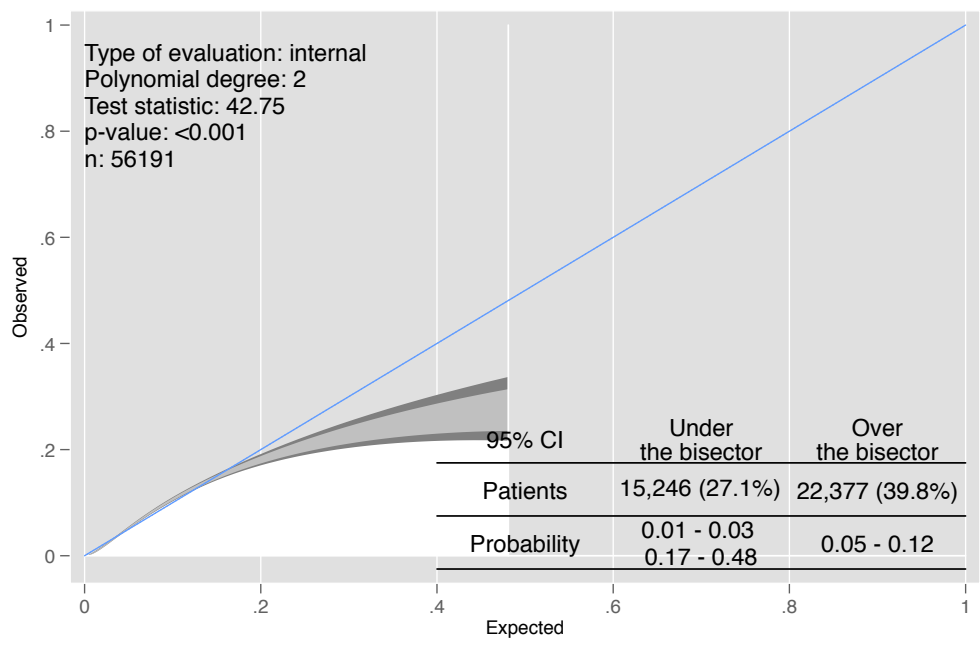


Figure D.6: Calibration belt plot for the existing NHFD model with ASA substituted for Elixhauser Co-morbidity Index.

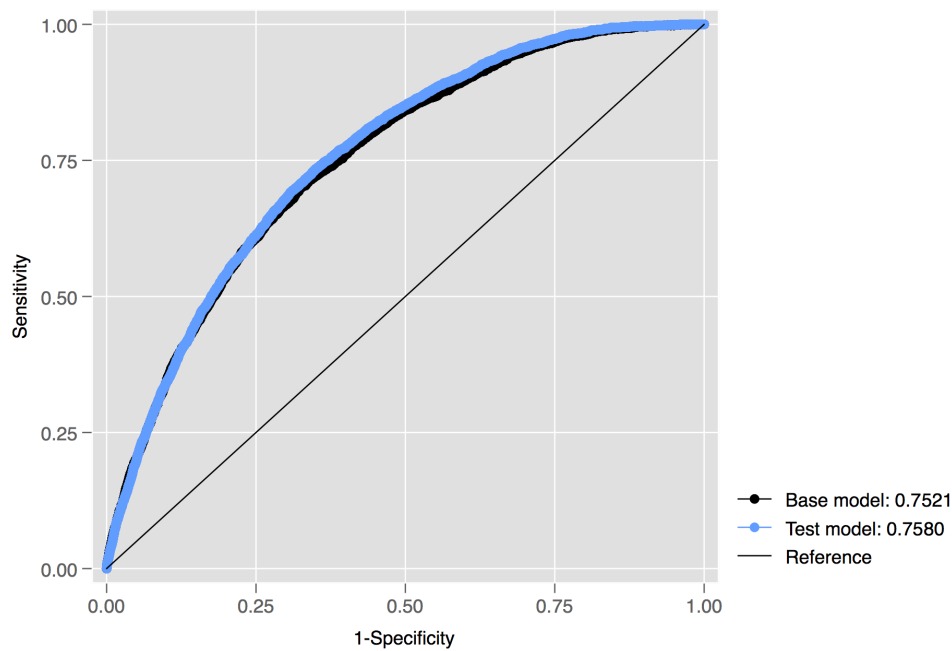


Figure D.7: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with the Charlson Co-morbidity Index.

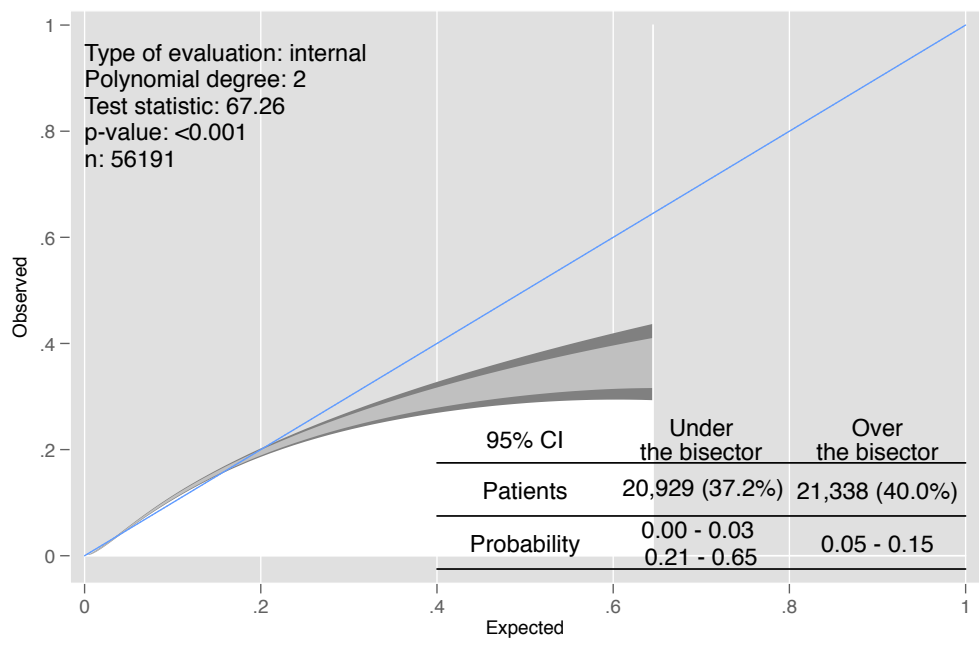


Figure D.8: Calibration belt plot for the existing NHFD model supplemented by the Charlson Co-morbidity Index.

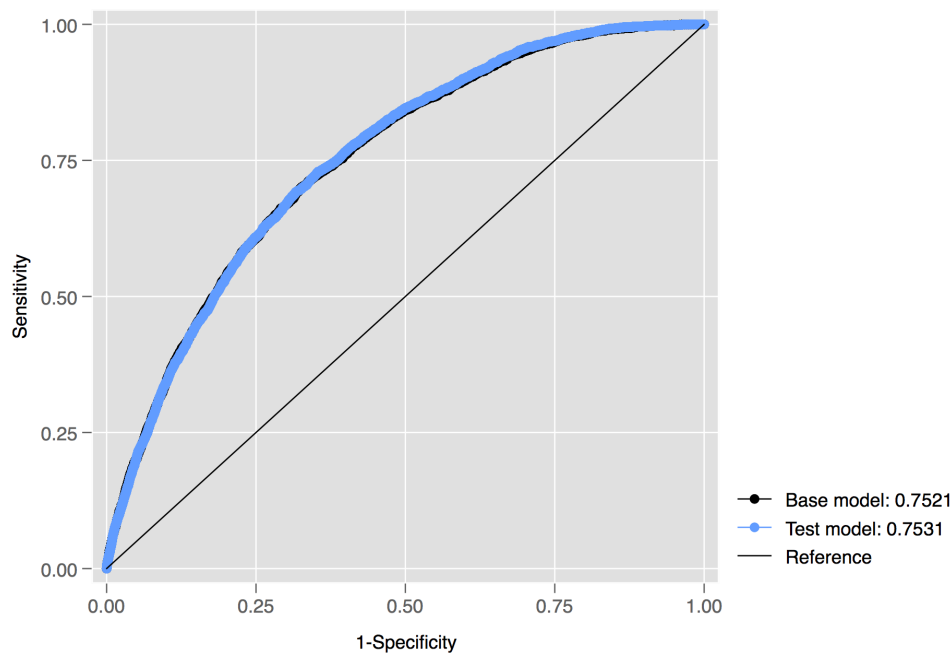


Figure D.9: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with the Elixhauser Co-morbidity Index.

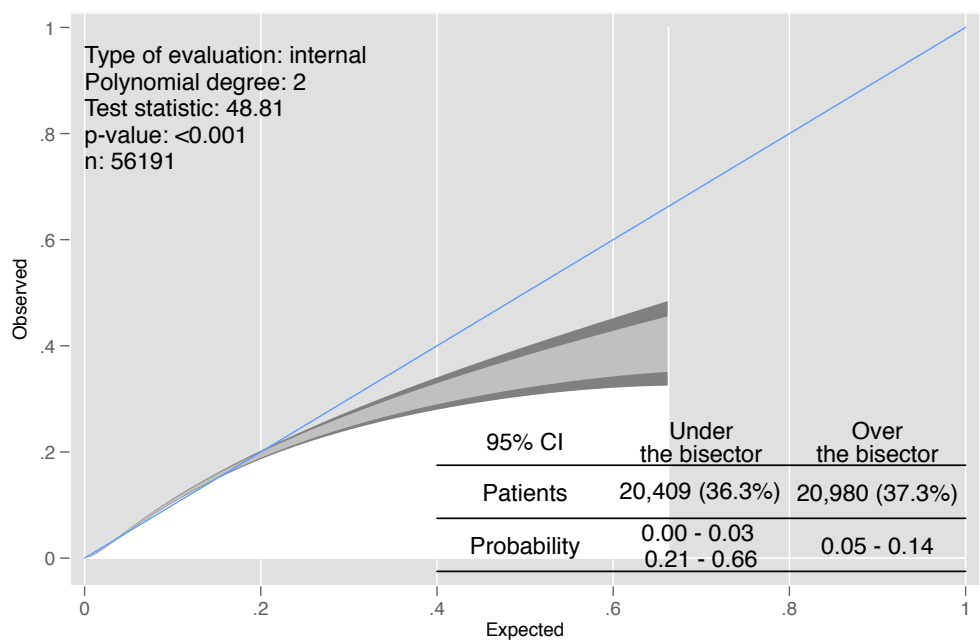


Figure D.10: Calibration belt plot for the existing NHFD model supplemented by the Elixhauser Co-morbidity Index.

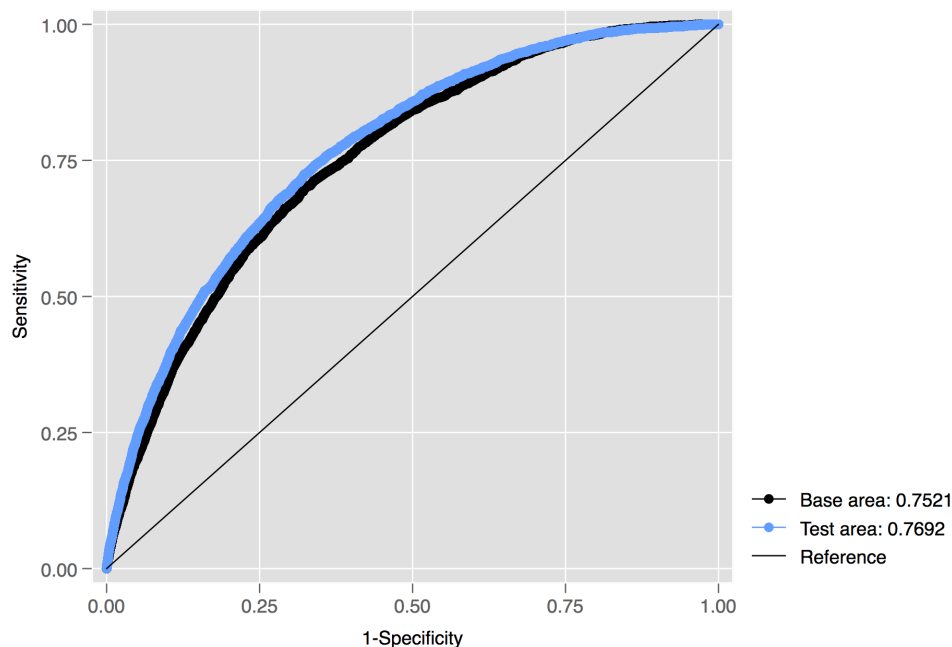


Figure D.11: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with Elixhauser Co-morbidity Index and Hospital Frailty Risk Score

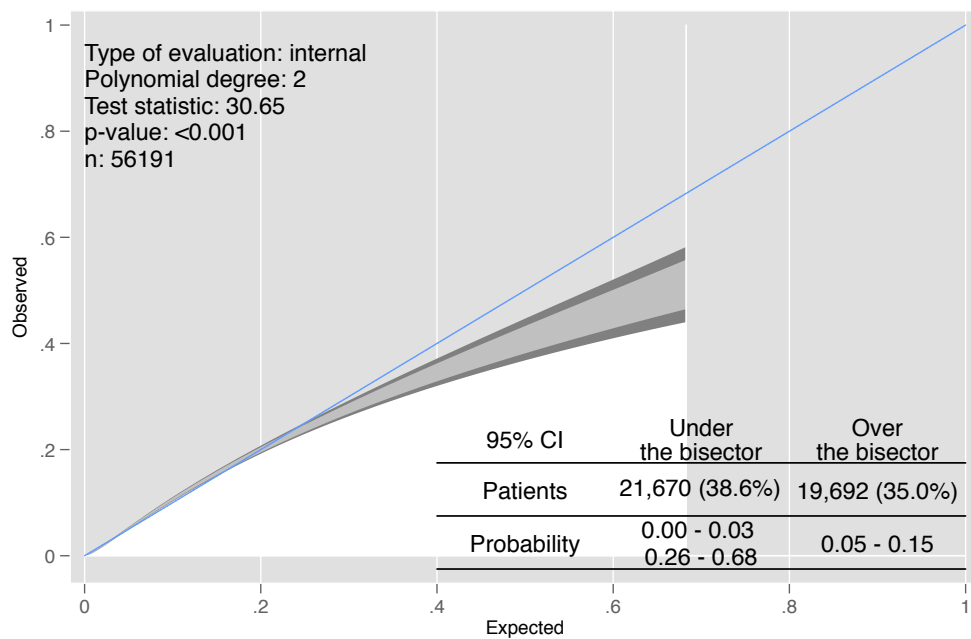


Figure D.12: Calibration belt plot for the existing NHFD model supplemented with the Elixhauser Co-morbidity Index and Hospital Frailty Risk Score.

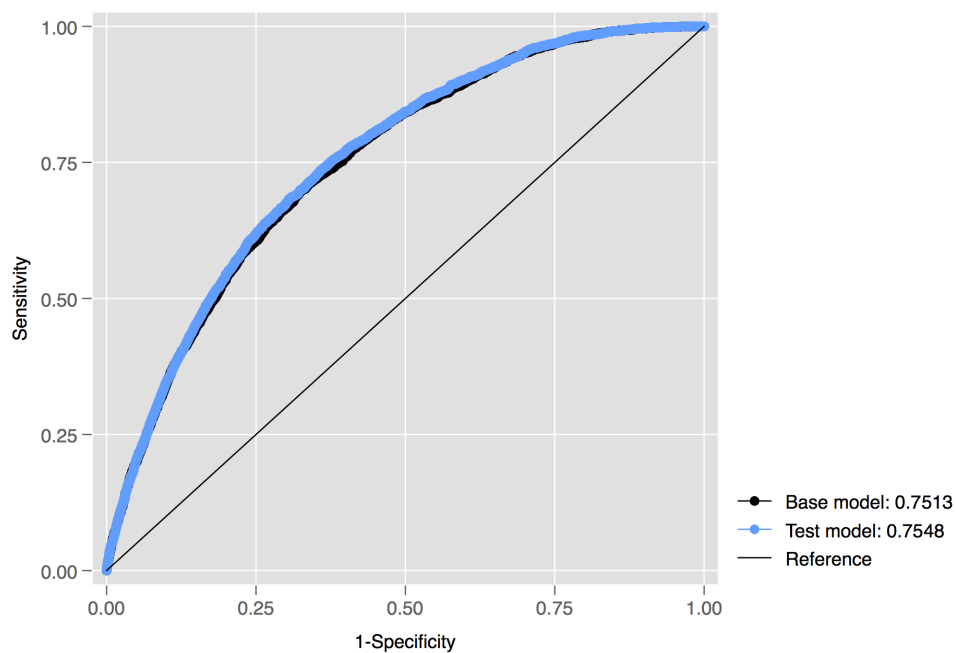


Figure D.13: Receiver Operating Characteristic curves showing the NHFD model versus the model supplemented with the Abbreviated Mental Test Score

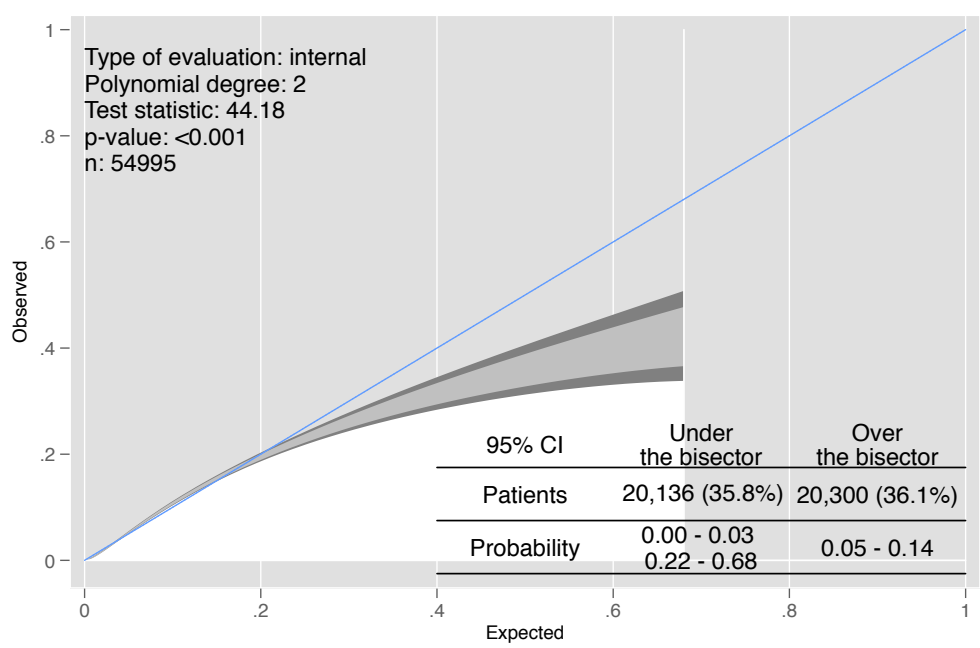


Figure D.14: Calibration belt plot for the existing NHFD model supplemented with the Abbreviated Mental Test Score.

Appendix E

Chapter 4 - Best Practice Tariff criteria

In order to achieve the BPT, the following criteria must be satisfied:

- Age ≥ 60 at admission AND;
- Valid NHS number AND;
- Timely operation (time to surgery <36 hours from arrival) AND;
- ALL of the following:

Admissions from April 2010

- Admitted under named orthopaedic surgeon
- Admitted under named geriatrician
- Admitted using a joint assessment protocol
- Timely geriatrician assessment (<72 hours)
- Rehabilitation assessment
- Specialist falls assessment
- Bone therapy assessment

Admissions from April 2012

- Admitted under named orthopaedic surgeon
- Admitted under named geriatrician
- Admitted using a joint assessment protocol
- Timely geriatrician assessment (<72 hours)
- Rehabilitation assessment
- Specialist falls assessment
- Bone therapy assessment
- Documentation of pre-operative AMTS
- Documentation of post-operative AMTS

Admissions from April 2017

- Timely geriatrician assessment (<72 hours)
- Specialist falls assessment
- Bone therapy assessment
- Documentation of pre-op AMTS
- Delirium assessment
- Physiotherapist assessment
- Nutrition assessment

Appendix F

Chapter 5 - Mobility score reconciliation

Version 7 mobility variables

WalkInside Walking ability indoors

- 0 Regularly walked without aids
- 1 Regularly walked with one aid
- 2 Regularly walked with two aids or frame
- 3 Wheelchair or bedbound
- 4 Unknown
- Missing

WalkOutside Walking ability outside

- 0 Regularly walked without aids
- 1 Regularly walked with one aid
- 2 Regularly walked with two aids or frame

- 3 Wheelchair or bedbound
- 4 Never goes outdoors
- 5 Electric buggy
- 6 Unknown
- Missing

AccompInside Accompanied to walk inside

- 0 No
- 1 Yes
- 2 Wheelchair or bedbound
- 3 Unknown
- Missing

AccompOutside Accompanied to walk outside

- 0 No
- 1 Yes
- 2 Wheelchair or bedbound
- 3 Unknown
- Missing

Version 8 mobility variables

Mobility Pre-fracture walking ability

- 1 Freely mobile without aids
- 2 Mobile outdoors with one aid
- 3 Mobile outdoors with two aids or frame
- 4 Some indoor mobility but never goes outside without help
- 5 Unknown
- 6 No functional mobility

Re-coding of version 7**Mobility** Pre-fracture walking ability

- 1 Freely mobile without aids
-
- WalkOutside 0
- 2 Mobile outdoors with one aid
-
- WalkOutside 1
- 3 Mobile outdoors with two aids or frame
-
- WalkOutside 2
- 4 Some indoor mobility but never goes outside without help
-
- $((\text{WalkInside} = 0 \text{ or } 1 \text{ or } 2) + (\text{AccompOutside} = 1 \text{ or } 2 \text{ or } 4)) \text{ OR } ((\text{WalkInside} = 0 \text{ or } 1 \text{ or } 2) + (\text{WalkOutdoors} = 3 \text{ or } 4 \text{ or } 5))$

- 5 No functional mobility
-
- WalkOutside 3
- 6 Unknown

Appendix G

Chapter 6 - Diagnostic codes

Table G.1: Charlson Co-morbidity Index scores for ICD-10 diagnostic codes

Category	ICD-10 codes
Score = 1 Myocardial infarction Congestive heart failure Peripheral vascular disease Dementia Cerebrovascular disease Chronic lung disease Connective tissue disease Peptic ulcer Chronic liver disease Diabetes	I21, I22 I50.0 I70-I73 F00-F03 I60-67 J41-47 M05-06; M08; M15-19; M35-36 K25-28 K70.0, K76.0, K76.1 E10-E14, 4th digit X,0,1,9
Score = 2 Hemiplegia Moderate or severe kidney disease Diabetes with end organ damage Tumour Leukaemia Lymphoma	G81 N17-19 E10-E14, 4th digit 2-8; N083 C00-C76; C80; C88; C90.0, C90.2; C96; C97 D00-49 C90.1; C91-95 C81-85
Score = 3 Moderate or severe liver disease	K70-76 without K70.0, K76.0, K76.1
Score = 6 Metastatic cancer AIDS	C77-79 B20-B23

Table G.2: OPCS-4 codes to define prosthetic hip dislocation*

Group	OPCS4 code	Description
A	T84.020	Dislocation of internal right hip prosthesis
	T84.020	Dislocation of internal right hip prosthesis – initial encounter
	AT84.020D	Dislocation of internal right hip prosthesis – subsequent encounter
	T84.020S	Dislocation of internal right hip prosthesis – sequela
	T84.021	Dislocation of internal left hip prosthesis
	T84.021A	Dislocation of internal left hip prosthesis – initial encounter
	T84.021D	Dislocation of internal left hip prosthesis – subsequent encounter
	T84.021S	Dislocation of internal left hip prosthesis - sequela
B	W396	Closed reduction of dislocated total prosthetic replacement of hip joint
	W485	Closed reduction of dislocated prosthetic replacement of head of femur
C	W65	Primary open reduction of traumatic dislocation of joint
	W658	Other specified primary open reduction of traumatic dislocation of joint
	W659	Unspecified primary open reduction of traumatic dislocation of joint
	W66	Primary closed reduction of traumatic dislocation of joint
	W662	Primary closed reduction of traumatic dislocation of joint and skeletal traction NEC
	W668	Other specified primary closed reduction of traumatic dislocation of joint
	W669	Unspecified primary closed reduction of traumatic dislocation of joint
	W67	Secondary reduction of traumatic dislocation of joint
	W676	Remanipulation of traumatic dislocation of joint
	W678	Other specified secondary reduction of traumatic dislocation of joint
	W679	Unspecified secondary reduction of traumatic dislocation of joint
D	Z843	Hip joint
	Z761	Head of femur
	Z756	Acetabulum

*Dislocation defined as (Group A OR Group B) OR (Group C AND Group D)

Table G.3: OPCS-4 codes to define hip revision**

Group	OPCS-4 code	Description
A	W370	Total prosthetic replacement of hip joint using cement, Conversion from previous cemented total prosthetic replacement of hip joint
	W372	Conversion to total prosthetic replacement of hip joint using cement
	W373	Revision of total prosthetic replacement of hip joint using cement
	W374	Revision of one component of total prosthetic replacement of hip joint using cement
	W380	Total prosthetic replacement of hip joint not using cement, Conversion from previous uncemented total prosthetic replacement of hip joint
	W382	Conversion to total prosthetic replacement of hip joint not using cement
	W383	Revision of total prosthetic replacement of hip joint not using cement
	W384	Revision of one component of total prosthetic replacement of hip joint not using cement
	W392	Conversion to total prosthetic replacement of hip joint NEC
	W393	Revision of total prosthetic replacement of hip joint NEC
	W395	Revision of one component of total prosthetic replacement of hip joint NEC
	W462	Conversion to prosthetic replacement of head of femur using cement
	W472	Conversion to prosthetic replacement of head of femur not using cement
	W482	Conversion to prosthetic replacement of head of femur NEC
	W932	Conversion to hybrid prosthetic replacement of hip joint using cemented acetabular component
	W933	Revision of hybrid prosthetic replacement of hip joint using cemented acetabular component
	W940	Conversion from previous hybrid prosthetic replacement of hip joint using cemented femoral component
	W942	Conversion to hybrid prosthetic replacement of hip joint using cemented femoral component
	W943	Revision of hybrid prosthetic replacement of hip joint using cemented femoral component
	W952	Conversion to hybrid prosthetic replacement of hip joint using cement NEC
	W953	Revision of hybrid prosthetic replacement of hip joint using cement NEC
	W954	Attention to hybrid prosthetic replacement of hip joint using cement NEC
B	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
	W572	Primary excision arthroplasty of joint NEC
	W574	Conversion to excision arthroplasty of joint
	W582	Revision of resurfacing arthroplasty of joint

Table G.4: OPCS-4 codes to define hip revision** continued

Group	OPCS-4 code	Description
C	W394	Attention to total prosthetic replacement of hip joint NEC
D	W544	Attention to prosthetic replacement of articulation of bone NEC
E	Z843	Hip joint
	Z761	Head of femur
	Z756	Acetabulum
F	Y032	Renewal of prosthesis in organ NOC
	Y037	Removal of prosthesis from organ NOC

***Revision defined as Group A OR (Group B AND Group E) OR (Group C AND Group F)
OR (Group D AND Group E AND Group F)*

Appendix H

Chapter 6 - Systematic review

Table H.1: Reasons for excluding potentially relevant studies

Study	Reason for exclusion
Dorr*	Included all patients with displaced subcapital fractures – no clear exclusion criteria.
Skinner*	Included all patients with displaced subcapital fractures – no exclusions based on cognitive or mobility status.
Ravikumar*	Included all patients with displaced subcapital fractures – no exclusions based on cognitive or mobility status.
Mouzopoulos*	Included all patients with displaced subcapital fractures – no exclusions based on cognitive or mobility status.
Deng	Included all patients with displaced intracapsular fractures – no exclusions based on cognitive or mobility status.
Van den Bekerom	Included patients aged >70 and only excluded those that could not consent to participation and/or were bedbound.
Lai	Included all patients with displaced intracapsular fractures – no exclusions based on cognitive or mobility status.
Li	Included all patients with displaced intracapsular fractures – no exclusions based on cognitive or mobility status.
Pang	Could not be obtained**.

*Quasi-randomised controlled trials

**Unable to obtain study despite attempts to contact authors, assistance from the China Centre Library at the University of Oxford, requests to The British Library, and attempts to source these articles from collaborating libraries in both the UK and the USA.

Table H.2: Characteristics of included studies

Study	Setting		Participants		Intervention		
	Country	Centres	Total	Inclusion criteria*	Hemiarthroplasty Total	Type	Total hip arthroplasty Total
Baker 2006	UK	1	81	Age >60; normal AMTS; ability to walk >0.8km independently.	41	Direct lateral approach. Cemented femoral component with an Endo Femoral Head (Zimmer).	40 Direct lateral approach. Cemented femoral component with an all-polyethylene cemented acetabular cup (Zimmer).
Keating 2006	UK	11	138	Normal cognitive function (AMTS >6); independently mobile; recruiting surgeon believes both interventions acceptable.	69	Approach and prosthesis at discretion of the operating surgeon.	69 Approach and prosthesis at discretion of the operating surgeon.
Blomfeldt 2007	Sweden	1	120	Age 70-90 years; absence of severe cognitive dysfunction assessed using the Short Portable Mental Status Questionnaire; independent walking ability.	60	Modified Hardinge approach. Cemented Exeter femoral component.	60 Modified Hardinge approach. Cemented Exeter femoral component with a bipolar head or an OGEE cemented acetabular component.
Macaulay 2008	USA	5	40	Age >50; independently mobile; score >23/30 on the Folstein Mini Mental State Examination.	23	Approach and prosthesis at discretion of the operating surgeon.	17 Approach and prosthesis at discretion of the operating surgeon.
Cadossi 2013	Italy	1	83	Age >70, walking independently without aids.	41	Direct lateral approach. Cemented or uncemented Exeter femoral component with a bipolar femoral head.	47 Direct lateral approach. Uncemented Conus stem with a large-diameter femoral head and a polycarbonate-urethane (PCU) acetabular component.
*In addition to displaced intracapsular hip fractures, which was an inclusion criterion common to all included trials.							

Table H.3: Risk of bias assessments for included studies

Study	Type of bias	Domain	Assessment	Reason
Baker 2006	Selection bias	Random sequence generation	Low	Sealed envelopes
	Performance bias	Allocation concealment	High	Sealed envelopes without additional precautions
	Detection bias	Blinding of participants	High	Not blinded
	Attrition bias	Blinding of outcome assessment	High	Not blinded
	Reporting bias	Incomplete outcome data	Low	1/41 patients declined follow-up
Keating 2006	Reporting bias	Selective reporting	Unclear	No protocol published
	Selection bias	Random sequence generation	Low	Computer-based telephone randomization
	Performance bias	Allocation concealment	Low	Computer-based telephone randomization
	Detection bias	Blinding of participants	High	Not blinded
	Attrition bias	Blinding of outcome assessment	High	Not blinded
Blomfeldt 2007	Reporting bias	Incomplete outcome data	Unclear	Not stated
	Reporting bias	Selective reporting	Unclear	No protocol published
	Selection bias	Random sequence generation	Low	Sealed envelopes
	Performance bias	Allocation concealment	High	Sealed envelopes without additional precautions
	Detection bias	Blinding of participants	High	Not blinded
Macaulay 2008	Attrition bias	Blinding of outcome assessment	High	Not blinded
	Reporting bias	Incomplete outcome data	Low	2/120 patients lost to follow-up
	Reporting bias	Selective reporting	Unclear	No protocol published
	Selection bias	Random sequence generation	Low	Sealed envelopes
	Performance bias	Allocation concealment	Low	Sealed envelopes with additional precautions
Cadossi 2013	Detection bias	Blinding of participants	High	Not blinded
	Attrition bias	Blinding of outcome assessment	High	Not blinded
	Reporting bias	Incomplete outcome data	Unclear	Not stated
	Reporting bias	Selective reporting	Unclear	No protocol published
	Selection bias	Random sequence generation	Low	Sealed envelopes
Cadossi 2013	Performance bias	Allocation concealment	High	Sealed envelopes without additional precautions
	Detection bias	Blinding of participants	High	Not blinded
	Attrition bias	Blinding of outcome assessment	High	Not blinded
	Reporting bias	Incomplete outcome data	Low	All patients followed up
	Reporting bias	Selective reporting	Unclear	No protocol published

Appendix I

Chapter 6 - Propensity score matching

Table I.1: Characteristics of the unmatched population

	Hemiarthroplasty	Total hip arthroplasty	Total
Age*	83 (78-87)	73 (68-78)	79 (73-85)
Sex**			
Male	3,607 (23.1%)	2,178 (21.4%)	5,785 (22.5%)
Female	11,991 (76.9%)	7,980 (78.6%)	19,971 (77.5%)
ASA*	2 (2-2)	2 (2-2)	2 (2-2)
Pre-injury mobility**			
Independently mobile	10,190 (65.3%)	9,264 (91.2%)	19,454 (75.5%)
Mobile indoors with one aid	5,408 (34.7%)	894 (8.8%)	6,302 (24.5%)
AMTS*	10 (9-10)	10 (10-10)	10 (10-10)
Admission source**			
Own home	15,298 (98.1%)	10,089 (99.3%)	25,387 (98.6%)
Rehabilitation unit	19 (0.1%)	2 (0.0%)	21 (0.1%)
Residential/nursing home	173 (1.1%)	24 (0.2%)	197 (0.8%)
Acute hospital	108 (0.7%)	43 (0.4%)	151 (0.6%)

*Median (interquartile range)

**Number (percentage)

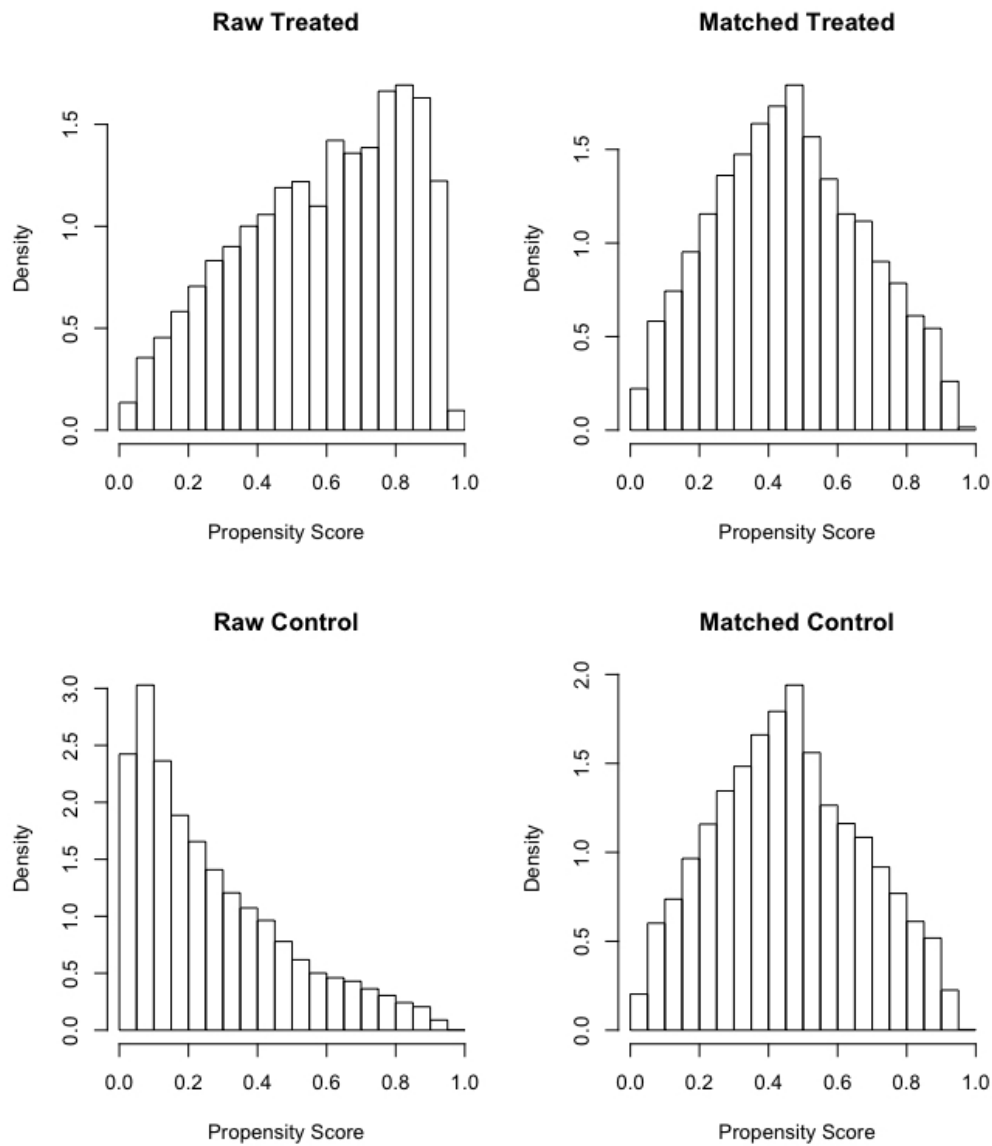


Figure I.1: Histograms showing the distribution of propensity scores before and after matching.

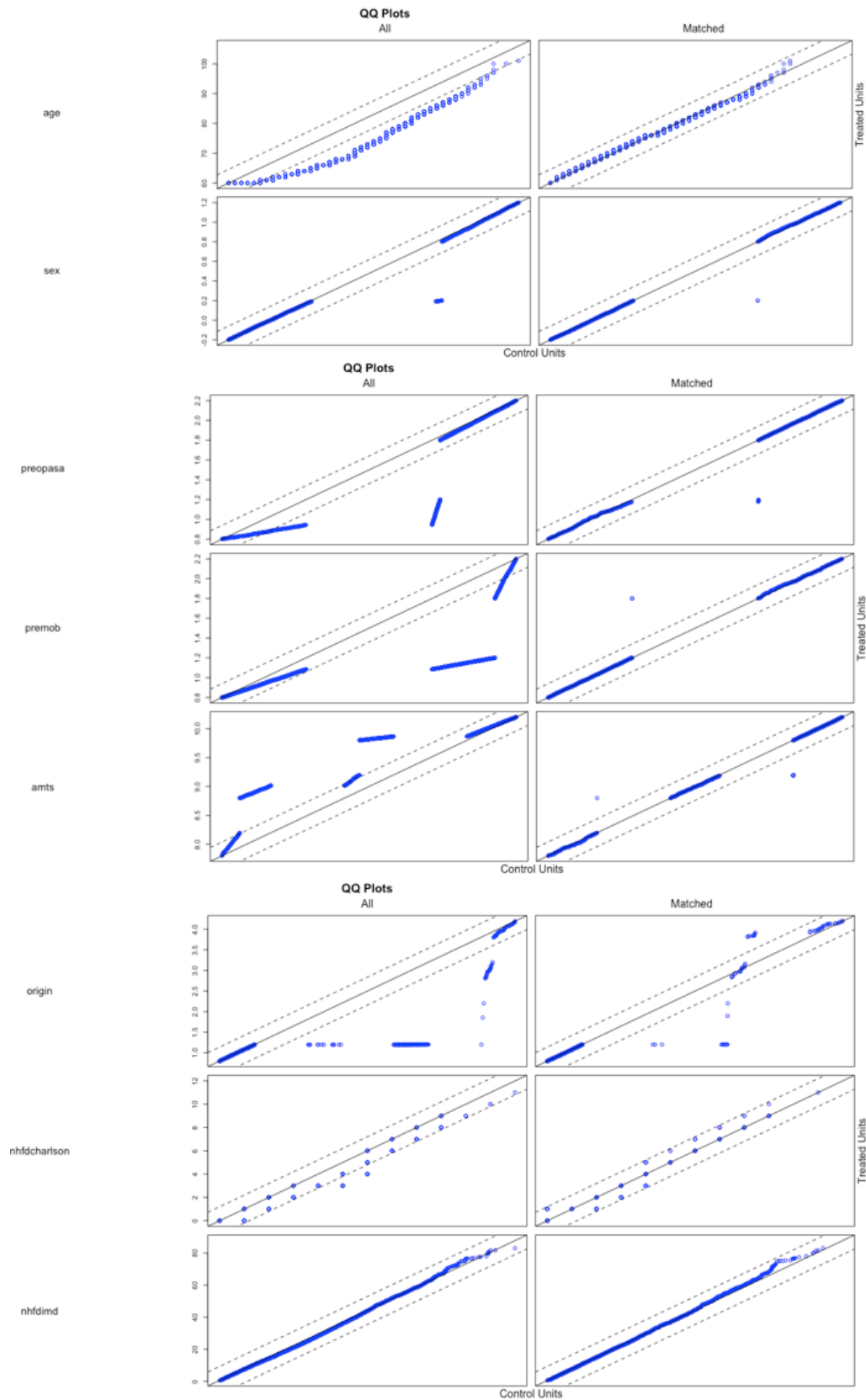


Figure I.2: Quantile-quantile plots of co-variables between the two groups before and after matching. Data from populations with the same empirical distribution will lie along the 45 degree reference line.

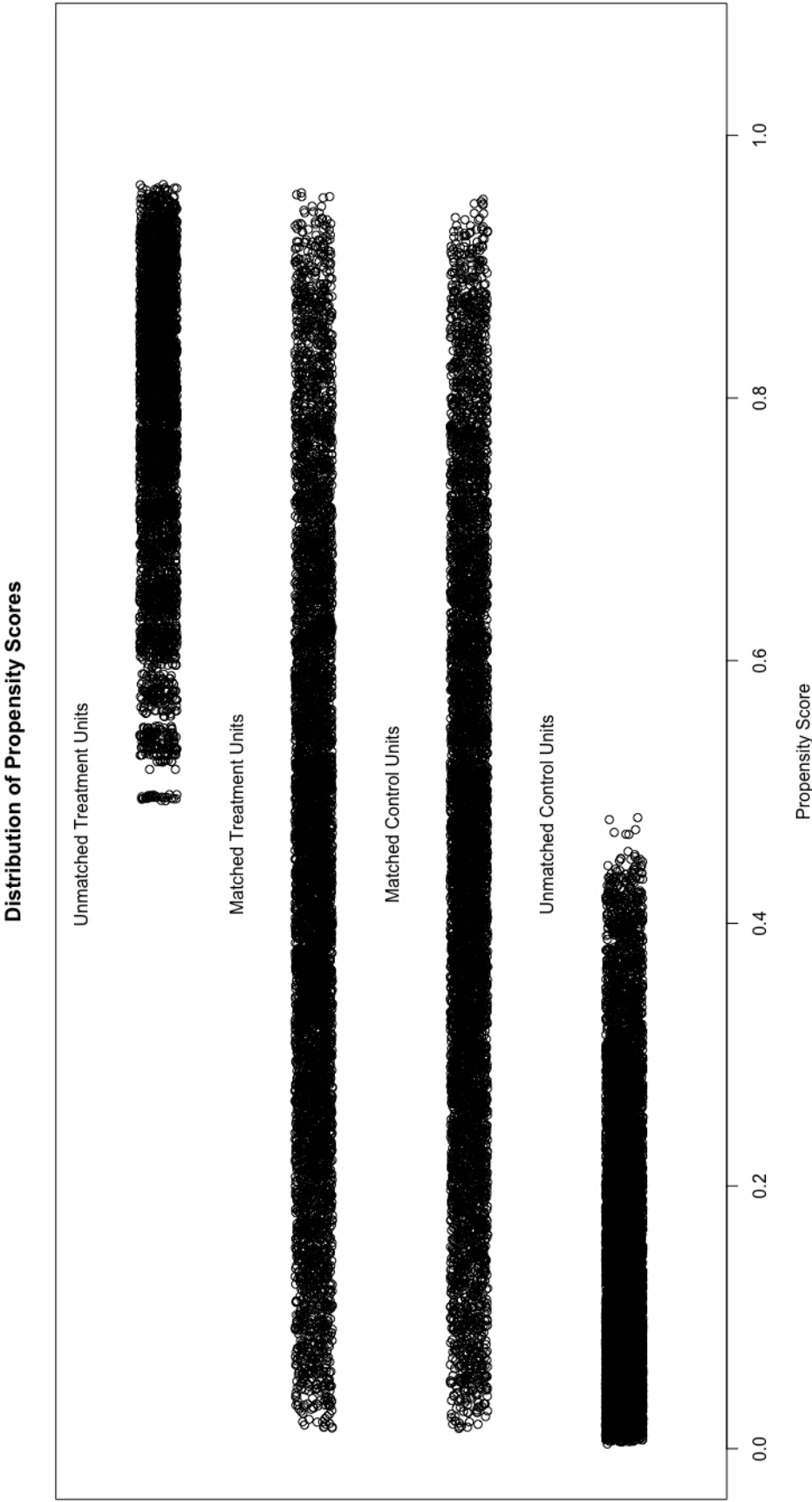


Figure I.3: A jitter plot showing the overall distribution of propensity scores for both matched and unmatched records.

Appendix J

Conclusion - Data access obstacles

This section is intended to highlight the nature of obstacles that may be faced by future doctoral students hoping to coordinate bespoke data linkages in England.

J.1 Inability to link primary care data

I first approached the CPRD in October 2015. They indicated that such a linkage was possible but subject to delays caused by their designated third party, the Health and Social Care Information Centre (HSCIC) (now NHS Digital). Over the next couple of years, I spoke with the CPRD on a number of occasions but was advised that - due to staffing issues - I could not submit an Independent Scientific Advisory Committee (ISAC) application. At each point, it was suggested that this period of waiting was temporary as - for example - they had advertised a post for more staff to join the linkages team. It was however made clear that ISAC approval was necessary before the process of submitting a Data Access Request Service (DARS) application to NHS Digital could begin. I was finally invited to submit a protocol for pre-ISAC review in September 2017 where a number of issues were raised:

- There were concerns about whether the CAG approval (necessary for NHFD linkage under s.251 NHS Act 2006) was sufficiently explicit to support a CPRD linkage. It was later accepted that this was sufficient.

- In addition to gaining a favourable decision at the ISAC and IGARD, a bespoke data sharing agreement would need to be drawn up between the CPRD and HQIP (as a 3rd party data controller).
- The cost of data access (estimated in 2015 to be up to £40,000) was thought to have increased in the interim.

These obstacles were overcome and I submitted an ISAC application in November 2017, which was returned in January 2018 with 3.5 pages of questions and requests to revise the study protocol. However, changes to the protocol would have required a new application to HQIP. As it did not feel that I had made any tangible progress in over two years - and most of my DPhil work could not start until I had access to the linked dataset - I reluctantly abandoned my plan to pursue a bespoke CPRD linkage.

J.2 Delays linking English secondary care data

Having developed concerns about the progress of the CPRD application, I had submitted a DARS application concurrently to NHS Digital. Following approval of the protocol by HQIP, this application (NIC-61090-T9Y0G) was submitted to NHS Digital on 22nd November 2016 and triaged as a “tier 3” application with an “estimated timescale of 60 working days”. The linked dataset (NHFD-HES-ONS) was received by the University of Oxford on 26th February 2018, i.e. 461 days later.

During this period, responsibility for civil registration death data passed from the ONS to NHS Digital. For some considerable time, neither organization seemed to know who could lawfully approve use of mortality data. I underwent ONS data governance training, gained “approved researcher” status under s.39(5) Statistics and Registration Service Act 2007, and had my projects approved by the ONS Microdata Release Panel. None of these steps were likely required by the time that the data was actually released.

The arrival of the linked dataset in February 2018 left ten months in which to

complete the bulk of the analytical work required for my DPhil. Unfortunately, the arrival of the data did not mark the end of my data access adventures because:

- Within a month, my project was flagged to undergo a NHS Digital security review. During this process, it became apparent that our own departmental protocols had changed during the long period over which my data access applications were being considered. Large datasets were now required to be stored in a secure computing room at the Botnar Research Centre. However, all my data applications (approved by HQIP, NHS Digital, and the ONS) specified that the data would be stored securely at the Kadoorie Centre for Critical Care Research, which is on a different site. It took some time (and many discussions) to unpick the conflict between our own information governance policies and the details agreed with the various data controllers.
- As the data had taken so long to arrive, the data sharing agreement expired only three months later. Extending the agreement required payment of a further fee and another DARS application. This re-application was further complicated as the EU GDPR had come into effect in the UK on 25th May. This change in the information governance landscape meant having to re-establish the lawful basis for data processing as well as fulfilling additional criteria, such as maintaining a study website.
- Four months after the data arrived, I received an email from NHS Digital on a Friday evening stating that type 2 patient opt-outs had not been properly applied to the dataset and that I should “securely destroy the data”. However, it was not clear that my existing data approvals (e.g. from HQIP) were sufficient to re-create the dataset. This situation took a number of months to resolve and introduced considerable further uncertainty while I was working to analyze the data.

J.3 Access to data from Scotland

By way of comparison, I submitted an application for SMR01 data on 9th March 2018. It was quickly established that the project in Chapter 4 on page 101 could be undertaken without patient-level data if sufficient analyst time could be freed to create an aggregate dataset. A single analyst from the MSk Audit created a file with aggregate data (based on patient-level records), cleaned the data, and provided a bespoke data dictionary. This data - which required considerable processing - was released on 28th May 2018 at no cost to the research team.

Appendix K

Published works



Cochrane
Library

Cochrane Database of Systematic Reviews

Impact of public release of performance data on the behaviour of healthcare consumers and providers (Review)

Metcalfe D, Rios Diaz AJ, Olufajo OA, Massa MS, Ketelaar NABM, Flottorp SA, Perry DC

Metcalfe D, Rios Diaz AJ, Olufajo OA, Massa MS, Ketelaar NABM, Flottorp SA, Perry DC.

Impact of public release of performance data on the behaviour of healthcare consumers and providers.

Cochrane Database of Systematic Reviews 2018, Issue 9. Art. No.: CD004538.

DOI: 10.1002/14651858.CD004538.pub3.

www.cochranelibrary.com

TABLE OF CONTENTS

HEADER	1
ABSTRACT	1
PLAIN LANGUAGE SUMMARY	2
SUMMARY OF FINDINGS FOR THE MAIN COMPARISON	4
BACKGROUND	6
OBJECTIVES	7
METHODS	7
Figure 1.	9
RESULTS	11
Figure 2.	13
Figure 3.	14
DISCUSSION	18
AUTHORS' CONCLUSIONS	21
ACKNOWLEDGEMENTS	21
REFERENCES	22
CHARACTERISTICS OF STUDIES	25
ADDITIONAL TABLES	46
APPENDICES	56
WHAT'S NEW	67
HISTORY	68
CONTRIBUTIONS OF AUTHORS	68
DECLARATIONS OF INTEREST	68
SOURCES OF SUPPORT	68
DIFFERENCES BETWEEN PROTOCOL AND REVIEW	69
INDEX TERMS	69

[Intervention Review]

Impact of public release of performance data on the behaviour of healthcare consumers and providers

David Metcalfe¹, Arturo J Rios Diaz², Olubode A Olufajo³, M. Sofia Massa⁴, Nicole ABM Ketelaar⁵, Signe A. Flottorp^{6,7}, Daniel C Perry¹

¹Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK.

²Department of Surgery, Thomas Jefferson University Hospital, Philadelphia, PA, USA. ³Department of Surgery, Howard-Harvard Health Sciences Outcomes Research Center Howard University College of Medicine, Washington, DC, USA. ⁴Nuffield Department of Population Health, University of Oxford, Oxford, UK. ⁵Social Work Research Group, Saxion University of Applied Sciences, Enschede, Netherlands. ⁶Norwegian Institute of Public Health, Oslo, Norway. ⁷Institute of Health and Society, University of Oslo, Oslo, Norway

Contact address: David Metcalfe, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK. d.metcalfe@doctors.org.uk.

Editorial group: Cochrane Effective Practice and Organisation of Care Group.

Publication status and date: New search for studies and content updated (no change to conclusions), published in Issue 9, 2018.

Citation: Metcalfe D, Rios Diaz AJ, Olufajo OA, Massa MS, Ketelaar NABM, Flottorp SA, Perry DC. Impact of public release of performance data on the behaviour of healthcare consumers and providers. *Cochrane Database of Systematic Reviews* 2018, Issue 9. Art. No.: CD004538. DOI: 10.1002/14651858.CD004538.pub3.

Copyright © 2018 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

ABSTRACT

Background

It is becoming increasingly common to publish information about the quality and performance of healthcare organisations and individual professionals. However, we do not know how this information is used, or the extent to which such reporting leads to quality improvement by changing the behaviour of healthcare consumers, providers, and purchasers.

Objectives

To estimate the effects of public release of performance data, from any source, on changing the healthcare utilisation behaviour of healthcare consumers, providers (professionals and organisations), and purchasers of care. In addition, we sought to estimate the effects on healthcare provider performance, patient outcomes, and staff morale.

Search methods

We searched CENTRAL, MEDLINE, Embase, and two trials registers on 26 June 2017. We checked reference lists of all included studies to identify additional studies.

Selection criteria

We searched for randomised or non-randomised trials, interrupted time series, and controlled before-after studies of the effects of publicly releasing data regarding any aspect of the performance of healthcare organisations or professionals. Each study had to report at least one main outcome related to selecting or changing care.

Data collection and analysis

Two review authors independently screened studies for eligibility and extracted data. For each study, we extracted data about the target groups (healthcare consumers, healthcare providers, and healthcare purchasers), performance data, main outcomes (choice of healthcare provider, and improvement by means of changes in care), and other outcomes (awareness, attitude, knowledge of performance data, and costs). Given the substantial degree of clinical and methodological heterogeneity between the studies, we presented the findings for each policy in a structured format, but did not undertake a meta-analysis.

Impact of public release of performance data on the behaviour of healthcare consumers and providers (Review)
Copyright © 2018 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

I

Main results

We included 12 studies that analysed data from more than 7570 providers (e.g. professionals and organisations), and a further 3,333,386 clinical encounters (e.g. patient referrals, prescriptions). We included four cluster-randomised trials, one cluster-non-randomised trial, six interrupted time series studies, and one controlled before-after study. Eight studies were undertaken in the USA, and one each in Canada, Korea, China, and The Netherlands. Four studies examined the effect of public release of performance data on consumer healthcare choices, and four on improving quality.

There was low-certainty evidence that public release of performance data may make little or no difference to long-term healthcare utilisation by healthcare consumers (3 studies; 18,294 insurance plan beneficiaries), or providers (4 studies; 3,000,000 births, and 67 healthcare providers), or to provider performance (1 study; 82 providers). However, there was also low-certainty evidence to suggest that public release of performance data may slightly improve some patient outcomes (5 studies, 315,092 hospitalisations, and 7502 providers). There was low-certainty evidence from a single study to suggest that public release of performance data may have differential effects on disadvantaged populations. There was no evidence about effects on healthcare utilisation decisions by purchasers, or adverse effects.

Authors' conclusions

The existing evidence base is inadequate to directly inform policy and practice. Further studies should consider whether public release of performance data can improve patient outcomes, as well as healthcare processes.

PLAIN LANGUAGE SUMMARY

Can the public release of performance data in health care influence the behaviour of consumers, healthcare providers, and organisations?

What is the aim of this review?

The aim was to find out if publicly releasing information about the performance of healthcare providers (e.g. hospitals and individual professionals) has a measurable influence on changing the behaviour of consumers, providers, and purchasers of care. We also sought to determine whether this affected the performance of healthcare providers, patient outcomes, and staff morale.

Key messages

Public release of performance data may lead to little or no difference in healthcare choices (made by either consumers or providers), or provider performance. However, it may slightly improve outcomes for patients.

What was studied in the review?

Healthcare providers are increasingly expected to inform the public on how well they are performing. However, it is not yet known whether public release of performance data has a measurable influence on patients' choice of healthcare services, or whether it can truly drive improvements in the quality of health care.

What are the main results of the review?

The authors searched the literature for studies evaluating the effects of publicly releasing healthcare performance information, and found 12 relevant studies that analysed data from more than 7570 providers, and a further 3,333,386 clinical encounters, e.g. individual patients.

There was low-certainty evidence that public release of performance data may lead to little or no difference in the services that patients choose to access, the decisions taken by healthcare providers, or overall provider performance. There was low-certainty evidence suggesting that some patient outcomes may slightly improve following public release of performance data, but that this might have less of an effect on the behaviour of disadvantaged populations. There was no evidence relating to healthcare utilisation decisions by purchasers, or adverse effects.

Although a number of the studies were individually well conducted, there were limitations: in particular, the evidence base varied substantially in terms of setting (e.g. United States or Korea), health condition (e.g. heart attack or hip replacement), type of performance data (e.g. process or patient outcome), and the mode of data publication (e.g. mail shot or poster). Their findings were also inconsistent, with some reporting changes attributed to public release of information, and others reporting no such changes.

How up-to-date is this review?

The review authors searched for studies that had been published up to June 2017.

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON [\[Explanation\]](#)

People: Insurance plan beneficiaries, birthing mothers, GPs Settings (countries and clinical settings): United States, Canada, South Korea, Netherlands, China / Community, primary care and hospitals Intervention: Public release of performance data Comparison: No public reporting			
Outcomes	Impact	No of clinical encounters (studies)	Certainty of the evidence (GRADE)*
Changes in healthcare utilisation by consumers	Public release of performance data may make little or no difference to long-term healthcare utilisation by consumers. However, two studies (one cNRT and one ITS) found that some population subgroups might be influenced by public release of performance data	18,294 insurance plan beneficiaries ^a (3: 1 cRT, 1 cNRT, 1 ITS)	⊕⊕○○ low
Changes in healthcare decisions taken by healthcare providers (professionals and organisations)	Public release of performance data may make little or no difference to decisions taken by healthcare professionals. Two studies (2 cRTs) found that some decisions might be affected by public release of performance data. One study (ITS) found that decisions might be influenced by the initial release of data, but that subsequent releases might have less impact	3,000,000 births ^b and 67 healthcare providers (4: 2 RTs, 2 ITS)	⊕⊕○○ low ^c
Changes in the healthcare utilisation decisions of purchasers	No studies reported this outcome.	-	-
Changes in provider performance	Public release of performance data may make little or no difference to objective measures of provider performance	82 healthcare providers (1 cRT)	⊕⊕○○ low ^d
Changes in patient outcome	Public release of performance data may slightly improve patient outcomes	315,092 hospitalisations and 7503 healthcare providers (5: 1 RT, 3 ITS, 1 CBA)	⊕⊕○○ low ^e

Adverse effects	No studies reported this outcome.	-	-
Impact on equity	Public release of performance data may have a greater effect on provider choice among advantaged populations	Unknown (1 ITS)	⊕⊕○○ low

EPOC adapted statements for GRADE Working Group grades of evidence

High-certainty. This research provides a very good indication of the likely effect. The likelihood that the effect will be substantially different[†] is low.

Moderate-certainty. This research provides a good indication of the likely effect. The likelihood that the effect will be substantially different[†] is moderate.

Low-certainty. This research provides some indication of the likely effect. However, the likelihood that it will be substantially different[†] is high.

Very low-certainty. This research does not provide a reliable indication of the likely effect. The likelihood that the effect will be substantially different[†] is very high.

[†] Substantially different = a large enough difference that it might affect a decision

^a Number was based only on [Farley 2002a](#) and [Farley 2002b](#) studies, as the total number of cases analysed in [Romano 2004](#) was unclear

^b Number of participants in [Jang 2011](#) (3,000,000) estimated from data presented in [Chung 2014](#)

^c Downgraded one level for inconsistency as effect shown by [Zhang 2016](#), but not [IkkesheJang 2011](#), [Ikkesheim 2013](#), or [Flett 201511](#)

^d Downgraded two levels for risk of bias, as there was attrition of participating hospitals, evidence of contamination of the intervention across intervention and control hospitals, and blinding was not possible given the nature of the intervention

^e Downgraded two levels for inconsistency, as there was marked disagreement between studies, with two showing improvements in patient outcome ([Liu Tu 2009](#); [Liu 20179](#)), and three showing no such improvements ([DeVoRinke 2015](#); [DeVore 2016](#); [Joynt 201615](#))

cluster-randomised trial (cRT); cluster-non-randomised trial (cNRT); controlled before-after (CBA) study; interrupted time series (ITS) study; randomised trial (RT)

BACKGROUND

It is becoming increasingly common to release information about the performance of healthcare systems into the public domain. In the present era of accountability, cost-effectiveness, quality improvement, and demand-driven healthcare systems, decision makers such as governments, regulators, purchaser and provider organisations, health professionals, and consumers of health care are becoming more interested in measuring performance (Smith 2009). Such measurements may be presented in consumer reports, provider profiles, or report cards. It is not always clear who the information users are or what the release of data is expected to achieve. However, it is often assumed that the information will influence the behaviours of various stakeholders, and so ultimately lead to health system improvements (Berwick 2003; Smith 2009; Campanella 2016).

One study has conceptualised public reporting of performance data as (1) supporting patient choice, (2) improving accountability, and (3) allowing providers to benchmark their performance against others (Greenhalgh 2018).

Publication of performance data can support patient choice by helping them to identify the highest performing providers. However, there are many barriers to patient use of performance data (Canaway 2017). These include the complexity of the performance data (Hibbard 2010), lack of skills to comprehend and use performance data (Hibbard 2007; Canaway 2017; Canaway 2018), and the way data are presented (Damman 2010; Canaway 2017; Canaway 2018). Such barriers might negate the impact of choice, and even reduce equity in health care. Consumers from poorer backgrounds and with lower educational levels may be less able to choose, and less able to afford travel to better performing, but more distant, providers (Aggarwal 2017; Moscelli 2017). There is also evidence that patients often do not use published performance data when making healthcare choices (Greenhalgh 2018).

Improved accountability may be achieved by encouraging providers to focus on quality issues, as they know that performance measures will be published (Fung 2008; Hendriks 2009). This in turn, may stimulate quality improvements, particularly as providers can see their own performance against that of other clinicians and hospitals. Similarly, patients who preferentially choose high-quality health care might help drive improvements, by concentrating resources with the best performing providers (Hibbard 2009; Kolstad 2009; Werner 2009).

Other proposed goals for performance measurements have been linked to controlling costs (Berwick 1990; Sirio 1996), regulating the overall healthcare system (Rosenthal 1998; Schut 2005), and influencing the decisions of healthcare purchasers (Brook 1994; Hibbard 1997; Mukamel 1998).

Professional concerns to public release of performance data often relate to the validity of both the performance measures them-

selves, and comparisons between health providers (Sherman 2013; Kiernan 2015; Burns 2016;). There are concerns that failure to adequately adjust for case mix differences might lead to providers that treat higher-risk patients being labelled as poor performers, or to providers preferentially selecting lower-risk patients (Wasfy 2015; Burns 2016; Shahian 2017; Wadhera 2017). In healthcare systems where providers charge for their services, the 'better' performing providers might feel empowered to increase charges, thereby restricting access to better care (Mukamel 1998). An additional risk is that publication of performance data may lead to improved reporting, without necessarily improving performance. It has been argued that the care processes that are easiest to measure are often those that are least important in a quality improvement context, and can result in the de-prioritisation of other tasks (Loeb 2004).

Description of the intervention

Public release of performance data is the release of information about the quality of care, so that patients and consumers can better decide what health care they wish to select, and healthcare professionals and organisations can better decide what to provide, improve, or purchase. This mechanism excludes the use of auditing and feedback as a tool for improving professional practice and healthcare outcomes, which has been reviewed elsewhere (Ivers 2012).

How the intervention might work

Public release of performance data may change individual or organisational behaviour through a number of mechanisms. The goal of improving quality of health care can be achieved through a selection pathway or a change pathway (Berwick 2003). Consumers, patients, and purchaser organisations that are in a position to do so, can select the best healthcare professionals and organisations. This type of selection will not change the quality of the delivered care by itself, but it can be a stimulus for quality improvement. Importantly, such changes might be attenuated by the limited choice that patients have in many cases, e.g. in the case of emergencies, the need to access specialised care that is only available in few centres, or because of resource limitations (Aggarwal 2017; Moscelli 2017). In a change pathway, healthcare professionals and organisations can improve performance by changing their work procedures or professional culture, and organisations can make structural changes.

Why it is important to do this review

Some systematic reviews have suggested positive effects of publicly releasing performance data, but included a broad range of study designs (Marshall 2000; Shekelle 2008; Fung 2008; Faber 2009). This study (which is the first update of Ketelaar 2011) aimed to

review the evidence for the impact of such interventions using more stringent selection criteria.

OBJECTIVES

To estimate the effects of publicly releasing performance data on changing the healthcare utilisation behaviour of healthcare consumers, providers (professionals and organisations), and purchasers of care. In addition, we sought to estimate the effects on healthcare provider performance, patient outcomes, and staff morale.

METHODS

Criteria for considering studies for this review

Types of studies

- Randomised trials, including cluster-randomised trials
- Non-randomised trials, including cluster-non-randomised trials, which use non-random methods of allocation, such as alternation or allocation by case note number
- Controlled before-after studies, with at least two intervention sites and two control sites that are chosen for similarity of main outcome measures at baseline
- Interrupted time series studies, with at least three data points before and three data points after the intervention

We included non-randomised studies in anticipation of a lack of randomised trials, but also because some interventions might not be appropriate for a trial (e.g. randomising participants to not receive important information that might affect their healthcare choices), and others might have a variable effect over time that is best observed by an alternative study design, such as an interrupted time series.

Types of participants

Patients or other healthcare consumers and healthcare providers, including organisations (e.g. hospitals), without any restriction by type of healthcare professional, provider, setting, or purchaser.

Types of interventions

We included interventions that contained the following elements:

- Performance data about any aspect of the healthcare organisations or individuals, including process measures (e.g. waiting times), healthcare outcomes (e.g. mortality), structure measures (e.g. presence of waiting rooms), consumer or patient

experiences (e.g. Consumer Assessment of Healthcare Providers and System (CAHPS) data), with or without expert or peer-assessed measures, e.g. certification, accreditation, and quality ratings given by colleagues. Performance data were included if prepared and released by any organisation, such as the government, insurers, consumer organisations, or providers. We excluded studies that did not evaluate publication of performance data concerning process measures, healthcare outcomes, structure measure, consumer or patient experiences, or expert or peer-assessed measures.

- The release of performance data into the public domain in written or electronic form without regard to any minimum degree of accessibility. For example, this could include a report available in a publicly accessible library, as well as active dissemination directly to consumers through personal mailings.

Comparators

The following comparisons were planned:

1. Public release of performance data compared to settings in which data were not released to the public
2. Different modes of releasing performance data to the public

Types of outcome measures

Primary outcomes

We planned the primary outcome measures according to two key aims of publicly releasing performance data.

1. Improvement by selection

- Changes in healthcare utilisation by consumers
 - Objective measures of changing consumer behaviour, such as increased use of a specific healthcare provider
- Changes in healthcare decisions taken by healthcare providers (professionals and organisations)
 - Objective measures of changing healthcare provider behaviour, such as changes to drug prescribing
- Changes in the healthcare utilisation decisions of purchasers
 - Objective measures of changing purchaser behaviour, such as increased or decreased funding for services

2. Improvement by changes in care

- Changes in provider performance
 - Objective changes, such as reaching the correct diagnosis or time to treatment
 - Including measures that were made both public and others that were not
- Changes in patient outcome
 - Objective changes, such as mortality or patient-reported outcome measures
- Changes in staff morale
 - Using a previously validated assessment tool

Secondary outcomes

We considered unintended and adverse effects or harms, and any potential impact on equity (e.g. differential effects between advantaged and disadvantaged populations), and awareness, knowledge, attitude, or costs.

We excluded studies that reported awareness, attitude, perspectives, and knowledge of performance data and cost data in the absence of objective measures of decision behaviour, provider performance or patient outcomes.

Search methods for identification of studies

Electronic searches

We searched the Database of Abstracts of Reviews of Effects (DARE) for primary studies included in related systematic reviews.

We searched the following databases on 26 June 2017:

- Cochrane Central Register of Controlled Trials (CENTRAL; 2017, Issue 5) in the Cochrane Library;
- MEDLINE Ovid (including Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Versions);
- Embase Ovid.

The Cochrane Effective Practice and Organisation of Care (EPOC) Information Specialist developed the search strategies in consultation with the authors. Search strategies are comprised of keywords and controlled vocabulary terms. We applied no language or time limits. We searched all databases from database start date to 26 June 2017.

Searching other resources

Trial Registries

- International Clinical Trials Registry Platform (ICTRP), World Health Organization (WHO) www.who.int/ictrp/en/ (searched 26 June 2017)
- ClinicalTrials.gov, US National Institutes of Health (NIH) clinicaltrials.gov/ (searched 26 June 2017)

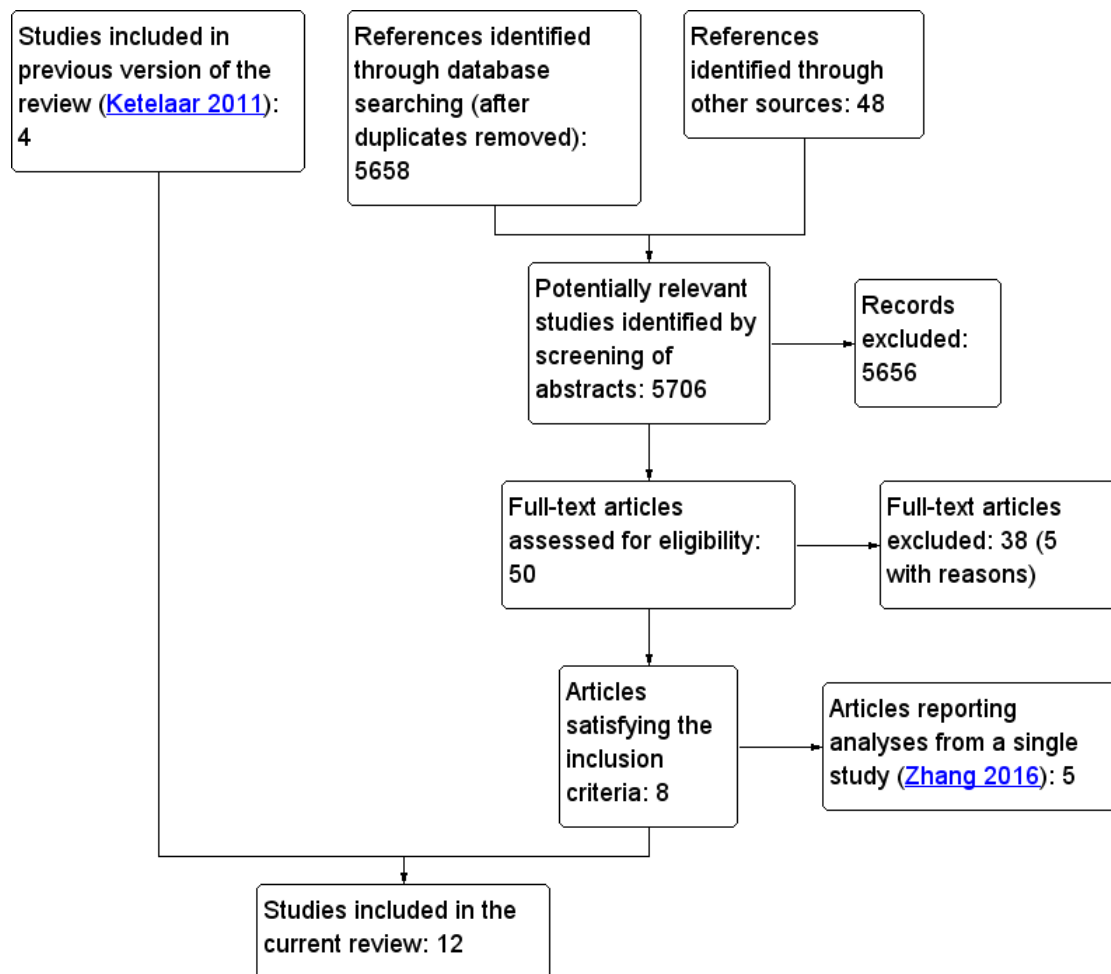
We manually searched the reference lists of all included studies. We provided all search strategies used in [Appendix 1](#).

Data collection and analysis

Selection of studies

We downloaded all titles and abstracts retrieved in the electronic search to a reference management database. We removed the duplicates, and two review authors then independently examined the remaining references. All review authors recorded their assessments of abstracts with points: '0' for exclusion, '1' for doubtful and '2' for inclusion. Two review authors (DM, ARD) independently rated each abstract; therefore, a minimum score of zero, and a maximum score of four was possible. Abstracts with a combined score of zero or one were excluded. Studies with a combined score of three or four were included. Two review authors resolved the fate of studies with a combined score of two by discussion. A third review author (OO) adjudicated on any disagreements that remained unresolved. [Figure 1](#) shows the PRISMA flow diagram that accounts for exclusion of all items received by the search strategy.

Figure 1. Flowchart for study selection



Data extraction and management

Two authors (DM, OO) independently extracted the data about the study design, patient and provider characteristics, interventions, outcome measures, and healthcare choices to a form specially designed for our review. Disagreements were resolved by discussion, and we accepted the judgement of a third author (ARD) in the event of continued disagreement.

Assessment of risk of bias in included studies

We assessed risk of bias by applying the guidance from the *Cochrane Handbook for Systematic Reviews of Interventions*, which recommends using the following items: (i) adequate sequence generation, (ii) concealment of allocation, (iii) blinding, (iv) incomplete outcome data, (v) selective reporting, and (vi) no risk of bias from

other sources (Higgins 2011). However, we deviated from this guidance: we used three additional criteria that are specified by the Cochrane Effective Practice and Organisation of Care Group: (vii) baseline characteristic similarity, (viii) baseline outcome similarity, and (ix) adequate protection against contamination (EPOC 2013). We used these nine standard criteria for randomised trials, non-randomised trials, and controlled before-after studies. We used seven criteria for interrupted time series studies, and applied these as recommended by EPOC 2013: (i) the intervention is independent of other changes, (ii) the shape of the intervention effect is pre-specified, (iii) the intervention is unlikely to affect data collection, (iv) knowledge of the allocated interventions is adequately prevented during the study, (v) the outcome data are incomplete, (vi) reporting is not selective, and (vii) there is no risk of bias from other sources. Two review authors (DM, ARD) inde-

pendently reached judgements about risk of bias using the guidance provided by Higgins 2011 and EPOC 2013, and resolved disagreements by discussion. A third review author (OO) dealt with any disagreements that the two review authors could not resolve.

Measures of treatment effect

In order to standardise reporting of effect sizes, we re-analysed data from individual studies to ensure that randomised trials and controlled before-after studies could be reported as relative effects. Interrupted time series were reported as change in level and change in slope. We described the methods used for re-analysing and presenting these data in Data synthesis.

Unit of analysis issues

We noted whether randomised trials randomised patients or healthcare providers. If analysis did not allow for clustering of patients within healthcare providers, we recorded a unit of analysis error, because such analyses tend to overestimate the precision of the treatment effect. In the event of a unit of analysis error and insufficient data to account for clustering, we did not report P values or confidence intervals.

Dealing with missing data

In the event of important missing data, we contacted the authors of individual studies. As described in Data synthesis, we electronically extracted missing interrupted time series data that were presented in graphs.

Assessment of heterogeneity

There were substantial differences between the policies and interventions described. There were also differences between the settings, in terms of culture and health system delivery. Although some studies evaluated similar interventions, there were still important clinical and methodological differences. As statistical tests for heterogeneity lack power when few studies are included, we elected not to calculate average effects across studies, or to estimate statistical heterogeneity (Schroll 2011).

Assessment of reporting biases

We did not present funnel plots as we did not undertake a meta-analysis and there were not more than 10 studies contributing to any individual analysis (Higgins 2011).

Data synthesis

We followed the EPOC recommendations with regard to analysing data from individual studies and meta-analysis (EPOC 2013). We expressed the findings from controlled before-after studies as relative effects. To achieve this, we reported continuous variables as relative change in outcome measures, adjusted for baseline differences. We undertook absolute difference-in-difference analyses that were adjusted for differences in the postintervention control group using: ((postintervention intervention group - postintervention control group) - (preintervention intervention - pre-intervention control))/postintervention control. For ease of comparison with the findings of controlled before-after studies, we reported the findings of randomised and non-randomised trials using the same difference-in-difference analysis.

Interrupted time series are typically reported using regression analysis, such as autoregressive integrated moving average (ARIMA) analysis. Pursuant to the EPOC recommendations, we present outcomes along two dimensions: change in level and change in slope (EPOC 2013). The former represents the immediate effect of the intervention as measured by the difference between the fitted value for the first post-intervention time point and the predicted outcome at the same point, based only on an extrapolation of the pre-intervention slope. Change in slope is an expression of any longer-term effect of the intervention. We decided to use a similar method to the change in level, but a later follow-up period, e.g. six months.

In the event that appropriate interrupted time series analyses were not reported but that data were presented graphically, we read values from graphs using Plot Digitizer v2.6.8 (Huwaldt 2004). We extracted 'actual' data points from all studies and only planned to use lines of best fit in the event that true points were not available. A segmented time series model ($Y(t) = B0 + B1 \cdot \text{preslope} + B2 \cdot \text{postslope} + B3 \cdot \text{intervention} + e(t)$) was specified, in which $Y(t)$ was the outcome in month t . Preslope is a continuous variable that indicates time from the beginning of the study until the end of the pre-intervention phase, after which it was coded as a constant. Postslope is assigned the value 0 until after the intervention takes place, after which it is coded sequentially from 1 (i.e. 1, 2, 3). Intervention is assigned the value 0 pre-intervention and 1 in the postintervention time period. In this model, $B1$ estimates the pre-intervention slope, $B2$ the postintervention slope, and $B3$ the change in level, i.e. the difference between the first postintervention time point and the extrapolated first postintervention time point had the pre-intervention line continued into the postintervention period. The difference in slope was determined using $B2 - B1$.

We reported effects at 3, 6, 9, 12, and 24 months postintervention when the data were available. Given the substantial degree of clinical and methodological heterogeneity between the studies, we presented the findings for each policy in a structured format, but did not undertake a meta-analysis.

Summary of findings

We summarised the findings of the main intervention comparisons in a 'Summary of findings' table to illustrate the certainty of the evidence. One review author (DM) categorised the certainty of the evidence as high, moderate, low, or very low, using the five GRADE domains (study limitations, consistency of effect, imprecision, indirectness, and publication bias (Guyatt 2011)). We undertook this pursuant to Chapter 12 of the *Cochrane Handbook for Systematic Reviews of Interventions* and worksheets created by EPOC (Higgins 2011; EPOC 2013). All other co-authors checked these judgments, and resolved disagreements through discussion. When ratings were up- or down-graded, we justified these decisions using footnotes in Appendix 2 and Summary of findings for the main comparison. Standardised statements for reporting effects and certainty of evidence were selected, based on the GRADE assessments for each outcome, and used throughout the review (EPOC 2017). The seven outcomes reported in Summary of findings for the main comparison are:

- Changes in healthcare utilisation by consumers
- Changes in healthcare decisions taken by healthcare providers (professionals and organisations)
- Changes in the healthcare utilisation decisions by healthcare purchasers
- Changes in provider performance
- Changes in patient outcome
- Adverse effects
- Impact on equity

Subgroup analysis and investigation of heterogeneity

As described in Data synthesis, we presented the findings of individual studies in a structured format rather than attempting meta-analysis, given the substantial heterogeneity between the studies. Therefore, it was not possible to undertake subgroup analyses.

Sensitivity analysis

In the absence of a formal meta-analysis, we did not undertake any sensitivity analyses.

RESULTS

Description of studies

The included studies are summarised in Table 1 and described fully in Characteristics of included studies. A number of studies that narrowly failed to satisfy our selection criteria are described in Characteristics of excluded studies.

Results of the search

The electronic searches for this update retrieved 5658 individual items; a further 48 were identified from other sources, e.g. manual searching of reference lists. We excluded 5656 items because the titles and abstracts did not meet our inclusion criteria. We retrieved the full-text versions of the remaining 50 articles; 38 of these did not satisfy the inclusion criteria; five with reasons, see (Characteristics of excluded studies). Five of the remaining 16 articles reported separate analyses of a single cluster randomised trial, and so we treated them as a single study for the purposes of this review (Zhang 2016). Therefore, we included 12 studies in the review. As described in Data synthesis, we did not undertake formal meta-analyses due to substantial inter-study heterogeneity. We presented the study flow chart in Figure 1 (Moher 1999).

Included studies

We included 12 studies that comprised more than 7570 providers (e.g. professionals and organisations) and a further 3,333,386 clinical encounters (e.g. patient referrals, prescriptions). There were four cluster randomised trials (Farley 2002a; Tu 2009; Ikkersheim 2013; Zhang 2016), one cluster-non-randomised trial (Farley 2002b), six interrupted time series studies (Romano 2004; Jang 2011; Flett 2015; DeVore 2016; Joynt 2016; Liu 2017), and one controlled before-after study (Rinke 2015). Eight were conducted in the USA (Farley 2002a; Farley 2002b; Romano 2004; Flett 2015; Rinke 2015; DeVore 2016; Joynt 2016; Liu 2017), and one each in Canada (Tu 2009), the Netherlands (Ikkersheim 2013), Korea (Jang 2011), and China (Zhang 2016).

Three studies focused on changes in the healthcare utilisation decisions of consumers (Farley 2002a; Farley 2002b; Romano 2004), four of providers (Jang 2011; Ikkersheim 2013; Flett 2015; Zhang 2016), and none of purchasers. Two studies reported data on changes to provider performance (Tu 2009; Rinke 2015), five on patient outcomes (Tu 2009; Flett 2015; DeVore 2016; Joynt 2016; Liu 2017), and none on staff morale. No study explicitly reported adverse events as a separate outcome, or gave particular consideration to effects on equitable health care.

Three US studies examined the effect of a single suite of interventions (i.e. laws mandating public reporting of healthcare-associated infections in the United States), which were introduced by some state legislatures between 2006 and 2009 (Flett 2015; Rinke 2015; Liu 2017). Liu 2017 examined the effect of mandatory reporting on central line-associated bloodstream infection rates in adult intensive care units. They undertook an interrupted time series study using data from hospitals contributing to the National Healthcare Safety Network between 2006 and 2012. States that did not introduce mandatory reporting were used to control for secular trends through a difference-in-difference analysis. The other two studies focused their analyses on healthcare-associated infections in paediatric inpatients (Flett 2015; Rinke 2015). Rinke 2015 sought to determine whether mandatory central line-asso-

ciated bloodstream infection public reporting was associated with a reduction in a specific paediatric safety indicator (PDI12, i.e. selected infections due to medical care), which is defined using diagnosis codes on hospital discharge. They undertook a controlled before-after study using the Kids' Inpatient Database, which is one of a suite of administrative healthcare databases coordinated by the Healthcare Cost and Utilization Project at the US Agency for Healthcare Research and Quality. [Flett 2015](#) did not examine patient outcomes, but aimed to test the hypothesis that clinicians in hospitals that are required to report central line-associated bloodstream infections would modify their behaviour by sending fewer blood culture tests or prescribing longer courses of antibiotics. They undertook an interrupted time series using data from the Pediatric Health Information System, which is a collaborative venture between children's hospitals that is used for clinical audit and quality improvement. The data were analysed using generalised linear mixed-effects models with auto-correlated residuals to compare central line-associated bloodstream infections adjusted rate ratios before and after implementation of mandatory reporting laws.

Two US studies studied the effect of providing information about plan performance on choice of insurance plan by new Medicaid beneficiaries ([Farley 2002a](#); [Farley 2002b](#)). [Farley 2002a](#) was a cluster-randomised trial, using data from new Medicaid beneficiaries in Iowa. Under Iowa Medicaid, new enrollees were automatically assigned, by default, to one of four private health maintenance organisations or the Medicaid primary care case management programme. They were sent a packet of information about their specific health plan and benefits under Medicaid. The control group received the standard packet of information and the intervention group received this, plus an additional report that described the performance of each health plan, along domains such as 'overall health care rating', and 'personal doctor rating'. The authors used multinomial logistic regression to model the odds of new beneficiaries electing to continue with or change their allocated plan. In [Farley 2002b](#), the same author team undertook a cluster-non-randomised trial to evaluate the same performance reports on beneficiary choice within the New Jersey Medicaid programme. The study design was very similar to [Farley 2002a](#), in terms of control and intervention groups, although this was technically a non-randomised trial, because participants were allocated according to the last digit of their Medicaid case ID number. The objective outcome measure reported was the effect of performance reports on Medicaid beneficiary plan choices.

The other three US studies each examined the impact of different public reporting initiatives on patient outcomes ([Romano 2004](#); [DeVore 2016](#); [Joynt 2016](#)). Two used Medicare claims data, and so confined their analyses to the Medicare population, i.e. those aged 65 years or older ([DeVore 2016](#); [Joynt 2016](#)). [DeVore 2016](#) undertook an interrupted time series to study the effect on 30-day re-admissions, of publicly reporting risk-adjusted hospital re-admission rates for patients with selected conditions (acute my-

ocardial infarction, heart failure, and pneumonia) on the Hospital Compare website. [Joynt 2016](#) reported an interrupted time series with a similar study design to [DeVore 2016](#), but examined the impact on mortality rates, of public reporting of mortality (for patients with the same three selected conditions) on the Hospital Compare website. They used hierarchical modelling to compare 30-day mortality in the pre- and postreporting periods. The final US study presented an interrupted time series based on the California Hospital Outcomes Project in California and the Cardiac Surgery Reporting System in New York ([Romano 2004](#)). This study evaluated the effects of publishing report cards on trends in hospital volumes for specific diagnoses, i.e. coronary artery bypass surgery mortality in New York, and both acute myocardial infarction and postdissection complications in California. The interrupted time series examined hospital case volumes, determined using administrative data sets in each state (the California Patient Discharge Data Set and the New York Statewide Planning and Research Cooperative System) before and after the publication of reports that identified hospitals as performance outliers. These reports were published by the California Hospital Outcomes Project and the New York Cardiac Surgery Reporting System.

There were three cluster-randomised trials outside the US; one each in Canada ([Tu 2009](#)), the Netherlands ([Ikkersheim 2013](#)), and China ([Zhang 2016](#)). In Canada, [Tu 2009](#) evaluated the public release of performance data about 12 care quality indicators for acute myocardial infarction and six for congestive heart failure in 86 hospitals. Participating hospitals were randomised to either early (January 2004) or delayed (September 2005) publication of performance report cards. The performance data were provided to individual hospitals, and then publicised both online and through popular media, with coverage achieved through television, radio, and newspapers. The outcomes reported by this study were any change in hospital performance, measured using the 18 care quality indicators. The cluster-randomised trial in the Netherlands randomised 26 GPs to receive either individualised hospital report cards (65.4%), or to a control group (34.6%) that did not receive this information ([Ikkersheim 2013](#)). The study then captured individual patient referrals (for breast cancer, cataract surgery, and hip or knee replacement) to one of four hospitals in the region, using an electronic referral system. [Zhang 2016](#) undertook a cluster-randomised trial in Hubei Province, south central China. They matched 20 primary care providers within a single city, based on similar organisational characteristics. In this matched-pair cluster-randomised trial, half the providers were randomised to public reporting of injection prescribing, by way of league tables that were posted on outpatient bulletin boards. Performance data were also disseminated to both local health authorities and the leaders of hospitals in the intervention group. The outcomes were the percentage of prescriptions requiring antibiotics, percentage requiring intravenous antibiotics, and the average expenditure per prescription.

Finally, a single interrupted time series study was undertaken in

Seoul, South Korea by [Jang 2011](#). In this study, the intervention was public release of data (online and in media releases) about caesarean section rates for 1194 institutions across the country. These rates were publicised as part of a series of public releases, which were not described in detail. The outcome was change in risk-adjusted institutional caesarean section rates over the whole study period, and after each public release of data.

Excluded studies

In total, we excluded 38 studies after assessing full copies of the papers. The main reasons for exclusion were: ineligible study design (24), interventions did not contain process measures, health care outcomes, structure measures, consumer or patient experiences, expert- or peer-assessed measures (8), no objective outcome data were recorded or available for one or both arms (3), the study

was about hypothetical choices (3). We listed selected studies that readers might reasonably have expected to find included in this review in the '[Characteristics of excluded studies](#)' table.

Risk of bias in included studies

The included studies were rated on different risk of bias items as appropriate for each study design (randomised trial, non-randomised trial, controlled before-after, or interrupted time series). We described this in [Assessment of risk of bias in included studies](#), but in summary, we rated randomised trials, non-randomised trials, and controlled before-after studies using the same nine criteria, and used seven criteria for interrupted time series studies. We showed the results of these risk of bias assessments in the '[Characteristics of included studies](#)' tables and summarised them in both [Figure 2](#) and [Figure 3](#).

Figure 2. Risk of bias graph: review authors' judgements about each risk of bias item, presented as percentages across all included studies. The blank spaces represent risk of bias criteria that were not applicable to the study design.

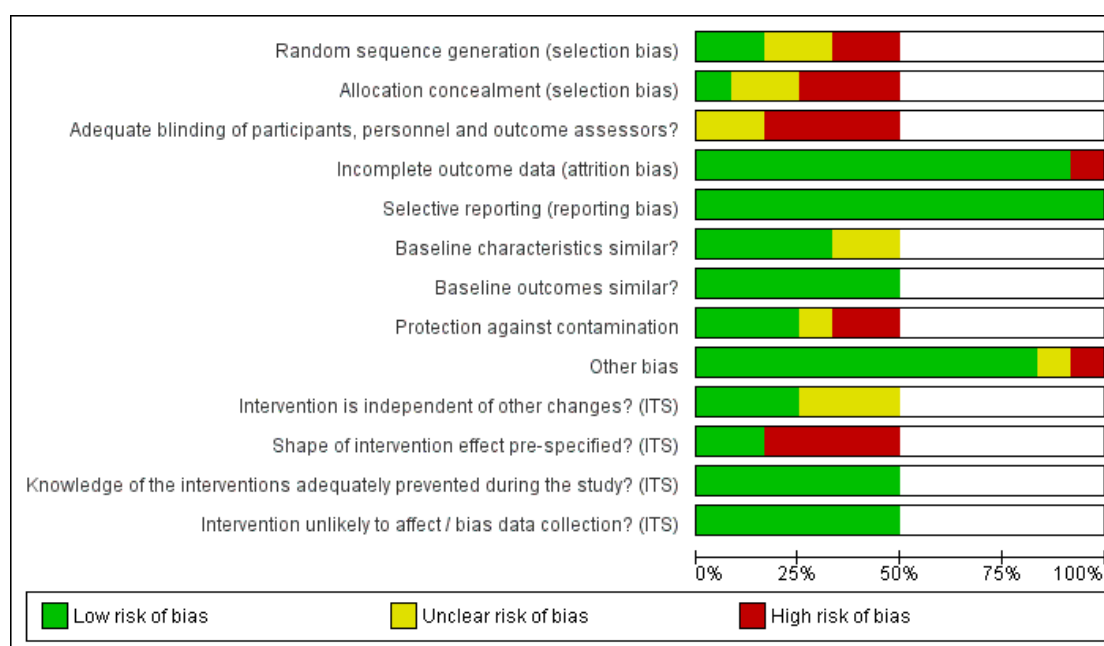


Figure 3. Risk of bias summary: review authors' judgements about each risk of bias item for each included study. The blank cells represent risk of bias criteria that were not applicable to the study design.

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Adequate blinding of participants, personnel and outcome assessors?	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Baseline characteristics similar?	Baseline outcomes similar?	Protection against contamination	Other bias	Intervention is independent of other changes? (ITS)	Shape of intervention effect pre-specified? (ITS)	Knowledge of the interventions adequately prevented during the study? (ITS)	Intervention unlikely to affect / bias data collection? (ITS)
DeVore 2016				+	+				+	?	-	+	+
Farley 2002a	?	?	?	+	+	?	+	+	+				
Farley 2002b	-	-	?	+	+	?	+	+	+				
Flett 2015				+	+				?	+	-	+	+
Ikkersheim 2013	?	?	-	+	+	+	+	+	+				
Jang 2011				+	+				+	?	+	+	+
Joynt 2016				+	+				+	+	-	+	+
Liu 2017				+	+				+	+	-	+	+
Rinke 2015	-	-	-	+	+	+	+	-	+				
Romano 2004				+	+				-	?	+	+	+
Tu 2009	+	+	-	-	+	+	+	-	+				
Zhang 2016	+	-	-	+	+	+	+	?	+				

Allocation

The extent of possible selection bias due to the random sequence generation process was unclear in two studies, because the precise method of random sequence generation was not described (Farley 2002a; Ikkersheim 2013). Two studies were at high risk, as Rinke 2015 was a controlled before-after study, and Farley 2002b was a cluster-non-randomised trial, and so used a non-random method of sequence generation. We judged risk of selection bias as low for Zhang 2016 who 'flipped a coin to randomly assign' paired primary care institutions, and Tu 2009 who employed a dedicated study statistician to implement a stratified randomisation process. We made the same judgements for allocation concealment as for random sequence generation, except for Zhang 2016, which was judged to be at high risk for allocation concealment given their use of a coin flip.

Blinding

Although hospitals and healthcare providers could not be blinded to their allocated groups, individual participants were unlikely to have been aware that a study was taking place. No study explicitly contacted individual patients or members of the public to inform them about the research question, intervention, or measured outcomes. For this reason, two studies were considered to be at unclear risk, as it was not stated whether individuals in those trials were informed that a study was taking place (Farley 2002a; Farley 2002b). Four studies were at high risk, because providers were likely to know that a study was taking place, and it was not possible to blind them to their group allocation (Tu 2009; Ikkersheim 2013; Rinke 2015; Zhang 2016).

Incomplete outcome data

We judged 11/12 included studies to be at low risk of attrition bias, because these studies based their outcomes on routinely collected administrative data, e.g. electronic prescriptions or hospital referrals. Only Tu 2009 was judged to be at high risk of bias, because five randomised hospitals withdrew due to resource constraints; one after randomisation and four during follow-up. Although only a small proportion (5.8%) of the hospitals randomised in this cluster-randomised trial withdrew, it is plausible that poorly performing institutions would be more likely to withdraw than those with average or high performance.

Selective reporting

Only Tu 2009 registered a trial protocol with ClinicalTrials.gov (NCT00187460) in advance of undertaking the study. All outcomes described in this protocol were presented in the final report, which also included all-cause mortality as an additional outcome.

Therefore, we judged it to be at low risk of reporting bias. Although Zhang 2016 presented a trial protocol, this was published in March 2015, eighteen months after the intervention began in October 2013. None of the remaining ten studies registered a protocol in advance of randomisation (randomised and non-randomised trials) or data analysis (interrupted time series and controlled before-after series).

Other potential sources of bias

As outlined in the 'Assessment of risk of bias in included studies' section, the four cluster-randomised trials (Farley 2002a; Tu 2009; Ikkersheim 2013; Zhang 2016), cluster-non-randomised trial (Farley 2002b), and controlled before-after study (Rinke 2015), were assessed for bias in terms of baseline characteristics, baseline outcome measures, and protection against contamination. In addition, we assessed these sources of bias for the six interrupted time series studies: intervention is independent of other changes, shape of the intervention is prespecified, intervention is unlikely to affect data collection, and knowledge of the allocated interventions is adequately prevented during the study (Romano 2004; Jang 2011; Flett 2015; DeVore 2016; Joynt 2016; Liu 2017).

Baseline characteristics

We considered four studies to be at low risk of bias for baseline characteristics because the intervention and control groups were shown to be similar (Tu 2009; Ikkersheim 2013; Rinke 2015; Zhang 2016). Two studies did not report baseline characteristics, and we considered them to be at unclear risk of bias (Farley 2002a; Farley 2002b).

Baseline outcome measures

All six interrupted time series studies presented baseline outcome measures that differed between the intervention and control groups. However, all six also used appropriate statistical techniques, including multivariable regression (Farley 2002b; Tu 2009; Ikkersheim 2013; Rinke 2015; Zhang 2016), and difference-in-differences analyses (Tu 2009; Rinke 2015; Zhang 2016) to account for differences in baseline between the groups. They were therefore all considered to be at low risk of bias from this source.

Protection against contamination

We judged three studies to be at low risk of contamination, either because they randomised healthcare professionals (Ikkersheim 2013), or because their intervention was sent by post, and so unlikely to reach individuals in the control group (Farley 2002a; Farley 2002b).

We assessed two studies to be at high risk. The authors of [Tu 2009](#) stated that several hospitals in the delayed feedback group reported that they also initiated quality improvement activities after becoming aware that performance measures were due to be released publicly. As this was not quantified, it was difficult to determine the degree to which hospitals in the control group modified their activities in anticipation of having to publicly release performance data. We also assessed [Rinke 2015](#) at high risk because hospitals in states that did not mandate healthcare-associated infection reporting might still have modified their practice, given that such laws were being introduced elsewhere in the USA.

We judged [Zhang 2016](#) to be at unclear risk, because no specific efforts were taken to protect against contamination. However, it is not certain that their intervention (posters on bulletin boards in outpatient areas of intervention organisations) would necessarily have influenced behaviour in control institutions.

Intervention independent of other changes

In three interrupted time series studies, it was unclear whether the intervention occurred independently of other changes over time, or whether the outcome was influenced by other confounding variables and events during the study period ([Romano 2004](#); [Jang 2011](#); [DeVore 2016](#)). We judged the remaining three interrupted time series studies to be at low risk of bias. In the two studies that examined public reporting of healthcare-associated infections, this was because they analysed data from a number of states that introduced legislation at different times ([Flett 2015](#); [Liu 2017](#)). We judged [Joynt 2016](#) to be at low risk, because they did not demonstrate a substantial change in the postintervention period, so this was unlikely to be attributable to other factors.

Shape of intervention effect prespecified

Two interrupted time series studies prespecified the shape of the intervention effect, so we assessed both to be at low risk of bias in this domain ([Romano 2004](#); [Jang 2011](#)). The remaining four interrupted time series studies did not, and we judged them to be at high risk.

Knowledge of the allocated interventions adequately prevented during the study

All six interrupted time series studies reported objective outcome measures, so we judged them to be at low risk of bias for this domain.

Intervention unlikely to affect data collection

The intervention was unlikely to affect data collection in any of the six interrupted time series studies, as all were undertaken retrospectively, using routinely collected data. In all cases, the methods

of data collection were the same before and after the intervention. Therefore, we judged all six studies to be at low risk of bias.

Effects of interventions

See: [Summary of findings for the main comparison Public reporting of performance data versus no public reporting](#)

The studies included in this review used a wide range of different interventions, which are described in the 'Characteristics of included studies' tables. We presented the effect sizes reported by each outcome and study in [Table 2](#), [Table 3](#), [Table 4](#), and [Table 5](#), together with the relative effects, for ease of comparison between different study designs and outcome measures. We also provided a 'Summary of findings' table, together with our decisions on how we determined levels of certainty ([Summary of findings for the main comparison](#); [Appendix 2](#)).

Primary outcomes

Changes in healthcare utilisation by consumers

This review provided an indication of the likely effect of public release of performance data on healthcare utilisation by consumers. There was low-certainty evidence from three studies that public release of performance data may make little or no difference to long-term healthcare utilisation by consumers. Two studies included data from over 18,294 insurance beneficiaries ([Farley 2002a](#); [Farley 2002b](#)), and it was unclear how many consumers were analysed by [Romano 2004](#).

There was low-certainty evidence from one study that public release of performance data can lead to small and transient effects on healthcare utilisation behaviour by consumers ([Romano 2004](#)). This study analysed hospital patient volumes following implementation of the California Hospital Outcomes Project, which classified acute hospitals as better, worse or neither better nor worse than expected, based on the adjusted-mortality of patients with acute myocardial infarction, or undergoing disectomy. They found that hospitals, which were high performing for adjusted mortality from acute myocardial infarction, received higher volumes of acute myocardial infarction than expected in the third and fourth quarters after publication of the California Hospital Outcomes Project, although there was no measurable effect in the early period following publication. Similarly, inconsistent trends were observed amongst disectomy patients; the only reported association was between high performing (low complication) hospitals and volume of patients with lumbar disectomy. However, this effect size was very small (less than one additional patient per month per hospital), and so may not have been an important effect. Performance data from New York was released as part of the Cardiac Surgery Reporting System. [Romano 2004](#) analysed Cardiac Surgery Reporting System data from New York, and found that high performing (low mortality) hospitals received a higher number of cases in the

month following publication of a report (74.5 actual cases versus 61.1 expected). In the six months following designation as a high performance outlier, hospitals admitted 24 (22%) additional patients for coronary artery bypass surgery, and within two months after designation as a low performance outlier, hospitals treated 11 (16%) fewer patients. However, all volume effects had disappeared within three months of data publication.

There was low-certainty evidence that suggested that public release of performance data might effect the behaviour of specific subgroups. For example, [Farley 2002b](#) reported that the subgroup of enrollees who actually read the Consumer Assessment of Healthcare Providers and Systems report chose plans with higher standardised Consumer Assessment of Healthcare Providers and Systems ratings than those in the control group (2.58 versus 1.81, $P < 0.01$). Similarly, [Romano 2004](#) found that the only detectable changes in hospital volume were among patients undergoing coronary artery bypass grafting in New York, and this change was entirely driven by patients who identified as 'white and other race'. They did not find evidence that black or Hispanic patient volumes were affected by designating a hospital as a high coronary artery bypass graft mortality outlier.

It is possible that restrictions on patient choice might act as an effect modifier ([Aggarwal 2017](#); [Moscelli 2017](#)). However, the interventions in [Farley 2002a](#) and [Farley 2002b](#) were presented as 'true' choices, since new insurance beneficiaries should not have been limited by concerns around cost and distance. Similarly, [Romano 2004](#) studied hospital choice amongst elective surgical populations seeking treatment at hospitals within a single city.

Changes in healthcare decisions taken by healthcare providers (professionals and organisations)

This review provides some indication of the likely effect of public release of performance data on decision making by healthcare professionals. There was low-certainty evidence from four studies that public release of performance data may make little or no difference to decisions taken by healthcare professionals. These studies included three million births ([Jang 2011](#)), and 67 healthcare providers ([Ikkersheim 2013](#); [Flett 2015](#); [Zhang 2016](#)).

Two studies reported modest effects on some outcomes. [Ikkersheim 2013](#) did not find any clear affect on referral patterns following public release of data about cataract surgery, or hip and knee replacement. However, there was a small effect on referrals for breast cancer, with general practitioners in the intervention group referring 1.0% more cases ($P = 0.01$) to hospitals per incremental percentage point on the report card scale of medical effectiveness. Similarly, [Zhang 2016](#) found that the effect of displaying prescription performance data in outpatient areas varied across outcomes and disease groups. Public release of performance data did not change the number of prescriptions containing antibiotics in the bronchitis group, two or more antibiotics in the gastritis group, injections in the hypertension group, or antibiotic injections

in the bronchitis and hypertension groups. Similarly, the average prescription cost did not change for patients with hypertension. However, public release of performance data did appear to reduce prescriptions containing antibiotics for gastritis (intervention effect -12.7%, $P < 0.001$), two or more antibiotics for gastritis (-3.8%, $P = 0.005$), injections for gastritis (-10.6%, $P < 0.001$), and antibiotic injections for gastritis (-10.7%, $P < 0.001$). Average antibiotic prescription cost fell for patients with bronchitis (-7.9%, $P < 0.001$) and gastritis (-5.7%, $P = 0.005$). These mixed findings were also complicated by evidence that public release of prescribing data increased prescriptions containing antibiotics for patients with hypertension (intervention effect 2.0%, $P = 0.08$), and injections for bronchitis (2.0%, $P = 0.012$).

One study found that the first public release of hospital caesarean section rate data may have slightly reduced the number of patients undergoing this procedure (-0.8%, $P < 0.01$), and that this persisted until the end of the study, 20 months later. However, further public releases of data did not exhibit any further effect on caesarean section rates ([Jang 2011](#)).

Finally, [Flett 2015](#) did not find any evidence that mandatory public reporting of central line-associated bloodstream infections had any effect on blood culture testing or antibiotic utilisation in paediatric and neonatal intensive care units in the United States.

Changes in the healthcare utilisation decisions of purchasers

We found no evidence on the effect of public release of performance data on this outcome.

Changes in provider performance

This review provides some indication of the likely effect of public release of performance data on healthcare provider performance. There was low-certainty evidence from one study that public release of performance data may make little or no difference to objective measures of provider performance. [Tu 2009](#) included data from 82 healthcare providers.

[Tu 2009](#) found that a media campaign and release of hospital performance data online had no effect on 11 of 12 acute myocardial infarction process-of-care quality indicators. The twelfth acute myocardial infarction quality indicator (fibrinolytics given prior to transfer to the Coronary Care Unit or Intensive Care Unit) increased by 5.8% ($P = 0.02$), although no statistical correction was made for multiple hypothesis testing. Similarly, public release of performance data did not clearly effect five of six congestive heart failure quality indicators, although the sixth (Angiotensin-Converting Enzyme (ACE) inhibitor or Angiotensin Receptor Blocker (ARB) for left ventricular dysfunction) increased by 5.9% ($P = 0.02$). Neither the acute myocardial infarction nor congestive heart failure composite process-of-care quality indicators improved following the public release of performance data.

The main outcomes in two studies described above, are sometimes considered evidence of provider performance ([Jang 2011](#); [Zhang](#)

2016). However, as these outcomes (caesarean section and antibiotic prescribing) may be appropriate clinical decisions, they are not direct evidence of poor performance, so we have considered them under 'Changes in healthcare decisions taken by healthcare providers (professionals and organisations)' instead of 'Provider performance'.

Changes in patient outcome

Low-certainty evidence from five studies suggested that public release of performance data may slightly improve patient outcomes. We graded the certainty as low, because the evidence was mixed, with two studies reporting improvements (Tu 2009; Liu 2017), and three finding no evidence of improved patient outcomes (Rinke 2015; DeVore 2016; Joynt 2016). These five studies included 7503 healthcare providers and 315,092 hospitalisations. Two studies reported that patient outcomes were not changed by publication of hospital-level quality metrics on Hospital Compare, which is a website run by the Centers for Medicare & Medicaid Services. DeVore 2016 did not find any evidence that publication of hospital re-admission rates had an effect on 30-day re-admissions for patients with myocardial infarction, heart failure, or pneumonia. Similarly, Joynt 2016 reported a very small slowing in a pre-existing trend (change 0.13% per quarter; 95% CI 0.12% to 0.14%) towards reduced 30-day mortality following publication of mortality rates on Hospital Compare.

Rinke 2015 did not find any evidence that mandatory hospital reporting of central line-associated blood stream infections had any effect on the rate of paediatric central line-associated bloodstream infections. However, Liu 2017 reported a 34% reduction (incidence rate ratio 0.66, $P < 0.001$) in adult central line-associated bloodstream infections after mandatory reporting, when compared with the 25-month period before each state introduced legislation. This discrepancy between the findings of Rinke 2015 and Liu 2017 might reflect a genuine difference in terms of impact on children and adult central line-associated bloodstream infection rates. Importantly, both studies found that central line-associated bloodstream infection rates declined across the USA during their study period, including in states that did not introduce mandatory reporting. It is unclear whether public release of performance data in some states contributed to this national decline, even within states that did not introduce mandatory reporting. Tu 2009 found that public release of hospital performance data online and through the media was associated with a 2.5% reduction in 30-day mortality ($P = 0.045$) for patients with acute myocardial infarction, although no such effect was observed in patients with congestive heart failure.

Changes in staff morale

We found no evidence on the effect of public release of performance data on this outcome.

Secondary outcomes

Unintended and adverse effects or harms

We found no evidence on the effect of public release of performance data on this outcome.

Impact on equity

Low-certainty evidence from one study suggested that public release of performance data may have different effects on advantaged and disadvantaged populations (Romano 2004). As described in 'Changes in healthcare utilisation by consumers', this study reported that patients who identified as white and other race in New York might have been influenced by publicly released hospital mortality rates when choosing a hospital in which to undergo coronary artery bypass grafting. However, this same effect was not observed in black or Hispanic patients undergoing the same procedure at hospitals in New York.

Other outcome measures

Two studies reported on awareness, knowledge of performance data, attitude, and cost data (Farley 2002b; Ikkersheim 2013). Farley 2002b reported secondary outcomes as a result of a survey, although this was disseminated using a 3:1 ratio, and the results were further complicated by low response rates. Ikkersheim 2013 undertook semi-structured interviews with 17 GPs but these were largely focused on the specific intervention (report cards) and the findings were poorly reported. Therefore, we decided to exclude these results, and did not report these outcomes.

DISCUSSION

Summary of main results

Changes in healthcare utilisation by consumers

Changes in healthcare utilisation are one of the two key ways in which public release of performance data might improve healthcare quality (Berwick 2003). However, only three studies addressed the impact on healthcare utilisation decisions by consumers (Farley 2002a; Farley 2002b; Romano 2004). We judged that they provided low-certainty evidence of little or no effect. There were consistent results from two studies that showed some consumers may engage with published performance data, and change their healthcare choices accordingly; this group was too small to register an effect in the population as a whole (Farley 2002b; Romano 2004).

Changes in healthcare decisions taken by healthcare providers (professionals and organisations)

There was low-certainty evidence with mixed findings from four studies, which reported either modest effects (Jang 2011; Ikkersheim 2013; Zhang 2016), or no effect (Flett 2015), on healthcare decisions taken by healthcare providers. Two studies found evidence that public release of performance data had modest effects on some of the healthcare decisions taken by healthcare providers, but not all of the decisions measured (Ikkersheim 2013; Zhang 2016). One study found that the first public release of data had a small but sustained effect on caesarean rates, but that subsequent releases did not affect the rate any further (Jang 2011).

Changes in provider performance

There was low-certainty evidence from one study that informed conclusions about the effect of public release of performance data on provider performance. A single randomised trial addressed this question, and found that 2/18 (11.1%) of measured processes appeared to improve in the intervention hospitals (Tu 2009). However, as no correction was made for multiple hypothesis testing (Bender 2001), this did not provide convincing evidence that provider performance was affected by public release of performance data.

Changes in patient outcome

Low-certainty evidence showed that five studies that included patient outcomes had inconsistent findings, with two reporting improvements (Tu 2009; Liu 2017), and three reporting no difference (Rinke 2015; DeVore 2016; Joynt 2016).

Impact on equity

Only one study undertook a subgroup analysis to identify differential effects of public release of performance data (Romano 2004). Low-certainty evidence from one study reported that white and other race patients, undergoing coronary artery bypass grafting in New York, may have been influenced by publicly released mortality rates. However, this finding was not reproduced among black and Hispanic patients. Although Farley 2002b did not study equity directly, their finding that only consumers who read the Consumer Assessment of Healthcare Providers and Systems report were influenced, raises the possibility that some groups (e.g. those with greater rates of literacy) might be preferentially influenced by public release of performance data.

Other outcomes

There were no studies that considered the effect of public release of performance data on changes in the healthcare utilisation decisions of purchasers, changes in staff morale, or adverse effects.

Two studies reported on awareness, knowledge of performance data, attitude, and cost data but we did not include the data due to concerns about reporting and high attrition bias.

Overall completeness and applicability of evidence

There are many systems around the world that include public release of performance data. However, only a small proportion were represented in this review, so it is likely that most have either not been evaluated, or were subject only to low-quality studies. It is notable that some interventions have been evaluated more robustly than others, with two studies in this review considering the Centers for Medicare & Medicaid Services website Hospital Care (DeVore 2016; Joynt 2016), and three, the introduction of state-based mandatory reporting of central line-associated blood stream infections (Flett 2015; Rinke 2015; Liu 2017). Similarly, the majority of the studies included in this review (9/12, 75%) were based in North America, with no representation from South America, Africa, or Australasia. Therefore, it is likely that a small number of initiatives have attracted a disproportionate number of studies, and there is clearly work that needs to be done to robustly evaluate similar interventions in other settings. There was also insufficient evidence to draw any conclusions about the healthcare utilisation decisions of purchasers, staff morale, or adverse effects. The applicability of the evidence was also limited by considerable heterogeneity in interventions. For example, it was possible that the freedom of patients to choose healthcare providers was curtailed in some cases, which might have acted as an effect modifier that explains some of the differences in findings between included studies. However, only three studies included interventions that might lead to improved consumer selection, and consumer choice would not obviously have been restricted by considerations around distance and cost (Farley 2002a; Farley 2002b; Romano 2004). These studies suggested that those engaging with publicly reported performance data (Farley 2002b), and those from privileged backgrounds (Romano 2004), might be more likely to modify their choice of healthcare provider. This raises the possibility that lack of education and health literacy might restrict patient choice, and act as an effect modifier in some cases.

The three studies that took place in the USA involved only a small proportion of the numerous major reporting systems available. We included one new study from Canada, which was published after the latest systematic reviews by Fung 2008, Shekelle 2008, and Faber 2009 (Tu 2009). We excluded many of the more recent studies, because they did not have a rigorous study design, or did not report the defined primary outcome measures. The studies we included evaluated interventions that used data that might have been originally collected for a purpose other than influencing behaviour or improving outcomes. It is possible that custom-made interventions, using data collected for the specific purpose of influencing behaviour or improving outcomes, would have a greater

impact. However, the lack of such interventions in the literature highlighted the fact that their delivery may be excessively resource intensive, and that future initiatives aimed at public release of performance data will continue to draw on data initially collected for a different purpose.

Despite evidence that secondary outcome measures (e.g. awareness, attitude, knowledge of performance data) are crucial, since public reporting can only change behaviour if the target population (healthcare consumers, providers, or purchasers of care) understands the information, these measures were lacking in the included studies (Hibbard 2010). Therefore, it was difficult to explain the lack of effect. For example, Faber 2009 found that the effect of performance data was higher for those who understand the information, which might be consistent with the evidence from Farley 2002b. Damman 2011 showed that comparative performance information was complex, and consumers had difficulties in interpreting and using performance data. However, it is notable that this review did not find that healthcare providers (who might be in a better position to interpret such data) were necessarily influenced more than consumers.

Certainty of the evidence

We deemed the certainty of the evidence that examined the effect of public release of performance data on a number of outcomes to be low. These outcomes were:

- Changes in healthcare utilisation by consumers;
- Changes in healthcare utilisation by providers (organisations and professionals); and
- Changes in patient outcome.

Only 4/12 included studies (33.3%) were randomised trials, so the evidence for these outcomes was partly informed by non-randomised study designs. However, the use of EPOC study design criteria ensured that all included observational studies took considerable steps to minimise the risk of bias (EPOC 2013). There was also considerable heterogeneity in the settings, outcomes, and modes of public release, and inconsistent effects reported between studies.

We also judged the certainty of the evidence that examined the effect on changes in provider performance to be low. Although this outcome was informed by a single randomised trial, we had concerns about risk of bias in the following items: (1) allocation concealment, (2) adequate blinding of participants, personnel and outcome assessors, and (3) protection against contamination (Tu 2009). It is also uncertain whether the findings of a single randomised trial, in a narrowly defined patient group, within one region of Canada, can be generalised to other settings.

Due to lack of evidence, we were unable to draw any conclusions about the following primary outcomes:

- Changes in the healthcare utilisation decisions of purchasers;
- Changes in staff morale.

In terms of secondary outcomes, there were no studies that set out to consider adverse effects or harms. We deemed the evidence for any potential impact on equity to be low, as it was based on a subgroup analysis from a single interrupted time series study (Romano 2004).

Potential biases in the review process

Although our search was comprehensive, we could not exclude the possibility of having missed relevant studies. However, we minimised this risk by asking an Information Specialist to help design and implement the search strategy, and ensured that two review authors independently examined all items retrieved from our search. We also ensured that data extraction and 'Risk of bias' assessments were independently undertaken by two review authors. Although the GRADE assessments were determined by a single author (DM), these were checked by all review authors, and disagreements resolved through discussion. These steps ensured that potential biases in the review processes were mitigated as much as possible. However, this stringent approach to study collection also meant excluding most of the studies that have evaluated public release of performance data in other settings, and using a range of study designs. It was possible that this approach biased our review against settings that were less likely to deliver studies that satisfied the EPOC inclusion criteria, and this might have accounted for the over-representation of studies from North America, Europe, and Asia. It might also have led to the exclusion of studies (e.g. those utilising qualitative designs) that contained important information about the impact of public release of performance data. However, it was necessary to limit our review to studies that were at the lowest possible risk of bias, to maximise the certainty of its findings. There may nevertheless be scope for future reviews to synthesise evidence from studies using a broader range of designs.

Agreements and disagreements with other studies or reviews

Our systematic literature search and a further PubMed search of studies citing an earlier version of this review (Ketelaar 2011), identified three relevant systematic reviews (Fung 2008; Faber 2009; Campanella 2016). Our review agreed with these earlier publications that previous studies were limited by risk of bias, inconsistent findings, and heterogeneity of interventions, healthcare settings, and outcomes.

Faber 2009 considered public release of performance data on consumer choice, and concluded that there was only evidence to support an effect on the small subgroup of participants that actively engaged with the published performance data. This was consistent with our findings, and those of Fung 2008.

Campanella 2016 attempted a meta-analysis of data from ten studies, and reported improved mortality (risk ratio 0.85, 95% confi-

dence interval 0.79 to 0.92). However, this finding was reported in the context of very high heterogeneity ($P < 0.0001$; $I^2 = 100\%$). The authors limited their meta-analysis to studies that reported sufficient data, and excluded those with inappropriate study designs, or those that were at high risk of bias. Our review only considered studies that proffered the highest certainty of evidence, and did not consider a meta-analysis appropriate in view of the considerable degree of heterogeneity between studies (see [Assessment of heterogeneity](#)). Instead, our findings were consistent with those of [Fung 2008](#), which concluded that “studies of the effect of public reporting on outcomes provide mixed signals, and the usefulness of public reporting in improving patient safety and patient-centeredness remains unknown, because few studies assessed these end points”.

AUTHORS' CONCLUSIONS

Implications for practice

The existing evidence base on the effects of public release of performance data on changing behaviour of healthcare decision makers was inadequate to directly inform practice.

Implications for research

In order to understand the effectiveness of the public release of performance data, we need more longitudinal studies with robust evaluation designs. In particular, the evidence base would benefit from more studies that consider whether public release of performance data can improve patient outcomes, rather than simply healthcare processes. In this review, only one of the included studies reported data on patient outcomes ([Tu 2009](#)). Further work should also specifically consider whether public release of performance data might result in adverse effects or harms.

Unfortunately, most studies were unable to guarantee that disseminated performance data actually reached its intended audience, i.e. that lack of effect was not simply a result of failed exposure to the intervention. Importantly, [Farley 2002b](#) reported evidence to suggest that the subgroup of patients that read the reports sent by post were influenced when choosing a health insurance programme. Therefore, future studies should consider carefully how they might maximise the number of people exposed to their intervention, and whether this can be quantified. However, the effect of public release of performance information in the 'real world' is likely to be limited by difficulties in reaching its intended audience ([Hibbard 2007](#); [Damman 2010](#); [Aggarwal 2017](#); [Canaway 2017](#); [Moscelli 2017](#); [Canaway 2018](#); [Greenhalgh 2018](#)). Therefore, the need to ensure that performance data reach those who are intended to be influenced, needs to be balanced against the risk of reducing study validity by creating artificial conditions that cannot be replicated when the intervention is used in practice.

Berwick's model suggests that public release of performance data may improve quality of care by means of a pathway of change or selection ([Berwick 2003](#)). The studies we included focused exclusively on either one or the other of these pathways. In addition, one assumption underlying public release of performance data is that provider choice is a rational decision, i.e. consumers prefer the healthcare provider or health plan that is rated as the best. However, there is little evidence to confirm this assumption ([Faber 2009](#); [Kolstad 2009](#)), although a number of other factors are known to influence consumer choice, e.g. established relationships with local physicians, health plans ([Schwartz 2005](#); [Hibbard 2009](#)), hospitals, distance, and opinions of friends, and family ([Harris 2008](#); [The King's Fund 2010](#)). Similarly, [Ikkersheim 2013](#) found that decisions taken by healthcare professionals were often informed by their personal preferences, experience of, and communication with other providers, and personal relationships with other professionals. These factors influenced hospital referral decisions even when professionals were provided with objective performance data. Future studies may wish to consider the mechanism(s) by which public release of performance data can effect change, as well as whether such changes can be demonstrated empirically.

ACKNOWLEDGEMENTS

The author team are grateful for the assistance of Sasha Shepperd (UK Co-ordinating Editor), Gillian Leng and Luciana Ballini (Editors), Julia Worswick (Managing Editor), and Paul Miller (Information Specialist), all of whom work at the Cochrane Effective Practice and Organisation of Care (EPOC) group. We are also grateful to Vicki Pennick at the Cochrane Editorial Unit for copy editing support.

We would like to acknowledge the authors of the original protocol (Phil Alderson and Sandy Oliver), who formulated the idea for this review in 2003, and the authors of a previous version of this review (Marjan Faber, Liv Rygh, Katherine Deane, and Martin Eccles) published in 2011 ([Ketelaar 2011](#)). We would also like to acknowledge the following contributors to the earlier version of this review: Craig Ramsay (EPOC Statistical Editor), Jan Ogaard-Jensen (Norwegian Knowledge Centre for the Health Services), Fiona Beyer (Newcastle University), and Alice Tillema (Radboud University Nijmegen Medical Centre). Finally, we acknowledge the reviewers who have contributed to the development of this review: Donna Farley, Denise O'Connor, Phil Anderson, Andrew Rix, Federica Davolio, Faiza Coleman-Sala, Newton Opiyo, Gregg M Gascon, and Cristobal Cuadrado.

Support was provided by the National Institute for Health Research via Cochrane Infrastructure funding to the Effective Practice and Organisation of Care Group. The views and opinions expressed therein are those of the authors and do not necessarily

reflect those of the Systematic Reviews Programme, NIHR, NHS or the Department of Health.

REFERENCES

References to studies included in this review

DeVore 2016 *{published data only}*

DeVore AD, Hammill BG, Hardy NC, Eapen ZJ, Peterson ED, Hernandez AF. Has public reporting of hospital readmission rates affected patient outcomes? Analysis of Medicare claims data. *Journal of the American College of Cardiology* 2016;**67**(8):963–72. PUBMED: 26916487]

Farley 2002a *{published data only}*

* Farley DO, Elliott MN, Short PF, Damiano P, Kanouse DE, Hays RD. Effect of CAPHS Performance Information on health plan choices by Iowa Medicaid. *Medical Care Research and Review* 2002;**59**(3):319–36. PUBMED: 12205831]

Farley 2002b *{published data only}*

* Farley DO, Short PF, Elliott MN, Kanouse DE, Brown JA, Hays RD. Effect of CAPHS health plan performance information on plan choices by New Jersey. *Health Services Research* 2002;**37**(4):985–1007. PUBMED: 12205831]

Flett 2015 *{published data only}*

Flett KB, Ozonoff A, Graham DA, Sandora TJ, Priebe GP. Impact of mandatory public reporting of central line-associated bloodstream infections on blood culture and antibiotic utilization in pediatric and neonatal intensive care units. *Infection Control and Hospital Epidemiology* 2015;**36**(8):878–85. PUBMED: 25913602]

Ikkersheim 2013 *{published data only}*

Ikkersheim D, Koolman X. The use of quality information by general practitioners: does it alter choices? A randomized clustered study. *BMC Family Practice* 2013;**14**:95. PUBMED: 23824745]

Jang 2011 *{published data only}*

Jang WM, Eun SJ, Lee C, Kim Y. Effect of repeated public releases on cesarean section rates. *Journal of Preventative Medicine and Public Health* 2011;**44**(1):2–8. PUBMED: 21483217]

Joynt 2016 *{published data only}*

Joynt KE, Orav EJ, Zheng J, Jha AK. Public reporting of mortality rates for hospitalized Medicare patients and trends in mortality for reported conditions. *Annals of Internal Medicine* 2016;**165**(3):153–60. PUBMED: 27239794]

Liu 2017 *{published data only}*

Liu H, Herzig CTA, Dick AW, Furuya EY, Larson E, Reagan J, et al. Impact of state reporting laws on central line-associated bloodstream infection rates in U.S. adult intensive care units. *Health Services Research* 2017;**52**(3): 1079–98. PUBMED: 27451968]

Rinke 2015 *{published data only}*

Rinke ML, Bundy DG, Abdullah F, Colantuoni E, Zhang Y, Miller MR. State-mandated hospital infection reporting

is not associated with decreased pediatric health care-associated infections. *Journal of Patient Safety* 2015;**11**: 123–34. PUBMED: 24681422]

Romano 2004 *{published data only}*

* Romano PS, Zhou H. Do well-publicized risk-adjusted outcomes reports affect hospital volume?. *Medical Care* 2004;**42**(4):367–77. PUBMED: 15076814]

Tu 2009 *{published data only}*

Tu JV, Donovan LR, Douglas SL, Wang JT, Austin PC, Alter DA, et al. Effectiveness of public report cards for improving the quality of cardiac care. The EFFECT study: a randomized trial. *JAMA* 2009;**302**(21):2330–7. PUBMED: 19923205]

Zhang 2016 *{published data only}*

Du X, Wang D, Wang X, Yang S, Zhang X. Exploring the transparency mechanism and evaluating the effect of public reporting on prescription: a protocol for a cluster randomized controlled trial. *BMC Public Health* 2015;**21**(15):277. PUBMED: 25881035]

Liu C, Zhang X, Wan J. Public reporting influences antibiotic and injection prescription in primary care: a segmented regression analysis. *Journal of Evaluation in Clinical Practice* 2015;**21**(4):597–603. PUBMED: 25902726]

* Liu C, Zhang X, Wang X, Wan J, Zhong F. Does public reporting influence antibiotic and injection prescribing to all patients? A cluster-randomized matched-pair trial in China. *Medicine (Baltimore)* 2016;**95**(26):e3965. PUBMED: 27367995]

Tang Y, Liu C, Zhang X. Public reporting as a prescriptions quality improvement measure in primary care settings in China: variations in effects associated with diagnoses. *Scientific Reports* 2016;**6**:39361. PUBMED: 27996026]

Wang X, Tang Y, Zhang X, Yin X, Du X, Zhang X. Effect of publicly reporting performance data of medicine use on injection use: a quasi-experimental study. *PLOS One* 2014;**9**(10):e109594. PUBMED: 25313853]

Yang L, Liu C, Wang L, Yin X, Zhang X. Public reporting improves antibiotic prescribing for upper respiratory tract infections in primary care: a matched-pair cluster-randomized trial in China. *Health Research Policy and Systems* 2014;**12**:61. PUBMED: 25304996]

References to studies excluded from this review

Cavender 2015 *{published data only}*

Cavender MA, Joynt KE, Parzynski CS, Resnic FS, Rumsfeld JS, Moscucci M, et al. State mandated public reporting and outcomes of percutaneous coronary intervention in the United States. *American Journal of Cardiology* 2015;**115**(11):1494–501. PUBMED: 25891991]

Moscucci 2005 {published data only}

Moscucci M, Eagle KA, Share D, Smith D, De Franco AC, O'Donnell M, et al. Public reporting and case selection for percutaneous coronary interventions: an analysis from two large multicenter percutaneous coronary intervention databases. *Journal of the American College of Cardiology* 2005;**45**(11):1759–65. PUBMED: 15936602]

Paris 2013 {published data only}

Paris B, Arahoud T, Asche C, Amundson G. Lessons from voluntary reporting of Illinois hospital employee seasonal influenza vaccination rates (2009-2013). *Value in Health* 2013;**16**(3):A96.

Park 2011 {published data only}

Park J, Konetzka RT, Werner RM. Performing well on nursing home report cards: does it pay off?. *Health Services Research* 2011;**46**(2):531–54. PUBMED: 21029093]

Saratzis 2017 {published data only}

Saratzis A, Thatcher A, Bath MF, Sidloff DA, Bown MJ, Shakespeare J, et al. Reporting individual surgeon outcomes does not lead to risk aversion in abdominal aortic aneurysm surgery. *Annals of the Royal College of Surgeons of England* 2017;**99**(2):161–5. PUBMED: 28071950]

Additional references**Aggarwal 2017**

Aggarwal A, Lewis D, Mason M, Sullivan R, van der Meulen J. Patient mobility for elective secondary health care services in response to patient choice policies: a systematic review. *Medical Care Research and Review: MCRR* 2017;**74**(4):379–403. [PUBMED: 27357394]

Bender 2001

Bender R, Lange S. Adjusting for multiple testing - when and how?. *Journal of Clinical Epidemiology* 2001;**54**(4):343–9. [PUBMED: 11297884]

Berwick 1990

Berwick DM, Wald DL. Hospital leaders' opinions of the HCFA mortality data. *JAMA* 1990;**263**(2):247–9.

Berwick 2003

Berwick D, Jamer B, Coye M. Connections between quality measurements and improvement. *Medical Care* 2003;**41**(1):130–8.

Brook 1994

Brook RH. Health care reform is on the way: do we want to compete on quality?. *Annals of Internal Medicine* 1994;**120**(1):84–6.

Burns 2016

Burns EM, Pettengell C, Athanasiou T, Darzi A. Understanding the strengths and weaknesses of public reporting of surgeon-specific outcome data. *Health Affairs* 2016;**35**(3):415–21.

Campanella 2016

Campanella P, Vukovic V, Parente P, Sulejmani A, Ricciardi W, Specchia ML. The impact of public reporting on clinical outcomes: a systematic review and meta-analysis. *BMC Health Services Research* 2016;**16**:296.

Canaway 2017

Canaway R, Bismark M, Dunt D, Kelaher M. Perceived barriers to effective implementation of public reporting of hospital performance data in Australia: a qualitative study. *BMC Health Services Research* 2017;**17**(1):391. [PUBMED: 28592277]

Canaway 2018

Canaway R, Mismark M, Dunt D, Prang KH, Kelaher M. "What is meant by public?": stakeholder views on strengthening impacts of public reporting of hospital performance data. *Social Science & Medicine* 2018;**202**:143–50. [PUBMED: 29524870]

Chung 2014

Chung SH, Seol HJ, Choi YS, Oh SY, Kim A, Bae CW. Changes in the cesarean section rate in Korea (1982-2012) and a review of the associated factors. *Journal of Korean Medical Science* 2014;**29**(10):1341–52. [PUBMED: 25368486]

Damman 2010

Damman OC, Van den Hengel YK, Van Loon AJ, Rademakers J. An international comparison of Web-based reporting about healthcare quality: content analysis. *Journal of Medical Internet Research* 2010;**13**(12):e8.

Damman 2011

Damman OC, Hendrik M, Rademakers J, Spreeuwenberg P, Delnoij DM, Groenewegen PP. Consumers interpretation and use of comparative information on the quality of health care: the effect of presentation approaches. *Health Expectations* 25 May 2011 Epub ahead of print]. DOI: 10.1111/j.1369-7625.2011.00671.x

EPOC 2013

Effective Practice, Organisation of Care (EPOC). Analysis in EPOC reviews. EPOC resources for review authors. Available at: epoc.cochrane.org/epoc-resources-review-authors (accessed 22 August 2018).

EPOC 2017

Effective Practice, Organisation of Care (EPOC). Reporting the effects of an intervention in EPOC reviews. Available at: epoc.cochrane.org/epoc-resources-review-authors (accessed 22 August 2018).

Faber 2009

Faber M, Bosch M, Wollersheim H, Leatherman S, Grol R. Public reporting in health care: how do consumers use quality of care information? A systematic review. *Medical Care* 2009;**47**(1):1–8.

Fung 2008

Fung C, Yee-Wei L, Soeren M, Damberg C, Shekelle P. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Annals of Internal Medicine* 2008;**148**(2):111–23.

Greenhalgh 2018

Greenhalgh J, Dalkin S, Gibbons E, Wright J, Valderas JM, Meads D, et al. How do aggregated patient-reported outcome measures data stimulate health care improvement? A realist synthesis. *Journal of Health Services Research & Policy* 2018;**23**(1):57–65. [PUBMED: 29260592]

Harris 2008

Harris KM, Beeuwkes Buntin M, The RAND Cooperation. *Research Synthesis Report. Choosing a healthcare provider: the role of quality information*. Princeton: Robert Wood Johnson Foundation, 2008.

Hendriks 2009

Hendriks M, Spreeuwenberg P, Rademakers J, Delnoij DM. Dutch healthcare reform: did it result in performance improvement of health plans? A comparison of consumer experiences over time. *BMC Health Services Research* 2009; **9**:167. DOI: 10.1186/1472-6963-9-167

Hibbard 1997

Hibbard JH, Jewett JJ, Legnini MW, Tusler M. Choosing a health plan: do large employers use the data?. *Health Affairs* 1997; **16**(6):172–80.

Hibbard 2007

Hibbard JH, Peters E, Dixon A, Tusler M. Consumers competencies and the use of comparative quality information: it isn't just about literacy. *Medical Care Research and Review* 2007; **64**(4):379–94.

Hibbard 2009

Hibbard JH. Using systematic measurement to target consumer activation strategies. *Medical Care Research and Review* 2009; **66**(1):9S–27S.

Hibbard 2010

Hibbard JH, Greene J, Daniel D. What is quality anyway? Performance reports that clearly communicate to consumers the meaning of quality of care. *Medical Care Research and Review* 2010; **67**(3):275–93.

Higgins 2011

Higgins JPT, Green S, editor(s). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* (updated March 2011). The Cochrane Collaboration, 2011. Available from handbook.cochrane.org.

Huwaldt 2004 [Computer program]

Huwaldt JA, Steinhurst S, Paul G. Plot Digitizer. Version accessed 27 October 2015. USA: Huwaldt JA; SourceForge, 2004.

Ivers 2012

Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews* 2012, Issue 6. DOI: 10.1002/14651858.CD000259.pub3; PUBMED: 22696318

Kiernan 2015

Kiernan F, Rahman F. Measuring surgical performance: a risky game?. *Surgeon* 2015; **13**(4):213–7.

Kolstad 2009

Kolstad JT, Chernew ME. Quality and consumer decision making in the market for health insurance and health care services. *Medical Care Research and Review* 2009; **66**(Suppl 1):28S–52S.

Liu 2015

Liu C, Zhang X, Wan J. Public reporting influences antibiotic and injection prescription in primary care:

a segmented regression analysis. *Journal of Evaluation in Clinical Practice* 2015; **21**(4):597–603. [PUBMED: 25902726]

Liu 2016

Liu C, Zhang X, Wang X, Zhang X, Wan J, Zhong F. Does public reporting influence antibiotic and injection prescribing to all patients? A cluster-randomized matched-pair trial in China. *Medicine (Baltimore)* 2016; **95**(26):e3965. [PUBMED: 27367995]

Loeb 2004

Loeb JM. The current state of performance measurement in health care. *International Journal for Quality in Health Care* 2004; **16**(Suppl 1):i5–9.

Marshall 2000

Marshall MN, Shekelle PG, Leatherman S, Brook RH. The public release of performance data: what do we expect to gain? A review of the evidence. *JAMA* 2000; **283**(14):1866–74. [PUBMED: 10770149]

Moher 1999

Moher D, Cook DJ, Eastwood S, Olkin I, Drummon R, Stroup DF, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999; **354**(9193):1896–900.

Moscelli 2017

Moscelli G, Siciliani L, Gutacker N, Cookson R. Socioeconomic inequality of access to healthcare: Does choice explain the gradient?. *Journal of Health Economics* 23 June 2017 Epub ahead of print]. DOI: 10.1016/j.jhealeco.2017.06.005; PUBMED: 28935158

Mukamel 1998

Mukamel DB, Mushlin AI. Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports. *Medical Care* 1998; **36**(7):945–54.

Rosenthal 1998

Rosenthal GE, Hammar PJ, Way LE, Shipley SA, Doner D, Wojtala B, et al. Using hospital performance data in quality improvement: the Cleveland Health Quality Choice experience. *Joint Commission Journal of Quality Improvement* 1998; **24**(7):347–60.

Schroll 2011

Schroll JP, Moustgaard R, Gøtzsche PC. Dealing with substantial heterogeneity in Cochrane reviews. Cross-sectional study. *BMC Medical Research Methodology* 2011; **11**:22. [PUBMED: 21349195]

Schut 2005

Schut FT, van de Ven WP. Rationing and competition in the Dutch health-care system. *Health Economics* 2005; **14**(Suppl 1):S59–74.

Schwartz 2005

Schwartz LM, Woloshin S, Birkmeyer JD. How do elderly patients decide where to go for major surgery? Telephone

- interview survey. *BMJ* 2005;**331**(7520):821. DOI: 10.1136/bmj.38614.449016.DE
- Shahian 2017**
Shahian DM, Jacobs JP, Badhwar V, D'Agostino RS, Bavaria JE, Prager RL. Risk aversion and public reporting. Part 1: observations from cardiac surgery and interventional cardiology. *Annals of Thoracic Surgery* 2017;**104**(6): 2093–101. [PUBMED: 29100643]
- Shekelle 2008**
Shekelle PG, Lim Y-W, Mattke S, Damberg C, Southern California Evidence-based Practice Centre, RAND Corporation. *Does public release of performance results improve quality of care? A systematic review*. London: The Health Foundation, 2008.
- Sherman 2013**
Sherman KL, Gordon EJ, Mahvi DM, Chung J, Bentrem DJ, Holl JL, et al. Surgeons' perceptions of public reporting of hospital and individual surgeon quality. *Medical Care* 2013;**51**(12):1069–75.
- Sirio 1996**
Sirio CA, McGee JL. Public reporting of clinical outcomes - the data needs of health care stakeholders. *American Journal of Medical Quality* 1996;**11**(1):S78–81.
- Smith 2009**
Smith PC, Mossialos E, Papanicolas I, Leatherman S. *Performance measurement for health system improvement. Experiences, Challenges and Prospects*. Cambridge: Cambridge University Press, 2009.
- Tang 2016**
Tang Y, Liu C, Zhang X. Public reporting as a prescriptions quality improvement measure in primary care settings in China: variations in effects associated with diagnoses. *Scientific Reports* 2016;**6**:39361. [PUBMED: 27996026]
- The King's Fund 2010**
Dixon A, Roberson R, Appleby J, Burge P, Devlin N, Magee H. *Patient Choice. How patients choose and how providers respond*. London: The King's Fund, 2010.
- Wadhwa 2017**
Wahera RK, Anderson JD, Yeh RW. High-risk percutaneous coronary intervention in public reporting states: the evidence, exclusion of critically ill patients, and implications. *Current Heart Failure Reports* 2017;**14**(6):514–8. [PUBMED: 29101664]
- Wang 2014**
Wang X, Tang Y, Zhang X, Yin X, Du X, Zhang X. Effect of publicly reporting performance data of medicine use on injection use: a quasi-experimental study. *PLoS One* 2014;**9**(10):e109594. [PUBMED: 25313853]
- Wasfy 2015**
Wasfy JH, Borden WB, Secemsky EA, McCabe JM, Yeh RW. Public reporting in cardiovascular medicine: accountability, unintended consequences, and promise for improvement. *Circulation* 2015;**131**(17):1518–27. [PUBMED: 25918041]
- Werner 2009**
Werner RM, Konetzka RM, Stuart EA, Norton EC, Polsky D, Park J. Impact of public reporting on quality of postacute care. *Health Services Research* 2009;**44**(4):1169–87.
- Yang 2014**
Yang L, Liu C, Wang L, Yin X, Zhang X. Public reporting improves antibiotic prescribing for upper respiratory tract infections in primary care: a matched-pair cluster-randomized trial in China. *Health Research Policy and Systems* 2014;**12**:61. [PUBMED: 25304996]
- Zhang 2018 [pers comm]**
Zhang X (School of Medicine and Health Management, Tongji Medical College, HuaZhong University of Science and Technology, Wuhan, Hubei Province, China). [personal communication]. David Metcalfe (Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences [NDORMS], University of Oxford, UK) 15 March 2018.

References to other published versions of this review

- Ketelaar 2011**
Ketelaar NA, Faber MJ, Flottorp S, Rygh LH, Deane KH, Eccles MP. Public release of performance data in changing the behaviour of healthcare consumers, professionals or organisations. *Cochrane Database of Systematic Reviews* 2011, Issue 11. DOI: 10.1002/14651858.CD004538
* Indicates the major publication for the study

Table 2. Changes in the healthcare utilisation decisions of consumers (Continued)

	As- signed to low- rated HMO (1 option)			0.1	0	23.7	0.0042
	Proportion choosing plan	Farley 2002b	cNRT	0.01	0	0.69	0.0145

cluster-randomised trial (cRT); cluster-non-randomised trial (cNRT); health maintenance organization (HMO)

Table 3. Changes in the healthcare utilisation decisions of healthcare providers (professionals and organisations)

Interven- tion	Outcome	Study	Type of study	Absolute post-in- tervention difference	Absolute pre-inter- vention difference	Post- interven- tion level in control	Relative effect
Public re- porting of injection prescrib- ing rates in outpatient areas	Average ex- penditure per pre- scription	Zhang 2016	cRT	3.4	2.2	41.2	0.0291
	Percent- age of pre- scriptions requiring antibiotics			4.6	6.1	62.8	-0.0249
	Percent- age of pre- scriptions requiring combined antibiotics			2.1	4.1	18.6	-0.1083
	Percent- age of pre- scriptions requiring injections			9.0	13.2	64.9	-0.0643
	Average ex- penditure per pre- scription			7.2	6.9	44.3	0.0070

Table 3. Changes in the healthcare utilisation decisions of healthcare providers (professionals and organisations) (Continued)

Mandatory public reporting of health-care-associated infections	Pediatric quality indicator per 1000 eligible discharges	Rinke 2015	CBA	0.6	0.5	1.0	0.1000		
Intervention	Outcome	Study	Type of study	Absolute level effect (95% CI)	Relative change at 3 months (95% CI)	Relative change at 6 months (95% CI)	Relative change at 9 months (95% CI)	Relative change at 12 months (95% CI)	Relative change at 24 months (95% CI)
Repeated public release of hospital caesarean section rates	Caesarean section rate	Jang 2011	ITS	-0.52 (-0.77 to -0.26)	-0.04 (-0.23 to 0.18)	-1.49 (-2.55 to -0.40)	-2.92 (-4.50 to -1.30)	-4.34 (-6.61 to -1.95)	-
Mandatory public reporting of health-care-associated infections	PICU blood cultures	Flett 2015	ITS	7.48 (1.09 to 13.87)	6.21 (-2.84 to 17.10)	9.90 (-0.45 to 22.64)	13.87 (1.42 to 29.82)	18.17 (2.90 to 38.77)	22.87 (4.11 to 49.86)
	PICU antibiotics			7.29 (4.46 to 10.12)	-0.11 (-2.03 to 1.89)	1.61 (-0.45 to 3.75)	3.36 (0.96 to 5.87)	5.15 (2.26 to 8.20)	6.98 (2.50 to 10.70)
	NICU antibiotics			-5.79 (-9.17 to -2.42)	8.12 (4.11 to 12.46)	6.06 (2.08 to 10.35)	4.05 (-0.35 to 8.85)	1.90 (-3.17 to 7.53)	-0.36 (-6.25 to 6.33)
	NICU blood cultures			-1.14 (-1.90 to -0.39)	2.49 (-0.51 to 5.67)	1.06 (-2.07 to 4.39)	-0.42 (-3.93 to 3.36)	-1.95 (-6.02 to 2.49)	-3.53 (-8.26 to 1.72)

cluster-randomised trial (cRT); controlled before-after (CBA) study; 95% confidence interval (95% CI); interrupted time series (ITS) study; neonatal intensive care unit (NICU); paediatric intensive care unit (PICU)

Table 4. Changes in provider performance

Intervention	Outcome	Study	Type of study	Absolute post-intervention difference	Absolute pre-intervention difference	Postintervention level in control group	Relative effect
--------------	---------	-------	---------------	---------------------------------------	--------------------------------------	---	-----------------

Table 4. Changes in provider performance (Continued)

Public release of a range of quality indicators	All AMI processes	Tu 2009	cRT	2.0	0.9	65.6	0.0168
	Use of standard admission orders			6.1	0.7	72.5	0.0745
	Left ventricular function assessment			2.9	6.3	49.8	-0.0683
	Lipid test < 24 hours arrival			3.8	1.6	51.1	0.0431
	Fibrinolytics < 30 mins after arrival			2.6	3.1	45.7	-0.0109
	Fibrinolytics decided by ED physician			2.0	4.4	84.3	-0.0285
	Fibrinolytics prior to transfer to CCU			3.8	2.9	95.7	0.0094
	Aspirin < 6 hours arrival			5.5	3.1	82.6	0.0291
	B blockers < 12 hours arrival			2.4	3.9	73.7	-0.0204
	Aspirin at discharge			0.9	0.0	84.0	0.0107
	B blockers at discharge			0.6	0.0	85.6	0.0070
	ACEi, ARB for LV dysfunction			4.7	3.4	81.7	0.0159
	Statin at discharge			0.3	0.2	85.5	0.0012
	All CHF processes			1.0	3.0	54.6	-0.0366

Table 4. Changes in provider performance (Continued)

	LVF assessment			2.7	4.5	55.2	-0.0326
	Daily weights recorded			1.3	0.3	24.0	0.0417
	Counselling on > 1 aspect of CHF			0.9	1.7	55.3	-0.0145
	ACEi, ARB for LV dysfunction			6.3	1.7	92.4	0.0498
	B blockers for LV dysfunction			4.0	1.7	71.7	0.0321
	Warfarin for AF			0.6	3.1	64.2	-0.0389

atrial fibrillation (AF); acute myocardial infarction (AMI); angiotensin-converting enzyme inhibitor (ACEi); angiotensin-2 receptor blockers (ARB); beta-adrenergic blocking agents (B blockers); cluster-randomised trial (cRT); coronary care unit (CCU); congestive heart failure (CHF); emergency department (ED); left ventricular (LV); left ventricular failure (LVF); minutes (mins)

Table 5. Changes in patient outcome

Intervention	Outcome	Study	Type of study	Absolute postintervention difference	Absolute pre-intervention difference	Postintervention level in control group	Relative effect
Public release of a range of quality indicators	AMI 30-day mortality	Tu 2009	cRT	2.4	0.5	9.8	0.1939
	AMI 1-year mortality			3.1	1	19.4	0.1082
	STEMI 30-day mortality			3.1	0.4	8.3	0.3253
	STEMI 1-year mortality			3.9	1.2	13.5	0.2000

Table 5. Changes in patient outcome (Continued)

	NSTEMI 30-day mortality			2.3	0.3	10.5	0.1905	
	NSTEMI 1-year mortality			3	0.9	22.6	0.0929	
	CHF 30-day mortality			1	0.9	9.6	0.0104	
	CHF 1-year mortality			2.6	0.6	30.3	0.0660	
	CHF and LV dysfunction 30-day mortality			0.9	0.6	8.5	0.0353	
	CHF and LV dysfunction 1-year mortality			6.3	1.8	26.3	0.1711	
Mandatory reporting of healthcare-associated infections	Pediatric quality indicator per 1000 eligible discharges	Rinke 2015	CBA	0.6	0.5	1	0.1000	
Intervention	Outcome	Study*	Type of study	Absolute level effect (95% CI)	Relative change at 4 months (95% CI)	Relative change at 8 months (95% CI)	Relative change at 12 months (95% CI)	Relative change at 24 months (95% CI)
Hospital quality process and outcome metrics reported on a public website	30-day risk-adjusted mortality	Joynt 2016	ITS	0.12 (0.03 to 0.21)	1.57 (-4.28 to 8.18)	-2.47 (-8.20 to 4.03)	3.71 (-3.25 to 11.74)	7.18 (-1.91 to 18.13)
Public reporting	30-day re-admission	DeVore 2016	ITS	0.00 (0.00 to 0.00)	-2.04 (-8.56 to 5.48)	-1.36 (-7.92 to 6.20)	-0.69 (-7.34 to 7.00)	0.72 (-6.32 to 8.90)

Table 5. Changes in patient outcome (Continued)

of risk-standardised hospital re-admission rates	(AMI)							
30-day re-admission (heart failure)				0.00 (0.00 to 0.00)	-1.39 (-4.17 to 1.56)	-1.84 (-4.59 to 1.08)	-1.88 (-4.68 to 1.10)	-2.78 (-6.42 to 1.15)
30-day re-admission (pneumonia)				0.00 (0.00 to 0.00)	-4.44 (-13.61 to 6.91)	-5.07 (-14.17 to 6.20)	-5.69 (-14.71 to 5.47)	-7.45 (-18.10 to 6.37)
30-day re-admission (COPD)				0.00 (0.00 to 0.00)	-6.66 (-11.42 to -1.37)	-0.76 (-6.11 to 5.23)	-7.64 (-12.31 to -2.44)	-9.06 (-13.62 to -4.00)
30-day re-admission (diabetes)				0.00 (-0.00 to 0.01)	-0.65 (-13.66 to 16.96)	0.00 (-13.13 to 17.81)	0.65 (-12.44 to 18.35)	1.98 (-13.57 to 24.36)
30-day mortality (AMI)				0.00 (0.00 to 0.00)	34.38 (2.71 to 94.32)	35.83 (2.79 to 100.17)	37.38 (2.88 to 106.67)	43.06 (3.20 to 133.08)
30-day mortality (heart failure)				0.00 (0.00 to 0.00)	6.04 (-5.86 to 21.37)	13.78 (-0.56 to 32.94)	9.98 (-3.46 to 27.77)	13.31 (-0.54 to 31.64)
30-day mortality (pneumonia)				0.00 (0.00 to 0.00)	-3.96 (-23.10 to 27.85)	-3.72 (-16.70 to 14.05)	2.94 (-18.04 to 19.00)	-3.84 (-22.51 to 26.69)
30-day mortality (COPD)				0.00 (0.00 to 0.00)	20.89 (5.51 to 41.52)	21.63 (5.68 to 43.24)	20.99 (5.54 to 41.75)	22.00 (5.77 to 44.13)
30-day mortality (diabetes)				0.00 (0.00 to 0.00)	-14.73 (-34.83 to 23.29)	-15.10 (-35.48 to 24.12)	-14.78 (-34.92 to 23.40)	-19.39 (-42.65 to 35.66)

Acute Myocardial Infarction (AMI); ST-Elevation Myocardial Infarction (STEMI); Non-ST-Elevation Myocardial Infarction (NSTEMI); Congestive Heart Failure (CHF); Left Ventricular (LV); Chronic Obstructive Pulmonary Disease (COPD); Cluster Randomised Trial (cRT); Controlled Before-After (CBA) study; Interrupted Time Series (ITS) study; 95% Confidence Interval (95% CI)

* [Joynt 2016](#) and [DeVore 2016](#) provided outcomes in quarters rather than months and so have been presented as 4- and 8-months rather than the pre-specified 3- and 6-months.

■ TRAUMA

Pay for performance and hip fracture outcomes

AN INTERRUPTED TIME SERIES AND DIFFERENCE-IN-DIFFERENCES ANALYSIS IN ENGLAND AND SCOTLAND

D. Metcalfe,
C. K. Zogg,
A. Judge,
D. C. Perry,
B. Gabbe,
K. Willett,
M. L. Costa

From University
of Oxford, Oxford,
United Kingdom

Aims

Hip fractures are associated with high morbidity, mortality, and costs. One strategy for improving outcomes is to incentivize hospitals to provide better quality of care. We aimed to determine whether a pay-for-performance initiative affected hip fracture outcomes in England by using Scotland, which did not participate in the scheme, as a control.

Materials and Methods

We undertook an interrupted time series study with data from all patients aged more than 60 years with a hip fracture in England (2000 to 2018) using the Hospital Episode Statistics Admitted Patient Care (HES APC) data set linked to national death registrations. Difference-in-differences (DID) analysis incorporating equivalent data from the Scottish Morbidity Record was used to control for secular trends. The outcomes were 30-day and 365-day mortality, 30-day re-admission, time to operation, and acute length of stay.

Results

There were 1 037 860 patients with a hip fracture in England and 116 594 in Scotland. Both 30-day (DID -1.7%; 95% confidence interval (CI) -2.0 to -1.2) and 365-day (-1.9%; 95% CI -2.5 to -1.3) mortality fell in England post-intervention when compared with outcomes in Scotland. There were 7600 fewer deaths between 2010 and 2016 that could be attributed to interventions driven by pay-for-performance. A pre-existing annual trend towards increased 30-day re-admissions in England was halted post-intervention. Significant reductions were observed in the time to operation and length of stay.

Conclusion

This study provides evidence that a pay-for-performance programme improved the outcomes after a hip fracture in England.

Cite this article: *Bone Joint J* 2019;101-B:1015–1023.

Hip fracture is a leading cause of death and disability among the elderly worldwide.^{1,2} The incidence is rising as populations age, and there are now over 1.6 million hip fractures globally each year.¹ In the United Kingdom alone, there are 70 000 cases annually at a cost of £2 billion.³

Pay-for-performance initiatives are increasingly used to improve outcomes.^{4,6} These schemes link healthcare payments to quality metrics in order to incentivize providers to improve the quality or efficiency of care.⁷ There is mixed evidence about whether these initiatives can truly drive improvements in healthcare.^{4,6} There is evidence that they can modestly improve care. However, few pay-for-performance schemes have been shown to positively affect outcomes.⁶

A national clinical audit was established in England and Wales in 2007 with the aim of improving hip fracture outcomes.⁸ This programme included a National Hip Fracture Database (NHFD) and support for local clinical teams to improve the quality of care provided to elderly patients with a hip fracture. In 2010, the NHFD was the basis for a pay-for-performance initiative, called the 'Best Practice Tariff' (BPT). The BPT scheme paid hospitals a supplement for each patient whose care satisfied six clinical standards, such as surgery within 36 hours.⁹ Cases satisfying these standards, which have evolved over time (Supplementary Table i), were identified from data submitted to the NHFD. Importantly, Scottish hospitals did not participate in the NHFD and were not subject to the BPT.

Correspondence should be sent to D. Metcalfe; email: david.metcalfe@ndorms.ox.ac.uk

©2019 Author(s) et al
doi:10.1302/0301-620X.101B8.
BJJ-2019-0173.R1 \$2.00

Bone Joint J
2019;101-B:1015–1023.

This study aimed: first to quantify any effect of the NHFD and BPT on the outcomes of hip fractures in England using data from Scotland to control for secular trends; and second to estimate the effect of introducing pay-for-performance for hip fractures in Scotland.

Materials and Methods

This study was a natural experiment using interrupted time series¹⁰ and difference-in-differences (DID) analysis.¹¹ It relied on national data from two sources in order to conduct quasi-experimental modelling of temporal trends. Changes in England, where the NHFD/BPT was introduced, were analyzed as an ‘exposed’ group and those in Scotland as a ‘control’. Given the countries’ geographical proximity, cultural similarities, and common political union within the United Kingdom, it was anticipated that secular changes in Scotland would closely mimic those in England had the NHFD/BPT not been implemented.¹²

Data for England were abstracted from the Hospital Episode Statistics Admitted Patient Care (HES APC) data set¹³ linked to Office for National Statistics (ONS) death certificate registrations. Data for Scotland were abstracted from Scottish Morbidity Records (SMR01).¹⁴

The HES APC data set is managed by NHS Digital and collects data on all admissions to National Health Service (NHS) hospitals, as well as those treated in private hospitals but funded by the NHS.¹³ Approximately 98% of hospital activity in England is funded by the NHS.¹⁵ It is unlikely that many elderly patients with a hip fracture were treated in the private sector during the study period. The HES APC data does not include information about Emergency Department attendances that do not lead to admission.¹³

The ONS holds data on all deaths registered in England and Wales. All English deaths should be captured, although registration could be delayed in cases referred to a coroner for post-mortem or inquest. In 2016, upwards of 96% of deaths were registered within the year that they occurred.¹⁶

The SMR collects administrative data on episodes of inpatient care provided by all hospitals in Scotland. It is managed by the Information Services Division (ISD) for NHS National Services Scotland and is linked directly to Scottish death certificate data. ISD Scotland estimate that the SMR01 captures 99% of admissions to hospitals in Scotland.¹⁷

The United Kingdom is a unitary state composed of four countries: England, Scotland, Wales, and Northern Ireland. Comprehensive publicly funded healthcare is freely available throughout the United Kingdom under the NHS. Provision of healthcare under the auspices of NHS Scotland was devolved to the Scottish Parliament by the Scotland Act 1998.¹⁸

All adults aged more than 60 years were included in the analysis if they were treated for a hip fracture in England or Scotland with inpatient admission dates between January 2000 and December 2016 and had complete follow-up information for a period of one year following admission (2000 to 2017). No additional exclusion criteria were applied. Patients had to present with a primary International Classification for Diseases, Tenth Revision (ICD-10) diagnostic code¹⁹ on admission consistent with: S72.0 (“fracture of neck of femur”),

S71.1 (“pertrochanteric fracture”), or S72.2 (“subtrochanteric fracture”).

The principal aim of the study was to determine the effect of introducing a pay-for-performance initiative on outcomes for elderly patients with a hip fracture. However, the Hip Fracture BPT was only feasible once a framework had been established for capturing high-quality clinical audit data. This framework was provided by the National Hip Fracture Database (NHFD), which was launched three years previously. We therefore examined the effect: first of the introduction of the NHFD from January 2007; second of the introduction of the Hip Fracture BPT from April 2010; and third of the combined effect of the NHFD/BPT intervention.

The NHFD was launched in 2007 and captures data on most adults aged more than 60 years with a hip fracture treated in England, Wales, and Northern Ireland.⁸ All acute hospitals treating hip fractures in England, Wales, and Northern Ireland contribute data to the NHFD.⁸ In addition to publicly reporting hospital-level outcomes in an annual report, the NHFD provides an online platform, through which clinical teams can visualize their outcomes and performance and compare them with national clinical standards.⁹ These standards have changed over time (Supplementary Table i).

All NHS hospitals are reimbursed by a system of tariffs based on an adjusted formula applied to the “reference costs” returned by NHS organizations that estimate the cost of treating patients the previous year. In order to achieve the Hip Fracture BPT, hospitals must satisfy all the criteria shown in Supplementary Table i. The NHFD reports patient-level compliance with the national standards to the local Clinical Commissioning Group (CCG),²⁰ which makes a quarterly correction payment to individual hospitals.

The primary outcome was 30-day mortality. Secondary outcomes included 60-, 90-, and 365-day mortality as well as 30-, 60-, and 90-day re-admission, time to operation (defined as binary early time to the operating room of less than or more than two days), and acute length of stay (LOS) in days.

Statistical analysis. Differences in demographic and clinical variables were compared between countries before and after the introduction of the NHFD and BPT in 2007 to 2010, respectively, in order to visualize potential differences between groups. The ‘pre-intervention’ period was defined as 1 January 2000 to 31 December 2006 and the ‘post-intervention’ period was defined as 1 May 2010 to 1 February 2018. Patients admitted in the period between 1 January 2007 and 30 April 2010 were incorporated into stepwise analyses examining changes before and after establishment of the NHFD and the BPT. Covariate information, presented in Table I, was largely consistent between groups with subtle time-consistent differences in admitted hip fracture patients in England and Scotland.

Changes in hip fracture outcomes in England and Scotland were first visualized graphically by month in order to detect obvious changes and ensure the existence of pre-intervention parallel trends, which is a requirement of quasi-experimental DID analysis. These visualizations included scatter plots with locally weighted smoothing (LOWESS) lines, which are smooth lines created using regression analysis to help visualize trends over time.

Table I. Differences in demographic parameters before and after implementation in England and Scotland

Variable	Overall		Pre-intervention: January 2000 to December 2006		Post-intervention: April 2010 to December 2016	
	England	Scotland	England	Scotland	England	Scotland
Hip fractures, n (%)	1 037 860 (89.9)	116 594 (10.1)	391 697 (89.4)	46 404 (10.6)	446 098 (90.3)	47 730 (9.7)
Age, n (%)						
60 to 64 yrs	30 454 (2.9)	5071 (4.3)	10 225 (2.6)	1889 (4.1)	13 692 (3.1)	2115 (4.4)
65 to 69 yrs	49 178 (4.7)	7646 (6.6)	17 336 (4.4)	2978 (6.4)	23 178 (5.2)	3176 (6.7)
70 to 74 yrs	83 750 (8.1)	12 308 (10.6)	33 538 (8.6)	5236 (11.3)	34 752 (7.8)	4767 (10.0)
75 to 79 yrs	152 778 (14.7)	19 502 (16.7)	63 205 (16.1)	8170 (17.6)	60 507 (13.6)	7605 (15.9)
80 to 84 yrs	240 553 (23.2)	26 397 (22.6)	96 160 (24.5)	10 642 (22.9)	97 993 (22.0)	10 691 (22.4)
85 to 89 yrs	259 565 (25.0)	25 767 (22.1)	92 498 (23.6)	9813 (21.1)	113 854 (25.5)	10 825 (22.7)
≥ 90 yrs	221 582 (21.3)	19 903 (17.1)	78 735 (20.1)	7676 (16.5)	102 122 (22.9)	8551 (17.9)
Sex, n (%)						
Male	256 703 (24.7)	29 141 (25.0)	84 411 (21.6)	10 426 (22.5)	112 404 (25.2)	12 927 (27.1)
Female	781 157 (75.3)	87 453 (75.0)	307 286 (78.4)	35 978 (77.5)	323 694 (72.6)	34 803 (72.9)
Multiple deprivation index, n (%)						
Least deprived 10%	94 045 (9.1)	N/A	32 649 (8.3)	N/A	42 881 (9.6)	N/A
Less deprived 10% to 20%	105 647 (10.2)	N/A	38 843 (9.9)	N/A	46 275 (10.4)	N/A
Less deprived 20% to 30%	109 502 (10.6)	N/A	40 488 (10.3)	N/A	48 010 (10.8)	N/A
Less deprived 30% to 40%	113 209 (10.9)	N/A	41 918 (10.7)	N/A	49 311 (11.1)	N/A
Less deprived 40% to 50%	115 604 (11.1)	N/A	43 150 (11.0)	N/A	50 517 (11.3)	N/A
More deprived 40% to 50%	112 110 (10.8)	N/A	42 733 (10.9)	N/A	48 052 (10.8)	N/A
More deprived 30% to 40%	104 486 (10.1)	N/A	39 906 (10.2)	N/A	44 443 (10.0)	N/A
More deprived 20% to 30%	99 736 (9.6)	N/A	38 543 (9.8)	N/A	41 805 (9.4)	N/A
More deprived 10% to 20%	92 699 (8.9)	N/A	36 962 (9.4)	N/A	37 803 (8.5)	N/A
Most deprived 10%	90 822 (8.8)	N/A	36 505 (9.3)	N/A	37 001 (8.3)	N/A
Charlson Comorbidity Index, n (%)	One-year look back	Hospital reported	One-year look back	Hospital reported	One-year look back	Hospital reported
0	465 976 (44.9)	102 874 (88.2)	229 389 (58.6)	40 917 (88.2)	143 451 (32.2)	42 178 (88.4)
1	309 067 (29.8)	9331 (8.0)	104 123 (26.6)	3805 (8.2)	140 969 (31.6)	3774 (7.9)
2	139 411 (13.4)	3583 (3.1)	36 722 (9.4)	1387 (3.0)	77 399 (17.4)	1444 (3.0)
≥ 3	123 406 (11.9)	806 (0.7)	21 463 (5.5)	295 (0.6)	84 279 (18.9)	334 (0.7)

N/A, not applicable

The quantitative assessment of before-and-after changes was also undertaken for English data using interrupted time series analysis (ITSA). ITSA was used to contextualize the main DID results for mortality and to account for changes in English outcomes not reported in Scottish data, such as time to operation. ITSA functions by fitting linear regression models to observations from the pre- and post-intervention periods. For the purposes of ITSA analysis, longitudinal patient-level data were aggregated into monthly bins for each month of the year and were plotted by month as the proportion (or indicated quantile of initial LOS) of each outcome of interest. Analyses were based on 84 pre-intervention points (patients admitted between January 2000 and December 2006) and 81 post-intervention points (patients admitted between April 2010 and December 2016). Models estimated the pre-intervention trend ('pre-intervention annual change'), change in level immediately following the intervention ('instant change') and change in post-intervention trend ('post-intervention annual change'). The presence of autocorrelation was tested using the Durbin-Watson test. The extent to which the intercept of the post-intervention model deviates from the anticipated pre-intervention trend is assumed to represent an instantaneous causal effect of the intervention taking place during the same period of time. Ongoing changes during the post-intervention period, the post-intervention slope,

can sometimes be observed as a marked and maintained change from pre-intervention trends.

Differences in outcomes for mortality for England and Scotland were further compared using DID regression. This quasi-experimental technique functions by fitting linear models to temporally aggregated data from the pre- and post-intervention periods. It includes coefficients for intervention group (e.g. England vs Scotland), time period (e.g. pre- vs post-intervention) and an interaction term between an intervention group and a period of time. The magnitude and direction of the interaction term (the so-called DID between temporal changes within each country) is assumed to represent the causal effect.¹²

Software: StataIC v.15.0 (StataCorp, College Station, Texas) was used for all statistical analyses. Panel data for ITSA were constructed from the admission-level master data set using collapse commands. They were analyzed using the ITSA module²¹ in Stata. The DIFF module²² was used for linear DID regression in order to obtain p-values and country-specific and time period-specific tabulations for tables; 95% confidence intervals (CI) were obtained by manually fitted versions of the same models.

The use of HES data for this project was approved by the NHS Digital Independent Group Advising on the Release of Data. NHS Digital undertook the linkage to ONS data and created mortality flags at defined timepoints. Pseudo-anonymized

Table II. Interrupted time-series analysis (ITSA) results before and after implementation among adults aged more than 60 years in England, 2000 to 2016. ITSA compared differences in English temporal trends before (January 2000 to December 2006) and after (April 2010 to December 2016) combined policy implementation in 2007 to 2010

	Annual trend pre-implementation	95% CI	Instant change	95% CI	p-value	Annual trend post-implementation	95% CI
Mortality, %							
30-day	0.0	-0.1 to 0.2	2.6	-3.4 to -1.7	< 0.001	-0.2	-0.2 to -0.1
60-day	0.1	0.0 to 0.3	-4.3	-5.5 to -3.1	< 0.001	-0.2	-0.4 to -0.1
90-day	0.2	0.0 to 0.4	-5.4	-6.8 to -4.1	< 0.001	-0.2	-0.4 to -0.1
365-day	0.2	0.1 to 0.4	-5.3	-6.3 to -4.2	< 0.001	-0.1	-0.2 to 0.0
Re-admission, %							
30-day	0.4	0.3 to 0.5	-1.3	-2.2 to -0.5	0.003	-0.2	-0.4 to -0.1
60-day	0.7	0.5 to 0.8	-1.4	-2.3 to -0.5	0.002	-0.1	-0.1 to 0.0
90-day	0.8	0.6 to 0.9	-1.2	-2.1 to -0.2	0.015	0.0	-0.1 to 0.1
Length of stay							
50th percentile (median)	-0.1	-0.2 to 0.0	-2.8	-3.5 to -2.1	< 0.001	-0.4	-0.5 to -0.3
60th percentile	-0.1	-0.2 to 0.0	-3.8	-4.6 to -3.1	< 0.001	-0.5	-0.7 to -0.4
70th percentile	-0.1	-0.3 to 0.0	-5.4	-6.4 to -0.3	< 0.001	-0.7	-0.8 to -0.5
80th percentile	-0.3	-0.5 to -0.1	-7.3	-8.8 to -5.8	< 0.001	-1.0	-1.2 to -0.8
Early time to theatre, %	-0.6	-0.9 to -0.4	15.4	13.7 to 17.0	< 0.001	0.7	0.5 to 0.8

CI, confidence interval

Table III. Difference-in-difference (DID) results before and after Best Practice Tariff (BPT) implementation among adults aged more than 60 years, 2000 to 2016. DID compared differences in mortality trends before (January 2000 to December 2006) and after (April 2010 to December 2016) combined policy implementation in 2007 to 2010 in England *versus* Scotland, 'intervention overall'. They also broke down contributions to differences in mortality before (January 2000 to December 2016) and after (January 2008 to March 2010) introduction of the National Hip Fracture Database (NHFD) and clinical audit alone in 2007, 'NHFD introduced'; and before (January 2008 to March 2010) and after (April 2010 to December 2016) formal introduction of payment penalties under the Best Practice Tariff alone in 2010, 'BPT introduced'

	Scotland			England			NHFD introduced			BPT introduced			Intervention overall		
	2000 to 2006	2007 to 2009	2010 to 2016	2000 to 2006	2007 to 2009	2010 to 2016	DID	95% CI	p-value	DID	95% CI	p-value	DID	95% CI	p-value
Mortality, %															
30-day	9.8	8.8	8.5	9.8	8.7	6.8	-0.1	0.5 to 0.5	0.987	1.6	2.1 to 1.2	< 0.001	1.7	2.0 to 1.2	< 0.001
60-day	15.4	13.8	13.1	15.6	14.0	11.4	0.0	0.6 to 0.6	0.994	1.9	2.5 to 1.3	< 0.001	1.9	2.3 to 1.4	< 0.001
90-day	18.9	17.3	16.4	19.4	17.6	14.7	-0.2	0.9 to 0.5	0.581	2.0	2.6 to 1.3	< 0.001	2.2	2.6 to 1.6	< 0.001
365-day	32.0	30.4	29.8	31.9	30.2	27.8	-0.1	1.0 to 0.6	0.688	1.8	2.5 to 1.0	< 0.001	1.9	2.5 to 1.3	< 0.001
Re-admission, %															
30-day	11.1	11.9	12.2	20.5	20.0	21.0	-1.3	1.9 to 0.6	< 0.001	0.7	0.1 to 1.4	0.051	0.6%	1.1 to 0.1	0.038
Length of stay															
50th percentile	20.0	21.0	20.0	17.0	14.0	12.0	-4.0	-4.4 to -3.6	< 0.001	-1.0	-1.4 to -0.6	< 0.001	-5.0	-5.3 to -4.7	< 0.001
60th percentile	28.0	29.0	26.0	21.0	18.0	15.0	-4.0	-4.4 to -3.6	< 0.001	0.0	-0.1 to 0.1	0.999	-4.0	-4.2 to -3.8	< 0.001
70th percentile	38.0	39.0	36.0	28.0	22.0	19.0	-7.0	-7.5 to -6.5	< 0.001	0.0	-0.2 to 0.2	0.999	-7.0	-7.2 to -6.8	< 0.001
80th percentile	53.0	54.0	50.0	37.0	30.0	25.0	-8.0	-9.1 to -6.9	< 0.001	-1.0	-1.7 to -0.3	0.008	-9.0	-9.7 to -8.3	< 0.001

CI, confidence interval

data were then transmitted to researchers at the University of Oxford. The Information Services Division of National Services Scotland provided pseudo-anonymized records from the Scottish Morbidity Record Scheme. Ethical approval was not sought in line with GAFReC guidance.²³ Personal data were processed under Articles 6(1)(f) and 9(2)(f) of the General Data Protection Regulation (EU) 2016/679.

This research was undertaken independently of the authors' funding bodies, which did not have any influence on the study design, analysis, data interpretation, or decision to publish.

Results

A total of 1 037 860 adults aged more than 60 years were admitted between 2000 and 2016 with a hip fracture in England, and

116 594 in Scotland. The demographic characteristics of these groups are shown in Table I. Table II presents the ITSA results of English data, and Table III shows the more detailed DID analyses for mortality that include Scotland as a comparison group.

Figure 1a shows that the pre-intervention trends in 30-day mortality were the same in England and Scotland; 30-day mortality trended downwards in both countries after the launch of the NHFD, although the decline was more pronounced in England. The diverging lines became more obvious after April 2010 and this continued until the data were censored in December 2016 (ITSA instant change following combined policy implementation -2.6 percentage points (95% CI -3.4 to -1.7); annual trend post-implementation -0.2 (-0.2 to -0.1). DID analysis

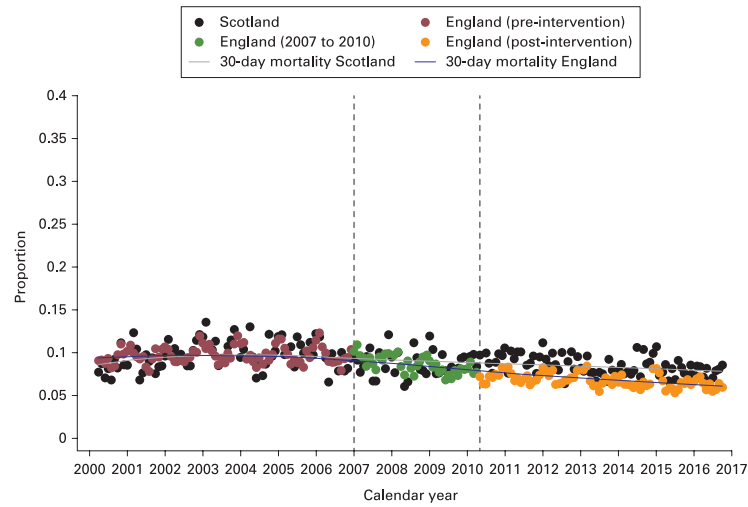


Fig. 1a

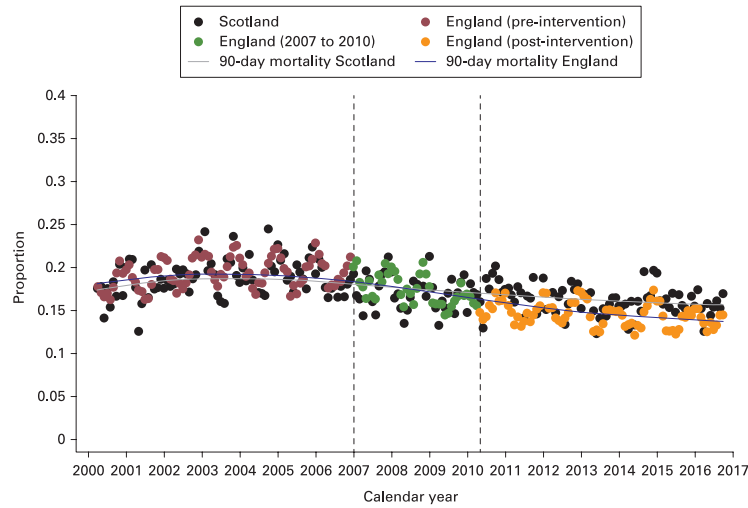


Fig. 1b

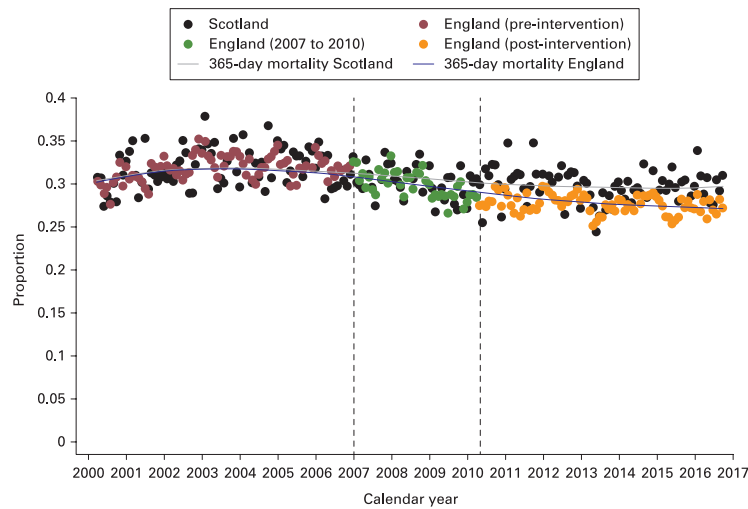


Fig. 1c

Charts showing monthly changes in a) 30-, b) 90- and c) 365-day mortality among adults aged more than 60 years, 2000 to 2016. Dashed lines represent introduction of the National Hip Fracture Database in January 2007 and the Best Practice Tariff in April 2010.

corroborated these findings, suggesting an overall reduction in 30-day mortality in England relative to Scotland of -1.7 percentage-points (95% CI -2.0 to -1.2). When stratified by each component of the intervention alone, the results suggest a modest reduction in 30-day mortality following NHFD introduction that did not reach significance (-0.1 (-0.5 to +0.5) percentage-points; $p = 0.987$) and a larger significant change of -1.6 percentage-points (-2.1 to -1.2) following introduction of the BPT. Between 2010 and 2016 in England (Fig. 1a), there were 7600 fewer deaths than expected within 30 days, following implementation of the BPT.

The effects on 60-day mortality showed the same direction and magnitude as on 30-day mortality (Fig. 1a). Figure 1b shows similar findings for 90-day mortality, although the effect of the BPT was more apparent at this time. DID analysis suggested that the combined intervention was associated with a change of -2.2 percentage-points (95% CI -2.6 to -1.6). However, this appeared to be driven entirely by the BPT: NHFD DID -0.2 percentage-points (95% CI -0.9 to -0.5) and BPT -2.0 (-2.6 to -1.3)).

Figure 1c shows that the effect on mortality at 365 days was similar to that for mortality at 30 and 60 days. Although a small (non-significant; $p = 0.688$) improvement was observed when the NHFD was introduced (DID -0.1 percentage point; 95% CI -1.0 to 0.6), the BPT was associated with a significant fall in 365-day mortality (-1.8; 95% CI -2.5 to -1.0; $p < 0.001$). The effect of the combined intervention on 365-day mortality was a change of -1.9 percentage-points (95% CI -2.5 to -1.3). Projection modelling (presented in Supplementary Figure a) suggests that were the BPT to be implemented in Scotland in 2019, upwards of 115 deaths could be prevented each year – a number totaling more than 1377 deaths by 2030.

Table II shows that re-admissions at all times (30, 60, and 90 days) were increasing steadily in England in the pre-implementation phase. The annual trend towards increasing 30-day re-admissions (0.4 percentage-points; 95% CI 0.3 to 0.5) was, however, reversed on implementation of the BPT (instant change -1.3 percentage points; 95% CI -2.2 to -0.5) and this decline continued each year subsequently (annual trend post-implementation -0.2 percentage points; 95% CI -0.4 to -0.1). Similar findings were observed for 60- and 90-day re-admissions, although the annual trend post-implementation did not change after the sudden fall associated with the BPT at these timepoints.

There was an annual trend towards fewer patients undergoing surgery within 36 hours in the pre-intervention period (annual trend -0.6 percentage-points; 95% CI -0.9 to -0.4). However, in the year following introduction of the NHFD/BPT, the proportion of patients reaching the operating theatre within this timeframe increased by an absolute value of 15.4 percentage-points (95% CI 13.7 to 17.0; Table II). This positive trend continued to increase by 0.7 percentage-points (95% CI 0.5 to 0.8) each year thereafter. Projection modelling (presented in Supplementary Figure a) suggests that were the BPT to be implemented in Scotland starting in 2019, upwards of 220 fewer re-admissions within 30 days would be expected among Scottish hip fracture patients by 2030.

The median LOS was declining modestly (annual trend -0.6 days; 95% CI -0.2 to 0.0) in the pre-intervention period. This

reduction increased following implementation of the NHFD/BPT (instant change following policy implementation -2.8 days; 95% CI -3.5 to -2.1; annual trend post-policy implementation: -0.4 days; 95% CI -0.5 to -0.3). The magnitude of these reductions increased in a stepwise manner with each ascending quantile (e.g. 60th, 70th, and 80th) so that the largest reductions were observed amongst the patients with greatest initial LOS (80th percentile instant change: -7.3 days; 95% CI -8.8 to -5.8; annual trend post-implementation -1.0 days, 95% CI -1.2 to -0.8).

Discussion

This study provides evidence that the BPT drove changes in practice that reduced mortality for elderly patients with a hip fracture in England by as many as 7600 fewer deaths within 30 days between 2010 and 2016. It also suggests that the BPT increased the proportion of patients receiving an operation within 36 hours, shortened LOS, and reduced re-admissions within 30, 60, and 90 days.

A number of small studies have reported improved compliance with measures that were associated with introduction of the NHFD²⁴ and BPT.²⁵⁻²⁷ The NHFD annual reports have also shown that English hospitals are increasingly achieving the hip fracture national clinical standards.⁸ One national study reported that mortality fell from 10.9% before the NHFD was launched to 8.5% afterwards.⁹ However, this study did not include a control population or analyze data after the BPT came into effect. Our data suggest that there was a gradual trend towards reduced mortality between 2007 and 2010 but that this was also apparent, albeit to a lesser extent, in Scotland, which did not participate in the NHFD. DID results restricted to the influence of the NHFD suggest that the differential trend between the two countries was not statistically significant following introduction of the NHFD in isolation. There was, however, a significant change leading up to full BPT implementation in 2010. DID results comparing BPT implementation alone between England and Scotland revealed a 2.0 percentage-point reduction in 90-day mortality (8.7% to 6.8% in England) that accounted for 90.9% of the overall effect (DID -2.2 percentage-points; 9.8% to 6.8% in England).

There are a number of changes that could account for improved outcomes over time across the United Kingdom, including publication of the BOA/BGS guidelines,²⁸ increasing recognition of the need for early surgery and postoperative rehabilitation,²⁹ and the emergence of orthogeriatrics as a medical subspecialty dedicated to caring for elderly patients with a fracture.^{30,31} It therefore seems unlikely that the NHFD alone accounted for the fall in mortality reported by Neuburger et al.⁹ Although our findings suggest that the NHFD might have had a small positive effect on English hip fracture outcomes, this was not statistically significant. However, implementation of the BPT was associated with a marked and sustained improvement in outcomes. It is nevertheless worth noting that the NHFD was a prerequisite for the choice of hip fracture outcomes as a target for pay-for-performance in England and so these two interventions are fundamentally linked.³² However, our data suggest that a system for rewarding best practice can improve outcomes beyond that of a voluntary audit of national clinical standards.

The improvements in LOS and re-admission suggest substantial resource savings attributable to the BPT in addition to reduced mortality.^{4,6} Importantly, the BPT itself did not require a substantial investment. It was initially set up as a payment of £445 (\$570), which was based on an estimate of the cost that an average hospital was likely to incur to provide additional operating capacity. However, the base tariff was reduced initially to adjust for compliance with the BPT criteria that was already present throughout the NHS. As a consequence of the falling base tariff, the BPT has accounted for a greater proportion of the overall payment to hospitals each year – from £445 (\$570) in 2010/11 to £890 (\$1141) in 2011/12, £1335 (\$1712) in 2012/13, and £1353 (\$1735) in 2016/17.³² As 100% compliance with the standards has not been achieved, the overall payment nationally by CCGs changed little during the first three years. Although we have not presented a formal health economic analysis, it is likely that the BPT delivered improved hip fracture care at reduced cost to NHS commissioners. There is a rectification process in the NHS of hospital trusts returning reference costs, for the delivery of care, to ensure alignment between tariff price and the average cost of delivery.

Previous evaluations of pay-for-performance initiatives have reported mixed findings.^{4,6} Many early studies focused on ‘Value-Based Purchasing’ (VBP), which is a strategy used by the Centers for Medicare and Medicaid Services (CMS) in the United States. The VBP programme withholds 2% of annual Medicare payments and allocates these to hospitals based on the quality of care, compliance with best clinical practice, and patient experience.³³ However, few studies have been able to demonstrate improvements in mortality or re-admissions that may be attributable to VBP.^{34–36} A similar scheme in the northwest region of England (‘Advancing Quality’) was found to have no long-term effect on 30-day mortality.³⁷ A number of explanations have been proposed for this finding.³⁸ First, the financial impact of the VBP is small (average \$213 000 bonus and \$1 200 000 penalty per hospital in 2015),³⁹ which might be insufficient to motivate changes in clinical pathways given that only a proportion of patients in the United States are funded through CMS. Second, there are 21 individual measures and improving these in isolation is unlikely to improve a hospital’s overall score.⁴⁰ Third, the financial reward for improvement is unclear until the end of the performance period because the scheme is designed to be cost-neutral and to transfer payments from low- to high-performers.³⁸ By contrast, the Hip Fracture BPT overcomes many of these criticisms as it has a simple design, focuses on a small number of high-value measures and carries a financial incentive that may be sufficient to motivate changes.⁴¹ An alternative explanation for the success of the BPT is that it was part of a more complex intervention that began with the national clinical audit and NHFD. There is evidence that some hospitals engaged with the NHFD from 2007, by designing quality improvement processes, aimed at improving their performance using data provided through online visual dashboards and in publicly accessible reports.^{8,24} It is possible that the BPT provided additional impetus that helped clinicians and hospital leaders to create business cases that justified local investment in hip fracture services. Our study provides evidence that, despite concerns about the success of other schemes, it is possible to improve hip

fracture outcomes through pay-for-performance. Further work should aim to identify the features that distinguish programmes that can demonstrably improve outcomes.

The apparent success of the Hip Fracture BPT in England could have policy implications for a number of countries. First, although there are 36 national clinical audits that are operated by the Healthcare Quality Improvement Program (HQIP)⁴² in England, there are only 21 BPTs that are used by NHS England to refine healthcare payments.⁴³ This suggests that there are further opportunities to extend pay-for-performance to other groups of patients in England. Second, our findings raise the possibility that the introduction of a comparable pay-for-performance initiative might reduce mortality following hip fracture in Scotland. Finally, this study might encourage policy makers outside the United Kingdom to consider implementing pay-for-performance programmes to improve hip fracture outcomes. For example, although the CMS coordinates health payments for most patients aged more than 65 years in the United States, the VBP programme does not yet extend to hip fractures. There are more than 200 000 hip fractures in the United States each year^{44,45} with a reported mortality of 5.2% at 30 days.⁴⁴ If the estimated benefits of the BPT in England were generalizable to the United States, the CMS expansion of pay-for-performance to elderly patients with a hip fracture could prevent as many as 3600 deaths per year.

The strengths of this study are the use of comprehensive national cohorts linked to death certificate registrations and a ‘control’ region, which overcame the limitations of earlier before-and-after studies. There are, however, a number of possible limitations to this approach. First, our study would still be vulnerable to confounding factors if another event had occurred at the same time as the NHFD/BPT but only affected outcomes in either England or Scotland.¹² We are not aware of any such events and the factors that are thought to have driven recent trends towards improved hip fracture outcomes, such as the rise of orthogeriatrics as a medical subspecialty,³¹ would be expected to have applied across the whole of the United Kingdom. Although there was a reconfiguration of major trauma services in England (but not Scotland) from April 2012, this did not have a measurable effect on the quality of hip fracture care.⁴⁶ Second, we did not have access to some variables, such as LOS, in Scottish data and so were limited to undertaking ITSA without a control region for these outcomes. As discussed above, the absence of a control can result in erroneous attribution of change to a single intervention.¹² It is, however, reassuring that the findings from ITSA for the other outcomes were consistent with those of the DID analyses. A lack of variables also restricted our ability to present baseline characteristics for hip fracture patients in England and Scotland. Some variables from SMR01, such as age, were categorized by the data owners to preserve the anonymity of patients and others, such as index of multiple deprivation (IMD), were available for England but not Scotland. However, this limitation is partly accounted for by the study design as there is no obvious reason why differences in patient characteristics should have changed between the pre-intervention period (when outcome trends were parallel) and post-intervention period (when trends diverged). Third, we selected outcomes that could be readily quantified using

administrative data. Although mortality and re-admissions are important quality metrics, other outcomes such as pain, mobility, and health-related quality of life might be more important to patients.⁴⁷ Finally, we focused on hip fracture outcomes and so could not determine whether the BPT had an effect on other groups of patients. Unintended consequences of the BPT could include the deprioritization of other elderly patients with lower limb injuries, such as of the distal femur or ankle, who share many vulnerabilities as those with a hip fracture.^{48,49} Alternatively, further benefits might extend to such patients (a so-called 'halo effect') as hospitals are likely to have invested in orthogeriatricians and dedicated trauma operating theatres in order to achieve the BPT. The effect of pay-for-performance on related groups patients should be a focus for future research.

In conclusion, this study provides evidence that the Hip Fracture BPT improved hip fracture outcomes in England. It is therefore possible that BPTs could improve outcomes and reduce costs in other disease groups. Policymakers and clinicians should support the controlled expansion of the BPT model to other clinical areas and health policy environments.



Take home message

- The Hip Fracture Best Practice Tariff (BPT) was associated with reduced mortality for elderly patients with a hip fracture in England.

- The BPT may also have driven changes that increased the proportion of patients receiving prompt surgery, shortened length of stay, and reduced hospital re-admissions.

Twitter

Follow D. Metcalfe @TraumaDataDoc

Follow C. K. Zogg @CherylZogg

Follow A. Judge @andyjudgeox

Follow D. C. Perry @MrDanPerry

Supplementary material



A table summarizing eligibility for payment of the Best Practice Tariff, as well as a figure showing the projected reductions in 30-day mortality, 365-day mortality, and 30-day re-admission were the Best Practice Tariff to be introduced in Scotland in 2019.

References

1. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos Int* 2006;17:1726–1733.
2. Rapp K, Buchele G, Dreinhofer K, et al. Epidemiology of hip fractures: systematic literature review of German data and an overview of the international literature. *Z Gerontol Geriatr* 2018;52:10–16.
3. Burge RT, Worley D, Johansen A, Bhattacharyya S, Bose U. The cost of osteoporotic fractures in the UK: projections for 2000–2010. *J Med Econ* 2001;4: 51–62.
4. Maynard A. The powers and pitfalls of payment for performance. *Health Econ* 2012;21:3–12.
5. O'Connor RJ, Neumann VC. Payment by results or payment by outcome? The history of measuring medicine. *J R Soc Med* 2006;99:226–231.
6. Ogundejji YK, Bland JM, Sheldon TA. The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes. *Health Policy* 2016;120:1141–1150.
7. Marshall L, Charlesworth A, Hurst J. The NHS payment system: evolving policy and emerging evidence. Nuffield Trust. 2014. <https://www.nuffieldtrust.org.uk/research/the-nhs-payment-system-evolving-policy-and-emerging-evidence> (date last accessed 2 April 2019).
8. No authors listed. Falls and Fragility Fracture Audit Programme (FFFAP). National Hip Fracture Database (NHFD) Annual Report 2017. National Hip Fracture Database. 2017. <https://www.nhfd.co.uk/2017report> (date last accessed 2 April 2019).
9. Neuburger J, Currie C, Wakeman R, et al. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in England: an external evaluation using time trends in non-audit data. *Med Care* 2015;53:686–691.
10. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health* 2012;66:1182–1186.
11. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* 2015;350:h2750.
12. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* 2014;312:2401–2402.
13. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017;46:1093–1093i.
14. No authors listed. SD Scotland (2010). Secondary care information collection. <https://www.isdscotland.org/Products-and-Services/Terminology-Services/Information-for-Clinicians/Secondary-Care-Information> (date last accessed 24 May 2019).
15. No authors listed. Healthcare across the UK: a comparison of the NHS in England, Scotland, Wales and Northern Ireland. National Audit Office. 2012. <https://www.nao.org.uk/report/healthcare-across-the-uk-a-comparison-of-the-nhs-in-england-scotland-wales-and-northern-ireland> (date last accessed 2 April 2019).
16. No authors listed. Impact of registration delays on mortality statistics. Office for National Statistics. 2016. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/impactofregistrationdelaysonmortalitystatistics> (date last accessed 2 April 2019).
17. No authors listed. SMR Completeness Estimates. Information Services Division (ISD) Scotland. 2018. <https://www.isdscotland.org/Products-and-Services/Data-Support-and-Monitoring/SMR-Completeness> (date last accessed 2 April 2019).
18. Robson K. The National Health Service in Scotland 2016. The Scottish Parliament. 2016. http://www.parliament.scot/ResearchBriefingsAndFactsheets/S5/SB_16-100_The_National_Health_Service_in_Scotland.pdf (date last accessed 23 April 2019).
19. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. WHO: Geneva. <https://www.who.int/classifications/icd/en/bluebook.pdf> (Date last accessed 24 May 2019).
20. No authors listed. 2019/20 National Tariff Payment System – A Consultation Notice: Annex Dtd. NHS England and NHS Improvement. https://improvement.nhs.uk/documents/484/Annex_Dtd_Best_practice_tariffs.pdf (date last accessed 24 May 2019).
21. Linden A. ITSA: Stata module to perform interrupted time series analysis for single and multiple groups. 2017; Statistical Software Components, Boston College Department of Economics, Massachusetts. <https://ideas.repec.org/c/boc/bocode/s457793.html> (date last accessed 24 May 2019).
22. Villa JM. DIFF: Stata module to perform differences in differences estimation. 2018; Statistical Software Components, Boston College Department of Economics, Massachusetts, USA. <https://ideas.repec.org/c/boc/bocode/s457083.html> (date last accessed 24 May 2019).
23. No authors listed. Governance arrangements for research ethics committees. NHS Health Research Authority. 2018. <https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/governance-arrangement-research-ethics-committees> (date last accessed 2 April 2019).
24. Patel NK, Sarraf KM, Joseph S, Lee C, Middleton FR. Implementing the National Hip Fracture Database: An audit of care. *Injury* 2013;44:1934–1939.
25. Chamberlain M, Pugh H. Improving inpatient care with the introduction of a hip fracture pathway. *BMJ Qual Improv Rep* 2015;4:w204075.w2786.
26. Lisk R, Yeong K. Reducing mortality from hip fractures: a systematic quality improvement programme. *BMJ Qual Improv Rep* 2014;3:w205006.w2103.
27. Oakley B, Nightingale J, Moran CG, Moppett IK. Does achieving the best practice tariff improve outcomes in hip fracture patients? An observational cohort study. *BMJ Open* 2017;7:e014190.
28. No authors listed. The care of patients with fragility fracture (Blue Book). British Geriatrics Society. 2007. <https://www.bgs.org.uk/resources/care-of-patients-with-fragility-fracture-blue-book> (date last accessed 2 April 2019).
29. Bretherton CP, Parker MJ. Early surgery for patients with a fracture of the hip decreases 30-day mortality. *Bone Joint J* 2015;97-B:104–108.

30. Hawley S, Javadi MK, Prieto-Alhambra D, et al. Clinical effectiveness of orthogeriatric and fracture liaison service models of care for hip fracture patients: population-based longitudinal study. *Age Ageing* 2016;45:236–242.
31. Sahota O, Currie C. Hip fracture care: all change. *Age Ageing* 2008;37:128–129.
32. Gerschlick B. Best Practice Tariffs. Country Background Note: United Kingdom (England). 2016. <https://www.oecd.org/els/health-systems/Better-Ways-to-Pay-for-Health-Care-Background-Note-England-Best-practice-tariffs.pdf> (date last accessed 24 May 2019).
33. No authors listed. Hospital Value-Based Purchasing. Centers for Medicare and Medicaid Services. 2017. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf (date last accessed 2 April 2019).
34. Chee TT, Ryan AM, Wasfy JH, Borden WB. Current state of Value-Based Purchasing programs. *Circulation* 2016;133:2197–2205.
35. Figueroa JF, Tsugawa Y, Zheng J, Orav EJ, Jha AK. Association between the Value-Based Purchasing pay for performance program and patient mortality in US hospitals: observational study. *BMJ* 2016;353:i2214.
36. Ryan AM, Krinsky S, Maurer KA, Dimick JB. Changes in Hospital quality associated with hospital value-based purchasing. *N Engl J Med* 2017;376:2358–2366.
37. Kristensen SR, Meacock R, Turner AJ, et al. Long-term effect of hospital pay for performance on mortality in England. *N Engl J Med* 2014;371:540–548.
38. Dalzell MD. Now might be the right time to kill hospital Value-Based Purchasing program. *Manag Care* 2017;26:14–16.
39. Rau J. 1,700 hospitals win quality bonuses from Medicare, but most will never collect. Kaiser Health News. 22 January 2015. <https://khn.org/news/1700-hospitals-win-quality-bonuses-from-medicare-but-most-will-never-collect> (date last accessed 24 May 2019).
40. No authors listed. Growth of population-based payments is not associated with a decrease in market-level cost growth, yet. Leavitt Partners. 2018. <https://leavittpartners.com/whitepaper/growth-of-population-based-payments-is-not-associated-with-a-decrease-in-market-level-cost-growth-yet/> (last accessed 2 April 2019).
41. Jha AK. Time to get serious about pay for performance. *JAMA* 2013;309:347–348.
42. No authors listed. A-Z of National Clinical Audits. Healthcare Quality Improvement Partnership. 2018. <https://www.hqip.org.uk/a-z-of-nca> (last accessed 2 April 2019).
43. No authors listed. 2017/18 and 2018/19 National Tariff Payment System. NHS England and NHS Improvement. 2016. <https://improvement.nhs.uk/resources/national-tariff-1719/> (last accessed 2 April 2019).
44. Brauer CA, Coca-Perrillon M, Cutler DM, Rosen AB. Incidence and mortality of hip fractures in the United States. *JAMA* 2009;302:1573–1579.
45. Michael Lewiecki E, Wright NC, Curtis JR, et al. Hip fracture trends in the United States, 2002 to 2015. *Osteoporos Int* 2018;29:717–722.
46. Metcalfe D, Gabbe BJ, Perry DC, et al. Quality of care for patients with a fracture of the hip in major trauma centres: a national observational study. *Bone Joint J* 2016;98-B:414–419.
47. Haywood KL, Griffin XL, Achten J, Costa ML. Developing a core outcome set for hip fracture trials. *Bone Joint J* 2014;96-B:1016–1023.
48. Lester HE, Hannon KL, Campbell SM. Identifying unintended consequences of quality indicators: a qualitative study. *BMJ Qual Saf* 2011;20:1057–1061.
49. Smith JR, Halliday R, Aquilina AL, Collaborative-Orthopaedic Trauma Society (OTS). Distal femoral fractures: the need to review the standard of care. *Injury* 2015;46:1084–1088.

Author information:

D. Metcalfe, MRCP, MRCS, MRCEM, Clinical Research Fellow in Musculoskeletal Trauma Surgery
 D. C. Perry, PhD, FRCS(Orth), Associate Professor of Orthopaedics & Trauma Surgery
 K. Willett, FRCS(Orth), Professor of Orthopaedic Trauma Surgery
 M. L. Costa, PhD, FRCS(Orth), Professor of Orthopaedic Trauma Surgery
 Oxford Trauma, Kadoorie Centre for Critical Care Research and Education, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), John Radcliffe Hospital, Oxford, UK.

C. K. Zogg, MSPH, MHS, MD-PhD Student, Yale School of Medicine, New Haven, Connecticut, USA.

A. Judge, PhD, Professor of Translational Statistics, Centre for Statistics in Medicine, NDORMS, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK; Musculoskeletal Research Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Southmead Hospital, Bristol, UK; National Institute for Health Research Bristol Biomedical Research Centre (NIHR Bristol BRC), University Hospitals Bristol NHS Foundation Trust, University of Bristol, Southmead Hospital, Bristol, UK; MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, UK.

B. Gabbe, PhD, Head of the Emergency and Trauma Research Unit, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia.

Author contributions:

D. Metcalfe: Designed the study, Analyzed and interpreted the data, Wrote the manuscript.

C. K. Zogg: Designed the study, Analyzed and interpreted the data, Edited the manuscript.

A. Judge: Designed the study, Interpreted the data, Edited the manuscript.

D. C. Perry: Designed the study, Interpreted the data, Edited the manuscript.

B. Gabbe: Designed the study, Interpreted the data, Edited the manuscript.

K. Willett: Designed the study, Interpreted the data, Edited the manuscript.

M. L. Costa: Designed the study, Interpreted the data, Edited the manuscript.
 D. Metcalfe and C. K. Zogg contributed equally and should be considered joint first authors.

Funding statement:

D. Metcalfe is funded by an Oxford-UCB Prize Fellowship in Biomedical Research (which funded this paper's open access status) and a Royal College of Surgeons of Edinburgh Small Pump Priming Grant. C. K. Zogg is supported by a National Institutes of Health (NIH) Medical Scientist Training Program Training Grant T32GM007205. C. K. Zogg is the principal investigator of a grant from the Emergency Medical Foundation and American College of Emergency Physicians entitled, 'Understanding Emergency Medicine Providers' Perceptions of the ACA in a Renewed Era of Healthcare Reform: National Survey and Qualitative Mixed-Methods Approach'. A. Judge is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. D. Perry is supported by a NIHR Clinician Scientist Fellowship (NIHR/CS/2014/14/012). All authors carried out this research independently of the funding bodies. The views expressed in this publication are those of the authors and do not necessarily reflect those of the NHS, the National Institute for Health Research, or the Department of Health and Social Care.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

Acknowledgements:

We are grateful to NHS Digital and the Information Services Division (ISD) Scotland for providing the data used for this study.

Open access statement:

This is an open-access article distributed under the terms of the Creative Commons Attribution licence (CC-BY-NC), which permits unrestricted use, distribution, and reproduction in any medium, but not for commercial gain, provided the original author and source are credited.

This article was primary edited by J. Scott.



Inequalities in use of total hip arthroplasty for hip fracture: population based study

Daniel C Perry,¹ David Metcalfe,² Xavier L Griffin,³ Matthew L Costa³

¹Institute of Translational Medicine, University of Liverpool, Liverpool, L12 2AP, UK

²Centre for Surgery and Public Health, Harvard Medical School, Boston, MA 02115, USA

³Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, The Kadoorie Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK

Correspondence to: D C Perry danperry@liverpool.ac.uk

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;353:i2021 <http://dx.doi.org/10.1136/bmj.i2021>

Accepted: 23 March 2016

ABSTRACT

OBJECTIVES

To determine whether the use of total hip arthroplasty (THA) among individuals with a displaced intracapsular fracture of the femoral neck is based on national guidelines or if there are systematic inequalities.

DESIGN

Observational cohort study using the National Hip Fracture Database (NHFD).

SETTING

All hospitals that treat adults with hip fractures in England, Wales, and Northern Ireland.

PARTICIPANTS

Patients within the national database (all aged ≥ 60) who received operative treatment for a non-pathological displaced intracapsular hip fracture from 1 July 2011 to 31 April 2015.

MAIN OUTCOME MEASURES

Provision of THA to patients considered eligible under criteria published by the National Institute for Health and Care Excellence (NICE).

RESULTS

114 119 patients with hip fracture were included, 11 683 (10.2%) of whom underwent THA. Of those who satisfied the NICE criteria, 32% (6780) received a THA. Of patients who underwent THA, 42% (4903) did not satisfy the NICE criteria. A recursive partitioning algorithm found that the NICE eligibility criteria did not optimally explain which patients underwent THA. A model with superior explanatory power drew distinctions that are not supported by NICE, which were an age cut off at 76 and a different ambulation cut off. Among patients who satisfied the NICE eligibility, the use of THA was less likely with higher age (odds ratio 0.88, 95% confidence interval 0.87 to 0.88), worsening abbreviated mental test scores (0.49 (0.41 to 0.58) for normal cognition v borderline cognitive

impairment)), worsening American Society of Anesthesiologists score (0.74, 0.66 to 0.84), male sex (0.85, 0.77 to 0.93), worsening ambulatory status (0.32, 0.28 to 0.35 for walking with a stick v independent ambulation), and fifths of worsening socioeconomic area deprivation (0.76 (0.66 to 0.88) for least v most deprived fifth). Patients receiving treatment during the working week were more likely to receive THA than at the weekend (0.90, 0.83 to 0.98).

CONCLUSIONS

There are wide disparities in the use of THA among individuals with hip fractures, and compliance with NICE guidance is poor. Patients with higher levels of socioeconomic deprivation and those who require surgery at the weekend were less likely to receive THA. Inconsistent compliance with NICE recommendations means that the optimal treatment for older adults with hip fractures can depend on where and when they present to hospital.

Introduction

There are over 70 000 hip fractures in the United Kingdom every year, with a combined health and social cost of £2bn (£2.5bn, \$2.8bn).¹ Demographic projections estimate that the annual incidence will increase to over 100 000 by 2020.² Mortality is high, with 8.5% of patients dying within 30 days after hip fracture.³

Several initiatives have been credited with improving outcomes in the UK.³ In 2004 the British Orthopaedic Association (BOA) and the British Geriatrics Society (BGS) established the National Hip Fracture Database (NHFD), with the aim of improving outcomes of hip fracture through continuous national clinical audit.⁴ The national database was supported by combined BOA/BGS clinical guidance⁵ and later by the best practice tariff for hip fracture, which rewards NHS organisations for meeting defined quality standards, including surgery within 36 hours after arrival at hospital.⁶ These initiatives have been associated with improved outcomes, including a fall in 30 day mortality from 10.9% in 2007 to 8.5% in 2011.³

Displaced intracapsular hip fractures are at high risk of painful non-union and so the recommended treatment is either hemiarthroplasty or total hip arthroplasty (THA).⁷⁻⁹ In hemiarthroplasty, the femoral head is replaced; in THA, both the femoral head and acetabulum are replaced. Although the risk-benefit profiles vary between these two operations, it has been shown that patients who undergo THA have better function and less need for revision surgery.⁷⁻¹¹ In June 2011, the National Institute for Health and Care Excellence (NICE) recommended that THA should be offered to patients with a displaced intracapsular hip fracture who are "(a) able to walk independently out of doors with no more than the

WHAT IS ALREADY KNOWN ON THIS TOPIC

A defined subset of patients with hip fracture achieve better functional outcomes with total hip arthroplasty (THA) than with hemiarthroplasty

NICE guidelines indicate which patients should be offered THA

WHAT THIS STUDY ADDS

Compliance with NICE guidelines is poor, and there is considerable variation between hospitals

Surgeons seem to apply different eligibility criteria than NICE

Socioeconomic deprivation and need for hip fracture surgery at the weekend are particular barriers to use of THA

Further efforts are necessary to improve the use of THA for eligible patients and reduce unexplained variation in care for older adults with hip fractures

use of a stick (b) not cognitively impaired and (c) medically fit for anaesthesia and the procedure.”⁸ The provision of THA is not explicitly included as a quality indicator within the NHFD and so the extent to which surgeons comply with this guideline is unknown.

From clinical experience, we hypothesised that there were inequalities in use of THA between hospitals. We identified whether the use of THA is based on factors that are consistent with national recommendations or if systematic inequalities exist with regards to the use of THA for hip fracture.

Methods

We carried out an observational study using data collected by the NHFD national clinical audit project. The study protocol was approved by the Healthcare Quality Improvement Partnership (HQIP) before data release, but research ethics committee approval was not sought for secondary analysis of administrative data in line with Governance Arrangements for Research Ethics Committee (GafREC) guidelines.¹²

Data source

The NHFD is commissioned by the Healthcare Quality Improvement Partnership and managed by the Royal College of Physicians as part of the Falls and Fragility Fracture Audit Programme (FFAP). It captures over 95% of hip fractures treated in England, Wales, Northern Ireland, and the Channel Islands. Data include patients' characteristics, fracture pattern, surgical interventions, and outcomes. These details are typically collected by specialist nurses within each hospital who provide continuity of care to patients with hip fractures and manage submissions to the NHFD. Data from patients aged under 60 are not captured within the database.

Inclusion criteria

This study included all patients aged ≥ 60 who presented to hospital from 1 July 2011 to 31 April 2015 with a displaced intracapsular hip fracture. We chose 1 July 2011 as one month after publication of NICE Clinical Guideline 124.⁸ Patients were excluded if their fracture was coded as “pathological” as this could represent a heterogeneous group that includes patients with disseminated cancer.

Variables and outcomes

Data cleaning involved several steps. Two patients had ages recorded as >115 (both >1000), which we recoded to exclude this variable. In 27 (0.01%) cases, the score of the abbreviated mental test (AMTS) was not recorded as an integer and so scores were rounded to the nearest integer. On 1 April 2014 the NHFD data collection tool was updated to record mobility differently within the revised database. Earlier data were therefore mapped onto the new version by using the algorithm shown in appendix 1. In the event of hospital trust reconfiguration (closure/merger), we used the hospital code at the time of data entry. As a consequence, some hospitals contributed data for only a few months before reconfiguration.

Variables extracted from the NHFD were age (whole years), sex, lower layer super output area (LSOA), date of admission, treating hospital, pre-morbid mobility, American Society of Anesthesiologists (ASA) classification score for physical status, and score on the abbreviated mental test. The physical status score ranges between 1 (healthy patient) and 5 (moribund patient not expected to survive for 24 hours with or without surgery). The abbreviated mental test is a test of 10 questions (such as “what is your age?”), which gives a score from 0 (zero answers correct) to 10 (all correct).

Deprivation scores for patients living in England were determined with the index of multiple deprivation, 2007. These scores reflect deprivation related to income, health and disability, employment, barriers to housing and services, living environment, education, and crime.¹³ Scores were generated from lower layer super output areas, which were then categorised into fifths of deprivation based on the population of the UK.

Day of the week was determined from the date of admission. In the UK, surgery for hip fracture usually takes place on the next available trauma operating list, which for most patients in the NHFD ($\geq 65\%$) is the day after admission. “Weekend” surgery was therefore identified by admission on a Friday or Saturday.

Hospital case volume was analysed by 10ths and defined by the number of people with displaced intracapsular fracture admitted to each centre over the study period.

Date of surgery was analysed as seven periods of six months (1 July 2011 to 31 December 2015) and one period of four months (1 Jan 2015 to 31 April 2015).

Statistical analysis

We determined compliance with guidelines with a decision tree ordered to mirror the NICE recommendations—that is, based on mobility (mobile outdoors with or without the use of a stick), cognition (defined as mental test score ≥ 8), and fitness for anaesthesia (defined as physical status score 1 or 2). Although the cut offs used for these two scores are not expressly published as part of the guideline, they have been used by NICE to monitor compliance with the guideline.¹⁴ A mental test score <8 has previously been shown to identify cognitive impairment¹⁵ and has been adopted as a threshold by the Royal College of Physicians of London.¹⁶ We determined the extent to which the NICE algorithm explained practice—that is, those individuals correctly classified as a percentage of the total.

We used recursive partitioning to determine the optimal decision tree that explains current practice—that is, to illustrate how the guidelines are being interpreted. Recursive partitioning is a statistical technique for multivariable analysis that models how variables are best organised to predict a given outcome (such as THA). Decision trees are built by identifying a variable that best splits the data into two groups. The partitioning process defines a cut off (split) for continuous or ordinal variables to enable the decision tree to correctly classify the maximum members of the population. Categorical

variables are similarly grouped to build a tree with the least error. This process is then applied separately to each subgroup and continues recursively until either a maximum number of steps are reached or no further improvement is possible.¹⁷

We undertook recursive partitioning using the “rpart” function in R. The tree was built with 10-fold cross validation and a negative complexity parameter to ensure that the maximum tree was built. Predictors included in the model were age, sex, mobility, cognition (AMTS), physical status (ASA score), fifth of index of multiple deprivation, and day of the week of admission. The tree was pruned with the complexity (“cp”) function of the smallest tree within one standard error of the best functioning tree—that is, the tree with the smallest error, which was confirmed graphically. We also used a pragmatic approach to consider the tree complexity and efficiency related to clinical practice.

Individuals who fulfilled the NICE criteria were further analysed to explore factors associated with undergoing THA. We constructed a recursive partitioning decision tree to differentiate between THA and no THA in this subgroup. The treating hospital was included as a factor variable, which allowed the partitioning algorithm to select optimal cut off points for best fit within the model.

We constructed a mixed effects logistic regression model to explore factors associated with the use of THA among patients who fulfilled the NICE criteria. Age, sex, date of surgery, cognition, and physical status were included as continuous predictors; and fifth of index of multiple deprivation and weekend surgery as categorical predictors. Weekend admission was then substituted for day of the week to explore this predictor further in a second analysis. Hospital case volume was included as a centre level fixed effect and the unique hospital identifier as a centre level random effect. We applied the same analysis to patients who did not fulfil the NICE criteria for THA to determine factors predictive of receiving a THA in this group.

Statistical analyses were performed with R and Stata version 14.0. $P < 0.05$ was adopted as the threshold for significance.

Patient involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community.

Results

In the 46 month period between 1 July 2011 and 31 April 2015, the NHFD recorded 248 013 patients with hip fracture. Of these, 114 119 satisfied the study criteria with a non-pathological displaced intracapsular hip fracture. Though 21 193 patients satisfied the NICE criteria to receive a THA (fig 1), only 11 683 within the NHFD underwent THA. Among these 11 683 patients, 4903 did not fulfil the NICE criteria.

The recursive partitioning algorithm identified 10 terminal nodes (nine splits) as the most predictive model, although this offered little improvement over five terminal nodes (four splits) (fig 2). The variable with the greatest importance was patient age, with a cut off age of 77 defining the initial split (fig 3). The mobility split occurred between patients who ambulate independently and those who required the use of a stick. The other important predictive variables were those recommended by NICE, with splits occurring as predicted at $ASA \geq 3$ for physical status and $AMTS \geq 8$ for cognition. With the decision tree, the explained practice across the dataset improved from 82.7% (NICE guidelines) to 90.4% (recursive model).

Among the 21 193 patients fulfilling the NICE eligibility criteria, the recursive partitioning algorithm identified 20 terminal nodes (19 splits) to be the most efficient, although after three splits (four terminal nodes), the complexity of the tree increased markedly with little associated gain in efficiency (fig 4). Again age was the most significant predictor, with aged 79 identifying

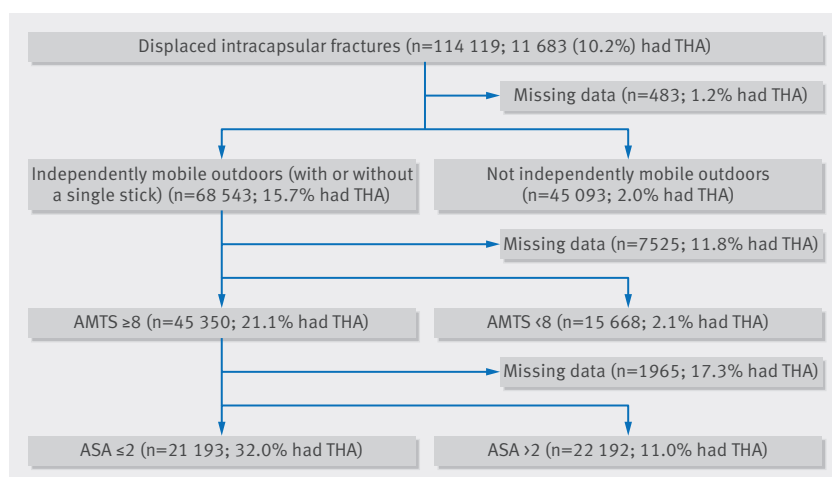


Fig 1 | Decision tree for total hip arthroplasty (THA) in displaced intracapsular fractures as per NICE guidelines. AMTS=abbreviated mental test score, ASA=American Society of Anesthesiologists score

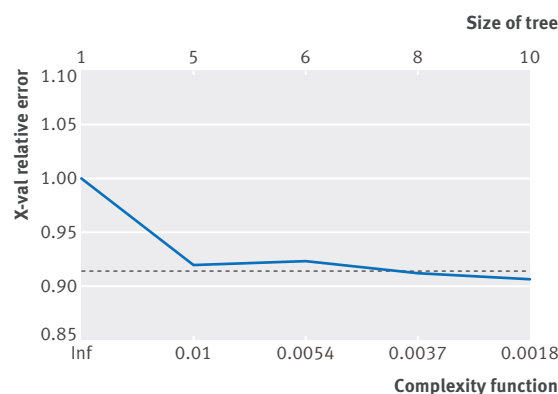


Fig 2 | Graph illustrating limited improvement in model using optimal tree size of 10 terminal nodes (lowest error), and more pragmatic tree with five nodes

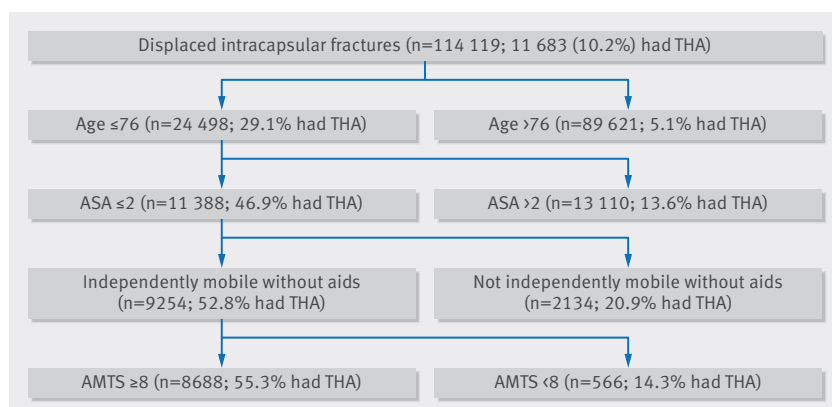


Fig 3 | Decision tree for total hip arthroplasty (THA) in displaced intracapsular fractures using recursive partitioning algorithm

the splitting point (fig 5). For patients aged ≥ 79 , the treating hospital was the next most important predictor (see appendix 2 for further details), followed by mobility (with or without the use of a stick). Hospital variation among individuals fulfilling the NICE guidelines was considerable (fig 6). Of the variation in practice, 77% could be explained using this recursive partitioning algorithm, compared with 32% by NICE guidelines alone.

Date of surgery showed that there was a progressive increase in the provision of THA for eligible individuals over the study period (table 1).

The logistic regression model (table 2) showed that 10ths of hospital volume did not affect THA (odds ratio 1.02, 95% confidence interval 0.97 to 1.08). Increasing age (0.88, 0.87 to 0.88), poorer cognition (AMTS) (0.49 (0.41 to 0.58) for 1.0 (ref) v borderline for cognitive impairment), and worsening physical status (ASA score) (0.74, 0.66 to 0.84), however, were associated with fewer procedures, as was male sex (0.85, 0.77 to 0.93). Admissions for surgery during the working week had the highest odds for receipt of THA (weekend admission 0.90, 0.83 to 0.98). There was a stepwise decrease in the odds of receiving THA with worsening

area deprivation, such that the most deprived fifth had the fewest procedures (0.76, 0.66 to 0.88).

We conducted a further analysis among individuals with a non-pathological displaced intracapsular fracture who did not fulfil the NICE eligibility criteria for THA ($n=92926$). Of these patients, 4903 underwent the procedure. With the same regression model, similar inequalities emerged. The receipt of THA outside the recommendations of NICE was least common among those with worse socioeconomic deprivation (odds ratio 0.64, 95% confidence interval 0.55 to 0.77), with a stepwise decrease from the most deprived fifth. Similarly, patients were less likely to receive THA outside the NICE guidelines when they were admitted at the weekend (0.89 (0.81 to 0.98) for all weekend, 0.87 (0.75 to 1.01) for Friday, 0.94 (0.81 to 1.09) for Saturday, 0.98 (0.84 to 1.15) for Sunday, 1.03 (0.89 to 1.19) for Monday, 1.01 (0.88 to 1.17) for Tuesday, 1.0 (reference) for Wednesday, and 1.01 (0.88 to 1.17) for Thursday.

Discussion

This observational study used a large national audit dataset and has shown that there is unexplained variation in the use of THA after a hip fracture. This surgery was influenced by several characteristics of patients, including age, sex, cognition (AMTS), physical status (ASA score), socioeconomic status, and mobility before the fracture. Other key determinants were the treating hospital and the day of the week of admission. The use of THA among eligible patients increased over the study period but remains both low and variable.

Compliance with NICE recommendations

NICE was established in 1999 to promote evidence based treatments and reduce unexplained variation in care across the NHS, the so called "postcode lottery."¹⁸ In June 2011, NICE recommended that THA should be offered to patients with a displaced intracapsular hip fracture who can walk independently outdoors (with no more than a single mobility aid), are cognitively intact, and are medically fit to undergo the operation. This guideline is consistent with a developing evidence base, which suggests that THA leads to better functional outcomes than hemiarthroplasty after hip fracture,⁷⁻¹¹ although a large scale intervention study is needed and is currently underway.¹⁹ Despite the NICE guideline, we found that variation in the use of THA persists across the NHS because of poor compliance with the guidelines. There was substantial variation in compliance (0.1-60%) between hospitals. As patient level predictors were unable to account for this variation, it is likely to reflect systematic differences in practice between centres.

The optimal recursive partitioning model suggested that surgeons might consider factors that could be relevant even if not strictly included within the NICE guidelines. For example, older patients were less likely to undergo THA, as were those who mobilised using a stick compared with those mobilising independently without aids. Although there is strong evidence that some patients with hip fracture benefit from THA,⁷⁻¹¹ its

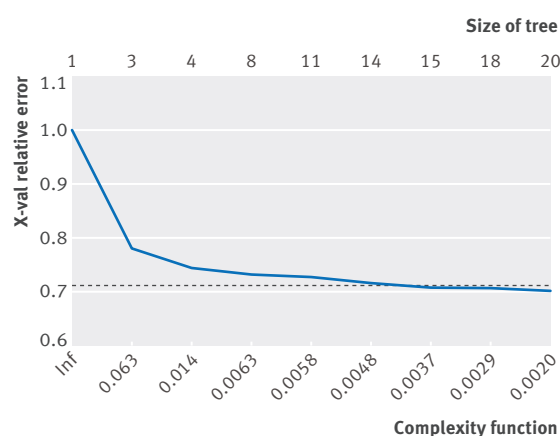


Fig 4 | Graph illustrating limited improvement in model using optimal tree size of 20 terminal nodes and simplified tree with four nodes

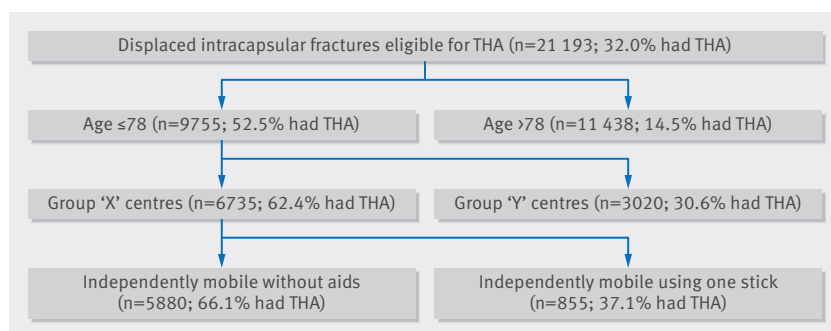


Fig 5 | Decision tree using recursive partitioning algorithm to indicate important predictors for total hip arthroplasty (THA) among individuals fulfilling NICE criteria for consideration of THA

precise indications are not well defined. Our model offers a glimpse into the collective judgment of orthopaedic surgeons and could be used to help inform the development of future NICE guidelines in the absence of higher level evidence. It is nevertheless concerning that deprivation was inversely associated with the use of THA. This observation persisted among patients who received a THA but did not meet the NICE guidelines, with deprived individuals least likely to inappropriately receive a THA. This is particularly important because NHS treatment is universally provided irrespective of ability to pay and “free at the point of use.” Challenging health inequalities is an ambition of initiatives aimed at increasing access to healthcare in other countries.²⁰ It is therefore important to understand reasons for socioeconomic inequalities that persist in public healthcare systems. There are many potential explanations for this observation, including patients’ preferences and confounding factors. It is also possible, however, that heuristic judgments about which patients are sufficiently “independent” to benefit from THA could be influenced by implicit surgeon bias. Social class biases have been shown to influence treatment decisions across a range of settings^{21–23} and could raise a barrier for patients who are otherwise eligible to undergo THA. This inverse association risks exacerbating health inequalities and

is a further reason to promote clear, evidence based, national guidelines.

Barriers to increased provision of THA

One potential obstacle to delivering THA for all eligible patients with hip fracture is the availability of experienced hip surgeons. It is widely accepted that patients undergoing elective THA by a low volume surgeon have greater risks of dislocation, need for revision surgery, postoperative complications, and death.^{24–28} For this reason, many orthopaedic surgeons do not perform THA for hip fracture if this operation is not part of their routine elective practice. The limited availability of suitably experienced hip surgeons might account for the reduced use of this procedure observed at weekends. This finding is important in the context of recent proposals to introduce seven day services across the NHS.²⁹ Although this discussion is principally framed around increased weekend mortality,^{30,31} timely access to THA for fracture might also need to be examined. Regionalisation of hip fracture services seems a plausible means of ensuring equal access to THA, by enabling specialist hip surgeons to support such a service every day. Dedicated hip fracture centres have already been successfully piloted in Germany.^{32,33} The potential benefits of regionalisation, however, would need to be weighed against competing considerations such as the desire of older adults to be treated close to their homes.

Strengths and limitations of study

The main strength of this study was its use of a dataset that captures almost every patient with hip fracture (>95%) treated in England, Wales, and Northern Ireland. There were variables that aligned closely with the NICE eligibility criteria, which permitted the recommended treatment algorithm to be mapped over the administrative data recorded within the NHFD.

The principal limitation was that the NHFD does not record individual patient comorbidities and so it was not possible to determine if specific comorbid diseases were associated with differences in the use of THA. Some of the variables in our analysis (such as age and deprivation) could simply represent a tendency towards a greater burden of comorbidity. The American Society of Anesthesiologists (ASA) score, however, has been shown to have equivalent or even greater predictive value for mortality and complications than standard

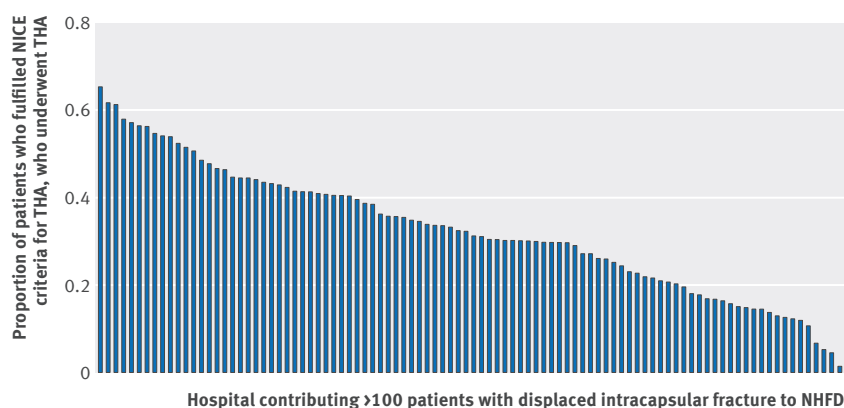


Fig 6 | Variation in number of total hip arthroplasty (THA) performed within each hospital as proportion of total number of individuals fulfilling NICE guidelines. Only hospitals that contributed >100 NICE eligible patients are included to minimise spurious data (n=96). Each bar represents one hospital

Table 1 | Proportion of eligible patients who underwent total hip arthroplasty (THA) by time period

Period	Individuals undergoing THA/individuals fulfilling NICE criteria for THA (%)
1 July 2011–31 Dec 2011 (6 months)	453/2020 (22)
1 Jan 2012–30th June 2012 (6 months)	649/2409 (27)
1 July 2012–31 Dec 2012 (6 months)	804/2703 (30)
1 Jan 2013–30th June 2013 (6 months)	942/3041 (31)
1 July 2013–31 Dec 2013 (6 months)	1007/3099 (32)
1 Jan 2014–30th June 2014 (6 months)	1104/3077 (36)
1 July 2014–31 Dec 2014 (6 months)	1160/3094 (37)
1 Jan 2015–30th April 2015 (4 months)	661/1750 (38)

Table 2 | Mixed effects logistic model showing odds of receiving total hip arthroplasty (THA) among those deemed eligible by NICE guidelines

Variable	OR (95% CI)	P value
Age (for each increasing year of life)	0.88 (0.87 to 0.88)	<0.001
AMTS:		
10 (maximum correct answers)	1.0 (ref)	—
9	0.69 (0.62 to 0.77)	<0.001
8 (borderline for cognitive impairment)	0.49 (0.41 to 0.58)	<0.001
ASA score:		
1 (healthy person)	1.0 (ref)	—
2 (mild systemic disease)	0.74 (0.66 to 0.84)	<0.001
Mobility:		
Walks independently without aids	1.0 (ref)	—
Walks with aid of one stick	0.32 (0.28 to 0.35)	<0.001
Sex:		
Female	1.0 (ref)	—
Male	0.85 (0.77 to 0.93)	0.002
Day of admission:		
Saturday	0.86 (0.75 to 0.99)	0.03
Sunday	0.88 (0.76 to 1.12)	0.09
Monday	0.93 (0.81 to 1.07)	0.32
Tuesday	0.94 (0.82 to 1.08)	0.41
Wednesday	1.00 (ref)	—
Thursday	1.00 (0.87 to 1.15)	0.97
Friday	0.85 (0.74 to 0.98)	0.03
Weekend admission*	0.90 (0.83 to 0.98)	0.01
Fifth of deprivation:		
1 (least deprived)	1.0 (ref)	—
2	0.98 (0.88 to 1.09)	0.68
3	0.92 (0.82 to 1.03)	0.13
4	0.82 (0.72 to 0.92)	0.001
5 (most deprived)	0.76 (0.66 to 0.88)	<0.001
Date of surgery (to most recent, increasing in 6 month intervals)	1.13 (1.10 to 1.15)	<0.001
Fracture volume of hospital (increasing in tenths)	1.02 (0.97 to 1.08)	0.46

*Defined as Friday or Saturday admission, as surgery most commonly occurs on day after admission. "Weekday" was therefore defined as Sunday-Thursday. "Weekend" included in logistic model as dichotomous variable and day of week excluded as collinear.

comorbidity measures, such as the Charlson comorbidity index).³⁴⁻³⁶ It is unlikely that patients assigned a score ≤ 2 (2="mild systemic disease") were medically unfit to undergo THA. The NHFD also does not include sufficient detail to understand clinical decision making at the individual patient level. For example, it is possible that THA was discussed with some patients and hemiarthroplasty was chosen after a balanced discussion of risk and benefit. The variation between hospitals in compliance with NICE guidelines, however, suggests that there is likely to be systematic differences with provision of THA.

Conclusion

Compliance with the NICE guidance on THA for hip fracture seems poor, with many apparently eligible patients not undergoing the procedure. There continues to be substantial variation in practice between hospitals, which is not readily explained by differences at the patient level. The limited use of THA among patients from deprived areas, the inappropriately high use among patients from more affluent areas, and inequalities in the provision of treatment at the weekend are particular concerns. Despite clear national guidelines,

it seems most likely that there are systematic differences with use of THA in hip fractures within this dataset. There have been substantial improvements in all of the quality indicators measured by the NHFD since its creation in 2004.³ The NHFD should consider reporting data on THA provision at the hospital level to help achieve greater consistency across the NHS.

Contributors: DCP designed the study, performed the analysis, and drafted the paper. DM contributed to the data analysis, interpretation of results, and draft manuscript. MLC and XLG contributed to the design of the study and interpretation of results and critically appraised the paper. All authors approved the final manuscript. DCP is guarantor.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf 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.

Ethical approval: The study protocol was approved by the Healthcare Quality Improvement Partnership (HQIP) before data release, but research ethics committee approval was not sought for secondary analysis of administrative data in line with Governance Arrangements for Research Ethics Committee (GAFREC) guidelines

Data sharing: Pursuant to the terms of our data sharing agreement with the National Hip Fracture Database no additional data can be made available by the authors.

Transparency: DP (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.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>.

- Royal College of Physicians. National Hip Fracture Database annual report 2014. London, 2014.
- Burge RT, Worley D, Johansen A, et al. The cost of osteoporotic fractures in the UK: projections for 2000-2020. *J Med Econ* 2001;4:51-62. doi:10.3111/200104051062.
- Neuburger J, Currie C, Wakeman R, et al. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in England: an external evaluation using time trends in non-audit data. *Med Care* 2015;53:686-91. doi:10.1097/MLR.0000000000000383.
- Sahota O, Currie C. Hip fracture care: all change. *Age Ageing* 2008;37:128-9. doi:10.1093/ageing/afn007.
- British Orthopaedic Association, British Geriatrics Society. *The Care of Patients with Fragility Fracture*. BOA, 2007.
- Department of Health. Payment by Results Guidance for 2013-14. DoH, 2013.
- Hopley C, Stengel D, Ekkernkamp A, Wich M. Primary total hip arthroplasty versus hemiarthroplasty for displaced intracapsular hip fractures in older patients: systematic review. *BMJ* 2010;340:c2332. doi:10.1136/bmj.c2332.
- National Institute for Health and Care Excellence. Hip fracture: the management of hip fracture in adults. NICE clinical guideline 124. NICE, 2011.
- Parker MJ, Gurusamy KS, Azegami S. Arthroplasties (with and without bone cement) for proximal femoral fractures in adults. *Cochrane Database Syst Rev* 2010;6:CD001706.
- Avery PP, Baker RP, Walton MJ, et al. Total hip replacement and hemiarthroplasty in mobile, independent patients with a displaced intracapsular fracture of the femoral neck: a seven- to ten-year follow-up report of a prospective randomised controlled trial. *J Bone Joint Surg Br* 2011;93:1045-8. doi:10.1302/0301-620X.93B8.27132.
- Yu L, Wang Y, Chen J. Total hip arthroplasty versus hemiarthroplasty for displaced femoral neck fractures: meta-analysis of randomized trials. *Clin Orthop Relat Res* 2012;470:2235-43. doi:10.1007/s11999-012-2293-8.
- Department of Health. *Governance arrangements for research ethics committees*. Department of Health, 2011.
- McLennan D, Barnes H, Noble M, et al. *The English Indices of Deprivation 2010*. Department for Communities and Local Government, 2011.

- 14 National Institute for Health and Care Excellence (NICE). Hip fracture: the management of hip fracture in adults [CG124]. Secondary Hip fracture: the management of hip fracture in adults [CG124] 2015. <https://www.nice.org.uk/guidance/cg124/uptake>.
- 15 MacKenzie DM, Copp P, Shaw RJ, Goodwin GM. Brief cognitive screening of the elderly: a comparison of the Mini-Mental State Examination (MMSE), Abbreviated Mental Test (AMT) and Mental Status Questionnaire (MSQ). *Psychol Med* 1996;26:427-30. doi:10.1017/S0033291700034826.
- 16 Young L, George J. *Guidelines for the diagnosis and management of delirium in the elderly*. British Geriatric Society, 1997.
- 17 Themeau TM, Atkinson EJ. *An introduction to recursive partitioning using rpart routines: The Comprehensive R Archive Network*. CRAN, 2015.
- 18 Vyawahare B, Hallas N, Brookes M, Taylor RS, Eldabe S. Impact of the National Institute for Health and Care Excellence (NICE) guidance on medical technology uptake: analysis of the uptake of spinal cord stimulation in England 2008-2012. *BMJ Open* 2014;4:e004182. doi:10.1136/bmjopen-2013-004182.
- 19 Bhandari M, Devereaux PJ, Einhorn TA, et al. HEALTH Investigators. Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial. *BMJ Open* 2015;5:e006263.
- 20 Abdus S, Mistry KB, Selden TM. Racial and Ethnic Disparities in Services and the Patient Protection and Affordable Care Act. *Am J Public Health* 2015;105(Suppl 5):S668-75. doi:10.2105/AJPH.2015.302892.
- 21 Haider AH, Schneider EB, Sriram N, et al. Unconscious race and social class bias among acute care surgical clinicians and clinical treatment decisions. *JAMA Surg* 2015;150:457-64. doi:10.1001/jamasurg.2014.4038.
- 22 Tamayo-Sarver JH, Dawson NV, Hinze SW, et al. The effect of race/ethnicity and desirable social characteristics on physicians' decisions to prescribe opioid analgesics. *Acad Emerg Med* 2003;10:1239-48. doi:10.1111/j.1553-2712.2003.tb00608.x.
- 23 Street RL Jr, O'Malley KJ, Cooper LA, Haidet P. Understanding concordance in patient-physician relationships: personal and ethnic dimensions of shared identity. *Ann Fam Med* 2008;6:198-205. doi:10.1370/afm.821.
- 24 Hedlundh U, Ahnfelt L, Hybbinette CH, Weckstrom J, Fredin H. Surgical experience related to dislocations after total hip arthroplasty. *J Bone Joint Surg Br* 1996;78:206-9.
- 25 Lavernia CJ, Guzman JF. Relationship of surgical volume to short-term mortality, morbidity, and hospital charges in arthroplasty. *J Arthroplasty* 1995;10:133-40. doi:10.1016/S0883-5403(05)80119-6.
- 26 Katz JN, Losina E, Barrett J, et al. Association between hospital and surgeon procedure volume and outcomes of total hip replacement in the United States medicare population. *J Bone Joint Surg Am* 2001;83-A:1622-9.
- 27 Losina E, Barrett J, Mahomed NN, Baron JA, Katz JN. Early failures of total hip replacement: effect of surgeon volume. *Arthritis Rheum* 2004;50:1338-43. doi:10.1002/art.20148.
- 28 Katz JN, Phillips CB, Baron JA, et al. Association of hospital and surgeon volume of total hip replacement with functional status and satisfaction three years following surgery. *Arthritis Rheum* 2003;48:560-8. doi:10.1002/art.10754.
- 29 Kleebauer A, Comerford C. Government commits to seven-day NHS. *Nurs Manag (Harrow)* 2015;22:6. doi:10.7748/nm.22.3.6.s2.
- 30 Freemantle N, Richardson M, Wood J, et al. Weekend hospitalization and additional risk of death: an analysis of inpatient data. *J R Soc Med* 2012;105:74-84. doi:10.1258/jrsm.2012.120009.
- 31 Keogh B. Should the NHS work at weekends as it does in the week? Yes. *BMJ* 2013;346:f621. doi:10.1136/bmj.f621.
- 32 Kelly M, Kates SL. Geriatrische Frakturzentren - verbesserte Patientenversorgung und ökonomische Vorteile: English Version. *Unfallchirurg* 2015.
- 33 Lau TW, Fang C, Leung F. The effectiveness of a geriatric hip fracture clinical pathway in reducing hospital and rehabilitation length of stay and improving short-term mortality rates. *Geriatr Orthop Surg Rehabil* 2013;4:3-9. doi:10.1177/2151458513484759.
- 34 Whitmore RG, Stephen JH, Vernick C, et al. ASA grade and Charlson Comorbidity Index of spinal surgery patients: correlation with complications and societal costs. *Spine J* 2014;14:31-8. doi:10.1016/j.spinee.2013.03.011.
- 35 Tan WP, Talbott VA, Leong QQ, Isenberg GA, Goldstein SD. American Society of Anesthesiologists class and Charlson's comorbidity index as predictors of postoperative colorectal anastomotic leak: a single-institution experience. *J Surg Res* 2013;184:115-9. doi:10.1016/j.jss.2013.05.039.
- 36 Dekker JW, Gooiker GA, van der Geest LG, et al. Use of different comorbidity scores for risk-adjustment in the evaluation of quality of colorectal cancer surgery: does it matter? *Eur J Surg Oncol* 2012;38:1071-8. doi:10.1016/j.ejso.2012.04.017.

© BMJ Publishing Group Ltd 2016

Appendix 1: Mobility scores

Appendix 2: Hospitals contributing to database

RESEARCH ARTICLE

Open Access



Total hip arthroplasty versus hemiarthroplasty for independently mobile older adults with intracapsular hip fractures

David Metcalfe^{1*} , Andrew Judge^{1,2}, Daniel C. Perry¹, Belinda Gabbe³, Cheryl K. Zogg⁴ and Matthew L. Costa¹

Abstract

Background: Displaced intracapsular hip fractures are typically treated with hemiarthroplasty (HA) or total hip arthroplasty (THA). A number of professional bodies recommend considering THA for patients that were independently mobile and cognitively intact before injury. The aim of this study was to compare the outcomes between HA and THA for independently mobile older adults with hip fractures.

Methods: A systematic review and meta-analysis of RCTs was undertaken alongside analysis of a propensity score matched national cohort of older adults (aged ≥ 60) with hip fractures. Participants were identified for the propensity score matched cohort from the National Hip Fracture Database (NHFD), which was linked to Hospital Episode Statistics (HES) and civil death registration data. The primary outcomes were 12-month dislocation, revision, and mortality. The secondary outcomes were length of stay, discharge home, unplanned re-admission, functional outcomes, and health-related quality of life.

Results: Five RCTs reported higher THA dislocation but this was not statistically significant (THA risk ratio [RR] 2.77, 95% CI 0.81 to 9.48). However, THA dislocation was significantly higher in the national observational dataset (sub-distribution hazard ratio [SHR] 1.73, 95% CI 1.24 to 2.41). Meta-analysis of data from four RCTs did not identify a significant difference in terms of revision (RR 1.52, 95% CI 0.56 to 4.14). However, THA revision was significantly lower in the national dataset (SHR 0.66, 95% CI 0.48 to 0.90). Meta-analysis of data from 5 RCTs suggested higher mortality amongst patients undergoing HA (RR 0.63, 95% CI 0.38 to 1.04), which was also observed within the national registry dataset (hazard ratio 0.45, 95% CI 0.37 to 0.54).

Conclusions: National clinical registries can provide important context when interpreting RCT data, which may alone be inadequate for comparing the safety profile of surgical interventions. These data suggest that THA is at significantly higher risk of dislocation but lower risk of revision within 12 months. The finding from both RCT and clinical registry data that THA is associated with lower 12-month mortality amongst the fittest patients with hip fractures requires urgent further study to determine whether or not this can be replicated in other balanced populations.

Keywords: Total hip replacement, Hemiarthroplasty, Hip fractures

* Correspondence: david.metcalfe@ndorms.ox.ac.uk

¹Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford OX3 9BU, UK
Full list of author information is available at the end of the article



Background

There are 70,000 hip fractures every year in the United Kingdom, with the total cost of care exceeding £2 billion per year. Mortality is high amongst these patients, with approximately 10% dying within 30 days of admission [1] and 30% within a year. Many survivors are unable to continue living independently and 4.5 million people worldwide are disabled every year by a hip fracture².

Most intracapsular hip fractures are displaced, such that the bone fragments are no longer in continuity. Displaced intracapsular fractures are either treated with hip hemiarthroplasty (HA), where the femoral head alone is replaced, or total hip arthroplasty (THA), where the femoral head and acetabulum are both replaced. Although HA is performed more frequently, a number of organisations (such as the American Academy of Orthopaedic Surgeons [AAOS] [2] and the UK National Institute for Health and Care Excellence [NICE] [3]) recommend offering THA to selected hip fracture patients owing to perceived functional benefits. NICE recommends offering THA to patients that (1) could walk independently before the fracture (2) are not cognitively impaired and (3) are medically fit for both anaesthesia and the procedure [3]. Despite this recommendation, an international survey of orthopaedic surgeons found that 73% favour HA [4], with studies demonstrating less than a third of eligible patients actually receive THA [5]. One explanation for this discrepancy is that the evidence in support of THA is mixed. A number of small randomised controlled trials have suggested that THA is associated with better functional outcomes, fewer wound infections, and reduced need for secondary procedures [6–9]. However, THA is also a more complex procedure that requires longer surgical time, is associated with greater blood loss, and has a higher risk of subsequent dislocation [10].

It is also uncertain whether the reported benefits for THA over HA [6–9] can be replicated beyond the controlled environment of clinical trials. For example, there is a clear association between THA outcome and surgeon volume [11] and it is likely that patients will be preferentially recruited to THA trials by experienced arthroplasty surgeons. It has been suggested that increasing the number of generalist surgeons providing THA will offset the benefits of this intervention for patients with hip fractures [2]. Similarly, there are concerns that the unavailability of appropriately trained arthroplasty surgeons might delay operative treatment. Surgical delays are thought to worsen outcomes for this vulnerable patient group [12, 13] and so might even worsen outcomes for patients selected to undergo THA. It is for these reasons that the “real world” effect of increasing use of THA in the hip fracture setting has been identified as a hip fracture research recommendation by the AAOS [2].

In this study we undertook an updated meta-analysis of RCTs and used data from a comprehensive national cohort of hip fractures to provide “real world” context to the existing trial literature. Our aim was to compare the outcomes between these two procedures for independently mobile older adults with hip fractures.

Methods

Systematic review and meta-analysis

A scoping review identified a number of previous systematic reviews that compared HA and THA for patients with displaced intracapsular hip fractures. We therefore employed a simplified search strategy using a modification of the method first proposed by Sampson et al. [14], which has been shown to be highly sensitive (median sensitivity 100%) for identifying RCTs when applied to systematic reviews with clinically focussed research questions [15]. We used a broad search strategy: (fracture* AND (“total hip” OR hemiarthroplasty) AND “systematic review”) to search three databases (Medline 1966–, EMBASE 1947–, and CINAHL 1982–) on 1st August 2018 to identify previous systematic reviews comparing HA and THA. The reference lists of all reviews were searched and the forward citation facility in PubMed used to identify trials published after each systematic review. Trial reference lists and citations were also searched for further studies. No language restrictions were applied. The full texts of all RCTs were then screened by two authors (DM and CZ) to identify those satisfying the following inclusion criteria. A single author (DM) evaluated studies published in Chinese with help from a Chinese-speaking health economist with experience of hip fracture research. The inclusion criteria were:

- A randomised or quasi-randomised controlled trial.
- Including patients predominantly aged ≥ 60 years with displaced intracapsular hip fractures.
- Excluding patients that had cognitive impairment or limited mobility before injury.
- Reporting dislocation, revision, mortality, unplanned re-admission, functional outcomes or health-related quality of life (using any validated scale).

Study characteristics and outcome data were extracted by one author (DM) and checked by a second (CZ). We planned to report all outcomes at 12-months for consistency. Two authors (DM and CZ) independently determined risk of bias using criteria recommended by the Cochrane Handbook [16] and resolved disagreements by consensus. These data were presented to guide judgements about the certainty of the evidence and not to determine eligibility for inclusion within meta-analyses. Data were pooled to estimate risk ratios (for categorical outcomes) and mean differences (for continuous outcomes)

using the DerSimonian and Laird method for random-effects meta-analysis as high levels of between-study heterogeneity were anticipated when pooling trials from different patient populations and healthcare settings [17]. Standardised mean differences were reported when studies reported the same outcome measured on difference scales. When studies did not provide standard deviations necessary to inform confidence intervals, these were calculated from absolute *p*-values [16]. Meta-analyses were undertaken using RevMan v.5.0 (Cochrane Collaboration, Vienna, Austria). The systematic review was reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [18] and the protocol registered prospectively in the PROSPERO database with reference CRD42018109415 [19].

Observational “real world” data

An observational study was undertaken using a comprehensive national cohort of older adults with displaced intracapsular hip fractures to extend and contextualise the existing RCT literature. Propensity score matching was used to mimic randomisation as far as is possible using observational data.

Data sources

The cohort was defined using the National Hip Fracture Database (NHFD) and patient records linked to administrative data (Hospital Episode Statistics) and civil death registrations.

National hip Fracture Database The National Hip Fracture Database (NHFD) is the largest hip fracture registry in the world. It is commissioned by the Healthcare Quality Improvement Partnership (HQIP) and captures data on almost all (> 95%) adults that are aged ≥ 60 years and admitted to hospital in England, Wales, or Northern Ireland with a proximal femoral fracture [20]. There were 177 hospitals contributing data to the NHFD in 2016 [21]. Data are collected by specialist nurses in each hospital and submitted through an online platform. Submissions are linked to hospital payments through the Hip Fracture Best Practice Tariff and so completeness of core variables is high.

Hospital episode statistics The Hospital Episode Statistics Admitted Patient Care (HES APC) dataset includes data on all admissions to National Health Service (NHS) hospitals or to independent sector providers that are funded by the NHS [22]. Approximately 99% of hospital activity in England is funded by the NHS [23] and so should be included within the HES APC. All activities are included that require a hospital bed (e.g. planned and emergency admissions) but outpatient and Emergency Department are excluded unless they lead to

admission. The dataset includes approximately 20 million episodes of care annually from around 450 individual NHS organisations [22].

Office for National Statistics The Office for National Statistics (ONS) captures data (including date and cause) on all registered deaths directly from civil registration records [24]. This dataset should therefore be complete except for the small number of cases referred to a coroner, which cannot be registered until coronial enquiries are complete and a death certificate has been issued.

Study population

The study period was 28th March 2011 until 4th January 2017. The start date was the earliest point at which the NHFD captured unique patient identifiers that could facilitate linkage to other datasets and the end date was chosen to facilitate 12 months follow-up. The inclusion criteria were those recommended by NICE [3]:

- All adults aged ≥ 60 .
- Displaced fracture of the femoral neck that was deemed unsuitable for internal fixation.
- Independently mobile or using a single stick before injury.
- Medically fit to undergo hip arthroplasty, defined as an American Society of Anaesthesiologists [ASA] grade ≤ 2 [5].
- Patients without substantial cognitive impairment, defined as an Abbreviated Mental Test Score (AMTS) ≥ 8 [5].

We excluded patients that presented to hospitals in Wales, Northern Ireland, and the Isle of Man as HES APC only captures data from hospitals in England. Cases were also excluded if they could not be positively matched to records within HES APC based on their NHS number, sex, date of birth, and full post-code.

Outcomes

The primary outcomes were dislocation, revision, and mortality within 12-months. The secondary outcomes were surgical delay, length of stay, discharge to own home, and re-admission within 30 days. Surgical delay, length of stay, and discharge destination were available directly from the NHFD. Revision operations were identified from HES APC and defined by OPCS v4 (OPCS4) procedure codes previously used in other studies and incorporating codes recommended for this purpose by the UK National Joint Registry [25] (Additional file 1). Dislocation OPCS4 codes were identified by manual searches using *disloc**, *manipula**, and *reduc** (Additional file 1).

Statistical analysis

Matching We calculated propensity scores that represented the estimated probability of each patient undergoing THA based on characteristics that are known to be associated with outcome in this population: age, sex, pre-injury mobility status, admission source, American Society of Anaesthesiologists (ASA) physical status grade, Charlson Co-morbidity Index (Additional file 1), Abbreviated Mental Test Score (AMTS), and Index of Multiple Deprivation (IMD) [26]. The model was otherwise specified iteratively to achieve the best possible match, as judged by visual inspection of the distribution of propensity scores after matching and plots of co-variables against propensity scores by treatment status. We also undertook post-estimation statistical checks [27], which included t-tests for differences in means and confirmation that the standardised mean difference for each co-variable between the groups was < 1% [28]. The final model utilised 1:1 nearest neighbour matching with a 0.02 calliper (as recommended by Austin [29]), no replacement, and the common support restriction. All subsequent descriptive, regression, and survival analyses were confined to the propensity score matched groups.

Descriptive statistics Categorical variables were compared using Chi-square tests and non-normally distributed continuous variables using the Kruskal-Wallis one-way analysis of variance test. Length of stay data were only analysed for the proportion of patients that were discharged alive from hospital to prevent left skew caused by early deaths.

Survival analysis Kaplan-Meier estimates were plotted with 95% confidence intervals for cumulative survival free from unplanned secondary procedures. The proportional hazards assumption was tested by visual examination and statistical assessment of the relationship between event time and Schoenfeld residuals. The proportional hazards assumption was satisfied and so we used Cox regression models fitted with our primary outcome (dislocation and/or revision) as the independent variable. Mortality is high in this population and so we undertook a sensitivity analysis using competing risks regression models with death specified as the competing risk. Competing risks regression models were also fitted for dislocation and revision as individual events. The co-variables for all regression models were those described above as the basis for propensity score matching, which include five of the six used routinely in the NHFD for case mix adjustment [30]. The sixth NHFD case mix co-variable (i.e. fracture type) was not used because only patients with displaced intracapsular hip fractures were

included in this study. Year of fracture was included as an ordinal variable within regression models to account for the possibility of changing outcomes over time.

Multivariable regression Multivariable logistic regression was used to adjust for residual imbalance between the two groups in respect of discharge to own home and 30-day re-admission. The co-variables were as specified above. Length of stay data conformed to a gamma distribution and so were adjusted using generalized linear models (GLM) together with post-estimation calculations of average marginal effects to yield predicted mean differences and 95% confidence intervals. Logistic regression and GLMs utilised cluster-robust standard errors and robust variance estimators [31] to account for the lack of independence between matched records [32].

Propensity score matching was achieved using the MatchIt application for R (R Foundation for Statistical Computing, Vienna, Austria). All subsequent analyses were undertaken using StataIC v.15 (StataCorp, College Station, TX, USA). Two tailed $p < 0.05$ was adopted a priori as the threshold for statistical significance.

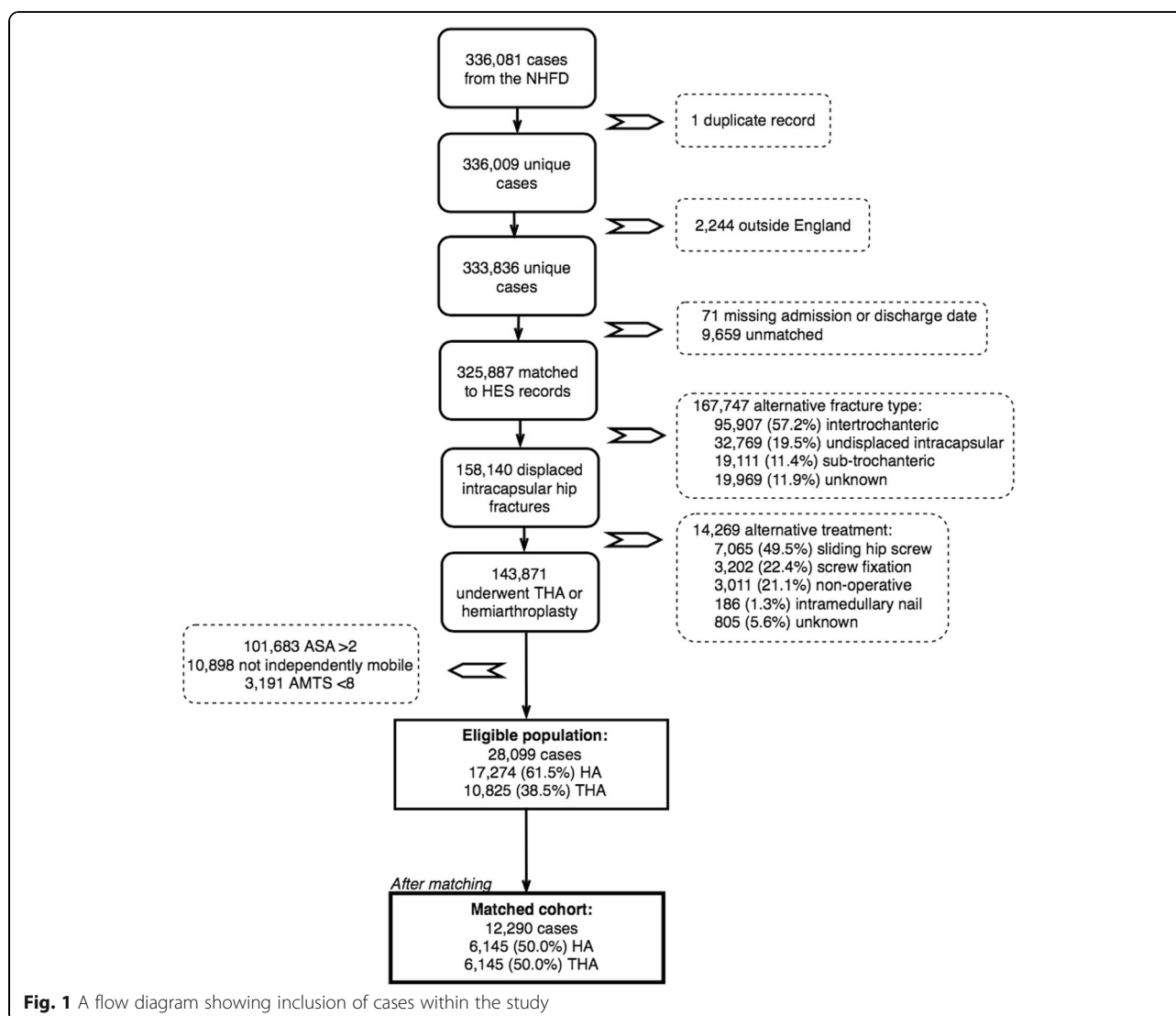
Results

Meta-analysis of randomised trials

There were 11 previous systematic reviews but none reported analyses limited to patients that were cognitively intact and independently mobile before injury (Additional file 2: Figure S1). The 11 earlier reviews included 16 trial reports, which presented data from 14 individual RCTs. Eight RCTs did not satisfy the restricted inclusion criteria of this systematic review, e.g. they did not exclude patients with cognitive impairment or limited mobility. One study could not be retrieved despite extensive attempts. The reasons for excluding each RCT are shown in Additional file 2: Table S1. Five randomised controlled trials satisfied the eligibility criteria for this review (Additional file 2: Table S2). Two were based in the UK [33, 34] and one each in Sweden [35], Italy [36], and the USA [9]. A further eligible RCT is on-going [37]. Characteristics of the RCTs and risk of bias assessments are described in Additional file 2. All the RCTs used adequate random sequence generation techniques and were judged to be at low risk of attrition bias as loss to follow-up was low. However, no RCT sought to blind patients, personnel, or outcome assessors.

Observational “real world” data

There were 143,871 patients with displaced intracapsular hip fractures that underwent HA or THA and could be matched to a record within HES APC (Fig. 1). 28,099 (19.5%) satisfied the pre-specified inclusion criteria, i.e. ASA < 2, AMTS \geq 8, and independently mobile. The groups initially varied considerably in terms of baseline



characteristics (Additional file 3). After propensity score matching, 12,290 cases were selected for further analysis. Table 1 shows that the baseline characteristics of the matched groups were similar. The distribution of propensity scores was also improved after matching (Additional file 3).

Primary outcomes

Dislocation

All five RCTs reported risk of dislocation. Although the pooled effect estimate suggested higher risk of dislocation amongst those undergoing THA, this was not significant (THA 9/233 [3.9%] versus HA 2/234 [0.9%], RR 2.77 [95% 0.81 to 9.48], Fig. 2). Within the propensity score matched cohort, those undergoing THA were significantly more likely to dislocate than those with HA (1.6% versus 0.9%, X^2 $p < 0.001$). This finding persisted when adjusting for co-variables in a competing

risks regression model (THA sub-distribution hazard ratio [SHR] 1.73, 95% CI 1.24 to 2.41, see Table 2).

Revision

All five RCTs reported risk of revision [9, 33–36]. The pooled effect estimate was initially in favour of HA, although this association was not statistically significant (HA 8/234 [3.4%] versus 15/233 [6.4%], RR 1.52 [95% CI 0.56 to 4.14], Fig. 3). The association also diminished when the data reported by Cadossi et al. [36] were excluded as these authors had trialled a non-standard THA prosthesis and reported an unusually high revision rate (HA 8/193 [4.1%] versus 9/186 [4.8%], RR 1.16 [95% CI 0.46 to 2.91]). However, within the propensity score matched cohort, a greater proportion of HA patients underwent revision surgery within the subsequent 12 months than THA (1.7% versus 1.1%, X^2 $p < 0.001$). This finding persisted

Table 1 Characteristics of the matched population

	Hemiarthroplasty	Total hip arthroplasty	Total	P
Age ^c	77 (72–81)	77 (73–81)	77 (73–81)	0.571 ^a
Sex ^d				
Male	1347 (21.9%)	1321 (21.5%)	2668 (21.7%)	
Female	4798 (78.1%)	4824 (78.5%)	9622 (78.3%)	0.569 ^b
ASA ^c	2 (2–2)	2 (2–2)	2 (2–2)	0.675 ^a
Pre-injury mobility ^d				
Independently mobile	5308 (86.7%)	5326 (86.7%)	10,634 (86.7%)	
Mobile indoors with one aid	837 (13.6%)	819 (13.3%)	1656 (13.5%)	0.634 ^b
AMTS ^c	10 (10–10)	10 (10–10)	10 (10–10)	0.457 ^a
Admission source ^d				
Own home	6071 (98.8%)	6092 (99.1%)	12,163 (99.0%)	
Rehabilitation unit	8 (0.1%)	2 (0.0%)	10 (0.1%)	
Residential/nursing home	37 (0.6%)	18 (0.3%)	55 (0.5%)	
Acute hospital	29 (0.5%)	33 (0.5%)	62 (0.5%)	0.015 ^b

*Median (interquartile range); **number (percentage); ^aKruskall-Wallis one-way analysis of variance; ^bChi2 test

when adjusting for co-variables in a competing risks regression model (THA SHR 0.66, 95% CI 0.48 to 0.90).

Mortality

Four RCTs reported mortality at 12 months and one at 6 months. A higher proportion of patients undergoing HA died (36/234, 15.4%) than those in the THA group (21/233, 9.0%, RR 0.63, 95% CI 0.38 to 1.04, Fig. 4). Within the propensity score matched cohort, 12-month mortality was higher in the HA group (5.4% versus 2.6%, X^2 $p < 0.001$) and this persisted within a multi-level flexible

parametric survival model (hazard ratio 0.45, 95% CI 0.37 to 0.54). Twelve-month mortality within the observational cohort is illustrated by a Kaplan-Meier plot in Fig. 5.

Secondary outcomes

Time to surgery

Two RCTs [33, 36] (164 patients) reported no difference in time to surgery between THA and HA (THA mean difference – 0.44 [95% CI – 0.93 to 0.05]). Within the propensity score matched cohort, patients underwent HA more promptly than THA (median

Table 2 Clinical outcomes for patients by operation

	Hemiarthroplasty	Total hip arthroplasty	P
Primary outcomes			
Dislocation (12 months)	57 (0.9%)	96 (1.6%)	0.002 ^a
	THA sub-distribution hazard ratio 1.73 (CI 1.24 to 2.41) ^b		
Revision (12 months)	106 (1.7%)	67 (1.1%)	< 0.001 ^a
	THA sub-distribution hazard ratio 0.66 (0.48 to 0.90) ^b		
Mortality (12 months)	58 (5.5%)	159 (2.6%)	< 0.001 ^a
	THA hazard ratio 0.45 (95% CI 0.37 to 0.54) ^c		
Secondary outcomes			
Surgical delay (hours) ^d	22.2 (17.8–29.0)	23.9 (18.9–40.6)	< 0.001 ^e
Length of stay (days) ^d	10 (7–15)	9 (7–13)	< 0.001 ^e
	THA predicted mean difference – 1.92 (95% CI – 2.30 to – 1.55) days ^f		
Discharge home	5017 (80.7%)	5519 (88.6%)	< 0.001 ^a
	THA adjusted odds ratio 1.77 (95% CI 1.58 to 1.99) ^g		
Re-admission (30-days)	361 (5.9%)	356 (5.8%)	0.847 ^a
	THA adjusted odds ratio 0.96 (95% CI 0.82 to 1.11) ^g		

^aChi² test; ^bCompeting risks regression model; ^cRoyston-Parmar flexible parametric model; ^dMedian (interquartile range); ^eKruskall-Wallis one-way analysis of variance; ^fPredicted mean difference from a generalized linear model; ^gMultivariable logistic regression model

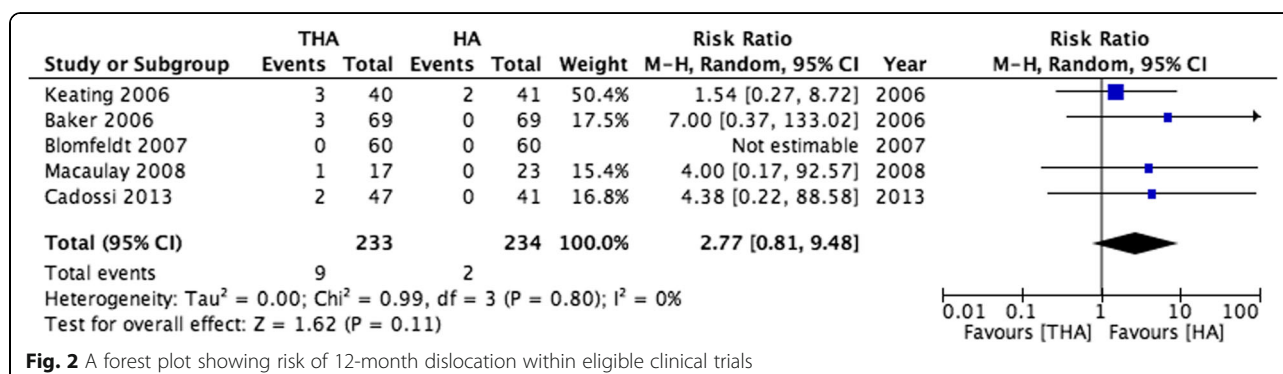


Fig. 2 A forest plot showing risk of 12-month dislocation within eligible clinical trials

22.2 [interquartile range (IQR) 17.8–29.0] hours versus 23.9, 18.9–40.6 h, Kruskal-Wallis $p < 0.001$).

Duration of surgery

All five RCTs (462 patients) reported surgical duration. Although THA took longer than HA, and this difference was statistically significant, the absolute effect was small (mean difference 15.0 [95% CI 6.4 to 23.7] minutes). Duration of surgery was not available from the propensity score matched cohort.

Length of stay

Two RCTs (123 patients) reported length of stay and there was no intervention effect on this outcome (THA mean difference 1.50 [95% CI 0.00 to 3.00] days). In the propensity score matched cohort, patients undergoing HA stayed in hospital longer than those undergoing THA (median 10 [IQR 7–15] versus 9 [7–13] days, Kruskal-Wallis, $p < 0.001$). When adjusting for co-variables within a generalised linear model, patients undergoing THA experienced a shorter length of stay (predicted mean difference -1.92 [95% CI -2.30 to -1.55] days).

Discharge destination

No RCT reported discharge destination as an outcome. Within the propensity score matched cohort, a smaller proportion of patients undergoing HA were discharged

to their own home than THA (80.7% versus 88.6%, $X^2 p < 0.001$). Within a multivariable logistic regression model, those undergoing THA also had higher odds of being discharged to their own home (adjusted odds ratio [aOR] 1.77, 95% CI 1.58 to 1.99).

30-day readmission

No RCT reported unplanned readmission to hospital as an outcome. Within the propensity score matched cohort, there was no statistically significant difference in 30-day re-admission between the two groups (HA 5.9% versus THA 5.8%, $X^2 p = 0.847$), and this finding persisted within a multivariable logistic regression model (aOR 0.96, 95% CI 0.82 to 1.11).

Hip functional outcomes

All five RCTs reported joint-specific functional outcomes measured at 12-months. Three studies used the Harris Hip Score [9, 35, 36] (234 patients) and one each used the Oxford Hip Score [33] (81 patients), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) [9] (40 patients), and a bespoke hip questionnaire [34] (138 patients). Higher scores on all of these measures reflect better outcomes except for the Oxford Hip Score in which a higher score represents worse function. There were no differences in terms of total score (THA standardised mean difference [SMD] 0.17 [95% CI -0.20 to 0.53]) or either the pain (-0.01

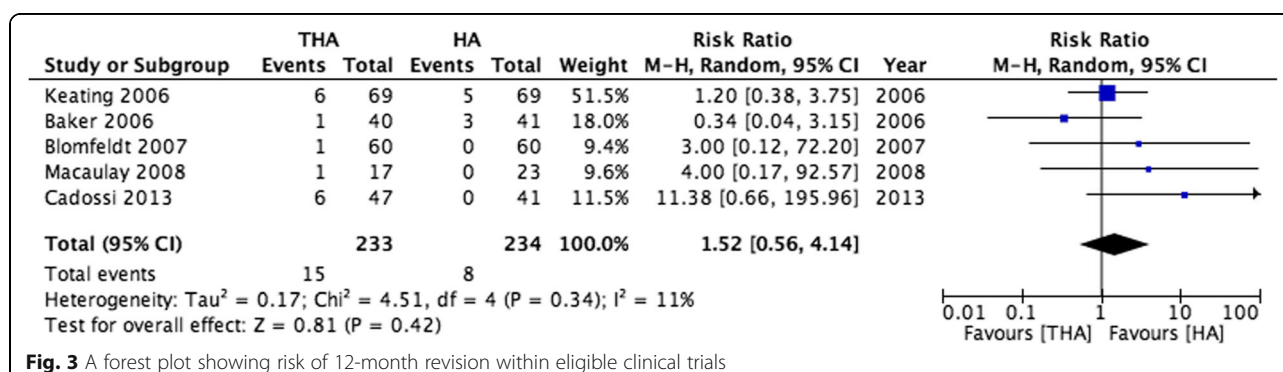


Fig. 3 A forest plot showing risk of 12-month revision within eligible clinical trials

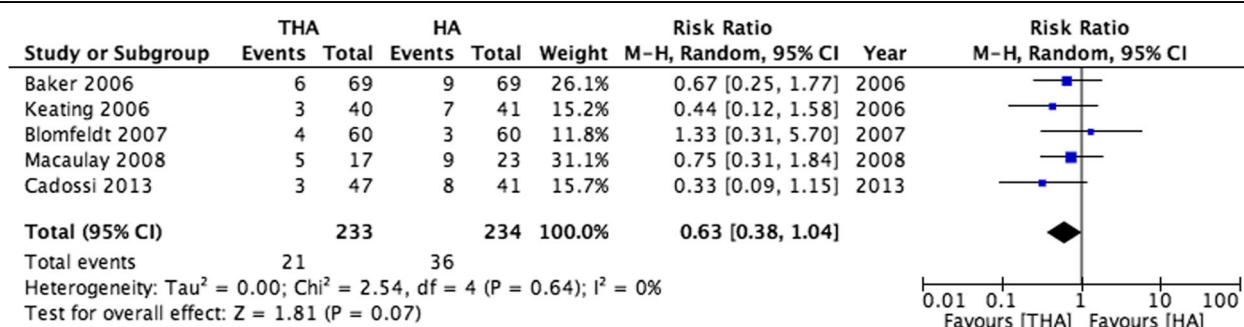


Fig. 4 A forest plot showing risk of 12-month mortality within eligible clinical trials

[−0.49 to 0.48]) or function (0.18 [−0.03 to 0.39]) domains. The only study using the Oxford Hip Score reported a difference between the groups in favour of THA (THA mean difference −3.50 [95% CI −6.66 to −0.34]). However, the only study reporting data from a “Timed Up and Go” (TUG) test [9] (40 patients) – which measures the time that it takes a patient to rise from a chair, walk three metres, turn around, walk back to the chair, and sit down – did not find a difference between the groups (THA mean difference −0.70 [95% CI −8.01 to 6.61] seconds).

Health-related quality of life

Two studies [9, 33] (121 patients) reported components of the Short Form (36) Health Survey (SF-36) and one the EQ-5D [34] (183 patients). There were no differences in the mental (THA mean difference 2.30 [95% CI −8.57 to 13.18]) or physical (2.98 [−0.89 to 6.85]) component

summary scores of the SF-36 or the EQ-5D (0.10 [0.00 to 0.20]) utility score.

Discussion

No previous meta-analysis has reported data limited to the fittest patients with hip fractures, which are the patients that national guidelines recommend should be considered for THA [2, 3]. This study identified five RCTs that compared HA and THA amongst independently mobile older adults with displaced intracapsular hip fractures [9, 33–36]. These trials were typically small (median 89 patients) single-centre studies that were limited by few events (pooled totals 11/467 [2.4%] dislocations, 23/467 [4.9%] revisions, and 57/467 [12.2%] deaths). No individual trial reported differences in outcomes and it is even possible that the pooled analyses were underpowered to detect important differences between the groups. We therefore analysed data from the largest available cohort of hip fracture patients and used

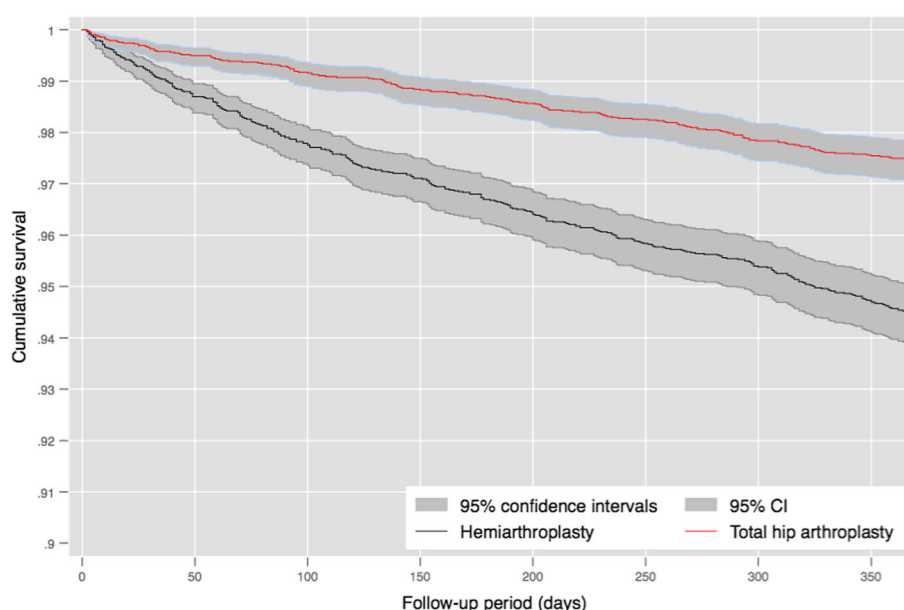


Fig. 5 Kaplan-Meier plot showing mortality for patients in the propensity score matched cohort

propensity score matching to replicate randomisation as far as is possible using observational data. The observational data confirmed the non-significant trend reported by RCTs that THA has a higher risk of 12-month dislocation. However, we found a 33% lower risk of 12-month revision for THA patients, which is contrary to the RCT finding of “no difference” between the groups observed in the RCTs.

Importantly, we identified a 58% lower risk of 12-month mortality for patients undergoing THA. Although this may reflect residual confounding, a similar association was evident from the meta-analysis of data from all five trials. One possibility is that the increased power available from the observational cohort has confirmed an association initially evident in the RCT data. This finding would however need to be replicated in further studies before it could be used to guide surgical decisions.

We also presented data that has not previously been reported by RCTs, including time to surgery, length of stay, discharge destination, and 30-day re-admission. Our study found that patients undergoing THA waited longer for an operation (approximately 1.7 h), although this delay is unlikely to be clinically significant. Although the AAOS have expressed concern that increased provision of THA might lead to operative delays [2], our study suggests that hospitals in England are providing THA within a timeframe that is comparable to HA. We found that THA was associated with a shorter length of stay (by approximately 1.9 days) and increased odds of discharge home. However, there was no difference between the groups in terms of 30-day re-admission.

There was mixed evidence from the RCTs as to whether or not functional outcomes or health-related quality of life vary between the groups at 12-months. The meta-analyses did not identify any statistically significant differences, although one study reported significantly better Oxford Hip Scores in the THA group [33]. There is however evidence to suggest that the functional benefits of THA become more pronounced over a number of years follow-up [7].

There is one on-going RCT [37] that might – either in isolation or when combined with data from previous trials – report sufficient events to identify differences between the two operations. However, the AAOS has expressed concern that the benefits of THA might not be generalisable beyond the controlled environment of clinical trials [2]. The RCTs identified in this study were all based in large academic centres and two [35, 36] specified that operations were only performed by experienced arthroplasty surgeons. Observational datasets can provide important context for RCT findings as they reflect “real world” practice in which operations may also be performed in smaller orthopaedic units, by generalist

orthopaedic surgeons, and by trainees. It is therefore reassuring that, although the propensity score matched cohort mirrored the RCT participants in terms of HA dislocation rate (both 0.9%), the THA dislocation rate was *lower* in the observational cohort than reported by trials (1.6% versus 3.9%). There were also fewer revisions identified in the propensity score matched cohort than were reported by the RCTs (THA 1.1% versus 1.7%; RCT 4.8% versus 4.1%). Although it is possible that some dislocations and revision procedures were not captured by the linked dataset, our findings are similar to those of a recent population-based study from Canada [11]. These authors reported findings that were the same in both magnitude and direction (THA dislocation 1.9% versus 0.8%; revision 0.4% versus 2.3%) as observed in our study. It is therefore possible that contemporary prostheses perform better (in terms of major hip complications) than those used in trials undertaken between 2006 and 2013. Our findings do not support the hypothesis that THAs undertaken outside RCTs are more prone to dislocation and early revision.

Limitations

There are a number of limitations to our approach. First, although extensive attempts were made to account for case-mix differences within the cohort study, it is possible that some findings were subject to residual confounding, which would be expected to bias findings against HA as surgeons are encouraged to reserve THA for the fittest patients. However, it is important that a similar signal was observed within the RCT data, which should be much more resistant to confounding. Second, as the NHFD was established to audit hip fracture care, it does not collect some variables (e.g. surgical approach) that might be found in a dedicated hip fracture registry. Surgical approach is known to be associated with dislocation [38] and this may be a further source of confounding. Third, coding errors are inevitable within the NHFD and HES. However, the NHFD has almost complete case capture and all re-admissions to hospitals in England over the subsequent 12 months should have been represented within HES. It is nevertheless possible that some events will not have recorded within HES. Although all arthroplasty revision procedures would have been within the context of an inpatient admission, some dislocations (e.g. those reduced and discharged home directly from the Emergency Department) might not have been captured by our study. Previous work in other surgical settings has found that OPCS4 codes in HES can reliably be used to identify some operations, although this can vary substantially between procedures [39]. However, a range of codes were used to define “revision surgery” and this selection might have influenced the findings. Nevertheless, our dislocation and revision

rates were reassuringly similar to those reported by a recent population-based study from Canada [11]. Finally, there is evidence that the functional and health-related quality of life benefits of THA only become apparent after a number of years [7]. This study sought to compare early complications and chose 12-month follow-up as a means of directly comparing RCT findings with those from a national cohort of comparable patients with hip fractures. It is however possible that our meta-analyses understated functional benefits of THA in this population.

Conclusion

This study found that concerns about increased provision of THA leading to clinically significant delays for older adults with hip fractures are unfounded. Similarly, there was not any evidence that dislocation or revision rates are higher in England outside the context of clinical trials. The finding of increased mortality amongst patients undergoing HA requires urgent further study to determine whether or not this can be replicated in other balanced populations.

Additional files

Additional file 1: Codes for defining Charlson co-morbidities* (DOCX 19 kb)

Additional file 2: Figure S1. PRISMA flow diagram showing identification of randomised and quasi-randomised controlled trials from previous systematic reviews. **Table S1.** Characteristics of excluded studies. **Table S2.** Characteristics of included studies. **Table S3.** Risk of bias assessments for included studies. (DOCX 321 kb)

Additional file 3: Table S1. Characteristics of the unmatched population. **Figure S1.** Histograms showing the distribution of propensity scores before and after matching. **Figure S2.** Quantile-quantile plots of co-variables between the two groups before and after matching. *Data from populations with the same empirical distribution will lie along the 45 degree reference line.* **Figure S3.** Co-variables plotted against propensity scores by treatment status. *If the two are identical, this indicates that the groups have the same mean for each value of the propensity score and so are well matched.* **Figure S4.** A jitter plot showing the overall distribution of propensity scores for both matched and unmatched records. (DOCX 689 kb)

Abbreviations

AAOS: American Academy of Orthopaedic Surgeons; AMTS: Abbreviated mental test score; aOR: Adjusted odds ratio; ASA: American Society of Anesthesiologists; CAG: Confidentiality advisory group; CCI: Charlson co-morbidity index; CI: Confidence interval; DARG: Data access request group; GAFReC: Governance arrangements for Research Ethics Committees; GLM: Generalised linear model; HA: Hemiarthroplasty; HES: APC Hospital Episode Statistics Admitted Patient Care; HES: Hospital Episode Statistics; HQIP: Healthcare quality improvement partnership; HR: Hazard ratio; IMD: Index of multiple deprivation; IQR: Interquartile range; NHFD: National hip fracture database; NHS: National Health Service; ONS: Office for National Statistics; OPCS: Office of Population Censuses and Surveys; PRISMA: Preferred reporting items for systematic reviews and meta-analyses; RCT: Randomised controlled trial; RR: Risk ratio; SHR: Standardized sub-distribution hazard ratio; SMD: Standardised mean difference; THA: Total hip arthroplasty; TUG: Timed up and go; WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

Acknowledgements

HQIP, NHS Digital, and the Office for National Statistics for supplying data. NHS Digital and Crown Informatics Ltd. for support with data linkage. Chris Boulton (Royal College of Physicians), Denise Pine (NHS Digital), and Tim Bunning (Crown Informatics Ltd) for managing data flows. The Bodleian Library (University of Oxford), The British Library, and the Georgetown University Library (USA). Dr. May Ee Png (University of Oxford) for assistance with evaluating studies published in Chinese.

Funding

Data access was funded by a Royal College of Surgeons of Edinburgh (RCSEd) Pump Priming Grant and an Oxford-UCB Prize Fellowship in Biomedical Research. Cheryl K. Zogg is supported by NIH Medical Scientist Training Program Training Grant T32GM007205. Andrew Judge is supported by the NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. Daniel Pery is supported by a National Institute for Health Research (NIHR) Clinician Scientist Fellowship (NIHR/CS/2014/14/012). All authors carried out this research independently of the funding bodies. The views expressed in this publication are those of the authors and do not necessarily reflect those of the NHS, the National Institute for Health Research, or the Department of Health and Social Care.

Availability of data and materials

Pursuant to the terms of our data sharing agreement with individual data controllers, we regret that no additional data can be made available by the authors.

Authors' contributions

DM, DP, and MC designed the study with advice from AJ, BG, and CZ. DM undertook the data analysis with support from AJ, DP, and CZ. DM interpreted the results and drafted the paper. AJ, DP, BG, CZ, and MC interpreted the results, and made critical revisions to the manuscript.

Ethics approval and consent to participate

Linkage of NHFD data was supported by the HQIP Data Access Request Group (DARG) and the Confidentiality Advisory Group (CAG) on behalf of the Secretary of State for Health and Social Care under s.251 NHS Act 2006 (CAG ref. 8–03(PR11)/2013). Access to HES APC was approved by the NHS Digital Independent Group Advising on the Release of Data (IGARD) and civil registration mortality data by the ONS Microdata Release Panel (MRP) under s.39(5) Statistics and Registration Service Act 2007. Formal research ethics committee approval was not required for secondary analysis of pseudonymised data in line with the NHS Health Research Authority Governance Arrangements for Research Ethics Committees (GAFREC) guidelines [40].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford OX3 9BU, UK.

²Musculoskeletal Research Unit, Translational Health Sciences, Bristol Medical School, University of Bristol, Learning and Research Building, Level 1, Southmead Hospital, Bristol BS10 5NB, UK. ³School of Public Health and Preventive Medicine, Monash University, Level 3, 553 St Kilda Road, Melbourne, VIC 3004, Australia. ⁴Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA.

Received: 15 November 2018 Accepted: 25 April 2019

Published online: 17 May 2019

References

1. Neuburger J, Currie C, Wakeman R, Tsang C, Plant F, De Stavola B, Cromwell DA, van der Meulen J. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in England: an external evaluation using time trends in non-audit data. *Med Care*. 2015;53(8):686–91.
2. Moderate evidence supports a benefit to total hip arthroplasty in properly selected patients with unstable (displaced) femoral neck fractures. 2015. [<http://www.orthoguidelines.org/guideline-detail?id=1239>].
3. National Institute for Health and Care Excellence (NICE). Hip fracture: management. In: *Clinical Guidelines*. London: NICE; 2017.
4. Bhandari M, Devereaux PJ, Tornetta P 3rd, Swiontkowski MF, Berry DJ, Haidukewych G, Schemitsch EH, Hanson BP, Koval K, Dirschl D, et al. Operative management of displaced femoral neck fractures in elderly patients. An international survey. *J Bone Joint Surg Am*. 2005;87(9):2122–30.
5. Perry DC, Metcalfe D, Griffin XL, Costa ML. Inequalities in use of total hip arthroplasty for hip fracture: population based study. *BMJ*. 2016;353:i2021.
6. Burgers PT, Van Geene AR, Van den Bekerom MP, Van Lieshout EM, Blom B, Aleem IS, Bhandari M, Poolman RW. Total hip arthroplasty versus hemiarthroplasty for displaced femoral neck fractures in the healthy elderly: a meta-analysis and systematic review of randomized trials. *Int Orthop*. 2012;36(8):1549–60.
7. Hedbeck CJ, Enocson A, Lapidus G, Blomfeldt R, Tornkvist H, Ponzer S, Tidermark J. Comparison of bipolar hemiarthroplasty with total hip arthroplasty for displaced femoral neck fractures: a concise four-year follow-up of a randomized trial. *J Bone Joint Surg Am*. 2011;93(5):445–50.
8. Keating JF, Grant A, Masson M, Scott NW, Forbes JF. Displaced intracapsular hip fractures in fit, older people: a randomised comparison of reduction and fixation, bipolar hemiarthroplasty and total hip arthroplasty. *Health Technol Assess*. 2005;9(41):iii–iv, ix–x, 1–65.
9. Macaulay W, Nellans KW, Lorio R, Garvin KL, Healy WL, Rosenwasser MP, Consortium D. Total hip arthroplasty is less painful at 12 months compared with hemiarthroplasty in treatment of displaced femoral neck fracture. *HSS J*. 2008;4(1):48–54.
10. Liao L, Zhao J, Su W, Ding X, Chen L, Luo S. A meta-analysis of total hip arthroplasty and hemiarthroplasty outcomes for displaced femoral neck fractures. *Arch Orthop Trauma Surg*. 2012;132(7):1021–9.
11. Ravi B, Jenkinson R, Austin PC, Croxford R, Wasserstein D, Escott B, Paterson JM, Kreder H, Hawker GA. Relation between surgeon volume and risk of complications after total hip arthroplasty: propensity score matched cohort study. *BMJ*. 2014;348:g3284.
12. Bretherton CP, Parker MJ. Early surgery for patients with a fracture of the hip decreases 30-day mortality. *Bone Joint J*. 2015;97-B(1):104–8.
13. Fu MC, Boddapati V, Gausden EB, Samuel AM, Russell LA, Lane JM. Surgery for a fracture of the hip within 24 hours of admission is independently associated with reduced short-term post-operative complications. *Bone Joint J*. 2017;99-B(9):1216–22.
14. Sampson M, Shojania KG, McGowan J, Daniel R, Rader T, Iansavichene AE, Ji J, Ansari MT, Moher D. Surveillance search techniques identified the need to update systematic reviews. *J Clin Epidemiol*. 2008;61(8):755–62.
15. Rice M, Ali MU, Fitzpatrick-Lewis D, Kenny M, Raina P, Sherifali D. Testing the effectiveness of simplified search strategies for updating systematic reviews. *J Clin Epidemiol*. 2017;88:148–53.
16. Higgins J, Green S. *Cochrane handbook for systematic reviews of interventions v.5.1.0*. Available from www.handbook.cochrane.org: the Cochrane collaboration; 2011.
17. DerSimonian R, Laird N. “Meta-analysis in clinical trials”. *Control Clin Trials*. 1986;7(3):177–
18. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;339:b2535.
19. Total hip arthroplasty versus hemiarthroplasty for independently mobile older adults with intracapsular hip fractures. 2018. [http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018109415].
20. Royal College of Physicians. National Hip Fracture Database (NHFD) Annual Report 2015. London: Royal College of Physicians of London; 2015.
21. Royal College of Physicians. National Hip Fracture Database (NHFD) Annual Report 2017. London: Royal College of Physicians of London; 2017.
22. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). *Int J Epidemiol*. 2017;46(4):1093–1093i.
23. National Audit Office. Healthcare across the UK: a comparison of the NHS in England, Scotland, Wales and Northern Ireland. London: National Audit Office; 2012.
24. Office for National Statistics. User Guide to Mortality Statistics. Swansea: Office for National Statistics; 2017.
25. Hip Revision OPCS4 Procedure Codes, 2012. [<http://www.njrcentre.org.uk/njrcentre/Portals/0/Documents/OPCS4%20Procedure%20Codes%20used%20in%20NJR%20Annual%20Report.pdf?ver=2012-02-15-165150-000>].
26. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006; 163(12):1149–56.
27. Leuven E, Sianesi B. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphic, and covariate imbalance testing. In: *Statistical Software Components S432001*. Boston: Boston College Department of Economics; 2012.
28. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083–107.
29. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51(1):171–84.
30. Tsang C, Boulton C, Burgon V, Johansen A, Wakeman R, Cromwell DA. Predicting 30-day mortality after hip fracture surgery: evaluation of the National hip Fracture Database case-mix adjustment model. *Bone Joint Res*. 2017;6(9):550–6.
31. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics*. 2000;56(2):645–6.
32. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32(16):2837–49.
33. Baker RP, Squires B, Gargan MF, Bannister GC. Total hip arthroplasty and hemiarthroplasty in mobile, independent patients with a displaced intracapsular fracture of the femoral neck. A randomized, controlled trial. *J Bone Joint Surg Am*. 2006;88(12):2583–9.
34. Keating JF, Grant A, Masson M, Scott NW, Forbes JF. Randomized comparison of reduction and fixation, bipolar hemiarthroplasty, and total hip arthroplasty. Treatment of displaced intracapsular hip fractures in healthy older patients. *J Bone Joint Surg Am*. 2006;88(2):249–60.
35. Blomfeldt R, Tornkvist H, Eriksson K, Soderqvist A, Ponzer S, Tidermark J. A randomised controlled trial comparing bipolar hemiarthroplasty with total hip replacement for displaced intracapsular fractures of the femoral neck in elderly patients. *J Bone Joint Surg Br*. 2007;89(2):160–5.
36. Cadossi M, Chiarello E, Savarino L, Tedesco G, Baldini N, Faldini C, Giannini S. A comparison of hemiarthroplasty with a novel polycarbonate-urethane acetabular component for displaced intracapsular fractures of the femoral neck: a randomised controlled trial in elderly patients. *Bone Joint J*. 2013;95-B(5):609–15.
37. Bhandari M, Devereaux PJ, Einhorn TA, Thabane L, Schemitsch EH, Koval KJ, Frihagen F, Poolman RW, Tetsworth K, Guerra-Farfan E, et al. Hip fracture evaluation with alternatives of total hip arthroplasty versus hemiarthroplasty (HEALTH): protocol for a multicentre randomised trial. *BMJ Open*. 2015;5(2): e006263.
38. Rogmark C, Leonardsson O. Hip arthroplasty for the treatment of displaced fractures of the femoral neck in elderly patients. *Bone Joint J*. 2016;98-B(3): 291–7.
39. Bortolussi G, McNulty D, Waheed H, Mawhinney JA, Freemantle N, Pagano D. Identifying cardiac surgery operations in hospital episode statistics administrative database, with an OPCS-based classification of procedures, validated against clinical data. *BMJ Open*. 2019;9(3):e023316.
40. Department of Health. Governance arrangements for research ethics committees: a harmonised edition. London: Department of Health and Social Care; 2011.